Adv. Appl. Prob. **39,** 141–161 (2007)
Printed in Northern Ireland

A NEW GENETIC ALGORITHM SPECIFICALLY BASED ON MUTATION AND SELECTION

L. RIGAL,* *Université de Provence*
L. TRUFFET,** *Ecole des Mines de Nantes*

Abstract

In this paper we propose a new genetic algorithm specifically based on mutation and selection in order to maximize a fitness function. This mutation–selection algorithm behaves as a gradient algorithm which converges to local maxima. In order to obtain convergence to global maxima we propose a new algorithm which is built by randomly perturbing the selection operator of the gradient-like algorithm. The perturbation is controlled by only one parameter: that which allows the selection pressure to be governed. We use the Markov model of the perturbed algorithm to prove its convergence to global maxima. The arguments used in the proofs are based on Freidlin and Wentzell's (1984) theory and large deviation techniques also applied in simulated annealing. Our main results are that (i) when the population size is greater than a critical value, the control of the selection pressure ensures the convergence to the global maxima of the fitness function, and (ii) the convergence also occurs when the population is the smallest possible, i.e. 1.

Keywords: Freidlin–Wentzell theory; evolutionary algorithm; stochastic optimization

2000 Mathematics Subject Classification: Primary 60F10
Secondary 60J10; 92D15

1. Introduction

We are interested in maximizing a fitness function f defined by $f: E \to [1, \infty)$, where E is a finite set. In many industrial engineering problems, the objective function has an obvious lower boundary. For instance, in reliability design problems the objective function (which is often the reliability of a system) is a positive function (see, e.g. [14]). Thus, in order to obtain an ad hoc fitness function f taking its values in the interval $[1, \infty)$, we just have to make a translation of the objective function.

Evolutionary algorithms are computation methods inspired by evolution. They have received much attention as optimization methods in the last two decades. These optimization methods are population-based algorithms which simulate natural evolution. A population is a set of individuals. Each of them represents a possible solution to the optimization problem. Applying genetic operators (the mutation process, the crossing over process, and the selection process) causes the populations to evolve. Three main algorithmic trends are based on evolutionary schemes: genetic algorithms (see [12] and [13]), evolution strategies (see [15]), and evolutionary programming (see [10]).

In the authors' opinion, a fourth technique should be added: parallel simulated annealing. The link between these algorithms and simulated annealing has been established through the

Received 14 June 2005; revision received 2 February 2006.
* Postal address: Laboratoire d'Analyse, Topologie, Probabilités (LATP/UMR 6632), Université de Provence, 39 rue F. Joliot-Curie, F-13453 Marseille cedex 13, France.
** Postal address: Ecole des Mines de Nantes & IRCCYN, 4 rue Alfred Kastler, BP 20722, 44307 Nantes cedex 3, France. Email address: laurent.truffet@emn.fr

concepts introduced by Catoni (in [1] and [2]) and further generalized by Trouvé (in [17] and [18]). Continuous-time versions of genetic algorithms have been analyzed following this simulated annealing-like approach by Del Moral and Miclo in [8]. Many authors have proposed this approach for evolutionary algorithms also. Here, we briefly discuss the works which are most related to ours.

- In [7] Davis introduced a cooling schedule in the mutation probability of a genetic algorithm. Unfortunately, this genetic algorithm converges to the set of uniform populations. To overcome this drawback Cerf [6] introduced a cooling schedule in the three processes (mutation, crossing over, and selection) of a genetic algorithm. This cooling schedule allows the mutation probability and the crossing over probability to be decreased. At the same time, this cooling schedule allows the selection pressure to be increased. Cerf introduced the new parameters (a, b, c) to control the interaction between these three perturbations. He proved the convergence of this genetic algorithm to the global maxima of the fitness function.

- In [5] Cerf simplified the principle of his genetic algorithm by eliminating the crossing over process. Thus, he obtained a mutation–selection algorithm. Nevertheless, this algorithm is controlled by the parameters a and c and the cooling schedule. In the model proposed by Cerf, the mutation operator is considered to be a random, vanishing perturbation. In our model, the mutation operator is integrated into the unperturbed process and is no longer considered to be a random, vanishing perturbation.

- In [9] François proved the convergence to the global maxima of the fitness function of an exploration/selection algorithm. The cooling schedule allows the number of offspring obtained by mutation at each generation to be governed. When the number of offspring decreases, the selection pressure increases. Thus, the selection operator and the exploration operator interact. The interest of this exploration/selection algorithm is that fewer arguments are involved in their analysis than are in Cerf's algorithms. Also, the model for the exploration operator in [9] is more general than ours, which are only based on binary coding.

In this paper we build a Markovian model of a new genetic algorithm by perturbing a process. This genetic algorithm is specifically based on mutation and selection. In contrast to Cerf's genetic algorithms, in our mutation–selection algorithm the perturbation is controlled by only one parameter. There exist at least two differences between our model and the model proposed by François in [9]. First, in the algorithm proposed by François, the exploration operator and the selection operator interact; in our algorithm the mutation operator and the selection operator are completely independent. Second, in the algorithm proposed by François the mutation process is a random perturbation which gradually vanishes, while in our algorithm the mutation process does not change during the search. However the mutation process allows the algorithm to explore new individuals. These new individuals could have better fitness values than the individuals which belong to the current population. Thus, for particular problems our algorithm, which continues to explore new solutions throughout the search, seems to be more efficient than the exploration/selection algorithm proposed by François. Nevertheless, a number of 'no free lunch' theorems establish that, for any algorithm, any elevated performance over one class of problems is exactly paid for in performance over another class (see [19]).

At the beginning of each step of our mutation–selection algorithm there are μ individuals. Each might generate a mutation, so midway through a step of the algorithm there are up

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

to 2μ individuals. By sampling among these individuals with replacement μ times, a new population of μ individuals is obtained. In the 'unperturbed version' of the algorithm the sampling is uniform, while in the 'perturbed version' of the algorithm the sampling is biased according to the fitness function raised to a power of ℓ, the perturbation parameter, in the same way as in [5].

In our genetic algorithm children are individuals which are created by applying the mutation operator to the parent's population. In Cerf's genetic algorithms children replace the parents, while in our algorithm the sampling is made among both parents and children. Consequently, we obtained a new cost function. Our algorithm belongs to the class of 'generalized simulated annealing' studied by Trouvé [17]. Trouvé's work allows us to deduce that our algorithm converges to a set $\mathcal{W}^*$. In order to determine the set $\mathcal{W}^*$, we follow the approach introduced by Freidlin and Wentzell in [11]. As in [4], the structure of the set of attractors of the unperturbed process is very rich. However, the 'unperturbed versions' of the algorithms are completely different. In our case the unperturbed version of the algorithm has the same behavior as a gradient algorithm. In Cerf's case the unperturbed version of the algorithm is a crossing over–selection algorithm in which the selection process preserves the diversity of the individuals present in the population. The attractors of our unperturbed process are particular subsets of populations. They form a partition of $\mathcal{S}$, the set of equifitness populations. They are divided into two groups: the stable attractors and the unstable attractors. There exists a group, $\mathcal{K}^*$, of stable attractors whose populations contain only the maxima of the fitness function. We study the communication cost between attractors. As a consequence, when μ is greater than a critical value μ^*, the set $\mathcal{W}^*$ is included in the set $\mathcal{K}^*$ (see Theorem 2). We also note that when $\mu = 1$ the set $\mathcal{W}^*$ is included in the set $\mathcal{K}^*$ (see Theorem 3).

This paper has the following structure. Sections 2–4 are devoted to the description of the model: the unperturbed process and the random perturbation. Section 5 describes the structure of the set of attractors of the unperturbed process. In Section 6 we prove the convergence of the new algorithm obtained by perturbing the gradient algorithm (see Theorem 2). The convergence is ensured if the population size is sufficiently large. However, in Section 7 we prove that the same algorithm converges to global extrema for population size equal to 1 (see Theorem 3).

