The Annals of Applied Probability
1996, Vol. 6, No. 3, 778–817

A NEW GENETIC ALGORITHM

BY RAPHAËL CERF

CNRS, Université d'Orsay

Here is a new genetic algorithm. It is built by randomly perturbing a two operator crossover-selection scheme. Three conditions of biological relevance are imposed on the crossover. A new selection mechanism is used, which has the decisive advantage of preserving the diversity of the individuals in the population. The attractors of the unperturbed process are particular equifitness subsets of populations endowed with a rich structure. The random vanishing perturbations are twofold: local perturbations of the individuals (mutations) and loosening of the selection pressure. When the population size is greater than a critical value which depends strongly on the optimization problem, their delicate asymptotic interaction ensures the convergence (possibly in finite time) of the population to the ideal attractor whose populations contain all the maxima of the fitness function. The process explores without respite the neighborhoods of the best points found so far (instead of focusing on a particular point) and finds simultaneously all the global maxima of the fitness function; it seems to be the first cooperative search procedure of this kind.

1. Introduction. In the realm of stochastic optimization, attention has focused essentially on two techniques during the last decade: simulated annealing [1] and evolutionary algorithms [12, 16].

The fundamental problem may be stated as follows: given a finite but huge space (the size precludes any exhaustive search procedure) endowed with a transition mechanism and a real-valued function defined on this space, determine the set of its global maxima or find at least suboptimal points, as fast as possible.

The theory of simulated annealing is now extensively developed and a great number of results describing the dynamics of this kind of algorithm in various settings are available [1–5, 8, 9, 11, 13–15, 17–20, 22, 24–29]. As far as we know, the most accurate work in this area has been achieved by Catoni, in the spirit of the Freidlin–Wentzell theory [2–5]. Nevertheless, simulated annealing presents a fundamental drawback: it is sequential in nature.

In an attempt to investigate the theoretical aspects of the parallelization of this algorithm, Trouvé carried out a systematic study (initiated by Hwang and Sheu [19]) of a broader class of algorithms he baptized "generalized simulated annealing" [25, 26, 28, 29]. As it turns out, this framework is also well adapted for evolutionary algorithms. Let us mention that several recent studies have been devoted to these kinds of processes. Holley, Kusuoka and Stroock [17, 18] have developed an approach leading to an estimation of the second eigenvalue of the infinitesimal generator of the Markov process. This method is particu-

Received April 1995; revised January 1996.
AMS 1991 *subject classifications*. 60F10, 60J10, 92D15.
Key words and phrases. Freidlin–Wentzell theory, genetic algorithms, stochastic optimization.

larly well suited to reversible situations, but it has also been recently extended to nonsymmetric annealing processes by Deuschel and Mazza [8] and is thus potentially applicable to genetic algorithms. One of our major goals is to show how the large deviations theory of Markov chains with rare transitions can be used to study the convergence of genetic algorithms.

In a first paper [6], we proposed a Markovian model of Holland's simple genetic algorithm which is built by randomly perturbing a very simple selection scheme: mutations and crossovers are considered as vanishing random perturbations. We proved that convergence to the global maxima of the fitness function becomes possible when the population size is greater than a critical value (which depends strongly on the optimization problem). Surprisingly, the crossover is not fundamental to ensure this convergence: the crucial point is the delicate asymptotic interaction between the local perturbations of the individuals (i.e., the mutations) and the selection pressure.

In a second paper [7], we used the concepts introduced by Catoni [2, 3, 5] and further generalized by Trouvé [26, 28, 29] to fathom more deeply the dynamics of the two operators mutation-selection algorithm when the population size becomes very large. The key result lies in the structure of the trajectories of populations joining two uniform populations; a small group of individuals sacrifice themselves in order to create an ideal path which is then followed by all other individuals. As a consequence, the various quantities associated with the algorithm (such as the communication cost, the virtual energy, the communication altitude, etc.) are affine functions of the population size. We proved that the hierarchy of cycles on the set of the uniform populations stabilizes. Furthermore, if the mutation kernel is symmetric, the limiting distribution is the uniform distribution over the set of the global maxima of the fitness function.

In this third paper, we introduce two major modifications to the previous schemes. The crossover is now integrated into the unperturbed process and is not considered any more as a random vanishing perturbation. Although this operator is not essential to ensure the desired convergence, it certainly increases the efficiency of the algorithm. Three conditions of biological relevance are imposed on the crossover.

1. When two identical individuals mate, they produce offspring identical to themselves.
2. There is always a nonzero probability that nothing happens during a crossover.
3. The two individuals of the mating pair play symmetric roles (the populations are asexual).

The first condition is essential for our algorithm to work for every fitness function; the second condition makes the analysis somewhat easier; the third condition is a natural symmetry assumption which could be removed.

Furthermore, we propose a new selection mechanism which has the decisive advantage of preserving the diversity of the individuals in the population.

The analysis of the algorithm follows the road opened by Freidlin and Wentzell [10]. Unlike the situations studied in [6, 7], the structure of the set of the attractors of the unperturbed process is very rich. These are particular subsets of populations and they stand in one-to-one correspondence with the equifitness subsets of the state space whose cardinality is less than the population size. When the population size is large enough, there is a unique ideal attractor whose populations contain all the maxima of the fitness function. We study the communication cost between attractors: the costs of bad transitions (either those which decrease the maximal fitness of the population or those which lose some peak fitness individuals) increase linearly with the population size, whereas the costs of good transitions (those which create some new peak fitness individuals) remain bounded. As a consequence, when the population size is greater than a critical value, the minimum of the virtual energy corresponds to the ideal attractor previously described. Therefore the sequence of the stationary measures (associated with a fixed level of intensity of the perturbations) concentrates on this attractor as the perturbations vanish. The remaining problem is to adapt carefully the rate of decrease of the perturbations in order to obtain an inhomogeneous Markov chain with the same limiting law. Besides, it is possible to ensure a stronger convergence; we may force the process to be forever trapped in the attraction basin of the ideal attractor after a finite number of transitions. Furthermore, when the population size is large, the cycles which do not contain the ideal attractor are reduced to one single attractor, and the optimal convergence exponent increases faster than an affine function of the population size. We show also how our general model specializes to the case where the state space is $\{0, 1\}^N$; we discuss the role of the crossover and compare the genetic algorithm with the parallel (independent) simulated annealing on a small numerical example.

One might wonder whether this asymptotic point of view is relevant for analyzing genetic algorithms. Indeed, the current state of this large deviations theory (even the sharp large deviations estimates of Catoni [3, 4]) does not yet provide probability bounds which can be effectively computed in a real problem. Therefore one does not know in practice when the process is near the asymptotic regime. Anyway, we hope that the paradigm we propose is one of the very first steps toward a complete theory. Asymptotic convergence is the least thing to require for such stochastic algorithms. Moreover, our model should shed some light on the true behavior of genetic algorithms. It provides also a tool to make theoretical comparisons; we are able to analyze the optimal convergence exponent for large population sizes (its rate of increase is a quantitative measurement of the intrinsic parallelism of genetic algorithms), to assess the impact of the crossover operator on it or to compare it with the one associated with parallel (independent) simulated annealing.

Finally, let us summarize the important aspects of this work. Our algorithm finds simultaneously all the global maxima of the objective function in finite time and thus solves completely the optimization problem; it seems to be the first of this kind. The cornerstone of this cooperative search procedure is the delicate asymptotic interaction between the mutations and our enhanced selection mechanism; the process explores simultaneously and without respite the neighborhoods of the best points found so far (instead of focusing on a particular point). Moreover, we hope to have found the right way of using the crossover operator.

This paper has the following structure. Sections 2–7 are devoted to the description of the model: the unperturbed process, its attractors and the random perturbations. The main results are presented in Section 8. The role of the crossover is analyzed in Section 9. In Section 10 we show how our model enters the class of generalized annealing processes. We then give technical results in Sections 11–14 and prove the main results in Section 15.

General conventions. The cardinality of a set X will be denoted indifferently $|X|$ or $\operatorname{card} X$ and its characteristic function 1_X. We adopt the usual conventions concerning empty sets:

$$\prod_{\varnothing} = 1, \qquad \sum_{\varnothing} = 0, \qquad \min \varnothing = +\infty, \qquad \max \varnothing = -\infty.$$

If s is a real number, $\lfloor s \rfloor$ denotes the unique integer such that $\lfloor s \rfloor \le s < \lfloor s \rfloor + 1$.

The Kronecker symbol $\delta(i, j)$ will be used to denote the identity matrix indexed by E:

$$\forall\, i, j \in E, \qquad \delta(i, j) = \begin{cases} 0, & \text{if } i \neq j, \\ 1, & \text{if } i = j. \end{cases}$$

For any integer r, the set of the permutations over $\{1 \cdots r\}$ is denoted by $\mathfrak{S}(r)$.

2. The fitness landscape. We recall that a Markov kernel k on a finite set X is a function $k(x, y)$ defined on $X \times X$ with values in $[0, 1]$ satisfying

$$\forall\, x \in X, \qquad \sum_{y \in X} k(x, y) = 1.$$

DEFINITION 2.1. An abstract fitness landscape consists of four objects (E, f, α, β), where:

(i) E, the search space, is a finite space of states;
(ii) f, the fitness function, is a positive nonconstant function defined on E;
(iii) α, the mutation kernel, is an irreducible Markov kernel on E;
(iv) β, the crossover kernel, is a Markov kernel on the set $E \times E$.

The points of E will be called individuals and will be mostly denoted by the letters i, j and e. The quantity $\alpha(i, j)$ determines the rate at which the individual i mutates to j. The irreducibility assumption on α, which means that

$$\begin{gathered} \forall\, i, j \in E, \ \exists\, e_1, \dots, e_r \in E, \qquad e_1 = i, e_r = j, \\ \forall\, k \in \{1 \cdots r-1\}, \qquad \alpha(e_k, e_{k+1}) > 0 \end{gathered}$$

is essential to ensure that the whole space E can be explored with the help of the mutation mechanism.

Similarly, the quantity $\beta((i_1, j_1), (i_2, j_2))$ is interpreted as the probability of producing the pair of individuals (i_2, j_2) by performing a crossover on the couple (i_1, j_1). Throughout this work, we impose the following conditions on the kernel β:

$$\forall\, i \in E, \qquad \beta((i,i),(i,i)) = 1, \tag{1}$$

$$\forall\, i, j \in E, \qquad \beta((i,j),(i,j)) > 0, \tag{2}$$

$$\forall\, i_1, j_1, i_2, j_2 \in E, \qquad \beta((i_1, j_1), (i_2, j_2)) = \beta((j_1, i_1), (j_2, i_2)). \tag{3}$$

These three conditions have a natural biological interpretation. Condition (1) states that when two identical individuals mate, they produce offspring identical to themselves. Condition (2) states that there is a nonzero probability that nothing happens during the crossover. Condition (3) states that the two individuals of the mating pair play symmetric roles (our population is asexual).

The set f^* of the global maxima of f is

$$f^* = \left\{ i \in E\colon\ f(i) = \max_{j \in E} f(j) \right\}$$

and $f(f^*)$ is the maximum value of f over E, that is, $\max_{j\in E} f(j)$. Symbols with an asterisk $(*)$ in superscript will denote sets realizing the minimum or the maximum of a particular functional. The main goal of a genetic algorithm is to locate the points of f^* or at least to find suboptimal points. The minimal and maximal variations of the fitness function are

$$\delta = \min\{|f(i) - f(j)|\colon\ i, j \in E,\ f(i) \neq f(j)\},$$
$$\Delta = \max\{|f(i) - f(j)|\colon\ i, j \in E\}.$$

Let λ be a positive real number. We put $f_\lambda = f^{-1}(\{\lambda\})$, which is the set of the points of the search space having fitness λ, and we define similarly $f_\lambda^+ = f^{-1}(]\lambda, \infty[)$, and $f_\lambda^- = f^{-1}([0, \lambda[)$. The maximal cardinality of a level set of the fitness function is $\Lambda = \max\{|f_\lambda|\colon\ \lambda \in \mathbb{R}_+^*\}$. If F is a subset of E, $f(F)$ (or sometimes simply fF) denotes the set $\{f(i)\colon\ i \in F\}$.

3. The population space. A key principle of genetic algorithms is to make a population of individuals evolve simultaneously in the fitness landscape. Let m be the population size of our algorithm. A genetic algorithm is a stochastic process with state space the population space E^m, the m-tuples of elements of E. That is, the points of the set E^m are called populations and will be mostly denoted by the letters x, y, z, u and v. We define a bracket operator $[\ ]$ on E^m with values in $\mathscr{P}(E)$, the set of all the subsets of E, by

$$x = (x_1, \dots, x_m) \in E^m \quad \mapsto \quad [x] = \{x_k\colon\ 1 \le k \le m\};$$

that is, $[x]$ is the set of all individuals present in the population x.

With f we associate a function $\widehat{f}$ defined on E^m by

$$\widehat{f}(x) = \widehat{f}(x_1, \dots, x_m) = \max\{f(x_k)\colon\ 1 \le k \le m\};$$

that is, $\widehat{f}(x)$ is the maximal fitness of the individuals of population x. In this way, the fitness landscape (E, f, α, β) induces a landscape structure $(E^m, \widehat{f})$ on population space E^m. A key issue is to understand this new landscape and to determine whether it is easier to optimize $(E^m, \widehat{f})$ than the naive m-fold copy of (E, f). The peculiarity of genetic algorithms is to introduce a cooperative dynamics on the search space.

We now introduce some notation describing the repartition of fitness values in a population or in a set of populations. For x in E^m, $\widehat{x}$ denotes the set of those elements of $[x]$ which realize the value $\widehat{f}(x)$:

$$\widehat{x} = \{x_k\colon\ 1 \le k \le m,\ f(x_k) = \widehat{f}(x)\}.$$

The previous definitions $[x]$, $\widehat{f}(x)$ and $\widehat{x}$ are extended to $\mathscr{P}(E^m)$, the set of all the subsets of E^m, in a natural way: if K is a subset of E^m, we have

$$\begin{aligned}
[K] &= \{x_k\colon\ 1 \le k \le m,\ x = (x_1, \dots, x_m) \in K\},\\
\widehat{f}(K) &= \max\{f(x_k)\colon\ 1 \le k \le m,\ x = (x_1, \dots, x_m) \in K\},\\
\widehat{K} &= \{x_k\colon\ 1 \le k \le m,\ x = (x_1, \dots, x_m) \in K,\ f(x_k) = \widehat{f}(K)\}.
\end{aligned}$$

For x in E^m, we let $x^\lambda = [x] \cap f_\lambda$ (the individuals of the population x having fitness λ) and we number the elements of the set x^λ,

$$x^\lambda = \left\{x_1^\lambda, \dots, x_{|x^\lambda|}^\lambda\right\},$$

in the order they appear in the sequence $(x_k, 1 \le k \le m)$. Equivalently, we have

$$\forall\ h, k,\ 1 \le h < k \le |x^\lambda|, \qquad \min\{r\colon\ x_r = x_h^\lambda\} < \min\{r\colon\ x_r = x_k^\lambda\}.$$

Such a numbering is clearly unique. Whenever λ does not belong to $f[x]$, the set x^λ is empty. For the special case $\lambda = \widehat{f}(x)$ (i.e., where λ is the maximal value of the fitness in the population x), we use the notation $\widehat{x} = \{\widehat{x}_1, \dots, \widehat{x}_{|\widehat{x}|}\}$ (i.e., we replace x^λ by $\widehat{x}$ in the preceding notation).

Finally, for x in E^m, we denote by $\lambda_1^x, \dots, \lambda_{|f[x]|}^x$ the $|f[x]|$ elements of the set $f[x]$ (i.e., the set of the fitness values observed in population x), where again the indexing respects the order of appearance of the elements of $f[x]$ in the sequence $(f(x_k),\ 1 \le k \le m)$, or equivalently

$$\forall\ h, k,\ 1 \le h < k \le |f[x]|, \qquad \min\{r\colon\ f(x_r) = \lambda_h^x\} < \min\{r\colon\ f(x_r) = \lambda_k^x\}.$$

When the context is unambiguous, we will drop the superscript in the above notation.

4. Example: the case $E = \{0, 1\}^N$. We specialize our general model to the case where the state space E is $\{0, 1\}^N$ ($N \in \mathbb{N}$). A point i of E is a word of length N over the alphabet $\{0, 1\}$ and is denoted $i = i_1 \cdots i_N$, where $i_k \in \{0, 1\}$. The Hamming distance $H(i, j)$ between two points i, j of E is the number of letters where i and j differ:

$$H(i, j) = \mathrm{card}\{k\colon\ 1 \le k \le m,\ i_k \ne j_k\}.$$

The mutation kernel α is defined by

$$\alpha(i, j) = \begin{cases} 0, & \text{if } H(i, j) > 1, \\ 1/N, & \text{if } H(i, j) = 1. \end{cases}$$

It is irreducible: the minimal number of transitions necessary to join two arbitrary points of E through the kernel α is N.

In order to build the crossover operator, we define now a cutting operator T_k for k in $\{0 \cdots N\}$; T_k maps $E \times E$ onto $E \times E$ and for i, j in E, we put $T_k(i, j) = (i', j')$, where

$$i' = i_1 \cdots i_k j_{k+1} \cdots j_N, \qquad j' = j_1 \cdots j_k i_{k+1} \cdots i_N.$$

Notice that T_0 only exchanges the two individuals of the mating pair, that is, $T_0(i, j) = (j, i)$, whereas T_N is the identity map over $E \times E$.

For any pairs (i, j) and (i', j') of $E \times E$, we put

$$C((i, j), (i', j')) = \mathrm{card}\{k\colon\ 1 \le k \le N,\ T_k(i, j) = (i', j')\}$$

and we define finally the crossover kernel β by

$$\beta((i, j), (i', j')) = \frac{C((i, j), (i', j')) + C((j, i), (i', j'))}{\sum_{(i'', j'') \in E \times E} C((i, j), (i'', j'')) + C((j, i), (i'', j''))}.$$

It is a straightforward matter to check that conditions (1), (2) and (3) are satisfied.

5. The unperturbed process (X_n^∞). We first describe the underlying process which drives the algorithm. When there is no random perturbation, the process under study is a Markov chain (X_n^∞) with state space E^m. The superscript ∞ reflects the fact that this process corresponds to the limit behavior of our model, when all perturbations vanish. The transition mechanism from X_n^∞ to X_{n+1}^∞ is decomposed in two stages:

$$X_n^\infty \xrightarrow{\text{crossover}} Z_n^\infty \xrightarrow{\text{selection}} X_{n+1}^\infty.$$

We describe now in detail the crossover and selection operators.

5.1. $X_n^\infty \to Z_n^\infty$: *crossover.* The phenomenon of crossover is modeled as a random operation on the couples formed by consecutive individuals of the population X_n^∞. This random operation is the one naturally associated with the crossover kernel β of the abstract fitness landscape (Definition 2.1). The transition probabilities from X_n^∞ to Z_n^∞ are

$$(4)\quad P(Z_n^\infty = z / X_n^\infty = x) = \delta_m(x_m, z_m) \prod_{1 \le k \le m/2} \beta((x_{2k-1}, x_{2k}), (z_{2k-1}, z_{2k})),$$

where $\delta_m(i, j) = \delta(i, j)$ if m is odd (the last individual of the population has no mating partner and remains unchanged after crossover) and $\delta_m(i, j) = 1$ if m is even.

Notice here a fundamental difference from the model studied in [6]: the crossover is now incorporated into the unperturbed process.

5.2. $Z_n^\infty \to X_{n+1}^\infty$: *selection.* We propose here an enhanced version of the selection mechanism used in the previous models [6, 7] which has the decisive advantage of preserving the diversity of the individuals present in the population. This mechanism may be described roughly as follows. Suppose $Z_n^\infty = z$. To build population $X_{n+1}^\infty = x$, we first select randomly with a uniform distribution a permutation σ of $\mathfrak{S}(|\widehat{z}|)$. We divide the set of indices $\{1, \dots, m\}$ in $|\widehat{z}| + 1$ parts of approximately the same size (around $\lfloor m/(|\widehat{z}| + 1)\rfloor$). The components of the rth part (for $1 \le r \le |\widehat{z}|$) are set equal to $\widehat{z}_{\sigma(r)}$. That is, roughly, for the indices k such that

$$1 + (r-1)\left\lfloor \frac{m}{|\widehat{z}|+1} \right\rfloor \le k \le r \left\lfloor \frac{m}{|\widehat{z}|+1} \right\rfloor,$$

we put $x_k = \widehat{z}_{\sigma(r)}$ (for $\lfloor\ \rfloor$, see the initial conventions). The components of the $(|\widehat{z}|+1)$th part are chosen independently with the uniform distribution on the set $\widehat{z}$.

We now explicate precisely this transition mechanism.

We first define a triangular array of integers $\tau(k, h)$, $0 \le k \le h+1$, $1 \le h \le m$, by

$$\forall\ h \in \{1, \dots, m\}, \qquad \tau(0, h) = 1, \qquad \tau(h+1, h) = m,$$

$$\forall\ k, h,\ 1 \le k \le h \le m, \qquad \tau(k, h) = 2k \left\lfloor \frac{m}{2(h+1)} \right\rfloor + 1.$$

Let x and z be two elements of E^m. If there exists a permutation σ of $\mathfrak{S}(|\widehat{z}|)$ such that

$$\forall\ h \in \{1, \dots, |\widehat{z}|\},\ \forall\ k,\ \tau(h-1, |\widehat{z}|) \le k < \tau(h, |\widehat{z}|), \qquad x_k = \widehat{z}_{\sigma(h)},$$

then

$$P\big(X_{n+1}^\infty = x \mid Z_n^\infty = z\big) = \frac{1}{|\widehat{z}|!} \prod_{\tau(|\widehat{z}|, |\widehat{z}|) \le k \le m} \frac{1_{\widehat{z}}(x_k)}{\operatorname{card} \widehat{z}}. \tag{5}$$

If no such permutation exists, then $P(X_{n+1}^\infty = x / Z_n^\infty = z) = 0$.

REMARK 1. In formula (5), the $|\widehat{z}|!$ stands for the choice of a random permutation belonging to $\mathfrak{S}(|\widehat{z}|)$, and the product corresponds to the choice of the components of x whose indices belong to the $(|\widehat{z}| + 1)$th part.

REMARK 2. The first $|\widehat{z}|$ parts of x have an even cardinality, so that the crossover can not act on a pair of individuals belonging to distinct parts. Since in addition each such part contains only one type of individual, condition (1) shows that the crossover operator will have no effect on the first $|\widehat{z}|$ parts of x. The main interest of the $(|\widehat{z}| + 1)$th part is to give the opportunity to distinct individuals of $|\widehat{z}|$ to mate without constraints.

REMARK 3. If $m < 2(|\widehat{z}| + 1)$, the first $|\widehat{z}|$ parts of x are empty, so that the components of x are chosen independently with uniform distribution on the set $\widehat{z}$. However, the dynamics of the algorithm becomes particularly interesting when m is large (as will be shown later), and such situations will not then occur.

6. The attractors of the chain (X_n^∞). Due to the selection mechanism, the populations generated by the Markov chain (X_n^∞) have a very specific structure. In fact, the Markov chain (X_n^∞) wanders through particular subsets of E^m which we call attractors. These subsets play the role of the attractors of the deterministic dynamical system in the Freidlin–Wentzell theory [10]. The aim of this section is to investigate the zoology of the attractors of the chain (X_n^∞) and to understand the dynamics of (X_n^∞) on these attractors.

DEFINITION 6.1. The attractors of the chain (X_n^∞) are the sets of populations K such that:

(i) $[K] = \widehat{K}$;

(ii) a population $x = (x_1, \dots, x_m)$ of E^m belongs to K if and only if

$$\forall\, r \in \{1 \cdots |\widehat{K}|\},\ \forall\, k, h, \qquad \tau(r-1, |\widehat{K}|) \le k, h < \tau(r, |\widehat{K}|) \ \Rightarrow\ x_k = x_h,$$

$$\{x_k\colon\ 1 \le k < \tau(|\widehat{K}|, |\widehat{K}|)\} = [K],$$

$$\{x_k\colon\ \tau(|\widehat{K}|, |\widehat{K}|) \le k \le m\} \subset [K].$$

The set of all the attractors is denoted by $\mathscr{K}$ [thus $\mathscr{K} \in \mathscr{P}(\mathscr{P}(E^m))$].

REMARK. Notice that property (ii) is an equivalence; that is, each population x satisfying the three conditions in (ii) has to belong to K whenever K is an attractor. Conversely, each population in K must fulfill these conditions.

Condition (i) implies that the populations of the attractors are equifitness populations (i.e., populations whose individuals all have the same fitness). More precisely, we have

$$\forall\, K \in \mathscr{K},\ \exists\, \lambda \in \mathbb{R}_+^*,\ \forall\, x \in K, \qquad [x] \subset f_\lambda.$$

That is, all the individuals belonging to the populations of a fixed attractor are in the same level set of f. We denote by $\mathscr{K}_\lambda$ (respectively, $\mathscr{K}_\lambda^+, \mathscr{K}_\lambda^-$) the set of attractors K such that $[K] \subset f_\lambda$ (respectively, $[K] \subset f_\lambda^+$, $[K] \subset f_\lambda^-$). For an attractor K, we denote by $f(K)$ the unique real number λ such that K belongs to $\mathscr{K}_\lambda$. We denote by $\mathscr{K}_*$ the attractors included in f^* (i.e., $\mathscr{K}_* = \mathscr{K}_{f(f^*)}$) and by K^* the unique attractor (it exists for $m \ge |f^*|$) such that $[K^*] = f^*$. We let also $\mathscr{K}_*^- = \mathscr{K}_{f(f^*)}^-$.

The transition mechanism implies that the process (X_n^∞) is instantaneously absorbed in the set of the populations which belong to attractors, that is,

$$\forall\, x \in E^m, \qquad P(\forall\, n \ge 1 \quad \exists\, K \in \mathscr{K} \quad X_n^\infty \in K \mid X_0^\infty = x) = 1.$$

Moreover, as a consequence of condition (2) on the crossover kernel β, we see that, for each attractor K and each population x in K, the probability

$P(X_{n+1}^\infty \in K/X_n^\infty = x)$ is positive so that the process has a nonzero probability of staying in an attractor.

We distinguish two kinds of attractors. An attractor K is said to be unstable if

$$\exists\, x \in K, \qquad P\left(X_{n+1}^\infty \in K \mid X_n^\infty = x\right) < 1.$$

An attractor K is said to be stable if

$$\forall\, x \in K, \qquad P\left(X_{n+1}^\infty \in K \mid X_n^\infty = x\right) = 1.$$

Notice that the bracket operator [ ] provides a one-to-one correspondence between the set $\mathscr{K}$ of all the attractors and the subset $\mathscr{E}$ of $\mathscr{P}(E)$ defined by

$$\mathscr{E} = \{F \subset E\colon\ |F| \le m,\ \widehat{F} = F\}.$$

In fact, if K is an attractor, for each x in K, we have $[x] = [K]$. That is, the bracket operator is constant over K and thus characterizes K. As a consequence, two distinct attractors do not intersect.

When m is small, the set $\mathscr{E}$ clearly depends on m, but for $m \ge \Lambda$, it does not depend on m any more, nor does the structure of the set of attractors $\mathscr{K}$; only the size of the populations changes, and the composition of the attractors is stabilized.

We use the crossover kernel β to build an operator on the set $\mathscr{P}(E)$. If F is a subset of E, we define

$$\beta(F) = \{i \in E\colon\ \exists\, (i', j', j) \in F \times F \times E,\ \beta((i', j'), (i, j)) > 0\}.$$

The operator $\widehat{\beta}$ is the composition of the operators β and the caret ($\widehat{\ }$). That is,

$$\widehat{\beta}(F) = \{i \in \beta(F)\colon\ f(i) = \widehat{f}(\beta(F))\}.$$

We have the following characterization of the stable attractors.

LEMMA 6.2. *An attractor K of $\mathscr{K}$ is stable if and only if $\widehat{\beta}([K]) = [K]$ and $m \ge 2(|\widehat{K}| + 1)$.*

PROOF. This is an immediate consequence of the transition mechanism of the process (X_n^∞). If the above conditions are satisfied, we have

$$P\left([X_{n+1}^\infty] = [K] \mid X_n^\infty \in K\right) = 1.$$

If $\widehat{\beta}([K]) \ne [K]$, there is a nonzero probability that the crossover creates new individuals not belonging to $[K]$ with a fitness greater than or equal to $\widehat{f}(K)$, so that X_{n+1}^∞ has a positive probability of leaving K. If $m < 2(|\widehat{K}| + 1)$, the selection mechanism does not guarantee the survival of all the individuals of $[K]$, so that

$$P\left([X_{n+1}^\infty] \subsetneq [K] \mid X_n^\infty \in K\right) > 0. \qquad \square$$

We define a partial relation $\prec_\infty$ on $\mathscr{K}$: for each pair K_1, K_2 of attractors, we have

$$K_1 \prec_\infty K_2 \quad \Leftrightarrow \quad \exists\, x \in K_1,\ \exists\, y \in K_2,\ \exists\, r \in \mathbb{N}, \quad P\left(X_{n+r}^\infty = y \mid X_n^\infty = x\right) > 0.$$

This relation is reflexive and transitive. In addition, the process X_n^∞ can leave an attractor only by creating with the crossover new individuals whose fitness is greater than or equal to the fitness of the starting attractor. Thus

$$K_1 \prec_\infty K_2 \quad \Rightarrow \quad [K_2] \subset \bigcup_{n=0}^{\infty} \underbrace{\widehat{\beta}(\cdots \widehat{\beta}(\widehat{\beta}}_{\widehat{\beta}, n \text{ times}}([K_1])))$$

and the fitness cannot decrease during such a transition:

$$K_1 \prec_\infty K_2, \qquad f(K_1) \neq f(K_2) \quad \Rightarrow \quad f(K_1) < f(K_2).$$

Suppose $m \geq 2(\Lambda + 1)$. The selection mechanism then never causes a loss of diversity within a level set of f so that

$$K_1 \prec_\infty K_2, \qquad f(K_1) = f(K_2) \quad \Rightarrow \quad [K_1] \subset [K_2]$$

and the relation $\prec_\infty$ is then a partial order on $\mathscr{K}$.