2. The population space

Our mutation–selection algorithm generates an initial population of size μ which evolves at each generation using the mutation and the selection operators successively. A population of size μ is a vector composed of μ individuals. In this paper we use a binary coding: an individual $e \in E$ is a binary string of length L. Thus, our mutation–selection algorithm is a stochastic process with state space

$$E^\mu = \{0, 1\}^{L\times\mu}.$$

The population which represents the μ parents will be denoted by the letters x and/or y. We denote by $[x]$ the set of all individuals present in the population $x \in E^\mu$:

$$x = (x_1, \ldots, x_\mu) \quad \Longrightarrow \quad [x] = \{x_j : 1 \le j \le \mu\}.$$

From the μ parents, λ offspring are generated using the mutation operator, where λ is a random variable such that $\lambda \le \mu$. To represent the offspring by a population with a fixed size μ, we model a population u composed of both λ offspring and $\mu - \lambda$ virtual individuals. A virtual individual is also called an 'empty individual', and is denoted by $e_\varnothing$. Generation

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

of a virtual individual $e_\varnothing$ means that the mutation operator has had no effect on the parent. Let $E_\varnothing = E \cup \{e_\varnothing\}$ be the extended solution space. We will denote the global population, of size 2μ, by $z = (x, u)$, which allows us to represent the μ parents and the λ offspring. For all global populations z of size 2μ, we denote by $[z]$ the set composed of both the μ parents and the λ offsprings present in z, i.e. we have the following mapping:

$$E^\mu \times E_\varnothing^\mu \to \mathcal{P}(E),$$
$$z = (z_1, \ldots, z_{2\mu}) \mapsto [z] = \{z_j \in E : 1 \le j \le 2\mu\},$$

where $\mathcal{P}(E)$ denotes the set of all subsets of E. For all $e \in E$, we denote by $z(e) = |\{1 \le j \le 2\mu : z_j = e\}|$ the number of occurrences of e in z. With f we associate a $[1, \infty]$-valued function $\hat{f}$ defined on $E^\mu \times E_\varnothing^\mu$ by

$$\hat{f}(z) = \max_{1 \le i \le 2\mu} f(z_i),$$

with the convention that $f(e_\varnothing) = 0$. That is, $\hat{f}$ is the maximal fitness of the μ parents and the λ offspring which comprise the global population z. The set composed of the best individuals from among both the parents and the offspring is the set $\hat{z}$ defined by

$$\hat{z} = \{z_i : f(z_i) = \hat{f}(z)\}.$$

With f we also associate another $[1, \infty]$-valued function denoted by $\hat{f}$, but defined on E^μ by

$$\hat{f}(x) = \max_{1 \le i \le \mu} f(x_i)$$

Because $f(e_\varnothing) = 0$ and $f \ge 1$, the two different functions denoted by $\hat{f}$ play the same role. That is, for a population v (in E^μ or in $E^\mu \times E_\varnothing^\mu$), $\hat{f}(v)$ is the maximal fitness of the individuals of v.

3. The unperturbed process

We first describe the underlying process which drives the algorithm. When there is no perturbation, the process under study is a Markov chain $(X_n^{+\infty})$ with state space E^μ. The superscript '$+\infty$' reflects the fact that this process corresponds to the limit behavior of our model when the perturbation vanishes. In fact, this unperturbed process is a mutation–selection algorithm which has the same behavior as the gradient algorithm. The transition mechanism from $(X_n^{+\infty})$ to $(X_{n+1}^{+\infty})$ is decomposed into two stages,

$$X_n^{+\infty} \xrightarrow{\text{mutation}} U_n^{+\infty} \quad \text{and} \quad (X_n^{+\infty}, U_n^{+\infty}) \xrightarrow{\text{selection}} X_{n+1}^{+\infty},$$

where $U_n^{+\infty}$ is the random population of size μ obtained after the mutation process. We now describe the mutation and selection operators in detail.

3.1. $X_n^{+\infty} \longrightarrow U_n^{+\infty}$: the mutation process

We define the $|E^\mu| \times |E_\varnothing^\mu|$ matrix α which describes the mutation process from $X_n^{+\infty}$ to $U_n^{+\infty}$ by

$$\alpha(x, u) = \mathrm{P}(U_n^{+\infty} = u \mid X_n^{+\infty} = x) \quad \text{for all } (x, u) \in E^\mu \times E_\varnothing^\mu.$$

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

The mutation process is applied independently to each individual of the population $X_n^{+\infty}$. Thus, we have

$$\alpha(x,u) = \alpha_E(x_1,u_1)\cdots\alpha_E(x_\mu,u_\mu),$$

where $\alpha_E(x_i,u_i)$ is the $|E|\times|E_\varnothing|$ transition matrix which describes the application of the process of mutation from $x_i \in E$ to $u_i \in E_\varnothing$, according to which with probability p_{m} a neighbor of x_i (namely u_i) is generated and with probability $1-p_{\mathrm{m}}$ an empty individual $e_\varnothing$ is generated. The individual u_i is said to be a neighbor of the individual x_i if $d(x_i,u_i)$, the number of bits of x_i and u_i which are different, is equal to 1. We remind the reader that d is the Hamming distance defined in, e.g. [16]. Thus,

$$\alpha_E(x_i,u_i) = \begin{cases} p_{\mathrm{m}}/L & \text{if } u_i \in E \text{ and } d(x_i,u_i)=1,\\ 1-p_{\mathrm{m}} & \text{if } u_i = e_\varnothing,\\ 0 & \text{otherwise.}\end{cases}$$

The mutation process is an exploration operator which allows the algorithm to visit the neighborhoods of λ different x_i, where the individuals x_i are chosen from among the parent population of size μ.

We remark that, for each couple of individuals $(x_i,y_i)\in E\times E$ with $x_i\neq y_i$, y_i could be generated by switching $d(x_i,y_i)$ bits of x_i. In other words, one way to generate y_i is successively to apply the mutation process $d(x_i,y_i)$ times starting from x_i. Let r_i be a positive integer such that $r_i = d(x_i,y_i)+1$ if $x_i\neq y_i$ and $r_i=3$ otherwise. Then,

$$\begin{aligned}&\text{for all } (x_i,y_i)\in E\times E,\ \text{there exists an } (e_i^1,\dots,e_i^{r_i})\in E^{r_i}\\ &\qquad\text{with } e_i^1 = x_i \text{ and } e_i^{r_i} = y_i \text{ such that } \textstyle\prod_{1\le k\le r_i-1}\alpha_E(e_i^k,e_i^{k+1})>0. \quad (1)\end{aligned}$$

3.2. $(X_n^{+\infty}, U_n^{+\infty}) \longrightarrow X_{n+1}^{+\infty}$: the selection process

When the perturbation vanishes, the selection pressure is equal to ∞. The selection process of the unperturbed Markov chain, $(X_n^{+\infty})$, is highly elitist. To form the population $X_{n+1}^{+\infty}$, only the best individuals of the global population $Z_n^{+\infty} = (X_n^{+\infty}, U_n^{+\infty})$ obtained by mutation are selected. If $Z_n^{+\infty} = z$ then the individuals of the population $X_{n+1}^{+\infty}$ are chosen, independently and randomly according to the uniform distribution, from among the elements of $\hat{z}$, i.e. for all $(z,y)\in(E^\mu\times E_\varnothing^\mu)\times E^\mu$,

$$\mathrm{P}(X_{n+1}^{+\infty} = y \mid Z_n^{+\infty} = z) =: \gamma_{+\infty}(z,y) = \prod_{i=1}^{\mu}\frac{\mathbf{1}_{\hat{z}}(y_i)z(y_i)}{|\hat{z}|},$$

where

$$\mathbf{1}_{\hat{z}}\colon E\to\{0,1\},\qquad e\mapsto\begin{cases}1 & \text{if } e\in\hat{z},\\ 0 & \text{otherwise.}\end{cases}$$

3.3. The convergence of the unperturbed process $(X_n^{+\infty})$

The transition mechanism of the process $(X_n^{+\infty})$ is given by

$$\mathrm{P}(X_{n+1}^{+\infty} = y \mid X_n^{+\infty} = x) = \sum_u \alpha(x,u)\gamma_{+\infty}((x,u),y).$$

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

Let $\mathcal{U}(x)$ be the set of the populations u generated by applying the mutation process to x, i.e.

$$\mathcal{U}(x) = \{u \in E_{\varnothing}^{\mu} : \alpha(x, u) > 0\}.$$

From the global population (x, u), the selection process chooses only the population y composed of the best individuals, i.e. such that

$$[y] \subseteq \widehat{(x, u)},$$

where $\widehat{(x, u)}$ denotes the set composed of the best individuals of the population (x, u). Then we deduce that

$$\mathrm{P}(X_{n+1}^{+\infty} = y \mid X_n^{+\infty} = x) > 0 \quad \Longleftrightarrow \quad [y] \subseteq \widehat{\mathcal{U}(x)}, \tag{2}$$

where $\widehat{\mathcal{U}(x)}$ is the set defined by

$$\widehat{\mathcal{U}(x)} = \bigcup_{u \in \mathcal{U}(x)} \widehat{(x, u)}.$$

In one step (of the algorithm), the process $(X_n^{+\infty})$ is trapped in the set of equifitness populations $\mathcal{S}$ defined by

$$\mathcal{S} = \{x \in E^{\mu} : f(x_1) = \cdots = f(x_{\mu})\}.$$

Let R be the equivalence relation defined on $\mathcal{S}$ by

$$\begin{aligned} xRy \quad \Longleftrightarrow \quad & \text{there exists a } k \in \mathbb{N} \setminus \{0, 1\} \text{ such that} \\ & (x = p^1, \ldots, p^k = y) \in \mathcal{S} \times \cdots \times \mathcal{S} \text{ for all } i \in \{1, \ldots, k-1\}, \\ & \text{where } [p^{i+1}] \subseteq \widehat{\mathcal{U}(p^i)} \text{ and } \hat{f}(p^1) = \cdots = \hat{f}(p^k). \end{aligned} \tag{3}$$

If xRy then there exists a path p leading from x to y (or from y to x) composed of populations which have the same fitness value. The equivalence relation R induces a partition of the set $\mathcal{S}$. Because the state space E^{μ} is finite, the number of equivalence classes is finite. Let r denote the number of equivalence classes. These equivalence classes, denoted by $\mathcal{C}_i$, $i \in \{1, \ldots, r\}$, are indexed such that

$$1 \leq \hat{f}(\mathcal{C}_1) \leq \cdots \leq \hat{f}(\mathcal{C}_r) = \max_{e \in E} f(e),$$

where $\hat{f}(\mathcal{C}_i) = \hat{f}(x)$ for all $i \in \{1, \ldots, r\}$ and all $x \in \mathcal{C}_i$.