If x is a population of E^m which belongs to a (necessarily unique) attractor K, we put $K(x) = K$. It follows from the very definition of the relation $\prec_\infty$ that

$$\forall\, n \geq 1, \qquad K(X_n^\infty) \ \prec_\infty \ K(X_{n+1}^\infty);$$

that is, the sequence $(K(X_n^\infty))_{n\geq 1}$ is increasing in the ordered set $(\mathscr{K}, \prec_\infty)$. Since $\mathscr{K}$ is finite, this sequence is stationary; with probability 1, the limit

$$\lim_{n\to\infty} K(X_n^\infty) = K_\infty$$

is a maximal element of $\mathscr{K}$ for the order $\prec_\infty$, and yet the maximal elements of $\mathscr{K}$ are precisely the stable attractors. Finally, let K be a stable attractor. For any populations x, y belonging to K, the transition probability $P(X_{n+1}^\infty = y \mid X_n^\infty = x)$ is independent of x and y (it is completely determined by the set $[K]$). As a consequence, the process (X_n^∞) admits a unique invariant probability measure on the attractor K, which is the uniform distribution.

REMARK. The main role of the crossover is to make some attractors unstable. Suppose for instance we define the kernel β_0 by

$$\forall\, i_1, j_1, i_2, j_2 \in E, \qquad \beta_0((i_1, j_1), (i_2, j_2)) = \tfrac{1}{2}\delta(i_1, i_2)\delta(j_1, j_2) + \tfrac{1}{2}\delta(i_1, j_2)\delta(j_1, i_2)$$

so that the crossover can only exchange the individuals of the mating pair and never creates new individuals. The corresponding algorithm is then a mutation–selection algorithm. In this case, all the attractors are stable whenever $m \geq 2(|\widehat{K}| + 1)$.

A numerical example on the space $E = \{0,1\}^3$. We consider the space $E = \{0,1\}^3$ and we define the fitness function f by

$$\begin{aligned} f(\{001, 011\}) &= \{1\},\\ f(\{010, 101\}) &= \{2\},\\ f(\{100\}) &= \{3\},\\ f(\{000, 110, 111\}) &= \{4\}. \end{aligned}$$

We take $m = 10$. There are 15 attractors. If F is an equifitness subset of E, we denote by $K(F)$ the associated attractor. The attractors $K(010, 101)$ and $K(000, 111)$ are unstable. All other attractors are stable. Actually, we have

$$\begin{aligned} &K(000, 111) \prec_\infty K(000, 110, 111),\\ &K(010, 101) \prec_\infty K(100),\\ &K(010, 101) \prec_\infty K(110). \end{aligned}$$

The ideal attractor $K^* = K(000, 110, 111)$ contains $3!\,3^4 = 486$ populations. A typical example of such a population is

$$(110, 110, 000, 000, 111, 111, 000, 111, 110, 000).$$

7. The perturbed Markov chain (X_n^l). The three operators of a genetic algorithm play different roles: the mutation tends to disperse the population over the space E, the crossover helps the information to spread quickly over the population and the selection tends to concentrate the population on the current best individual. Our point of view is to consider the mutation and the selection as random perturbations of very crude operators: random perturbations of the identity map for the mutations and random perturbations of the very strong selection mechanism of the chain (X_n^∞) for the selection. Within this framework we are able to carry out an analysis of the asymptotic dynamics of the algorithm when the perturbations vanish. It is of course questionable whether this paradigm is relevant in practice. A major remaining issue is to obtain operational probability bounds for real optimization problems. There are several such attempts in this direction for sequential simulated annealing. Hajek and Sasaki study a maximum matching problem [15] and Jerrum studies a maximum clique problem [20]. These authors analyze the rate of growth of the time necessary to reach a solution with a given accuracy when the size of the problem tends to infinity. Jerrum and Sinclair succeed in building polynomial-time algorithms designed to approximate the partition function of the Ising model [21]. Lundy and Mees describe situations where convergence is exponentially long and others where termination occurs in polynomial time with a good practical confidence [22]. Sinclair develops techniques to handle the multicommodity flow [23]. Sorkin investigates simulated annealing on fractal energy landscapes [24]. These interesting results rely on features which depend strongly on the problem under study and cannot be generalized easily. They are aimed at finding efficient implementations on real problems.

Nevertheless our abstract setting for genetic algorithms yields general conditions ensuring the convergence of the genetic algorithm for any fitness landscape. We also establish the existence of critical population sizes. The paradigm we propose is somewhat different from the practical implementations of genetic algorithms, where the parameters controlling the mutation and the selection are kept constant in time. On the one hand, our analysis might be interpreted as the analysis of these temporally homogeneous genetic algorithms for small values of the mutation rate and a strong selection pressure. The asymptotic dynamics therefore gives a sharpened and simplified picture of the true dynamics. On the other hand, the scheme of decreasing perturbations is a new procedure which should be tested in practice to see whether it succeeds in accelerating the convergence (in the same way that simulated annealing is an attempt to speed up the Metropolis algorithm).

We now describe the perturbed transition mechanism precisely. The previous Markov chain (X_n^∞) is randomly perturbed by two distinct mechanisms. The first one acts directly upon the population and mimics the phenomenon of mutation. The second one consists in loosening the selection of the individuals. The intensity of the perturbations is governed by an integer parameter l; as l goes to infinity, the perturbations progressively disappear. The transition mechanism of the perturbed Markov chain (X_n^l) is decomposed in three stages:

$$X_n^l \xrightarrow{\text{mutation}} Y_n^l \xrightarrow{\text{crossover}} Z_n^l \xrightarrow{\text{selection}} X_{n+1}^l.$$

7.1. $X_n^l \to Y_n^l$: *mutation.* The mutations are modeled by random independent perturbations of the individuals of the population X_n^l. These random perturbations are built with the help of the mutation kernel α of the abstract fitness landscape (Definition 2.1). Let a be a positive real number, which we call the mutation cost. Define

$$\alpha_l(i, j) = \begin{cases} \alpha(i, j)l^{-a}, & \text{if } i \neq j, \\ 1 - \sum_{e \neq i} \alpha(i, e)l^{-a}, & \text{if } i = j, \end{cases}$$

so that α_l is an irreducible Markov kernel on E.

The transition probabilities from X_n^l to Y_n^l are given by

$$P(Y_n^l = y \mid X_n^l = x) = \alpha_l(x_1, y_1) \cdots \alpha_l(x_m, y_m).$$

Note that

$$\lim_{l \to \infty} P(Y_n^l = y \mid X_n^l = x) = \delta(x_1, y_1) \cdots \delta(x_m, y_m); \tag{6}$$

that is, the mutations vanish when l goes to infinity.

7.2. $Y_n^l \to Z_n^l$: *crossover.* The crossover is not perturbed in any way: this stage is exactly the same as the passage from X_n^∞ to Z_n^∞ [formula (4)]. We define the crossover kernel β on $E^m \times E^m$ by

$$\forall\ y, z \in E^m, \qquad \beta(y, z) = P(Z_n^l = z \mid Y_n^l = y) = P(Z_n^\infty = z \mid X_n^\infty = y).$$

7.3. $Z_n^l \to X_{n+1}^l$: *selection.* We first describe the selection mechanism informally.

Suppose $Z_n^l = z$ and we wish to build the vector $X_{n+1}^l = x$. We first traverse $f[z]$; with each element λ of this set, we associate a sequence $\psi_1, \dots, \psi_{n_\lambda}$ obtained by reordering randomly (all orders being equally probable) the set $f^{-1}(\{\lambda\}) \cap f[z]$. The population x is then built in the following way: for each component x_k, $1 \le k \le m$, we draw a value λ under a distribution probability on the set $f[z]$ which is biased toward the high values. As l goes to infinity, this distribution concentrates on the value $\widehat{f}(z)$. With this value λ, we had previously associated a sequence $\psi_1, \dots, \psi_{n_\lambda}$. We divide the set $\{1 \cdots m\}$ into $n_\lambda + 1$ parts. If the index k under consideration belongs to the rth part, where $1 \le r \le n_\lambda$, we set $x_k = \psi_r$. If k belongs to the $(n_\lambda + 1)$th part, we choose x_k randomly and uniformly over the set $\{\psi_1, \dots, \psi_{n_\lambda}\}$.

We now describe this mechanism precisely. Suppose always $Z_n^l = z$. Let us recall some notation. The set $f[z]$ contains $|f[z]|$ distinct values $\lambda_1^z, \dots, \lambda_{|f[z]|}^z$. Throughout this section, when the context is unambiguous, we will drop the superscript z in this notation, so that λ_k will stand for λ_k^z. For each λ in $f[z]$, there are $|z^\lambda|$ distinct individuals in z whose fitness is equal to λ:

$$z^\lambda = \{z_1^\lambda, \dots, z_{|z^\lambda|}^\lambda\}.$$

The vector $X_{n+1}^l = x$ is built in the following way. For each h in $\{1 \cdots |f[z]|\}$, we select independently and randomly with the uniform distribution a permutation σ^h belonging to $\mathfrak{S}(|z^{\lambda_h}|) = \mathfrak{S}(|[z] \cap f_{\lambda_h}|)$. The law of each component x_k $1 \le k \le m$ of x is defined as follows: we randomly select a value λ in the set $f[z]$ with the distribution

$$\forall\ h \in \{1 \cdots |f[z]|\}, \qquad P(\lambda = \lambda_h) = \frac{\exp(c\lambda_h \ln l)}{\sum_{r=1}^{|f[z]|} \exp(c\lambda_r \ln l)},$$

where c is a positive real number, which we call the scaling parameter.

The value of x_k is then chosen in the set z^λ according to the value of the index k:

1. If $\tau(|z^\lambda|, |z^\lambda|) \le k \le \tau(|z^\lambda| + 1, |z^\lambda|) = m$, then x_k is chosen at random with the uniform distribution over the set $z^\lambda = [z] \cap f_\lambda$ (k lies in the last part of the set of indices).
2. If there exists an index r in $\{1 \cdots |z^\lambda|\}$ such that $\tau(r-1, |z^\lambda|) \le k < \tau(r, |z^\lambda|)$, then we put $x_k = z_{\sigma^h(r)}^\lambda$, where h is the unique integer in $\{1 \cdots |z^\lambda|\}$ satisfying $\lambda = \lambda_h$ (the index k lies in the rth part of the set of indices, where $1 \le r \le |z^\lambda|$).

REMARK. The selection mechanism at work in the chain (X_n^∞) is obtained as a particular case. The probability distribution on $f[z]$ is degenerate and assigns mass 1 to $\widehat{f}(z)$.

We give now an explicit expression for the transition probabilities. Let x and z belong to E^m. If $[x] \not\subset [z]$, we put $\sigma(z, x) = \varnothing$. Suppose $[x] \subset [z]$. Let

$\sigma(z, x)$ be the set of the $|f[z]|$-tuples of permutations

$$(\sigma^1, \dots, \sigma^{|f[z]|}) \in \mathfrak{S}(|z^{\lambda_1}|) \times \cdots \times \mathfrak{S}(|z^{\lambda_{|f[z]|}}|)$$

satisfying the following property:

> For each k in $\{1 \cdots m\}$, if h is the unique integer in $\{1 \cdots |f[z]|\}$ such that $\lambda_h = f(x_k)$, and if for some r in $\{1 \cdots |z^{\lambda_h}|\}$ we have $\tau(r-1, |z^{\lambda_h}|) \le k < \tau(r, |z^{\lambda_h}|)$, then $x_k = z^{\lambda_h}_{\sigma^h(r)}$.

The transition probabilities from Z_n^l to X_{n+1}^l are then given by the intuitive formula

$$\begin{aligned} &P\big(X_{n+1}^l = x \mid Z_n^l = z\big) \\ &\quad = \frac{|\sigma(z, x)|}{\prod_{\lambda \in f[z]} |z^\lambda|!} \bigg(\prod_{\lambda \in f[z]} \prod_{k=\tau(|z^\lambda|, |z^\lambda|)}^{m} \bigg(\frac{1_{z^\lambda}(x_k)}{|z^\lambda|} + 1 - 1_{z^\lambda}(x_k) \bigg) \bigg) \\ &\qquad \times \prod_{k=1}^{m} \frac{\exp(cf(x_k)\ln l)}{\sum_{\lambda \in f[z]} \exp(c\lambda \ln l)}. \end{aligned} \tag{7}$$

We define the kernel γ on $E^m \times E^m$ by the identity

$$P\big(X_{n+1}^l = x \mid Z_n^l = z\big) = \gamma(z, x) \prod_{k=1}^{m} \frac{\exp(cf(x_k)\ln l)}{\sum_{\lambda \in f[z]} \exp(c\lambda \ln l)}.$$

Only the last product in formula (7) depends upon l. It may be rewritten as

$$\frac{\exp(c \sum_{k=1}^{m} f(x_k) \ln l)}{\big(\sum_{\lambda \in f[z]} \exp(c\lambda \ln l)\big)^m}.$$

As l goes to infinity, this term is equivalent to

$$\exp\bigg(-c \bigg(m\widehat{f}(z) - \sum_{k=1}^{m} f(x_k) \bigg) \ln l \bigg),$$

which tends to zero whenever $[x]$ is not included in $\widehat{z}$. Thus

$$\lim_{l \to \infty} P\big(X_{n+1}^l = x \mid Z_n^l = z\big) = \frac{|\sigma(z, x)|}{\prod_{\lambda \in f[z]} |z^\lambda|!} \bigg(\prod_{k=1}^{m} 1_{\widehat{z}}(x_k) \bigg) |\widehat{z}|^{\tau(|\widehat{z}|, |\widehat{z}|) - 1 - m}.$$

Yet, for x such that $[x] \subset \widehat{z}$, we have either $\sigma(z, x) = \varnothing$ and $P(X_{n+1}^l = x / Z_n^l = z) = 0$, or

$$|\sigma(z, x)| = \prod_{\lambda \in f[z] \setminus \{\widehat{f}(z)\}} |z^\lambda|!,$$

in which case [see formula (5)]

$$\lim_{l \to \infty} P\big(X_{n+1}^l = x \mid Z_n^l = z\big) = \frac{|\widehat{z}|^{\tau(|\widehat{z}|, |\widehat{z}|) - 1 - m}}{|\widehat{z}|!} = P\big(X_{n+1}^\infty = x \mid Z_n^\infty = z\big); \tag{8}$$

that is, the limiting selection mechanism is the one used for the Markov chain (X_n^∞). Formulas (6) and (8) yield

$$\forall\ y, z \in E^m, \qquad \lim_{l\to\infty} P\big(X_{n+1}^l = z \mid X_n^l = y\big) = P\big(X_{n+1}^\infty = z \mid X_n^\infty = y\big)$$

so that the process (X_n^l) is a perturbation of the process (X_n^∞). A crucial point is that the perturbations really interact as l goes to infinity. More precisely, the rates of convergence in (6) and (8) should be logarithmically of the same order.

8. Convergence of the genetic algorithm. We state now our main results. The proofs are deferred to the remaining sections of the paper. Several critical quantities appear in the statements, such as the critical population size m^* and the critical heights H_1 and H_e^*. Explicit bounds on these quantities are obtained in the proofs of the results; however, these bounds involve some intricate constants associated with the abstract fitness landscape. For the sake of clarity, we state the results without introducing these constants. Let us point out that the proofs yield also a lot of information concerning the structure of the most probable trajectories of the process.

The first important result deals with the concentration of the equilibrium law of the algorithm on the ideal attractor K^*.

THEOREM 8.1 (Critical population size and limiting distribution). *There exists a critical population size m^* depending on the fitness landscape (E, f, α, β), the mutation cost a and the scaling parameter c, such that, when the population size m of the algorithm is greater than m^*, the limit of the sequence of the stationary measures of the Markov chains (X_n^l), $l \in \mathbb{N}$, as l goes to infinity, is the uniform distribution over the ideal attractor K^*; that is,*

$$\forall\ m \geq m^*,\ \forall\ x, y \in E^m \times K^*, \qquad \lim_{l\to\infty}\lim_{n\to\infty} P\big(X_n^l = y \mid X_0^l = x\big) = \frac{1}{|K^*|}.$$

We next state some key facts concerning the cycle decomposition when the population size is large.

THEOREM 8.2 (Structure of the cycles). *There exists another critical population size M^* such that, when the population size m of the algorithm is greater than M^*, each cycle over the set of attractors $\mathscr{K}$ not containing the attractor K^* is reduced to one single attractor K.*

In fact, for m large enough, the set of attractors does not depend on m any more (the attractors are in one-to-one correspondence with particular subsets of E; see Section 6). The trace of the limiting dynamics on the set of attractors stabilizes for m large and does not depend on m any more. This could be proved using the same technique as in [7], Section 7.

REMARK. Once more, this limiting structure is obtained as soon as the population size m is greater than a critical value, the other parameters being fixed. This structure of cycles is the most favorable one; the bad cycles, that is, those which slow down the convergence, are reduced to one single attractor.

In the next results, we suppose that the population size m is larger than m^*. Once we know that the sequence of the stationary measures of the Markov chains (X_n^l), $l \in \mathbb{N}$, concentrates on the ideal attractor K^*, the remaining problem is to build a process with the same limiting law. From now on, we consider the inhomogeneous algorithm; that is, the control parameter l is an increasing function of n and we deal with an inhomogeneous Markov chain $(X_n^{l(n)})_{n\in\mathbb{N}}$. We will suppress the superscript l in most notation. The challenging problem is to adapt the sequence $(l(n))_{n\in\mathbb{N}}$ in order to have

$$\forall\ x \in E^m, \qquad \lim_{n\to\infty} P(X_n \in K^* \mid X_0 = x) = 1$$

$[l(n)$ should not increase too fast$]$ and, simultaneously, to obtain the best rate of convergence $[l(n)$ should not increase too slowly$]$. The answer to this now classical problem is given by Catoni and Trouvé's results [2, 5, 28, 29].

The critical heights H_1 and H_e^*. For the definition and the properties of the height of exit H_e, we refer the reader to Trouvé's work [25, 26, 28, 29]. The crucial constant for the convergence of the algorithm is the critical height H_1 defined by

$$H_1 = \sup\{H_e(\pi)\colon\ \pi \text{ cycle not containing } K^*\}.$$

The rate of escape from the basin of attraction of K^* is

$$H_e^* = H_e(\{x \in E^m\colon\ f^* \subset [x]\}).$$

PROPOSITION 8.3. *The critical height H_1 is bounded as a function of m. The height H_e^* is greater than an affine function of m.*

The proof is in the Appendix.

We now restate in our context Trouvé's convergence result ([28], Theorem 2.22), which is an extension of a result by Hajek [13] for simulated annealing.

THEOREM 8.4. *Suppose m is larger than m^*. For all increasing sequences $l(n)$ going to infinity, we have the equivalence*

$$\sup_{x\in E^m} P(X_n \notin K^*/X_0 = x) \to 0 \quad \textit{as } n \to \infty \ \Leftrightarrow\ \sum_{n=0}^{\infty} l(n)^{-H_1} = \infty.$$

Furthermore, we may adapt the sequence $l(n)$ in order to be trapped in the basin of attraction of K^* after a finite number of transitions.

THEOREM 8.5. *Suppose m is large enough to have $m \geq m^*$ and $H_1 < H_e^*$. For all increasing sequences $l(n)$, we have the equivalence*

$$\forall\ x \in E^m, \qquad P(\exists\ N \quad \forall\ n \geq N \quad f^* \subset [X_n] \mid X_0 = x) = 1$$
$$\Leftrightarrow \quad \sum_{n=0}^{\infty} l(n)^{-H_1} = \infty, \qquad \sum_{n=0}^{\infty} l(n)^{-H_e^*} < \infty.$$

REMARK. Proposition 8.3 implies that for m large enough, we have $H_1 < H_e^*$ so that sequences $l(n)$ with the desired properties do exist.

The optimal rate of convergence. For the meaning and the properties of the optimal convergence exponent, we refer the reader to [2–5, 25–29]. We now restate Trouvé's result for the optimal convergence rate, which generalizes Catoni's work.

THEOREM 8.6. *There exist two strictly positive constants R_1 and R_2 such that for all $m \geq m^*$ and all n,*

$$\frac{R_1}{n^{\alpha_{\text{opt}}}} \leq \inf_{0 \leq l(1) \leq \cdots \leq l(n)} \max_{x \in E^m} P\big(X_n \notin K^* \mid X_0 = x\big) \leq \frac{R_2}{n^{\alpha_{\text{opt}}}}.$$

PROPOSITION 8.7. *The optimal convergence exponent α_{opt} is bounded between two affine strictly increasing functions of m. That is, we have*

$$0 < \liminf_{m \to \infty} \frac{\alpha_{\text{opt}}}{m} \leq \limsup_{m \to \infty} \frac{\alpha_{\text{opt}}}{m} < \infty.$$

The proof is in the Appendix.

The fact that the optimal convergence exponent α_{opt} increases linearly with m shows that our genetic algorithm is intrinsically parallel; it involves mostly local independent computations. This nice feature will be further discussed in the next section, where we compare the parallel (independent) simulated annealing with the genetic algorithm on a very simple example.

9. The role of the crossover. The crossover operator is integrated into the underlying process which drives the dynamics of the genetic algorithm. It is therefore essential to impose some stability conditions on it to make convergence to the global maxima possible for every fitness function. Our condition (1),

$$\forall\ i \in E, \qquad \beta((i,i),(i,i)) = 1,$$

is the simplest form of such a requirement.

PROPOSITION 9.1. *Suppose there exist i and e such that $i \neq e$ and*

$$\beta((i,i),(e,e)) > 0,$$
$$\forall\ j \in E \setminus \{i\}, \qquad \beta((j,j),(j,j)) = 1.$$

Then there exists a fitness function f such that, whatever the population size $m \geq 2$, the mutation cost a and the scaling factor c,

$$\forall\, x \in E^m, \qquad \lim_{l\to\infty} \lim_{n\to\infty} P\big([X_n^l] \subset f^* \mid X_0^l = x\big) = 0.$$

The proof is in the Appendix.

Instead of imposing a general condition like (1), one could impose a condition depending on a particular fitness function f. For instance, convergence can be ensured if

$$\forall\, i \in f^*, \qquad \beta((i,i),(i,i)) = 1$$

or even if $\beta(f^*) = f^*$ [the notation $\beta(f^*)$ is explained before Lemma 6.2].

Condition (2),

$$\forall\, i, j \in E, \qquad \beta((i,j),(i,j)) > 0,$$

is not as crucial as the first one. However, it makes the analysis of the set of attractors somewhat simpler. Without it, one should impose a further restriction in the definition of attractors (Definition 6.1) so that

$$\forall\, x \in K, \qquad P(X_{n+1}^\infty \in K \mid X_n^\infty = x) > 0;$$

otherwise the process might escape instantaneously from K with probability 1. An example of such a condition is

$$\forall\, x \in K,\ \forall\, i, j \in [x], \qquad \beta((i,j),(i,j)) > 0.$$

Condition (3),

$$\forall\, i_1, j_1, i_2, j_2 \in E, \qquad \beta((i_1,j_1),(i_2,j_2)) = \beta((j_1,i_1),(j_2,i_2)),$$

is superfluous and can be removed without affecting the analysis. However, it is a natural condition of symmetry.

An important issue is to understand the impact of the crossover on the speed of convergence of the algorithm. A natural direction for future research is to find how one should implement efficiently the crossover operator in order to increase significantly the optimal convergence exponent for a specific class of fitness functions. Let us see what happens on a small numerical example.

We consider the space $E = \{0,1\}^4$ endowed with the mutation and crossover kernels (α, β) defined in Section 4. We define the fitness function f by

$$\begin{aligned} f(\{0000\}) &= \{1000\},\\ f(\{1000, 0100, 0010, 0001\}) &= \{900\},\\ f(\{1100, 1010, 1001, 0110, 0101, 0011\}) &= \{800\},\\ f(\{1110, 1101, 1011, 0111\}) &= \{0\},\\ f(\{1111\}) &= \{1100\} \end{aligned}$$

so that f is a function of the number of digits of the individual equal to 1. Let β_0 be the identity crossover kernel. That is,

$$\forall\, i_1, j_1, i_2, j_2 \in E, \qquad \beta_0((i_1, j_1), (i_2, j_2)) = \tfrac{1}{2}\delta(i_1, i_2)\delta(j_1, j_2) + \tfrac{1}{2}\delta(i_1, j_2)\delta(j_1, i_2).$$

The algorithm with β_0 corresponds to a genetic algorithm without crossover, that is, a mutation–selection algorithm. Let us denote by $H_1(\beta_0)$ and $\alpha_{\mathrm{opt}}(\beta_0)$ the critical height and the optimal exponent associated with the algorithm running with the crossover β_0.

To compute H_1 we look at the trajectories of minimal cost which start from the attractor $K(0000)$ and end with the attractor $K(1111)$. With β_0 (i.e., no crossover) and a sufficiently large population size, the best way is to let an individual follow a mutation path such as

$$0000 \to 0100 \to 0101 \to 1101 \to 1111,$$

while all other individuals remain in 0000. Whenever this explorer reaches 1111, the whole population jumps to 1111. This trajectory requires four mutations and an antiselection cost of $c(3f(0000) - f(0100) - f(0101) - f(1101))$, so that its global cost is $H_1(\beta_0) = 4a + 1300c$. When a crossover mechanism is available, it is possible to build more efficient paths which avoid the difficult saddles. In our example, a good way to avoid the antiselection of an individual containing three digits equal to 1 (like 1101) is to let two individuals mutate from 0000 to 1100 and 0011 simultaneously and then to perform a crossover between them:

$$(0000, 0000) \to (1000, 0010) \to (1100, 0011) \to (1111, 0000).$$

This requires four mutations and an antiselection cost of $c(2f(0000) - f(1000) - f(0010))$ so that the global cost is $H_1(\beta) = 4a + 200c$, which is much less than $H_1(\beta_0)$.

We finally examine the rate of increase of the optimal convergence exponent when the population size is sufficiently large. We obtain analogously

$$\lim_{m\to\infty} \frac{\alpha_{\mathrm{opt}}(\beta)}{m} = \frac{\min(a, 100c)}{4a + 200c} > \lim_{m\to\infty} \frac{\alpha_{\mathrm{opt}}(\beta_0)}{m} = \frac{\min(a, 100c)}{4a + 1300c}.$$

Consider now m independent simulated annealing algorithms running over this fitness landscape (the moves of the particle are determined by the mutation kernel α and the simulated annealing does not make use of the crossover mechanism). We keep track of the best point found by the m algorithms. The optimal convergence exponent of this process is $m/10$. Trouvé proves that in general the optimal convergence exponent for parallel annealing based on periodically interacting searches is always worse than for independent multiple searches [27]. Introducing an interaction between simulated annealing algorithms may therefore damage the speed of convergence. We see that, for large populations, the optimal exponent of m independent simulated annealing is always better than the optimal exponent $\alpha_{\mathrm{opt}}(\beta_0)$ of the genetic algorithm without crossover. However, the situation changes radically whenever a crossover mechanism is available. Indeed, if we choose for instance $a = 100c$, we obtain $\alpha_{\mathrm{opt}}(\beta) \sim m/6$ which outperforms the m independent simulated annealing algorithms. An appropriate cooperation mechanism can therefore enhance significantly the speed of convergence.

10. Asymptotic expansion of $P(X^l_{n+1} = v \mid X^l_n = u)$. The aim of this section is to show how our model fits into the framework of the generalized simulated annealing. What we have to do is to study the asymptotic behavior of the transition matrix of (X^l_n) as l goes to infinity. By the very construction of the process (X^l_n), we have

$$\begin{aligned} &P(X^l_{n+1} = v \mid X^l_n = u) \\ &\quad = \sum_{y,z\in E^m} P(X^l_{n+1} = v \mid Z^l_n = z)P(Z^l_n = z \mid Y^l_n = y)P(Y^l_n = y \mid X^l_n = u). \end{aligned}$$

For each y, z in E^m,

$$\begin{aligned} &P(X^l_{n+1} = v \mid Z^l_n = z)P(Z^l_n = z \mid Y^l_n = y)P(Y^l_n = y \mid X^l_n = u) \\ &\quad \sim \alpha(u, y)\,\beta(y, z)\,\gamma(z, v) \\ &\qquad \times \exp\biggl(-\biggl(ad(u, y) + c\sum_{k=1}^{m}(\widehat{f}(z) - f(v_k))\biggr)\ln l\biggr) \quad \text{as } l \to \infty, \end{aligned} \tag{9}$$

where we note for u, y in E^m,

$$\alpha(u, y) = \prod_{k:\ u_k \neq y_k} \alpha(u_k, y_k)$$

and $d(u, y)$ is the Hamming distance between the vectors u and y. That is,

$$d(u, y) = \operatorname{card}\{k:\ 1 \le k \le m,\ u_k \neq y_k\}.$$

The above quantity (9) vanishes whenever $\alpha(u, y)\,\beta(y, z)\,\gamma(z, v) = 0$.

Let $\overline{D}_1(u, v)$ be the set of all four-tuples (u, y, z, v) satisfying $\alpha(u, y) \times \beta(y, z)\,\gamma(z, v) > 0$.

We define next the communication cost V_1 on $E^m \times E^m$ by

$$V_1(u, v) = \min_{(u,y,z,v)\in\overline{D}_1(u,v)} ad(u, y) + c\sum_{k=1}^{m}(\widehat{f}(z) - f(v_k))$$

and we denote by $\overline{D}^*_1(u, v)$ the elements of $\overline{D}_1(u, v)$ which realize the above minimum. Putting

$$q_1(u, v) = \sum_{(u,y,z,v)\in\overline{D}^*_1(u,v)} \alpha(u, y)\,\beta(y, z)\,\gamma(z, v),$$

we have

$$P(X^l_{n+1} = v \mid X^l_n = u) \sim q_1(u, v) \exp(-V_1(u, v) \ln l) \quad \text{as } l \to \infty.$$

Moreover, notice that for each u, v in E^m,

$$P(X^l_{n+1} = v \mid X^l_n = u) = 0 \;\Leftrightarrow\; V_1(u, v) = \infty \;\Leftrightarrow\; q_1(u, v) = 0.$$

We are now in the framework of the generalized simulated annealing studied by Trouvé [25–29]; that is, the transition probabilities of the process (X^l_n) form a family of Markov kernels on the space E^m indexed by l which is admissible for the communication kernel q_1 and the cost function V_1 ([28], Definition 2.1).