The equivalence (2) means that in one step the process $(X_n^{+\infty})$ is in a particular equivalence class. The Markov chain $(X_n^{+\infty})$ wanders through particular subsets of $\mathcal{S}$ which are the equivalence classes. We will say that the equivalence class $\mathcal{C}_i$ 'communicates' with the equivalence class $\mathcal{C}_j$ if and only if

$$\hat{f}(\mathcal{C}_i) < \hat{f}(\mathcal{C}_j) \quad \text{and} \quad \text{there exists a } (x, y) \in \mathcal{C}_i \times \mathcal{C}_j \text{ such that } [y] \subseteq \widehat{\mathcal{U}(x)}.$$

Assume that the equivalence class $\mathcal{C}_i$ communicates with another equivalence class. Because of the property of the matrix α_E stated in (1), if the Markov chain $(X_n^{+\infty})$ enters $\mathcal{C}_i$ then the exit time from $\mathcal{C}_i$ is finite. On the other hand, if we assume that $\mathcal{C}_i$ does not communicate with any other equivalence class, then $\mathcal{C}_i$ is an absorbing group of states of $(X_n^{+\infty})$. The equivalence classes $\mathcal{C}_i$ which do not communicate with any other equivalence class are the sets composed

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

of individuals which are the local maxima of the fitness function. These particular sets are denoted by $\mathcal{K}_j$, and indexed from 1 to τ (with $1 \le \tau \le r$) in increasing order of fitness value:

$$\hat{f}(\mathcal{K}_1) \le \cdots \le \hat{f}(\mathcal{K}_\tau).$$

The chain $(X_n^{+\infty})$ simulates the behavior of the gradient algorithm. Indeed, the Markov chain $(X_n^{+\infty})$ converges to the set, $\mathcal{K} = \bigcup_{j=1}^{\tau} \mathcal{K}_j$, of populations composed of local maxima of the fitness function; i.e.

$$\lim_{n\to\infty} \mathrm{P}(X_n^{+\infty} \in \mathcal{K} \mid X_0^{+\infty} = x) = 1 \quad \text{for all } x \in E^m.$$

4. The perturbed process

We build a stochastic mutation–selection algorithm which will be proved to converge to the set of global maxima of the fitness function f. This mutation–selection algorithm is obtained by randomly perturbing the Markov chain $(X_n^{+\infty})$. The perturbation acts only on the selection process. The mechanism of the perturbation consists in decreasing the selection pressure. The intensity of the perturbation is governed by a positive parameter ℓ, so we obtain a family of Markov chains (X_n^ℓ) indexed by ℓ. The transition from X_n^ℓ to X_{n+1}^ℓ includes two stages, respectively corresponding to the mutation and selection processes

$$X_n^\ell \xrightarrow{\text{mutation}} U_n^\ell \quad \text{and} \quad (X_n^\ell, U_n^\ell) \xrightarrow{\text{selection}} X_{n+1}^\ell.$$

4.1. $X_n^\ell \longrightarrow U_n^\ell$: the mutation process

The mutation process is not perturbed in any way: this stage is exactly the same as the passage from $X_n^{+\infty}$ to $U_n^{+\infty}$ described in Subsection 3.1. Thus, the mutation process from X_n^ℓ to U_n^ℓ is given by

$$\mathrm{P}(U_n^\ell = u \mid X_n^\ell = x) = \alpha(x, u) \quad \text{for all } (x, u) \in E^\mu \times E_\varnothing^\mu.$$

4.2. $(X_n^\ell, U_n^\ell) \longrightarrow X_{n+1}^\ell$: the selection process

We perturb the selection process of the Markov chain $(X_n^{+\infty})$. The selection operator of the perturbed Markov chain (X_n^ℓ) is built as a proportional selection process applied with the objective function

$$f_{\varnothing,\ell} \colon E_\varnothing \to \mathbb{R}^+, \qquad e \mapsto \begin{cases} \exp(\ln(f(e))\ell) & \text{if } e \in E, \\ 0 & \text{if } e = e_\varnothing. \end{cases}$$

If ℓ increases then so does the selection pressure, i.e. the process becomes more selective. The perturbed selection process allows us to choose μ independent individuals to form the next population X_{n+1}^ℓ from the μ parents and the λ offspring. Recall that the μ parents and λ offspring at the nth iteration are represented by the global population $Z_n^\ell = (X_n^\ell, U_n^\ell)$. Let $z = (z_1, z_2, \ldots, z_{2\mu}) \in E^\mu \times E_\varnothing^\mu$ be a vector of the global population Z_n^ℓ, and let $F_\ell(j, z)$ be the probability of selecting the individual z_j from the population $z = (z_1, z_2, \ldots, z_{2\mu})$. Thus, the selection process of the perturbed Markov chain (X_n^ℓ) is described by

$$\mathrm{P}(X_{n+1}^\ell = y \mid Z_n^\ell = z) =: \gamma_\ell(z, y) = \prod_{i=1}^{\mu} \sum_{\{j \colon z_j = y_i\}} F_\ell(j, z)$$

$$\text{for all } (z, y) \in (E^\mu \times E_\varnothing^\mu) \times E^\mu, \quad (4)$$

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

where the function F_ℓ is defined by

$$F_\ell : \{1, \dots, 2\mu\} \times E^\mu \times E_\varnothing^\mu \to \mathbb{R}^+, \qquad (j, z_1, \dots, z_{2\mu}) \mapsto \frac{f_{\varnothing,\ell}(z_j)}{\sum_{k=1}^{2\mu} f_{\varnothing,\ell}(z_k)}.$$

As the selection pressure tends to ∞, the selection process concentrates its choice on the best individuals of the population $Z_n^\ell = z$, which belong to the set $\hat{z}$, i.e.

$$\lim_{\ell\to\infty} F_\ell(j, z) = \frac{\mathbf{1}_{\hat{z}}(z_j)}{|\hat{z}|}.$$

Thus, when the perturbation vanishes the selection process of the perturbed Markov chain (X_n^ℓ) tends to the selection process of the unperturbed Markov chain $(X_n^{+\infty})$.

Lemma 1. *For all* $(z, y) \in (E^\mu \times E_\varnothing^\mu) \times E^\mu$,

$$\lim_{\ell\to\infty} \gamma_\ell(z, y) = \gamma_{+\infty}(z, y).$$

4.3. The vanishing perturbation

Considering all possible paths between the populations X_n^ℓ and X_{n+1}^ℓ, we can define the one-step transition probability of the Markov chain (X_n^ℓ) by

$$\mathrm{P}(X_{n+1}^\ell = y \mid X_n^\ell = x) = \sum_u \alpha(x, u)\gamma_\ell((x, u), y). \tag{5}$$

The perturbation appears only in the selection process; thus, Lemma 1 implies that

$$\lim_{\ell\to\infty} \mathrm{P}(X_{n+1}^\ell = y \mid X_n^\ell = x) = \mathrm{P}(X_{n+1}^{+\infty} = y \mid X_n^{+\infty} = x).$$

This means that the transition probabilities of (X_n^ℓ) converge to those of $(X_n^{+\infty})$ as ℓ tends to ∞. Thus, the Markov chain (X_n^ℓ) appears as a perturbation of the Markov chain $(X_n^{+\infty})$.

4.4. Convergence of the homogeneous algorithm

This subsection is devoted to the study of the behavior of the Markov chain (X_n^ℓ) for a fixed value of ℓ as n goes to ∞. For all $x \in E^\mu$ and all $u \in E_\varnothing^\mu$, $\gamma_\ell((x, u), x) > 0$ and $\sum_u \alpha(x, u) = 1$. Thus, we have

$$\mathrm{P}(X_{n+1}^\ell = x \mid X_n^\ell = x) = \sum_u \alpha(x, u)\gamma_\ell((x, u), x) > 0.$$

Consequently, (X_n^ℓ) is an aperiodic Markov chain. We remark that $\gamma_\ell(z, y) > 0$ if and only if $[y] \subseteq [z]$, and in applying the property of the transition matrix α_E stated in (1) we obtain the following result: for all $i \in \{1, \dots, \mu\}$ and $(x, y_i) \in E^\mu \times E$, there exists an $r_i \in \mathbb{N}$ such that

$$\mathrm{P}(X_{n+r_i}^\ell = (x_1, \dots, x_{i-1}, y_i, x_{i+1}, \dots, x_\mu) \mid X_n^\ell = (x_1, \dots, x_\mu)) > 0.$$

This result allows us to conclude that the chain (X_n^ℓ) is an aperiodic homogeneous and irreducible Markov chain. Thus, the chain (X_n^ℓ) admits a unique invariant probability measure, μ_ℓ, such that

$$\mu_\ell(y) = \lim_{n\to\infty} \mathrm{P}(X_n^\ell = y \mid X_0^\ell = x) > 0 \quad \text{for all } x \in E^\mu \text{ and all } y \in E^\mu.$$

This measure charges all populations of the space E^μ. Under the mutation–selection algorithm, and for a fixed value of the selection parameter ℓ, (X_n^ℓ) does not converge to the set of global maxima of the fitness function.

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

4.5. Asymptotic transition probabilities of the chain (X_n^ℓ)

Recall (5), which describes the transition matrix of (X_n^ℓ):

$$\mathrm{P}(X_{n+1}^\ell = y \mid X_n^\ell = x) = \sum_{u \in E_\varnothing^\mu} \alpha(x,u)\gamma_\ell((x,u),y).$$

Let $\mathcal{D}(x,y)$ denote the set of populations u which allow us to create population y from population x in only one iteration with a strictly positive transition probability, i.e.

$$\mathcal{D}(x,y) = \{u \in E_\varnothing^\mu : \alpha(x,u) > 0 \text{ and } [y] \subseteq [(x,u)]\}.$$

Considering all the paths between the populations X_n^ℓ and X_{n+1}^ℓ which have a strictly positive associated probability, we deduce that

$$\mathrm{P}(X_{n+1}^\ell = y \mid X_n^\ell = x) = \sum_{u \in \mathcal{D}(x,y)} \alpha(x,u)\gamma_\ell((x,u),y). \tag{6}$$

To determine the asymptotic transition probabilities of the perturbed Markov chain (X_n^ℓ), we just have to study the behavior of γ_ℓ as ℓ goes to ∞. The perturbation in the selection process appears in the same way as in [3, p. 49]. By translating the reasoning proposed in [3, p. 55] to our selection process, it is easy to see that, by definition of γ_ℓ (given by (4)),

$$\gamma_\ell(z,y) \sim \ddot{\gamma}(z,y)\exp(-V(z,y)\ell) \quad \text{as } \ell \to \infty, \text{ for all } (z,y) \in E^\mu \times E_\varnothing^\mu \times E^\mu.$$

Here

- $V(z,y) = \sum_{i=1}^\mu [\ln(\hat{f}(z)) - \ln(f(y_i))]$ represents the perturbation cost necessary to achieve all the antiselection, and
- $\ddot{\gamma}(z,y) = \prod_{i=1}^\mu z(y_i)/|\hat{z}|$ is the weight associated with the perturbation cost V.

By applying this result in (6), we obtain the following result, the proof of which is obvious.

Lemma 2. *We have*

$$\mathrm{P}(X_{n+1}^\ell = y \mid X_n^\ell = x) \sim C^*(x,y)\exp(-V^*(x,y)\ell) \quad \textit{as } \ell \to \infty,$$

where

- $V^*(x,y)$ *is the communication cost, defined on* $E^\mu \times E^\mu$ *by*

$$V^*(x,y) = \min_{u \in \mathcal{D}(x,y)} V((x,u),y),$$

- $\mathcal{D}^*(x,y)$ *is the set of the populations* $u \in \mathcal{D}(x,y)$ *which realize the minimum, i.e.*

$$\mathcal{D}^*(x,y) = \{u \in \mathcal{D}(x,y) : V((x,u),y) = V^*(x,y)\},$$

  and

- $C^*(x,y)$ *is the weighting associated with the communication cost* V^**, defined on* $E^\mu \times E^\mu$ *by*

$$C^*(x,y) = \sum_{u \in \mathcal{D}^*(x,y)} \alpha(x,u)\ddot{\gamma}((x,u),y).$$

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

We define the communication kernel q on $E^\mu \times E^\mu$ by

$$q(x, y) = \frac{C^*(x, y)}{\sum_{y \in E^\mu} C^*(x, y)}.$$

Moreover, we remark that, for each x, y in E^μ,

$$q(x, y) = 0 \quad \Longleftrightarrow \quad V^*(x, y) = \infty.$$

Thus, we are now in the framework of generalized simulated annealing studied by Trouvé in [17] and [18]. That is, the transition probabilities of the process (X_n^ℓ) form a family of Markov kernels on the space E^μ, indexed by ℓ, which is admissible for the communication kernel q and the cost function V^*. Our mutation–selection algorithm can be seen as an optimization algorithm minimizing the virtual energy (see Proposition 2.6 of [18]). We restate Definition 2.5 of [18], which defines the virtual energy $\mathcal{W}$ associated with the cost function V^*. To do so, we have to recall the notion of an $\mathcal{A}$-graph as defined in [11]. As there, $x \to y$ will denote the pair (x, y).

Definition 1. Let $\mathcal{A} \subseteq E^\mu$. We say that a set G of arrows $x \to y$ in $(E^\mu \setminus \mathcal{A}) \times E^\mu$ is an $\mathcal{A}$-graph if and only if

1. for all $x \in E^\mu \setminus \mathcal{A}$, there exists a unique $y \in E^\mu$ such that $x \to y \in G$, and
2. for all $x \in E^\mu \setminus \mathcal{A}$, there exists a path $x = p_1 \to p_2 \to \cdots \to p_r$ such that $p_k \to p_{k+1} \in G$, ending at a population in $\mathcal{A}$.

We denote by $G(\mathcal{A})$ the set of $\mathcal{A}$-graphs. Furthermore, for each $G \in G(\mathcal{A})$ we write

$$V^*(G) = \sum_{x \to y \in G} V^*(x, y).$$

We can now define the virtual energy as follows.

Definition 2. For all $y \in E^\mu$,

$$\mathcal{W}(y) = \inf_{G \in G(y)} V^*(G).$$

We also write

$$\mathcal{W}_{\min} = \inf_{y \in E^\mu} \mathcal{W}(y), \qquad \mathcal{W}^* = \{y \in E^\mu : \mathcal{W}(y) = \mathcal{W}_{\min}\}.$$

We now restate Trouvé's result for the convergence in the inhomogeneous case, where the parameter ℓ is an increasing function of n. Then we have an inhomogeneous Markov chain $(X_n^{\ell(n)})$.

Theorem 1. (Point 1 of Theorem 2.22 of [18].) *There exists a constant H_1, called the critical height, such that for all increasing sequences $\ell(n)$ we have*

$$\lim_{n \to \infty} \mathrm{P}(X_n^{\ell(n)} \notin \mathcal{W}^* \mid X_0 = x) = 0 \quad \textit{for all } x \in E^\mu \qquad \Longleftrightarrow \qquad \sum_{n \in \mathbb{N}} \exp(-H_1 \ell(n)) = \infty.$$

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

5. The virtual energy $\mathcal{W}$

As we have demonstrated in the last section, it is very important to determine the set $\mathcal{W}^*$ in order to describe the asymptotic convergence of the process $(X_n^{\ell(n)})$. In order to study the virtual energy $\mathcal{W}$, we follow the approach introduced by Freidlin and Wentzell [11]. They analyzed the structure of the set of attractors of the deterministic system. They distinguished between two kinds of attractors: the stable attractors and the unstable attractors. In our perturbation model, some particular subsets of populations play the same role as the attractors of the deterministic system. Thus, we call these particular subsets the attractors of the unperturbed process. By translating part of the theory proposed in [11] to our model, we will prove that the set $\mathcal{W}^*$ is included in the set of stable attractors.

5.1. The set of attractors

In this subsection we will prove that the attractors of the unperturbed process are the equivalence classes defined in Subsection 3.3. In order to define the set of attractors of the unperturbed process $(X_n^{+\infty})$, we use a new communication cost function, V'. To do so, we have to introduce the following definitions and notation.