11. The paths and their costs. If $\mathcal{S}$ is an arbitrary set, $\mathcal{S}^{(\mathbb{N})}$ denotes the set of paths in $\mathcal{S}$, that is, the set of finite sequences of elements of $\mathcal{S}$. A path s in $\mathcal{S}$ is denoted indifferently,

$$s = (s_1, \dots, s_r), \qquad s = (s^1, \dots, s^r) \quad \text{or} \quad s = s_1 \to \cdots \to s_r,$$

and its length is denoted $|s|$ (r in the above example). A path s in $\mathcal{S}$ is said to join two elements t_1 and t_2 if $s_1 = t_1$ and $s_{|s|} = t_2$; the set of all paths in $\mathcal{S}$ joining the points t_1 and t_2 is denoted $\mathcal{S}^{(\mathbb{N})}(t_1, t_2)$.

We will consider paths in the sets E, E^m and $\mathcal{P}(E)$. Paths in E^m will mostly be denoted by the letter p and paths in $\mathcal{P}(E)$ by the letter q.

By D^m we denote the paths in E^m which correspond to possible trajectories of the process (X^l_n), that is, the paths p in E^m satisfying

$$\forall\, k,\; 1 \le k < |p|, \qquad V_1(p_k, p_{k+1}) < \infty.$$

The V_1 cost of such a path is

$$V_1(p) = \sum_{k=1}^{|p|-1} V_1(p_k, p_{k+1}).$$

Notice that for the empty path (which has a null length), the cost is zero.

If p belongs to $E^{m(\mathbb{N})} \setminus D^m$, we put $V_1(p) = \infty$. Similarly, by $\overline{D}^m$ we denote the paths in E^m which correspond to possible trajectories for the whole process (the number of transitions r being variable)

$$X^l_n \to Y^l_n \to Z^l_n \to X^l_{n+1} \to Y^l_{n+1} \to \cdots \to Z^l_{n+r-1} \to X^l_{n+r};$$

that is, such a path p includes the intermediate populations Y^l_{n+k} and Z^l_{n+k}, $0 \le k < r$, has a length equal to $1 \bmod 3$ and satisfies

$$\forall\, k,\; 1 \le 3k < |p|, \qquad \alpha(p^{3k-2}, p^{3k-1})\beta(p^{3k-1}, p^{3k})\,\gamma(p^{3k}, p^{3k+1}) > 0.$$

The corresponding cost function $\overline{V}$ is defined by

$$\overline{V}(p) = \sum_{1 \le 3k < |p|} \left[a d(p^{3k-2}, p^{3k-1}) + c \sum_{h=1}^{m} \left(\widehat{f}(p^{3k}) - f(p^{3k+1}_h) \right) \right]$$

if the path p belongs to $\overline{D}^m$ (here p_h^{3k+1} is the hth component of the vector p^{3k+1}) and $\overline{V}(p) = \infty$ otherwise. We put also, for y, z in E^m,

$$D^m(y,z) = D^m \cap E^{m(\mathbb{N})}(y,z), \qquad D = \bigcup_{m\in\mathbb{N}^*} D^m,$$

and we define similarly $\overline{D}^m(y,z), \overline{D}$ (just replace D by $\overline{D}$ in the above formulas).

The bracket operator [ ] provides a natural projection from the set $\bigcup_{m\in\mathbb{N}^*} E^{m(\mathbb{N})}$ onto $\mathscr{P}(E)^{(\mathbb{N})}$; with each path $p = (p_1, \dots, p_r)$ in E^m we associate the path $[p] = ([p_1], \dots, [p_r])$ in $\mathscr{P}(E)$. Finally, we put $[\overline{D}] = \{[p]\colon\ p \in \overline{D}\}$.

12. The minimal communication cost and the virtual energy.

DEFINITION 12.1. We define the minimal communication cost V for y and z in E^m by

$$V(y,z) = \inf\{V_1(p)\colon\ p \in D^m(y,z)\}.$$

Notice that, for all x in E^m, we have $V(x,x) = 0$.

If F, G are two subsets of E^m, we define

$$\begin{aligned} V(F,G) &= \inf\{V_1(p)\colon\ p \in D^m,\ p^1 \in F,\ p^{|p|} \in G\} \\ &= \inf\{V(y,z)\colon\ y \in F,\ z \in G\}. \end{aligned}$$

Let g be a graph on E^m. The cost of g is

$$V(g) = \sum_{(x\to y)\in g} V(x,y).$$

This definition works also for a path p in E^m if we consider the path as a graph over E^m. Notice that for the empty graph (which has no arrows), the cost is zero.

We recall that an x-graph is a graph with no arrow starting from x and such that for any $y \neq x$ there exists a unique path in g leading from y to x. The set of all x-graphs is denoted by $G(x)$. For more details and notation concerning graphs, see [10], Chapter 6.

DEFINITION 12.2. The virtual energy W associated with the cost function V is defined by

$$\forall\ x \in E^m, \qquad W(x) = \min\{V(g)\colon\ g \in G(x)\}.$$

We put also for any subset F of E^m,

$$W(F) = \min\{W(x)\colon\ x \in F\}, \qquad W^* = \{x \in E^m\colon\ W(x) = W(E^m)\}.$$

Proposition 5.4 of [7] shows that this quantity could equivalently be defined through the cost function V_1 instead of V. The point is that the sequence of the stationary measures of the Markov chains (X_n^l), $l \in \mathbb{N}$, concentrates on the set W^* as l goes to infinity.

PROPOSITION 12.3 (Freidlin and Wentzell).

$$\forall\, x \in E^m, \qquad \lim_{l\to\infty} \lim_{n\to\infty} P\big(X_n^l \in W^* \mid X_0^l = x\big) = 1.$$

We face here the same difficulty as in the case of the mutation–selection algorithm: the size of the state space increases geometrically with the population size m. Anyway, we are mostly interested in the way the chain (X_n^l) visits the populations of the attractors $\mathscr{K}$. We use the same technique as in [7]. Since

$$\forall\, x \in E^m,\ \exists\, K \in \mathscr{K}, \qquad V(x, K) = 0,$$

then Theorems 5.8 and 7.3 of [7] show that the dynamics of the process may be studied by considering only the restrictions of the communication cost V and the virtual energy W to the set of the populations belonging to the attractors. Furthermore, we have

$$\forall\, K \in \mathscr{K},\ \forall\, x, y \in K, \qquad V(x, y) = 0$$

so that

$$\forall\, K_1, K_2 \in \mathscr{K},\ \forall\, x_1 \in K_1,\ \forall\, x_2 \in K_2, \qquad V(x_1, x_2) = V(K_1, K_2)$$

and similarly

$$\forall\, K \in \mathscr{K},\ \forall\, x \in K, \qquad W(x) = W(K).$$

The functionals W and V may thus be seen as defined on the set of attractors $\mathscr{K}$ rather than on the set of populations belonging to attractors.

For the sake of completeness, we recall also the definition of the communication altitude, as well as the construction of the hierarchy of cycles. These tools will only be used for Theorem 8.2. For the details, we refer the reader to [26, 28] as well as to [8], Section 4, where they are used to compute spectral estimates.

DEFINITION 12.4 ([28], Definition 2.15). The communication altitude $A(K_1, K_2)$ between two distinct attractors K_1 and K_2 is

$$A(K_1, K_2) = \inf\Big\{ \max_{1\le k<|p|} W(p_k) + V(p_k, p_{k+1})\colon\ p \in D^m,\ p^1 \in K_1,\ p^{|p|} \in K_2 \Big\}.$$

For any K in $\mathscr{K}$, we put $A(K, K) = W(K)$.

DEFINITION 12.5. Let $\lambda \in \mathbb{R}$. We define an equivalence relation $\mathscr{R}_\lambda$ on the set

$$W_\lambda = \{K \in \mathscr{K}\colon\ W(K) \le \lambda\}$$

by

$$\forall\, K_1, K_2 \in W_\lambda, \qquad K_1 \mathscr{R}_\lambda K_2 \iff A(K_1, K_2) \le \lambda.$$

PROPOSITION 12.6 (Trouvé [28], Proposition 2.20). *The set of cycles in $\mathscr{K}$ associated with the cost function V is*

$$\mathscr{C}(\mathscr{K}) = \bigcup_{\lambda\in\mathbb{R}^+} W_\lambda \,/\, \mathscr{R}_\lambda,$$

where $W_\lambda \,/\, \mathscr{R}_\lambda$ is the quotient set of the equivalence classes of W_λ for the relation $\mathscr{R}_\lambda$.

We will now study the costs of the transitions between the attractors. As it turns out, the costs of the bad transitions increase linearly with the population size m, whereas the costs of the good transitions remain bounded (as functions of m).

13. The costs of the bad transitions. We distinguish two kinds of bad transitions: those which lose some peak fitness individuals and those which decrease fitness.

LEMMA 13.1 (Loss of diversity in f^*). *Let K_1, K_2 be two elements of $\mathscr{K}$ such that $[K_1] \subset f^*$, $[K_2] \subset f^*$ (i.e., K_1, $K_2 \in \mathscr{K}_*$). We have*

$$V(K_1, K_2) \ge \left\lfloor \frac{m}{2(|f^*|+1)} \right\rfloor \min(a, c\delta^*)\,\mathrm{card}\big([K_1] \setminus [K_2]\big),$$

where $\delta^* = \min\{f(f^*) - f(i)\colon\ i \notin f^*\}$.

PROOF. Let K_1, K_2 be two attractors of $\mathscr{K}_*$. Put $[K_1] \setminus [K_2] = \{e_1, \dots, e_r\}$ and let p be a path joining the attractors K_1 and K_2, that is, $p^1 \in K_1$ and $p^{|p|} \in K_2$. We define for ι in $\{1 \cdots r\}$,

$$t(\iota) = \min\{k\colon\ 0 \le k < |p|,\ k \equiv 0 \ (\mathrm{mod}\ 3),\ e_\iota \notin [p^{k+1}] \cap [p^{k+2}] \cap [p^{k+3}]\};$$

that is, $t(\iota)$ is the last time equal to $0 \bmod 3$ before the disappearance of e_ι in path p.

We make the conventions $p^0 = p^1$ and $p^{|p|+1} = p^{|p|+2} = p^{|p|}$. The transition mechanism implies the following fact: for each ι in $\{1, \dots, r\}$, there exists h_ι in $\{1, \dots, |\widehat{p}^{\,t(\iota)}|\}$ such that

$$\forall\ h,\ \tau(h_\iota - 1, |\widehat{p}^{\,t(\iota)}|) \le h < \tau(h_\iota, |\widehat{p}^{\,t(\iota)}|), \qquad p_h^{t(\iota)+1} \in f^* \ \Rightarrow\ p_h^{t(\iota)+1} = e_\iota.$$

Therefore, the r sets indexed by ι in $\{1, \dots, r\}$,

$$\{t(\iota), t(\iota)+1, t(\iota)+2\} \times \{h\colon\ \tau(h_\iota - 1, |\widehat{p}^{\,t(\iota)}|) \le h < \tau(h_\iota, |\widehat{p}^{\,t(\iota)}|)\},$$

have pairwise empty intersections, and it follows that

$$\overline{V}(p) \ge \sum_{\iota=1}^{r} \sum_{h=\tau(h_\iota-1,|\widehat{p}^{\,t(\iota)}|)}^{\tau(h_\iota,|\widehat{p}^{\,t(\iota)}|)-1} \Big[c\,\big(\widehat{f}(p^{t(\iota)}) - f(p_h^{t(\iota)+1})\big) + a\,\big(1 - \delta(p_h^{t(\iota)+1},\ p_h^{t(\iota)+2})\big) \Big].$$

Let ι belong to $\{1, \dots, r\}$ and suppose

$$c\,\big(\widehat{f}(p^{t(\iota)}) - f(p_h^{t(\iota)+1})\big) + a\,\big(1 - \delta(p_h^{t(\iota)+1},\ p_h^{t(\iota)+2})\big) = 0$$

for some h satisfying $\tau(h_\iota - 1, |\widehat{p}^{\,t(\iota)}|) \le h < \tau(h_\iota, |\widehat{p}^{\,t(\iota)}|)$. This implies $p_h^{t(\iota)+1} = p_h^{t(\iota)+2} = e_\iota$. Suppose that

$$1 + \frac{1}{2}\big(\tau(h_\iota, |\widehat{p}^{\,t(\iota)}|) - \tau(h_\iota - 1, |\widehat{p}^{\,t(\iota)}|)\big) = \left\lfloor \frac{m}{2(|\widehat{p}^{\,t(\iota)}| + 1)} \right\rfloor + 1$$

such terms vanish. Necessarily, there exists an odd index h such that

$$\tau(h_\iota - 1, |\widehat{p}^{\,t(\iota)}|) \le h < \tau(h_\iota, |\widehat{p}^{\,t(\iota)}|)$$

and

$$p_h^{t(\iota)+1} = p_h^{t(\iota)+2} = p_{h+1}^{t(\iota)+1} = p_{h+1}^{t(\iota)+2} = e_\iota.$$

The crossover operator does not affect the pair $(p_h^{t(\iota)+2}, p_{h+1}^{t(\iota)+2})$ [by condition (1)] and we have also $p_h^{t(\iota)+3} = p_{h+1}^{t(\iota)+3} = e_\iota$. Thus the individual e_ι is present in the populations $p^{t(\iota)+1}, p^{t(\iota)+2}, p^{t(\iota)+3}$, which contradicts the definition of $t(\iota)$.

We have proved that less than $\lfloor m/(2(|\widehat{p}^{\,t(\iota)}| + 1))\rfloor$ terms vanish in each sum $(1 \le \iota \le r)$:

$$\sum_{h=\tau(h_\iota - 1, |\widehat{p}^{\,t(\iota)}|)}^{\tau(h_\iota, |\widehat{p}^{\,t(\iota)}|)-1} \Big[c\left(\widehat{f}(p^{t(\iota)}) - f(p_h^{t(\iota)+1})\right) + a\left(1 - \delta(p_h^{t(\iota)+1},\ p_h^{t(\iota)+2})\right)\Big].$$

Moreover, we know that $\widehat{f}(p^{t(\iota)}) = f(f^*)$ (since $e_\iota \in [p^{t(\iota)}]$) and each nonzero term is necessarily greater than $\min(a,\ c\delta^*)$, so that the above sum is greater than

$$\begin{aligned}&\left(\tau(h_\iota, |\widehat{p}^{\,t(\iota)}|) - \tau(h_\iota - 1, |\widehat{p}^{\,t(\iota)}|) - \left\lfloor \frac{m}{2(|\widehat{p}^{\,t(\iota)}| + 1)} \right\rfloor\right) \min(a,\ c\delta^*)\\ &\quad = \left\lfloor \frac{m}{2(|\widehat{p}^{\,t(\iota)}| + 1)} \right\rfloor \min(a,\ c\delta^*).\end{aligned}$$

Yet $\widehat{p}^{\,t(\iota)} \subset f^*$, so that $|\widehat{p}^{\,t(\iota)}| \le |f^*|$. Finally we obtain

$$\overline{V}(p) \ge \left\lfloor \frac{m}{2(|f^*| + 1)} \right\rfloor \min(a,\ c\delta^*)\, r$$

[where $r = \mathrm{card}([K_1] \setminus [K_2])$] and taking the infimum over all paths p joining the attractors K_1 and K_2 yields the desired inequality. □

To obtain a lower bound for the communication cost between two attractors, we will study the possible trajectories of a pair of contiguous individuals within a given path of populations joining the two attractors.