Definition 3. A path p in E^μ is a finite sequence of populations of size μ denoted by $p^1 \to \cdots \to p^r$. The length of the path p is denoted by $|p|$ and is equal to r.

A path p is said to join two populations x and y if $p^1 = x$ and $p^{|p|} = y$. By $\mathcal{D}^\mu$ we denote the set of paths in E^μ which correspond to possible trajectories of the process (X_n^ℓ), i.e. the paths p satisfying

$$V^*(p^k, p^{k+1}) < \infty \quad \text{for all } k \in \{1, \ldots, |p| - 1\}.$$

The set of all paths in $\mathcal{D}^\mu$ joining the populations x and y is denoted by $\mathcal{D}^\mu(x, y)$. The V^*-cost of $p \in \mathcal{D}^\mu$ is given by

$$V^*(p) = \sum_{k=1}^{|p|-1} V^*(p^k, p^{k+1}).$$

We define the communication cost function V' on $E^\mu \times E^\mu$ by

$$V'(x, y) = \inf_{p \in \mathcal{D}^\mu(x,y)} V^*(p) \quad \text{for all } (x, y) \in E^\mu \times E^\mu. \tag{7}$$

We can use the results of [11] to study the virtual energy by virtue of the following proposition.

Proposition 1. (Proposition 5.4 of [5].) *The virtual energy $\mathcal{W}$ associated with the communication cost function V^* coincides with the virtual energy associated with the communication cost function V'.*

We define the set of attractors of the unperturbed process $(X_n^{+\infty})$ in the same way as Freidlin and Wentzell defined the stable attractors of the deterministic system. They defined the stable attractors using the three properties stated in [11, p. 169]. In our case we will prove that the following properties are possessed by the r equivalence classes defined in Subsection 3.3.

Property 1. For all equivalence classes $\mathcal{C}_i$ we have

$$V'(x, y) = V'(y, x) = 0 \quad \text{for all } (x, y) \in \mathcal{C}_i \times \mathcal{C}_i. \tag{8}$$

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

Property 2. For all $(i, j) \in \{1, \ldots, r\} \times \{1, \ldots, r\}$ with $i \neq j$, we have

$$V'(x, y) \neq 0 \quad \text{or} \quad V'(y, x) \neq 0, \quad \text{for all } (x, y) \in \mathcal{C}_i \times \mathcal{C}_j.$$

Property 3. P(there exist $n_0 \geq 0$ and $j \in \{1, \ldots, r\}$ such that $X_n^{+\infty} \in \mathcal{C}_j$ for $n \geq n_0$) $= 1$.

In order to prove the first property, we demonstrate the following lemma.

Lemma 3. *$V^*(x, y) = 0$ if and only if $[y] \subseteq \widehat{\mathcal{U}(x)}$.*

Proof. Because $\mathcal{D}(x, y)$ is a finite set, we have

$$\begin{aligned} V^*(x, y) = 0 \quad &\Longleftrightarrow \quad \min_{u \in \mathcal{D}(x,y)} V((x, u), y) = 0 \\ &\Longleftrightarrow \quad \text{there exists a } u \in \mathcal{U}(x) \text{ such that} \\ &\qquad [y] \subseteq [(x, u)] \text{ and } \hat{f}((x, u)) = f(y_i) \text{ for all } i \in \{1, \ldots, \mu\} \\ &\Longleftrightarrow \quad \text{there exists a } u \in \mathcal{U}(x) \text{ such that } [y] \subseteq \widehat{(x, u)}. \end{aligned}$$

By using this lemma and the definition of the equivalence relation R (see (3)), for all equivalence classes $\mathcal{C}_i$ we have the following result:

$$\begin{aligned} &\text{for all } (x, y) \in \mathcal{C}_i \times \mathcal{C}_i, \text{ there exists a } k \in \mathbb{N} \setminus \{0, 1\} \text{ such that} \\ &\qquad (x = p^1, p^2, \ldots, p^{k-1}, p^k = y) \in \mathcal{S} \times \cdots \times \mathcal{S}, \\ &\qquad\qquad \text{where } V^*(p^{i+1}, p^i) = 0 \text{ for all } i \in \{1, \ldots, k-1\}. \end{aligned}$$

To conclude the proof of Property 1, we just have to use the definition of V' (see (7)) and remark that the cost, $V^*(p)$, associated with the path $p = (p^1 \to p^2 \to \cdots \to p^{k-1} \to p^k)$ is equal to 0.

In order to prove the second property, we demonstrate the following lemma.

Lemma 4. *For all $(x, y) \in E^\mu \times E^\mu$ such that $\hat{f}(x) \geq \hat{f}(y)$, we have*

$$V^*(x, y) \geq \mu(\ln(\hat{f}(x)) - \ln(\hat{f}(y))).$$

Proof. If $\mathcal{D}(x, y) = \varnothing$ then $V^*(x, y) = \infty$; otherwise, for all $u \in \mathcal{D}(x, y)$ we have $\hat{f}((x, u)) \geq \hat{f}(x)$ and, so,

$$\begin{aligned} V((x, u), y) &\geq \sum_{i=1}^{\mu} (\ln(\hat{f}(x)) - \ln(f(y_i))) \\ &\geq \mu(\ln(\hat{f}(x)) - \ln(\hat{f}(y))). \end{aligned}$$

In using the definition of V' stated in (7) and by applying the previous lemma, we obtain the following lemma.

Lemma 5. *For all $(j_1, j_2) \in \{1, \ldots, r\} \times \{1, \ldots, r\}$ such that $\hat{f}(\mathcal{C}_{j_1}) > \hat{f}(\mathcal{C}_{j_2})$, we have*

$$V'(x, y) \geq \mu(\ln(\hat{f}(\mathcal{C}_{j_1})) - \ln(\hat{f}(\mathcal{C}_{j_2}))) \quad \textit{for all } (x, y) \in \mathcal{C}_{j_1} \times \mathcal{C}_{j_2}.$$

Proof. See Appendix A.

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

We easily observe that if, for all $(i, j) \in \{1, \ldots, r\} \times \{1, \ldots, r\}$ with $i \neq j$, $\mathcal{C}_i$ communicates with $\mathcal{C}_j$, then $\hat{f}(\mathcal{C}_j) > \hat{f}(\mathcal{C}_i)$. Thus, by applying Lemma 5 we conclude the proof of Property 2.

Because the equivalence classes which do not communicate with any other equivalence class are the absorbing groups of states of the Markov chain $(X_n^{+\infty})$, the proof of Property 3 is obvious.

Thus, the equivalence classes are the attractors of the unperturbed process $(X_n^{+\infty})$.

5.2. The set of stable attractors

In this subsection it will be proved that

1. the sets $\mathcal{K}_j$, $j = 1, \ldots, \tau$, play the same role as the stable attractors of the deterministic system in [11], and
2. the sets $\mathcal{K}_j$, $j = 1, \ldots, \tau$, possess the same properties as the stable attractors of the deterministic system in [11].

The set $\mathcal{K}_j$ does not communicate with any other equivalence class. Thus,

$$V'(x, y) > 0 \quad \text{for all } (x, y) \in \mathcal{K}_j \times (E^\mu \setminus \mathcal{K}_j). \tag{9}$$

This property is similar to that which defines the stable attractors of the deterministic system in [11, p. 188]. Thus, we will say that the sets $\mathcal{K}_j$, $j = 1, \ldots, \tau$, are the stable attractors of the unperturbed process. We define the cost between attractors in the same way as in [11].

Definition 4. $V'(\mathcal{C}_i, \mathcal{C}_j) = \min_{(x,y)\in\mathcal{C}_i\times\mathcal{C}_j} V'(x, y)$.

From (8), we have

$$V'(\mathcal{C}_i, \mathcal{C}_j) = V'(x, y) \quad \text{for all } (x, y) \in \mathcal{C}_i \times \mathcal{C}_j. \tag{10}$$

Consequently, we give a lemma which is a rewriting of Lemma 4.2 of [11].

Lemma 6. *For all $x \in \mathcal{S}$, there exists a $j \in \{1, \ldots, \tau\}$ such that*

$$V'(x, y) = 0 \quad \textit{for all } y \in \mathcal{K}_j.$$

Proof. By applying Lemma 3, and by construction of the sets $\mathcal{K}_j$, $j = 1, \ldots, \tau$ (see Subsection 3.3), the proof of this lemma is obvious.

We define the virtual energy of an attractor in the same way as Freidlin and Wentzell [11, p. 185].

Definition 5. $\mathcal{W}(\mathcal{C}_i) = \min_{G\in G(i)} \sum_{m\to n\in G} V'(\mathcal{C}_m, \mathcal{C}_n)$.

By applying (10), we find that

$$\mathcal{W}(\mathcal{C}_i) = \mathcal{W}(x) \quad \text{for all } x \in \mathcal{C}_i.$$

Consequently, we give a lemma which is a rewriting of point (c) of Lemma 4.3 of [11], and which has a similar proof.

Lemma 7. *If $x \in \mathcal{S} \setminus \mathcal{K}$ then $\mathcal{W}(x) = \min_{y\in\mathcal{K}}(\mathcal{W}(y) + V'(y, x))$.*

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

5.3. $\mathcal{W}^* \subseteq \mathcal{K}$

In this subsection we will prove that any population which realizes the maximal value of the virtual energy $\mathcal{W}$ belongs to the set $\mathcal{K}$. (Recall that the set $\mathcal{K}$ is composed of the stable attractors.) First we will prove that each such population belongs to the set $\mathcal{S}$, which is composed of all the attractors. Let τ^ℓ be the first entrance time of the chain (X_n^ℓ) into $\mathcal{S}$:

$$\tau^\ell = \min\{n > 0 \colon X_n^\ell \in \mathcal{S}\}.$$

Also, let

$$s_\ell = \min_{x \in E^\mu} \sum_{\{y \in E^\mu \colon [y] \subseteq \hat{\mathcal{U}}(x)\}} \mathrm{P}(X_{n+1}^\ell = y \mid X_n^\ell = x).$$

Then, for all $x \in E^\mu$ and all $n \in \mathbb{N}$,

$$\mathrm{P}(X_{n+1}^\ell \in \mathcal{S} \mid X_n^\ell = x) \geq \sum_{\{y \in E^\mu \colon [y] \subseteq \hat{\mathcal{U}}(x)\}} \mathrm{P}(X_{n+1}^\ell = y \mid X_n^\ell = x) \geq s_\ell,$$

and we remark that, for all $x \in E^\mu$,

$$\lim_{\ell \to \infty} s_\ell = \sum_{\{y \in E^\mu \colon [y] \subseteq \hat{\mathcal{U}}(x)\}} \mathrm{P}(X_{n+1}^{+\infty} = y \mid X_n^{+\infty} = x) = 1.$$

From these equations, by applying the same reasoning as in [3, pp. 51–53], we deduce that

$$\begin{aligned} \lim_{\ell \to \infty} \mu_\ell(\mathcal{S}) &= 1, \\ \lim_{\ell \to \infty} \mu_\ell(E^\mu \setminus \mathcal{S}) &= 0. \end{aligned} \tag{11}$$

We denote by $\mu_{+\infty}$ the limit measure, i.e. such that

$$\mu_{+\infty}(x) = \lim_{\ell \to \infty} \mu_\ell(x) \quad \text{for all } x \in E^\mu.$$

By translating the reasoning described in [3, pp. 62–63], to our Markov chain (X_n^ℓ), we can show that

$$\mu_{+\infty}(x) \begin{cases} > 0 & \text{for all } x \in \mathcal{W}^*, \\ = 0 & \text{for all } x \notin \mathcal{W}^*. \end{cases}$$

By using the result in (11), we find that $\mathcal{W}^* \subseteq \mathcal{S}$. From this result, if we apply Lemma 7 and (9), we deduce that

$$\mathcal{W}^* \subseteq \mathcal{K}. \tag{12}$$

6. Convergence to the global maxima of f

Combined with (12), Theorem 1 means that if we choose the series $(\ell(n))_{n \geq 0}$ carefully, then the inhomogeneous Markov chain $(X_n^{\ell(n)})$ converges to the set $\mathcal{W}^* \subseteq \mathcal{K}$. In this case our mutation–selection algorithm converges to the set of local maxima. However, we would prefer to obtain an algorithm which converges to the global maxima. We remark that there exists a group, $\mathcal{K}^*$, of stable attractors which are composed only of global maxima. We denote by τ^* the smallest integer such that

$$\hat{f}(\mathcal{K}_j) = \max_{e \in E} f(e) \quad \text{for all } j \geq \tau^*.$$

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

Thus, we have $\mathcal{K}^* = \bigcup_{j=\tau^*}^{\tau} \mathcal{K}_j$. In this section we will prove our main result, namely that $\mathcal{W}^* \subseteq \mathcal{K}^*$ when the number of parents μ is greater than a critical value μ^*. To do so we study the communication cost between attractors. Lemma 5 demonstrates that the cost of bad transitions (transitions from x to y with $(x, y) \in \mathcal{K}^* \times (\mathcal{K} \setminus \mathcal{K}^*)$) increases linearly with μ. We prove that, as a consequence of Lemma 8, below, the cost of good transitions (transitions from x to y with $(x, y) \in (\mathcal{K} \setminus \mathcal{K}^*) \times \mathcal{K}^*$) remains bounded.

Lemma 8. *For all sets $\mathcal{K}_{j_1} \subseteq \mathcal{K}$ and $\mathcal{K}_{j_2} \subseteq \mathcal{K}^*$, for all $x \in \mathcal{K}_{j_1}$, and for all $y \in \mathcal{K}_{j_2}$, $V'(x, y)$ has an upper bound that does not depend on μ.*

Proof. See Appendix B.

We are now in position to establish the main result of the paper.

Theorem 2. *There exists a critical value μ^*, depending on the space E, the fitness function f, and the matrix transition α_E, such that*

$$\mu \geq \mu^* \quad \Longrightarrow \quad \mathcal{W}^* \subseteq \mathcal{K}^*.$$

Proof. For all $y \in \mathcal{K} \setminus \mathcal{K}^*$ and all $x \in \mathcal{K}^*$, by applying Lemma 5 we have

$$V^*(p) \geq \mu(\ln(\hat{f}(x)) - \ln(\hat{f}(y))) \quad \text{for all } p \in \mathcal{D}^{\mu}(x, y).$$

Thus,

$$\mathcal{W}(y) = \inf_{G \in G(y)} V^*(G) \geq \mu\Big(\ln\Big(\max_{e \in E} f(e)\Big) - \ln(\hat{f}(y))\Big).$$

We will prove that, for all $y \in \mathcal{K}^*$, there is a $G \in G(y)$ such that $V'(G) = \sum_{(z,z') \in G} V'(z, z')$ is independent of the value of μ. According to Lemma 6, for all $x \in E^{\mu} \setminus \mathcal{K}$ there exists an integer $j(x) \in \{1, \ldots, \tau\}$ such that

$$V'(x, z) = 0 \quad \text{for all } z \in \mathcal{K}_{j(x)}. \tag{13}$$

For all $j \in \{1, \ldots, \tau\}$, we choose a uniform population $z_j = (e_j, \ldots, e_j)$ which belongs to the stable attractor $\mathcal{K}_j$. For all $y \in \mathcal{K}_{j^*}$ and any $j^* \in \{\tau^*, \ldots, \tau\}$, we build the y-graph G as follows:

$$\begin{aligned} x \to z_{j(x)} &\in G, && \text{for all } x \in E^{\mu} \setminus \mathcal{K} \\ x \to z_j &\in G, && \text{for all } x \in \mathcal{K}_j \setminus \{z_j, y\} \\ z_j \to z_{j^*} \in G, \quad z_{j^*} \to y &\in G && \text{for all } j \in \{1, \ldots, \tau\} \setminus \{j^*\}. \end{aligned}$$

Then

$$\begin{aligned} V'(G) = &\sum_{x \in E^{\mu} \setminus \mathcal{K}} V'(x, z_{j(x)}) + \sum_{j \in \{1, \ldots, \tau\}} \sum_{x \in \mathcal{K}_j \setminus \{z_j, y\}} V'(x, z_j) \\ &+ \sum_{j \in \{1, \ldots, \tau\} \setminus \{j^*\}} V'(z_j, z_{j^*}) + V'(z_{j^*}, y) \end{aligned}$$

and, by (13), we have

$$\sum_{x \in E^{\mu} \setminus \mathcal{K}} V'(x, z_{j(x)}) = 0.$$

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

Property 1 yields

$$\sum_{j\in\{1,\dots,\tau\}} \sum_{x\in\mathcal{K}_j\setminus\{z_j,y\}} V'(x,z_j)=0, \qquad V'(z_{j^*},y)=0.$$

Thus,

$$V'(G)=\sum_{j\in\{1,\dots,\tau\}\setminus\{j^*\}} V'(z_j,z_{j^*}). \tag{14}$$

Using Lemma 8, we conclude that the cost $V'(G)$ is independent of the value of μ. Then, from Proposition 1, we have

$$\mathcal{W}(y)\le V'(G) \quad \text{for all } y\in\mathcal{K}^*.$$

The proof of this theorem allows us to remark that, for all $y\in\mathcal{K}\setminus\mathcal{K}^*$,

$$\mathcal{W}(y)\ge \mu\Big(\ln\Big(\max_{e\in E} f(e)\Big)-\ln(\hat{f}(\mathcal{K}))\Big).$$