DEFINITION 13.2 (Admissible paths). Let q be a path in $\mathscr{P}(E)$. We say that the pair (i_k, j_k), $1 \le k \le |q|$, of paths in E is admissible for q if:

(i) $\forall\ k \in \{1 \cdots |q|\},\ i_k \in q^k,\ j_k \in q^k$.

(ii) $\forall\, k,\ 1 \le 3k < |q|$:
Either $i_{3k-2} = i_{3k-1}$ or $\alpha(i_{3k-2}, i_{3k-1}) > 0$.
Either $j_{3k-2} = j_{3k-1}$ or $\alpha(j_{3k-2}, j_{3k-1}) > 0$.
$\beta((i_{3k-1}, j_{3k-1}), (i_{3k}, j_{3k})) > 0$.
$f(i_{3k+1}) = f(j_{3k+1}) \ \Rightarrow\ i_{3k+1} = j_{3k+1}$.

The set of all pairs of paths admissible for the path q is denoted by $\mathscr{A}(q)$.

We define the quantity $\rho(K_1, K_2)$ for K_1 and K_2 in $\mathscr{K}$ to be the infimum

$$\begin{aligned}\inf\Bigg\{ \sum_{1\le 3k<|q|} &\big[a\,(2 - \delta(i_{3k-2}, i_{3k-1}) - \delta(j_{3k-2}, j_{3k-1}))\\ &\quad + c\,\big(2\widehat{f}(q^{3k}) - f(i_{3k+1}) - f(j_{3k+1})\big)\big]\\ &\text{where } q \in [\overline{D}],\ q^1 = [K_1],\ q^{|q|} = [K_2],\\ &\qquad (i_k, j_k)_{1\le k\le |q|} \in \mathscr{A}(q),\ i_1 = j_1 \Bigg\}.\end{aligned}$$

We put

$$\rho = \min\{\rho(K_1, K_2)\colon\ K_1, K_2 \in \mathscr{K},\ \rho(K_1, K_2) > 0\}.$$

Since the quantities $\rho(K_1, K_2)$ are finite sums involving terms of the form a or $c(f(i) - f(j))$ [where $f(i) \ge f(j)$], we have $\rho \ge \min(a, c\delta) > 0$.

LEMMA 13.3. *Let K_1 and K_2 belong to $\mathscr{K}$. Suppose $\rho(K_1, K_2) = 0$. Then there exists a path q in the set $[\overline{D}]$ such that*

$$q^1 = [K_1],\ q^{|q|} = [K_2] \quad \text{and} \quad \widehat{f}(q^1) \le \widehat{f}(q^4) \le \cdots \le \widehat{f}(q^{|q|-3}) \le \widehat{f}(q^{|q|}).$$

That is, the sequence $(\widehat{f}(q^{3k+1}))$, $1 \le 3k+1 \le |q|$, is nondecreasing. In particular, we have $f(K_1) \le f(K_2)$.

COROLLARY 13.4. *Let K_1 and K_2 be two attractors such that $f(K_1) > f(K_2)$. Then $\rho(K_1, K_2) \ge \rho \ge \min(a, c\delta) > 0$.*

PROOF. Let K_1 and K_2 be as in the hypothesis of Lemma 13.3. By definition of $\rho(K_1, K_2)$, there exists a path q in $[\overline{D}]$ and a pair (i_k, j_k), $1 \le k \le |q|$, of paths in E admissible for q such that $q^1 = [K_1]$, $q^{|q|} = [K_2]$, $i_1 = j_1$ and

$$\begin{aligned}\sum_{1\le 3k<|q|} &\big[a\,(2 - \delta(i_{3k-2}, i_{3k-1}) - \delta(j_{3k-2}, j_{3k-1}))\\ &\quad + c\,\big(2\widehat{f}(q^{3k}) - f(i_{3k+1}) - f(j_{3k+1})\big)\big] = 0.\end{aligned}$$

This relation implies that, for each k, $1 \le 3k < |q|$, we have

$$i_{3k-2} = i_{3k-1}, \qquad j_{3k-2} = j_{3k-1}, \qquad f(i_{3k+1}) = f(j_{3k+1}) = \widehat{f}(q^{3k}). \tag{10}$$

Since the pair of paths (i_k, j_k) is admissible for the path q, the last of the above equalities implies that $i_{3k+1} = j_{3k+1}$ for each k, $1 \le 3k < |q|$. Notice also that $i_1 = j_1$. Moreover, the crossover has no effect when two identical individuals mate [condition (1)], so that

$$\forall\ k,\ 1 \le 3k < |q|, \qquad i_{3k} = j_{3k} = i_{3k-1} = j_{3k-1}. \tag{11}$$

Therefore, we conclude from (10) and (11) that

$$\forall\ k,\ 1 \le 3k < |q|, \qquad \widehat{f}(q^{3k+1}) = \widehat{f}(q^{3k}) \ge \widehat{f}(q^{3k-2})$$

and the sequence $(\widehat{f}(q^{3k+1}))$, $1 \le 3k+1 \le |q|$, is nondecreasing. □

LEMMA 13.5 (Lower bound for the communication cost). *Let K_1 and K_2 belong to $\mathscr{K}$. We have the inequality*

$$V(K_1, K_2) \ge \left(\frac{m}{4} - \frac{\Lambda+3}{2}\right)\rho(K_1, K_2).$$

(*We recall that* $\Lambda = \max\{|f_\lambda|:\ \lambda \in \mathbb{R}_+^*\}$.)

PROOF. Let p be a path in $\overline{D}$ joining the attractors K_1 and K_2 (i.e., $p^1 \in K_1$ and $p^{|p|} \in K_2$). We make the convention $p^0 = p^1$. Putting

$$\tau_k = \frac{1}{2}\big(\tau(|\widehat{p}^{\,3k-3}|, |\widehat{p}^{\,3k-3}|) - 1\big) = |\widehat{p}^{\,3k-3}|\left\lfloor \frac{m}{2(|\widehat{p}^{\,3k-3}|+1)}\right\rfloor, \qquad 1 \le 3k < |p|,$$

we have

$$\begin{aligned} \overline{V}(p) &= \sum_{1\le 3k<|p|}\ \sum_{h=1}^{m}\big[a\left(1 - \delta(p_h^{3k-2}, p_h^{3k-1})\right) + c\left(\widehat{f}(p^{3k}) - f(p_h^{3k+1})\right)\big] \\ &\ge \sum_{1\le 3k<|p|}\ \sum_{h=0}^{\tau_k-1}\big[a\left(2 - \delta(p_{2h+1}^{3k-2}, p_{2h+1}^{3k-1}) - \delta(p_{2h+2}^{3k-2}, p_{2h+2}^{3k-1})\right) \\ &\qquad + c\left(2\widehat{f}(p^{3k}) - f(p_{2h+1}^{3k+1}) - f(p_{2h+2}^{3k+1})\right)\big]. \end{aligned} \tag{12}$$

Let h belong to $\{1 \cdots \Lambda\}$. Let $m = 2(h+1)q + r$ be the Euclidean division of m by $2(h+1)$. We have [since $h \ge 1$ and $0 \le r < 2(h+1)$]

$$\begin{aligned} 2h\left\lfloor \frac{m}{2(h+1)}\right\rfloor &= 2hq = 2h\frac{m-r}{2(h+1)} = \frac{h}{h+1}(m-r) \\ &> \frac{1}{2}\big(m - 2(h+1)\big) \ge \frac{m}{2} - \Lambda - 1. \end{aligned}$$

For each k, $1 \le 3k < |p|$, the very definition of Λ implies $1 \le |\widehat{p}^{\,3k-3}| \le \Lambda$, whence

$$\tau(|\widehat{p}^{\,3k-3}|, |\widehat{p}^{\,3k-3}|) > \frac{m}{2} - \Lambda.$$

It follows from (12) that

$$\overline{V}(p) \geq \sum_{1\leq 3k<|p|} \sum_{0\leq h<\tau_k} \cdots = \sum_{1\leq 3k<|p|} \sum_{1\leq 2h+1<\tau(|\widehat{p}^{\,3k-3}|,|\widehat{p}^{\,3k-3}|)} \cdots$$
$$\geq \sum_{1\leq 3k<|p|} \sum_{1\leq 2h+1<m/2-\Lambda} \cdots = \sum_{1\leq 2h+1<m/2-\Lambda} \sum_{1\leq 3k<|p|} \cdots$$

[the ellipses stand for the summand in the last sum of (12)]. Yet, for each h such that $1 \leq 2h+1 < m/2-\Lambda$, the pair of paths (p^k_{2h+1}, p^k_{2h+2}), $1 \leq k \leq |p|$, is admissible for $[p]$ and satisfy also $p^1_{2h+1} = p^1_{2h+2}$; as a consequence, the sum

$$\sum_{1\leq 3k<|p|} a\left(2 - \delta\left(p^{3k-2}_{2h+1}, p^{3k-1}_{2h+1}\right) - \delta\left(p^{3k-2}_{2h+2}, p^{3k-1}_{2h+2}\right)\right)$$
$$+ c\left(2\widehat{f}(p^{3k}) - f\left(p^{3k+1}_{2h+1}\right) - f\left(p^{3k+1}_{2h+2}\right)\right)$$

is greater than $\rho(K_1, K_2)$ and finally $\overline{V}(p) \geq (m/4 - (\Lambda+3)/2)\,\rho(K_1, K_2)$. Taking the infimum over all the paths joining K_1 and K_2 gives the inequality of the lemma. $\square$

In the same flavor as the two preceding lemmas, we have the following proposition.

PROPOSITION 13.6. *Let K_1, K_2 be two elements of $\mathscr{K}$ such that $f(K_1) = f(K_2)$. Then*

$$V(K_1, K_2) \geq \min(a,\ c\delta) \min\left(\left\lfloor \frac{m}{2(\Lambda+1)} \right\rfloor \operatorname{card}\big([K_1] \setminus [K_2]\big),\ \frac{m}{4} - \frac{\Lambda+3}{2}\right).$$

PROOF. Let p be a path in E^m joining the attractors K_1 and K_2. We consider two cases.

Case 1. Suppose the sequence $(\widehat{f}(p^{3k+1}))$, $1 \leq 3k+1 \leq |p|$, is not nondecreasing; there exists an index h, $1 < 3h+1 \leq |p|$, such that $f(K_1) > \widehat{f}(p^{3h+1})$ and an attractor K_3 satisfying

$$V(p^{3h+1}, K_3) = 0, \qquad f(K_3) = \widehat{f}(p^{3h+1}).$$

Clearly $\overline{V}(p) \geq V(K_1, K_3)$ and Corollary 13.4 and Lemma 13.5 imply

$$V(K_1, K_3) \geq \min(a,\ c\delta)\left(\frac{m}{4} - \frac{\Lambda+3}{2}\right).$$

Case 2. Suppose the sequence $(\widehat{f}(p^{3k+1}))$, $1 \leq 3k+1 \leq |p|$, is nondecreasing. Then we have

$$\forall\, k,\ 1 \leq 3k+1 \leq |p|, \qquad \widehat{f}(p^{3k+1}) = f(K_1) = f(K_2) = \theta.$$

We modify (just for the purpose of the proof) the function f in the following way: for each i in E, if $f(i)$ is strictly greater than θ, we set $f(i) = \theta - \delta$. The attractors K_1 and K_2 then become elements of $\mathscr{K}_*$ and we are in a situation analogous to Lemma 13.1. Moreover, the cost of the path p, that is, the quantity

$$\overline{V}(p) = \sum_{1\le 3k<|p|}\left[ad(p^{3k-2}, p^{3k-1}) + c\sum_{h=1}^{m}\left(\widehat{f}(p^{3k}) - f(p_h^{3k+1})\right)\right],$$

decreases when evaluated with the new function f. Actually, for any indices h and k such that $1 \le 3k < |p|$, $1 \le h \le m$, the value $f(p_h^{3k+1})$ remains unchanged (since it was lower than or equal to θ) and the value $\widehat{f}(p^{3k})$, which was greater than or equal to θ, becomes equal to θ. [Since $\widehat{f}(p^{3k+1}) = \theta$ and $[p^{3k+1}] \subset [p^{3k}]$, the set $[p^{3k}]$ necessarily intersects the set f_θ.] Application of Lemma 13.1 yields

$$\overline{V}(p) \ge \left\lfloor \frac{m}{2(|f_\theta| + 1)} \right\rfloor \min(a, c\delta_\theta)\operatorname{card}\big([K_1] \setminus [K_2]\big),$$

where $\delta_\theta = \min\{\theta - f(i)\colon\ i \in E,\ f(i) < \theta\}$.

In both cases, the inequality of the lemma is satisfied. □

14. The costs of the good transitions. We distinguish also two kinds of good transitions: those which create some new peak fitness individuals and those which increase fitness.