If we use the fact that $|p|\le L+2$ in (20) (stated in the proof of Lemma 8, below), we find that, for all $y\in\mathcal{K}^*$ and all $x\in\mathcal{K}\setminus\mathcal{K}^*$,

$$V'(x,y)\le L\ln\Big(\max_{e\in E} f(e)\Big).$$

From (14), we deduce that

$$\mathcal{W}(y)\le \tau L\ln\Big(\max_{e\in E} f(e)\Big).$$

Thus, an upper bound for μ^* is

$$\frac{\tau L\ln(\max_{e\in E} f(e))}{\ln(\max_{e\in E} f(e))-\ln(\max_{\{\mathcal{K}:\,\mathcal{K}\cap\mathcal{K}^*=\varnothing\}}\hat{f}(\mathcal{K}))}.$$

If we use our mutation–selection algorithm with $\mu\ge\tau L/\varepsilon$, we obtain an approximate algorithm with a logarithmic relative error ε, defined by

$$\varepsilon=\frac{\ln(\max_{e\in E} f(e))-\ln(\max_{\{\mathcal{K}:\,\mathcal{K}\cap\mathcal{K}^*=\varnothing\}}\hat{f}(\mathcal{K}))}{\ln(\max_{e\in E} f(e))}.$$

This logarithmic relative error tells us how close the approximate solutions for the 'final' population are to the optimal solution.

7. The particular case $\mu=1$

Even though Theorem 2, the main result on convergence established in the previous section, is very important, it could be difficult, and sometimes impossible, to use it to solve particular real problems. Indeed, for some real problems, the value of the upper bound for μ^* found above could be very large. Therefore, our genetic algorithm will not be efficient in solving these real problems, because of computation time. To obtain a genetic algorithm which is more easy to apply to real problems, in this section we will prove the convergence of our genetic algorithm to the global maxima of the fitness function for the specific case in which $\mu=1$. This convergence is illustrated by the following theorem.

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

Theorem 3. *If $\mu = 1$ then $\mathcal{W}^* \subseteq \mathcal{K}^*$.*

Proof. To prove the theorem, we study the cost between two solutions, e^1 and e^2, to the optimization problem, where e^2 is a neighbor of e^1. Thus, the two solutions e^1 and e^2 differ by only one bit, i.e. $\alpha(e^1, e^2) > 0$. The cost V^* between e^1 and e^2 is

$$V^*(e^1, e^2) = \begin{cases} \ln(f(e^1)) - \ln(f(e^2)) & \text{if } f(e^1) > f(e^2), \\ 0 & \text{if } f(e^1) \leq f(e^2). \end{cases} \tag{15}$$

We will say that a path $p \in \mathcal{D}^1$ is 'decreasing' if and only if

$$\hat{f}(p^{k+1}) \leq \hat{f}(p^k) \quad \text{for all } k \in \{1, \ldots, |p| - 1\}.$$

On the contrary we will say that a path $p \in \mathcal{D}^1$ is 'increasing' if and only if

$$\hat{f}(p^{k+1}) \geq \hat{f}(p^k) \quad \text{for all } k \in \{1, \ldots, |p| - 1\}.$$

We consider a path $p = (p^1 \to p^2 \to p^3 \to \cdots \to p^{|p|})$ which belongs to $\mathcal{D}^1$. We study the case in which p is either a decreasing path or an increasing path. Using (15), we easily deduce the following result on the cost of the path p: if p is an increasing path then $V^*(p) = 0$, while if p is a decreasing path then $V^*(p) = \ln(f(p^{|p|})) - \ln(f(p^1))$.

We consider a path $p = (p^1 \to p^2 \to p^3 \to \cdots \to p^{|p|})$ which joins a global maximum y and a local maximum x. That is, $p \in \mathcal{D}^1(y, x)$ with $x \in \mathcal{K} \setminus \mathcal{K}^*$ and $y \in \mathcal{K}^*$. We remark that $\alpha(p^i, p^{i+1}) > 0$ if and only if $\alpha(p^{i+1}, p^i) > 0$. Thus, the path $p' = (p^{|p|} \to p^{|p|-1} \to p^{|p|-2} \to \cdots \to p^1)$ belongs to the set $\mathcal{D}^1(x, y)$. We call p' the 'inverse' path of p.

In the way described in Appendix A, below, we divide the path p into m paths denoted by $c_1, \ldots, c_m$. Each path c_i is an increasing path or a decreasing path. Obviously we have

$$V^*(p) = \sum_{j=1}^{m} V^*(c_j). \tag{16}$$

Thus, the path p' is divided into m paths $c'_m, \ldots, c'_1$, where c'_i is the inverse path of c_i, and we have

$$V^*(p') = \sum_{j=1}^{m} V^*(c'_j). \tag{17}$$

Using (16), (17), the result on the cost of an increasing or a decreasing path, and the fact that $f(x) < f(y)$, we remark that

$$V^*(p) > V^*(p').$$

Consequently, we have demonstrated the following result.

Lemma 9. *For all $x \in \mathcal{K} \setminus \mathcal{K}^*$ and all $y \in \mathcal{K}^*$,*

$$p \in \mathcal{D}^1(y, x) \qquad \Longrightarrow \qquad p' \in \mathcal{D}^1(x, y) \quad \textit{and} \quad V^*(p) > V^*(p').$$

We consider a local maximum x and a global maximum y, i.e. $x \in \mathcal{K} \setminus \mathcal{K}^*$ and $y \in \mathcal{K}^*$. The set G is an x-graph. We know that there exists a path $p \in \mathcal{D}^1(y, x)$ such that, for all i with $1 \leq i \leq |p| - 1$, $p^i \to p^{i+1}$ belongs to G. Thus, we build the y-graph

$$G' = (G \setminus \{p^i \to p^{i+1} : 1 \leq i \leq |p| - 1\}) \cup \{p^{i+1} \to p^i : 1 \leq i \leq |p| - 1\}.$$

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

We denote by p' the inverse path of p, and observe that, for all i with $1 \leq i \leq |p'| - 1$, $p'^i \to p'^{i+1}$ belongs to G'. Lemma 9 yields $V^*(p) > V^*(p')$, and we thus deduce that

$$V^*(G') = V^*(G) + V^*(p') - V^*(p) < V^*(G).$$

Consequently, we have proven that, for all $x \in \mathcal{K} \setminus \mathcal{K}^*$ and all $y \in \mathcal{K}^*$, if G is an x-graph then there exists a G' which is a y-graph such that

$$V^*(G') < V^*(G).$$

Finally, we have demonstrated that

$$\mathcal{W}(y) < \mathcal{W}(x) \quad \text{for all } x \in \mathcal{K} \setminus \mathcal{K}^* \text{ and all } y \in \mathcal{K}^*.$$

Thus, $\mathcal{W}^* \subseteq \mathcal{K}^*$.

8. Conclusion

In this paper we have proven that our new genetic algorithm converges to the maxima of the fitness function for a sufficiently large population (see Theorem 2) or for a population of size $\mu = 1$. Consequently, it would be interesting to make further theoretical studies in order to determine precisely the value of μ^*, which could, for some functions f, be conjectured to equal 1.

Appendix A. Proof of Lemma 5

For all $(x', y') \in E^\mu \times E^\mu$, we will say that a path $p \in \mathcal{D}^\mu(x', y')$ is 'decreasing' if and only if

$$\hat{f}(p^{k+1}) \leq \hat{f}(p^k) \quad \text{for all } k \in \{1, \ldots, |p| - 1\},$$

and we will say that a path $p \in \mathcal{D}^\mu(x', y')$ is 'increasing' if and only if

$$\hat{f}(p^{k+1}) \geq \hat{f}(p^k) \quad \text{for all } k \in \{1, \ldots, |p| - 1\}.$$

We choose a pair of populations $(x', y') \in E^\mu \times E^\mu$ such that $\hat{f}(x') \leq \hat{f}(y')$. According to Lemma 4, the cost of an increasing path p that joins the population x' and the population y' has the property that

$$\begin{aligned} V^*(p) &= \sum_{k=1}^{|p|-1} V^*(p^k, p^{k+1}) \\ &\geq \sum_{k=1}^{|p|-1} \mu(\ln(\hat{f}(p^k)) - \ln(\hat{f}(p^{k+1}))) \\ &= \mu(\ln(\hat{f}(x')) - \ln(\hat{f}(y'))). \end{aligned} \tag{18}$$

For $(x, y) \in \mathcal{C}_{j_1} \times \mathcal{C}_{j_2}$ such that $\hat{f}(\mathcal{C}_{j_1}) > \hat{f}(\mathcal{C}_{j_2})$, and for all $p \in \mathcal{D}^\mu(x, y)$, we divide the path p into m paths denoted by $c_1, \ldots, c_m$. To represent the m paths, we define $m + 1$ integers $k_1, \ldots, k_{m+1}$. These integers k_j allow us to choose a population p^{k_j} which belongs to the path $p = (p^1 \to \cdots \to p^{|p|})$. The integer k_{j+1} allows us to determine the first population of the path c_{j+1} which is equal to the last population of the path c_j. Thus, the path c_j is $p^{k_j} \to p^{k_j+1} \to p^{k_j+2} \to \cdots \to p^{k_{j+1}}$ and the path c_{j+1} is $p^{k_{j+1}} \to p^{k_{j+1}+1} \to p^{k_{j+1}+2} \to \cdots \to p^{k_{j+2}}$, and so on.

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

Then, to determine all the paths c_j, $j = 1, \ldots, m$, we consider the following procedure.

1. Let $k_1 = 1$; thus, $x = p_1$ is the first population of the path c_1.
2. Assume that c_i has been constructed for $i \leq j$ and then construct c_{j+1} by determining the integer $k_j + 1$ as follows.
   - If $\hat{f}(p^{k_j}) \geq \hat{f}(p^{k_j+1})$ then choose the integer k_{j+1} such that $p^{k_j} \to \cdots \to p^{k_{j+1}}$ is the decreasing path of maximum length.
   - If $\hat{f}(p^{k_j}) < \hat{f}(p^{k_j+1})$ then choose the integer k_{j+1} such that $p^{k_j} \to \cdots \to p^{k_{j+1}}$ is the increasing path of maximum length.

   The procedure ends when the population $p^{k_{j+1}}$ is equal to $p^{|p|} = y$.

Thus, the path p is divided into m' decreasing paths, $c_{j_1}, \ldots, c_{j_{m'}}$, and $m - m'$ increasing paths. Because the cost of a path is a positive real number, we have

$$V^*(p) = \sum_{j=1}^{m} V^*(c_j) \geq \sum_{i=1}^{m'} V^*(c_{j_i}).$$

Applying (18), we have

$$V^*(p) \geq \sum_{j \in \{1,\ldots,m\} \setminus \{j_1,\ldots,j_{m'}\}} V^*(c_j) \geq \sum_{j \in \{1,\ldots,m\} \setminus \{j_1,\ldots,j_{m'}\}} \mu(\ln(\hat{f}(p_{k_j})) - \ln(\hat{f}(p_{k_{j+1}}))).$$

If $j \in \{1, \ldots, m\} \setminus \{j_1, \ldots, j_{m'}\}$ then c_j is an increasing path and, thus,

$$\ln(\hat{f}(p_{k_{j+1}})) - \ln(\hat{f}(p_{k_j})) \geq 0.$$

By adding the terms $\ln(\hat{f}(p_{k_{j+1}})) - \ln(\hat{f}(p_{k_j}))$ if c_j corresponds to an increasing path, we obtain

$$\begin{aligned} V^*(p) &\geq \sum_{j=1}^{m} \mu(\ln(\hat{f}(p_{k_j})) - \ln(\hat{f}(p_{k_{j+1}}))) \\ &\quad + \sum_{j \in \{1,\ldots,m\} \setminus \{j_1,\ldots,j_{m'}\}} \mu(\ln(\hat{f}(p_{k_{j+1}})) - \ln(\hat{f}(p_{k_j}))) \\ &\geq \sum_{j=1}^{m} \mu(\ln(\hat{f}(p_{k_j})) - \ln(\hat{f}(p_{k_{j+1}}))) \\ &= \mu(\ln(\hat{f}(x)) - \ln(\hat{f}(y))). \end{aligned}$$

Appendix B. Proof of Lemma 8

For all $(x, y) \in \mathcal{K}_{j_1} \times \mathcal{K}_{j_2}$, by applying Lemma 3 we have

$$\begin{aligned} V^*((x_1, x_2, \ldots, x_\mu), (x_1, x_1, \ldots, x_1)) = 0, \\ V^*((y_1, y_2, \ldots, y_\mu), (y_1, y_1, \ldots, y_1)) = 0. \end{aligned} \tag{19}$$

https://doi.org/10.1239/aap/1175266473 Published online by Cambridge University Press

From (1), there exists a $(e^1, \dots, e^r) \in E^r$ with $e^1 = x_1$, $e^r = y_1$, and

$$\prod_{1 \le k \le r-1} \alpha_E(e^k, e^{k+1}) > 0.$$

We build the following path p of length $r+2$, which joins the populations x and y:

$$\begin{aligned}
p^1 &= (x_1, x_2, \dots, x_\mu),\\
p^2 &= (x_1, x_1, \dots, x_1),\\
p^k &= \begin{cases} (e^{k-1}, e^{k-1}, \dots, e^{k-1}) & \text{if } f(e^{k-1}) \ge f(p_2^{k-1}),\\ (e^{k-1}, p_2^{k-1}, \dots, p_\mu^{k-1}) & \text{if } f(e^{k-1}) < f(p_2^{k-1}), \end{cases} \qquad k \in \{3, \dots, r\},\\
p^{r+1} &= (y_1, y_1, \dots, y_1),\\
p^{r+2} &= (y_1, y_2, \dots, y_\mu).
\end{aligned}$$

From (19), we deduce that

$$V^*(p) = \sum_{k=2}^{|p|-2} V^*(p^k, p^{k+1}).$$

We remark that if $p^{k+1} = (e^k, \dots, e^k)$ then $V^*(p^k, p^{k+1}) = 0$ because $\alpha_E(e^{k-1}, e^k) > 0$, and that if $p^{k+1} = (e^k, p_2^k, \dots, p_\mu^k)$ then $V^*(p^k, p^{k+1}) = \ln(f(p_2^k)) - \ln(f(e^k))$. Thus,

$$V^*(p) = \sum_{k=2}^{|p|-2} V^*(p^k, p^{k+1}) = \sum_{k=2}^{|p|-2} \max(0, \ln(f(p_2^k)) - \ln(f(e^k))). \tag{20}$$

Hence, $V^*(p)$ is independent of μ and $V'(x, y) \le V^*(p)$, as required.

Acknowledgements

The authors thank Raphaël Cerf and Frédéric Lefevre for their useful advice and comments, which improved this work.

References

[1] Catoni, O. (1992). Large deviations for annealing. Doctoral Thesis, Université Paris XI.
[2] Catoni, O. (1992). Rough large deviations estimates for simulated annealing. Application to exponential schedules. *Ann. Prob.* **20,** 1109–1146.
[3] Cerf, R. (1994). Une théorie asymptotique des algorithmes génétiques. Doctoral Thesis, Université Montpellier II.
[4] Cerf, R. (1996). A new genetic algorithm. *Ann. Appl. Prob.* **6,** 778–817.
[5] Cerf, R. (1996). The dynamics of mutation-selection algorithms with large population sizes. *Ann. Inst. H. Poincaré* **32,** 455–508.
[6] Cerf, R. (1998). Asymptotic convergence of genetic algorithms. *Adv. Appl. Prob.* **30,** 521–550.
[7] Davis, T. E. and Principe, J. C. (1991). A simulated annealing like convergence theory for the simple genetic algorithm. In *Proc. Fourth Internat. Conf. Genetic Algorithms*, Morgan Kauffman, San Mateo, CA, pp. 174–181.
[8] Del Moral, P. and Miclo, L. (1999). On the convergence and the applications of the generalized simulated annealing. *SIAM J. Control Optimization* **37,** 1222–1250.
[9] François, O. (2002). Global optimization with exploration/selection algorithms and simulated annealing. *Ann. Appl. Prob.* **12,** 248–271.
[10] Fogel, L. J., Owens, A. J. and Walsh, M. J. (1996). *Artificial Intelligence Through Simulated Evolution*. John Wiley, New York.

[11] Freidlin, M. I. and Wentzell, A. D. (1984). *Random Perturbations of Dynamical Systems*. Springer, New York.
[12] Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA.
[13] Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI.
[14] Kuo, W., Prasad, V. R. and Tillman, F. A. (2000). *Optimal Reliability Design*. Cambridge University Press.
[15] Rechenberg, I. (1973). *Evolutionsstrategie: Optimierung Technischer Systeme nach Prinzipien der Biologischen Evolution*. Frommann-Holzboog, Stuttgart.
[16] Rudolph, G. (1994). Convergence analysis of canonical genetic algorithms. *IEEE Trans. Neural Networks* **5,** 96–101.
[17] Trouvé, A. (1993). Parallélisation massive du recuit simulé. Doctoral Thesis, Université Paris XI.
[18] Trouvé, A. (1996). Cycle decompositions and simulated annealing. *SIAM J. Control Optimization* **34,** 966–986.
[19] Wolpert, D. H. and Macready, W. G. (1997). No free lunch theorems for optimization. *IEEE Trans. Evolutionary Comput.* **1,** 67–82.