We note by R the minimal number of transitions necessary to join two arbitrary points of E through the kernel α. That is, R is the smallest integer satisfying

$$\forall\ i, j \in E,\ \exists\ r \le R,\ \exists\ e_1, \dots, e_{r+1} \in E \quad \text{such that} \quad e_1 = i, \qquad e_{r+1} = j,$$
$$\forall\ k \in \{1, \dots, r\}, \qquad \alpha(e_k, e_{k+1}) > 0.$$

LEMMA 14.1 (Increasing diversity in f^*). *Let*

$$V^* = \max_{i,\, j\in f^*} \min\left\{ar + c\sum_{k=1}^{r+1} f(f^*) - f(e_k)\colon\ e_1 = i,\ e_{r+1} = j,\ \prod_{k=1}^{r}\alpha(e_k, e_{k+1}) > 0\right\}.$$

We have

$$V^* \le aR + c(R-1)\Delta \le (a + c\Delta)|E| < \infty$$

and

$$\sup_{m\in\mathbb{N}^*}\ \max_{\substack{K_1\in\mathscr{K}_*\\ K_1\ne K^*}}\ \min_{\substack{K_2\in\mathscr{K}_*\\ [K_1]\subsetneq[K_2]}} V(K_1, K_2) \le V^*.$$

(*We recall that K^* is the unique attractor such that $[K^*] = f^*$.*)

PROOF. The inequalities $V^* \le aR + c(R-1)\Delta \le (a + c\Delta)|E| < \infty$ are straightforward. Let K_1 belong to $\mathscr{K}_* \setminus \{K^*\}$. There exists a point i in $[K_1]$, a point j in $f^* \setminus [K_1]$ and a path $e_1 \to \cdots \to e_{r+1}$ in E joining i and j such that

$$\forall\, k \in \{1,\dots,r\}, \qquad \alpha(e_k, e_{k+1}) > 0, \qquad \forall\, k \in \{2,\dots,r\}, \qquad e_k \notin f^*,$$

$$ar + c\sum_{k=1}^{r+1}\big[f(i) - f(e_k)\big] \le V^*.$$

Let p be a path in E^m defined by the extremal conditions

$$p^1 \in K_1 \text{ with } p^1_m = i \quad \text{and} \quad p^{r+1} \in K_2 \text{ where } K_2 \in \mathscr{K}_*,\ [K_2] = [K_1] \cup \{j\}.$$

(There exists a unique attractor K_2 satisfying $[K_2] = [K_1] \cup \{j\}$) and having for intermediate populations

$$\forall\, k \in \{2,\dots,r\},\ \forall\, h \in \{1,\dots,m-1\}, \qquad p^k_h = p^1_h, \qquad p^k_m = e_k.$$

In this path, the first $m-1$ components remain fixed and the last component follows the path $e_1 \to \cdots \to e_{r+1}$. Clearly $[K_1] \subsetneq [K_2]$ and $V(K_1, K_2) \le V(p) \le V^*$. $\square$

LEMMA 14.2 (Increasing fitness). *Let*

$$V^+ = \max_{i \in E \setminus f^*} \min \left\{ ar + c\sum_{k=1}^{r} |f(i) - f(e_k)|:\ e_1 = i,\ f(e_{r+1}) > f(i),\ \prod_{k=1}^{r} \alpha(e_k, e_{k+1}) > 0 \right\}.$$

We have

$$V^+ \le aR + c(R-1)\Delta \le (a + c\Delta)|E| < \infty$$

and

$$\sup_{m \in \mathbb{N}^*} \max_{K_1 \in \mathscr{K} \setminus \mathscr{K}_*} \min_{\substack{K_2 \in \mathscr{K} \\ f(K_2) > f(K_1)}} V(K_1, K_2) \le V^+. \tag{13}$$

PROOF. The proof is similar to the proof of the preceding lemma. Let K_1 belong to $\mathscr{K} \setminus \mathscr{K}_*$ and let i belong to $[K_1]$. By definition of V^+, there exists a path $e_1 \to \cdots \to e_{r+1}$ in E such that $e_1 = i$, $f(e_{r+1}) > f(i)$ and

$$\forall\, k \in \{1 \cdots r\}, \qquad \alpha(e_k, e_{k+1}) > 0, \qquad f(e_k) \le f(i),$$

$$ar + c\sum_{k=1}^{r} f(i) - f(e_k) \le V^+.$$

There exists a subdivision $k_1 = 1 < k_2 < \cdots < k_{s-1} < k_s = r+1$ of $\{1,\dots,r+1\}$ such that $f(e_{k_1}) = f(e_{k_2}) = \cdots = f(e_{k_{s-1}}) = f(i)$ and

$$\forall\, k \in \{1,\dots,r+1\} \setminus \{k_1,\dots,k_s\}, \qquad f(e_k) < f(i).$$

For each ι, $1 < \iota < s$, let K_ι be the unique attractor such that $[K_\iota] = [K_1] \cup \{e_{k_2},\dots,e_{k_\iota}\}$. Let K_s be the attractor $\{(e_{r+1},\dots,e_{r+1})\}$ (i.e., K_s contains only the population whose components are all equal to e_{r+1}). Clearly, the attractor K_s satisfies $f(K_s) > f(K_1)$.

Let p be a path in E^m such that, for each ι in $\{1 \cdots s\}$, p^{k_ι} belongs to K_ι and

$$\forall\ \iota \in \{1, \ldots, s-1\},\ \forall\ k \in \{k_\iota + 1, \ldots, k_{\iota+1} - 1\}, \qquad p_m^k = e_k,$$
$$\forall\ h \in \{1, \ldots, m-1\}, \qquad p_h^k = p_h^{k_\iota}.$$

(During the transition from K_ι to $K_{\iota+1}$, the first $m-1$ components remain fixed and the last component follows the mutation path $e_{k_\iota} \to \cdots \to e_{k_{\iota+1}}$.) The path p belongs to D^m and joins the attractors K_1 and K_s, and its cost is less than V^+. It follows that $V(K_1, K_s) \le V^+$. $\square$

15. The asymptotic dynamics of the algorithm. In this section, we put together the results of Sections 13 and 14 in order to prove Theorem 15.5, which implies Theorem 8.1. We give an explicit upper bound on the critical population size m^* in Corollary 15.6. We end the section with Theorem 15.7, which is Theorem 8.2 with an explicit upper bound on M^*.

The basic tools used to study the asymptotic dynamics of the process are the Freidlin–Wentzell graphs ([10], Chapter 6). For completeness, we recall the basic definition.

DEFINITION 15.1 (X-graphs). Let H be a finite set and let X be a subset of H. An X-graph is a graph consisting of arrows $h_1 \to h_2$ ($h_1 \in H \setminus X,\ h_2 \in H,\ h_1 \neq h_2$) satisfying:

(i) Every point of $H \setminus X$ is the initial point of exactly one arrow.
(ii) There are no closed cycles in the graph.

Condition (ii) may be replaced by:

(ii′) For any point h_1 of $H \setminus X$ there exists a sequence of arrows leading from h_1 to some point h_2 of X.

The set of X-graphs is denoted by $G(X)$ and we will use the letter g to denote a graph.

We will consider graphs on the set of attractors $\mathcal{K}$. For any graph g over $\mathcal{K}$, we define its cost by

$$V(g) = \sum_{(K_1 \to K_2) \in g} V(K_1, K_2).$$

If X and Y are two subsets of $\mathcal{K}$, we denote by $G_X(Y)$ the set of Y-graphs over $X \cup Y$. For instance, we have $G(X) = G_{\mathcal{K}}(X) = G_{\mathcal{K} \setminus X}(X)$. We define

$$W_X(Y) = \min_{g \in G_X(Y)} \sum_{(K_1 \to K_2) \in g} V(K_1, K_2) = \min_{g \in G_X(Y)} V(g).$$

We denote by $G_X^*(Y)$ the set of graphs in $G_X(Y)$ which realize this minimum. If g is a graph over $\mathscr{K}$, its restriction $g_{|\lambda}$ to the level λ is the graph

$$g_{|\lambda} = \{(K_1 \to K_2) \in g\colon\ K_1 \in \mathscr{K}_\lambda\}.$$

THEOREM 15.2 (Sufficient condition to ensure $W^* = \{K^*\}$). *If the inequality*

$$\begin{aligned}
&\sum_{\lambda \in f(E\backslash f^*)} W_{\mathscr{K}_\lambda}(\mathscr{K}_\lambda^+) + W_{\mathscr{K}_*}(K^*)\\
&\quad < \sum_{\lambda \in f(E\backslash f^*)} W_{\mathscr{K}_\lambda}(\mathscr{K}_\lambda^- \cup \mathscr{K}_\lambda^+)\\
&\qquad + \min\Big(\min_{K \in \mathscr{K}_* \backslash \{K^*\}} W_{\mathscr{K}_*}(\{K\} \cup \mathscr{K}_*^-),\\
&\qquad\qquad W_{\mathscr{K}_*}(\mathscr{K}_*^-) - \max_{K \in \mathscr{K}_*^-} \min_{K',\, f(K') \neq f(K)} V(K, K')\Big)
\end{aligned} \tag{14}$$

is satisfied, then the minimum of the virtual energy corresponds to the ideal attractor K^*; that is, $W^* = \{K^*\}$, and therefore

$$\forall\ x \in E^m, \qquad \lim_{l \to \infty} \lim_{n \to \infty} P\big([X_n^l] = f^* \mid X_0^l = x\big) = 1.$$

PROOF. Let g be a graph over $\mathscr{K}$. We decompose the sum $V(g)$ in the following way:

$$V(g) = \sum_{\lambda \in f(E)} \sum_{\substack{K_1 \in \mathscr{K}_\lambda \\ (K_1 \to K_2) \in g}} V(K_1, K_2) = \sum_{\lambda \in f(E)} \sum_{(K_1 \to K_2) \in g_{|\lambda}} V(K_1, K_2).$$

That is,

$$V(g) = \sum_{\lambda \in f(E)} V(g_{|\lambda}).$$

Suppose now that g is in $G(K)$ for some K in $\mathscr{K}$. Put $\theta = f(K)$. Then $g_{|\lambda}$ belongs to $G_{\mathscr{K}_\lambda}(\mathscr{K}_\lambda^- \cup \mathscr{K}_\lambda^+)$ whenever $\lambda \neq \theta$, whence

$$V(g_{|\lambda}) \geq W_{\mathscr{K}_\lambda}(\mathscr{K}_\lambda^- \cup \mathscr{K}_\lambda^+). \tag{15}$$

We consider two cases, depending upon the value of θ.

Case 1. $\theta = f(f^*)$. In this case, $g_{|\theta}$ belongs to $G_{\mathscr{K}_*}(\{K\} \cup \mathscr{K}_*^-)$ so that

$$V(g_{|\theta}) \geq W_{\mathscr{K}_*}(\{K\} \cup \mathscr{K}_*^-). \tag{16}$$

Summing inequalities (15) and (16) yields

$$V(g) \geq \sum_{\lambda \in f(E\backslash f^*)} W_{\mathscr{K}_\lambda}(\mathscr{K}_\lambda^- \cup \mathscr{K}_\lambda^+) + W_{\mathscr{K}_*}(\{K\} \cup \mathscr{K}_*^-).$$

Taking the minimum over all g in $G(K)$, we have

$$W(K) \geq \sum_{\lambda \in f(E\backslash f^*)} W_{\mathscr{K}_\lambda}(\mathscr{K}_\lambda^- \cup \mathscr{K}_\lambda^+) + W_{\mathscr{K}_*}(\{K\} \cup \mathscr{K}_*^-),$$

and taking again the minimum over all K in $\mathcal{K}_* \setminus \{K^*\}$, we obtain

$$\min_{K \in \mathcal{K}_* \setminus \{K^*\}} W(K) \geq \sum_{\lambda \in f(E \setminus f^*)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+) + \min_{K \in \mathcal{K}_* \setminus \{K^*\}} W_{\mathcal{K}_*}(\{K\} \cup \mathcal{K}_*^-). \tag{17}$$

Case 2. $\theta \neq f(f^*)$. Let K' be any element of $\mathcal{K}_\theta^- \cup \mathcal{K}_\theta^+$. Then $g_{|\theta} \cup \{(K \to K')\}$ is a graph belonging to $G_{\mathcal{K}_\theta}(\mathcal{K}_\theta^- \cup \mathcal{K}_\theta^+)$, whence

$$V(g_{|\theta}) \geq W_{\mathcal{K}_\theta}(\mathcal{K}_\theta^- \cup \mathcal{K}_\theta^+) - V(K, K').$$

This inequality being valid for all K' outside $\mathcal{K}_\theta$, we have

$$V(g_{|\theta}) \geq W_{\mathcal{K}_\theta}(\mathcal{K}_\theta^- \cup \mathcal{K}_\theta^+) - \min_{K',\, f(K') \neq f(K)} V(K, K'). \tag{18}$$

Summing inequalities (15) and (18) yields

$$V(g) \geq \sum_{\lambda \in f(E)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+) - \min_{K',\, f(K') \neq f(K)} V(K, K').$$

Taking the minimum over all g in $G(K)$, we obtain

$$W(K) \geq \sum_{\lambda \in f(E)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+) - \min_{K',\, f(K') \neq f(K)} V(K, K')$$

and taking again the minimum over all K in $\mathcal{K} \setminus \mathcal{K}_* = \mathcal{K}_*^-$, we have

$$\min_{K \in \mathcal{K}_*^-} W(K) \geq \sum_{\lambda \in f(E)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+) - \max_{K \in \mathcal{K}_*^-} \min_{K',\, f(K') \neq f(K)} V(K, K'). \tag{19}$$

Combining inequalities (17) and (19), we see that

$$\begin{aligned} \min_{K \neq K^*} W(K) \geq & \sum_{\lambda \in f(E \setminus f^*)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^- \cup \mathcal{K}_\lambda^+) \\ & + \min\Big(\min_{K \in \mathcal{K}_* \setminus \{K^*\}} W_{\mathcal{K}_*}(\{K\} \cup \mathcal{K}_*^-), \\ & \qquad W_{\mathcal{K}_*}(\mathcal{K}_*^-) - \max_{K \in \mathcal{K}_*^-} \min_{K',\, f(K') \neq f(K)} V(K, K') \Big). \end{aligned} \tag{20}$$

We build now a graph g of $G(K^*)$ which describes the most desirable dynamics of our algorithm. For each λ in $f(E \setminus f^*)$, we select a graph g_λ in the set $G^*_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^+)$. Let $g_{f(f^*)}$ be a graph of $G^*_{\mathcal{K}_*}(K^*)$. We define the graph g as the union of the graphs $(g_\lambda)_{\lambda \in f(E)}$:

$$(K_1 \to K_2) \in g \iff \exists\, \lambda \in f(E), \qquad (K_1 \to K_2) \in g_\lambda.$$

Clearly g belongs to $G(K^*)$. Furthermore, we have by construction

$$V(g) = \sum_{\lambda \in f(E)} V(g_{|\lambda}) = \sum_{\lambda \in f(E \setminus f^*)} W_{\mathcal{K}_\lambda}(\mathcal{K}_\lambda^+) + W_{\mathcal{K}_*}(K^*).$$

It follows that

$$W(K^*) \le \sum_{\lambda\in f(E\setminus f^*)} W_{\mathscr{K}_\lambda}(\mathscr{K}_\lambda^+) + W_{\mathscr{K}_*}(K^*). \tag{21}$$

Putting together inequalities (20) and (21) and the hypothesis (14) of the theorem, we see finally that $W^* = \{K^*\}$. $\square$

Of course, inequality (14) is strongly linked with the optimization problem; the quantities involved there are built with fitness function f, mutation kernel α, crossover kernel β, population size m and parameters a, c. However, this inequality is of little practical interest; we will now derive stronger and simpler conditions to ensure $W^* = \{K^*\}$.

COROLLARY 15.3. *Suppose that*

$$W_{\mathscr{K}_\lambda}(\mathscr{K}_\lambda^+) \le W_{\mathscr{K}_\lambda}(\mathscr{K}_\lambda^- \cup \mathscr{K}_\lambda^+) \quad \textit{for all } \lambda \textit{ in } f(E\setminus f^*), \tag{22}$$

$$W_{\mathscr{K}_*}(K^*) < W_{\mathscr{K}_*}(\mathscr{K}_*^-) - \max_{K\in\mathscr{K}_*^-} \min_{K',\, f(K')\ne f(K)} V(K, K'), \tag{23}$$

$$W_{\mathscr{K}_*}(K^*) < \min_{K\in\mathscr{K}_*\setminus\{K^*\}} W_{\mathscr{K}_*}(\{K\}\cup\mathscr{K}_*^-). \tag{24}$$

Then $W^* = \{K^*\}$.

PROOF. Clearly, inequalities (22), (23) and (24) imply inequality (14). $\square$

We will show that these inequalities hold when the population size m is sufficiently large. First, the left-hand side of (23) and (24) is bounded.

PROPOSITION 15.4. *The quantity* $W_{\mathscr{K}_*}(K^*)$ *is bounded as a function of* m:

$$\sup_{m\in\mathbb{N}^*} W_{\mathscr{K}_*}(K^*) < |\mathscr{K}_*|\, V^* = 2^{|f^*|}\, V^* < \infty. \tag{25}$$

PROOF. We build a graph g belonging to $G_{\mathscr{K}_*}(K^*)$ whose cost is less than $|\mathscr{K}_*|\, V^*$. For each K in $\mathscr{K}_*\setminus\{K^*\}$, there exists by Lemma 14.1 an attractor K' belonging to $\mathscr{K}_*$ such that $[K]\subsetneq[K']$ and $V(K, K') \le V^*$. We consider successively each attractor K of $\mathscr{K}_*\setminus\{K^*\}$ and we add such an arrow $(K \to K')$ to the graph. The resulting graph g belongs to $G_{\mathscr{K}_*}(K^*)$ and its cost is less than or equal to $(|\mathscr{K}_*| - 1)\, V^*$. $\square$

THEOREM 15.5. *Let* m *be an integer such that*

$$\rho\left(\frac{m}{4} - \frac{\Lambda+3}{2}\right) > \max_{K_1\in\mathscr{K}_*^-} \min_{K_2,\, f(K_2)>f(K_1)} V(K_1, K_2) + W_{\mathscr{K}_*}(K^*), \tag{26}$$

$$\left\lfloor \frac{m}{2(|f^*|+1)} \right\rfloor \min(a,\ c\delta^*) > W_{\mathscr{K}_*}(K^*). \tag{27}$$

For this integer, we have $W^* = \{K^*\}$ *and, in addition,*

$$\forall\, x \in E^m,\ \forall\, y \in K^*, \qquad \lim_{l\to\infty}\lim_{n\to\infty} P\big(X_n^l = y \mid X_0^l = x\big) = \frac{1}{|K^*|}.$$

COROLLARY 15.6. *There exists a critical population size* m^* *such that the conclusion of Theorem* 15.5 *holds for all* m *greater than* m^*. *In addition,* m^* *is smaller than*

$$\max\left(2(\Lambda+3) + \frac{4}{\rho}(2^{|f^*|}V^* + V^+),\ \frac{2^{|f^*|+1}V^*(|f^*|+1)}{\min(a, c\delta^*)}\right).$$

We recall that $\Lambda = \max\{|f_\lambda|\colon \lambda \in \mathbb{R}_+^*\}$, $\delta^* = \min\{f(f^*) - f(i)\colon i \notin f^*\}$, ρ *is introduced after Definition* 13.2, V^* *in Lemma* 14.1 *and* V^+ *in Lemma* 14.2. *Moreover, we have the crude estimates* $\max(V^*, V^+) \le aR + c(R-1)\Delta$ *and* $\rho \ge \min(a, c\delta)$.

PROOF. The corollary is a straightforward consequence of Theorem 15.5 together with inequalities (13) and (25). In particular, conditions (26) and (27) of Theorem 15.5 are fulfilled as soon as

$$\rho\left(\frac{m}{4} - \frac{\Lambda+3}{2}\right) \ge 2^{|f^*|}\, V^* + V^+,$$

$$\left\lfloor \frac{m}{2(|f^*|+1)} \right\rfloor \min(a,\ c\delta^*) \ge 2^{|f^*|}\, V^*. \qquad \square$$

PROOF OF THEOREM 15.5. We prove that if the integer m satisfies inequalities (26) and (27), then the set of inequalities (22), (23) and (24) hold.

Let λ be in $f(E \setminus f^*)$ and let g be a graph belonging to $G_{\mathscr{K}_\lambda}(\mathscr{K}_\lambda^- \cup \mathscr{K}_\lambda^+)$. This graph may contain a finite number of transitions from $\mathscr{K}_\lambda$ to $\mathscr{K}_\lambda^-$: $K_1 \to K_1', \dots, K_r \to K_r'$. The first inequality (26) implies that for each K_h, $1 \le h \le r$, there exists K_h'' in $\mathscr{K}_\lambda^+$ such that

$$\rho\left(\frac{m}{4} - \frac{\Lambda+3}{2}\right) > V(K_h, K_h'')$$

and Lemma 13.5 yields $V(K_h, K_h'') < V(K_h, K_h')$.

Let $\tilde{g}$ be the graph obtained from g by replacing the r arrows $K_1 \to K_1', \dots, K_r \to K_r'$ by $K_1 \to K_1'', \dots, K_r \to K_r''$. The graph $\tilde{g}$ is in the set $G_{\mathscr{K}_\lambda}(\mathscr{K}_\lambda^+)$ and satisfies $V(\tilde{g}) \le V(g)$. This construction being valid for any graph of $G_{\mathscr{K}_\lambda}(\mathscr{K}_\lambda^- \cup \mathscr{K}_\lambda^+)$, we have

$$W_{\mathscr{K}_\lambda}(\mathscr{K}_\lambda^+) \le W_{\mathscr{K}_\lambda}(\mathscr{K}_\lambda^- \cup \mathscr{K}_\lambda^+)$$

for all λ in $f(E \setminus f^*)$ and the first condition (22) is proved.

Concerning the second condition (23), we notice that any graph g belonging to $G_{\mathscr{K}_*}(\mathscr{K}_*^-)$ contains at least an arrow starting from $\mathscr{K}_*$ and ending in $\mathscr{K}_*^-$.

Thus

$$W_{\mathscr{K}_*}(\mathscr{K}_*^-) \geq \min_{\substack{K \in \mathscr{K}_* \\ K' \notin \mathscr{K}_*}} V(K, K')$$

and Lemma 13.5 implies

$$W_{\mathscr{K}_*}(\mathscr{K}_*^-) \geq \left(\frac{m}{4} - \frac{\Lambda + 3}{2}\right)\rho$$

which, together with (26), yields (23).

Similarly, any graph belonging to $G_{\mathscr{K}_*}(\{K\} \cup \mathscr{K}_*^-)$ (where $K \neq K^*$) contains either a transition from K^* to an attractor K of $\mathscr{K}_*$ or to an attractor K of $\mathscr{K}_*^-$, so that

$$\min_{\substack{K \in \mathscr{K}_* \\ K \neq K^*}} W_{\mathscr{K}_*}(\{K\} \cup \mathscr{K}_*^-) \geq \min\{V(K^*, K)\colon\ K \in \mathscr{K},\ K \neq K^*\}$$

and Lemmas 13.1 and 13.5 show that this quantity is greater than

$$\min\left(\left\lfloor \frac{m}{2(|f^*| + 1)} \right\rfloor \min(a,\ c\delta^*),\ \left(\frac{m}{4} - \frac{\Lambda + 3}{2}\right)\rho\right)$$

so that inequalities (26) and (27) imply the third and last condition (24). Corollary 15.3 implies that $W^* = \{K^*\}$. Thus the sequence of stationary measures of (X_n^l), $l \in \mathbb{N}$, concentrates on K^* as l goes to infinity (Proposition 12.3).

It remains now to prove that the limiting distribution is the uniform distribution over K^*. We could proceed as in [7] and use the representation formula of the stationary measure involving Freidlin–Wentzell graphs. However, this result is rather a consequence of the following fact: the virtual energy possesses a unique minimum at K^* and the unperturbed process (X_n^∞) admits a unique invariant probability measure on K^* (which is the uniform distribution over K^*). We are thus in a situation analogous to Theorem 4.2 of [10], Chapter 6. □

THEOREM 15.7 (Structure of the cycles). *Let m be such that*

$$\min\left(\left\lfloor \frac{m}{2(\Lambda + 1)} \right\rfloor,\ \left(\frac{m}{4} - \frac{\Lambda + 3}{2}\right)\right) \min(a,\ c\delta) > \max(V^+,\ V^*). \tag{28}$$

Then each cycle over the set of attractors $\mathscr{K}$ not containing the attractor K^* is reduced to one single attractor K.

PROOF. We define an order $\prec$ on the set of attractors $\mathscr{K}$ by

$$\forall\ K_1, K_2 \in \mathscr{K},\ \ K_1 \prec K_2$$
$$\Leftrightarrow\quad f(K_1) < f(K_2) \quad \text{or} \quad f(K_1) = f(K_2), \quad [K_1] \subset [K_2].$$

Notice that this order is an extension of the previous order $\prec_\infty$, that is,

$$\forall\ K_1, K_2 \in \mathscr{K}, \qquad K_1 \prec_\infty K_2 \quad \Rightarrow \quad K_1 \prec K_2.$$

Furthermore, the set $\mathscr{K}$ admits a greatest element which is precisely the attractor K^*:

$$\forall\ K \in \mathscr{K}, \qquad K \prec K^*.$$

Now let π be a cycle over $\mathscr{K}$ not containing K^*. Suppose π is not reduced to one attractor. Let K be a maximal element of π for the order $\prec$. Then, for each K' in π distinct from K, we have either $f(K') < f(K)$ or $f(K') = f(K)$, $[K] \not\subset [K']$. Proposition 13.6, Corollary 13.4 and Lemma 13.5 then imply

$$\forall\ K' \in \pi \setminus \{K\},\ \ V(K, K') \geq \min(a,\ c\delta) \min\left(\left\lfloor \frac{m}{2(\Lambda+1)} \right\rfloor, \frac{m}{4} - \frac{\Lambda+3}{2} \right).$$

Let K' be an element of $\pi \setminus \{K\}$ such that $V(K, K')$ is minimal. Then $A(K, K') = W(K) + V(K, K')$ (where A is the communication altitude), whence

$$A(K, K') \geq W(K) + \min(a,\ c\delta) \min\left(\left\lfloor \frac{m}{2(\Lambda+1)} \right\rfloor, \frac{m}{4} - \frac{\Lambda+3}{2} \right).$$

Yet Lemmas 14.1 and 14.2 show that there exists an attractor K'' such that

$$K \prec K'', \qquad K \neq K'', \qquad V(K, K'') \leq \max(V^+,\ V^*),$$

whence

$$A(K, K'') \leq W(K) + \max(V^+,\ V^*).$$

Inequality (28) yields $A(K, K'') < A(K, K')$ so that necessarily the attractor K'' belongs to the cycle π, but this contradicts the maximality of K in the ordered set $(\pi, \prec)$. $\square$

APPENDIX

Proof of Proposition 8.3. To obtain an upper bound on H_1, it is enough to consider only the cycles over the set of attractors $\mathscr{K}$. We also know that the attractors not containing K^* are reduced to one attractor. Therefore

$$H_1 \leq \max_{K \in \mathscr{K} \setminus \{K^*\}} \min\{V(K, K') \colon\ K' \in \mathscr{K},\ K' \neq K\}.$$

Lemmas 14.1 and 14.2 show that $H_1 \leq \max(V^+,\ V^*)$. Similarly, Proposition 1.33 of [26], Proposition 13.6 and Lemma 13.5 imply that H_e^* is greater than an affine increasing function of m. $\square$

Proof of Theorem 8.7. The definition of α_{opt} is ([28], Definition 2.21)

$$\alpha_{\text{opt}} = \min \left\{ \frac{W(\pi) - W(K^*)}{H_e(\pi)} \colon\ \pi \text{ is a cycle not containing } K^* \right\}.$$

For a cycle π not containing K^*, the quantity $H_e(\pi)$ remains bounded whereas the virtual energy $W(\pi)$ is greater than an affine strictly increasing function of m [a graph belonging to $G_{\mathscr{K}}(\pi)$ necessarily contains a bad transition].

Finally, the virtual energy of the ideal attractor $W(K^*)$ is bounded as a function of m. $\square$

PROOF OF PROPOSITION 9.1. Let f be constant over $E \setminus \{i\}$ and such that $f(i) > f(j)$ for all $j \neq i$. We denote by $K(j)$ the attractor associated to the point j [which consists simply of the uniform population $(j, \dots, j)$]. On the one hand we have $V(K(i), K(e)) = 0$ since a massive crossover event can transform the uniform population $(i, \dots, i)$ into the uniform population $(e, \dots, e)$. (Such an event does not involve the random perturbations and has a null cost.) On the other hand, the algorithm cannot escape from $K(e)$ without performing at least one mutation [since $\beta((e, e), (e, e)) = 1$, the crossover has no effect on $(e, \dots, e)$]. Therefore, $V(K(e), K) > 0$ for any attractor K. Let g be a graph in $G(K(i))$ realizing the value $W(K(i))$ (see the beginning of Section 15). We remove the arrow starting from $K(e)$ (which has a positive cost) and we add the arrow $K(i) \to K(e)$ (of null cost). This way we obtain a graph of $G(K(e))$ whose cost is strictly less than the cost of g. It follows that $W(K(e)) < W(K(i))$. Proposition 12.3 implies that the sequence of the stationary measures does not concentrate on $K(i)$ when the perturbations vanish. $\square$

Acknowledgments. I thank Alain Trouvé, Gérard Ben Arous and Alain Berlinet for useful discussions. I thank the referees for their comments and for pointing out interesting references.

REFERENCES

[1] AARTS, E. H. L. and VAN LAARHOVEN, P. J. M. (1987). *Simulated Annealing: Theory and Applications*. Reidel, Dordrecht.
[2] CATONI, O. (1992). Large deviations for annealing. Ph.D. dissertation, Univ. Paris XI.
[3] CATONI, O. (1991). Sharp large deviations estimates for simulated annealing algorithms. *Ann. Inst. H. Poincaré* **27** 291–383.
[4] CATONI, O. (1991). Applications of sharp large deviations estimates to optimal cooling schedules. *Ann. Inst. H. Poincaré* **27** 463–518.
[5] CATONI, O. (1992). Rough large deviations estimates for simulated annealing. Application to exponential schedules. *Ann. Probab.* **20** 1109–1146.
[6] CERF, R. (1993). Asymptotic convergence of genetic algorithms. Preprint.
[7] CERF, R. (1996). The dynamics of mutation–selection algorithms with large population sizes. *Ann. Inst. H. Poincaré.* **32** 455–508.
[8] DEUSCHEL, J. D. and MAZZA, C. (1994). L^2 convergence of time non-homogeneous Markov processes: I. Spectral estimates. *Ann. Appl. Probab.* **4** 1012–1056.
[9] FAIGLE, U. and KERN, W. (1991). Note on the convergence of simulated annealing algorithms. *SIAM J. Control Optim.* **29** 153–159.
[10] FREIDLIN, M. I. and WENTZELL, A. D. (1984). *Random Perturbations of Dynamical Systems*. Springer, New York.
[11] GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intelligence* **6** 721–741.
[12] GOLDBERG, D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA.
[13] HAJEK, B. (1988). Cooling schedules for optimal annealing. *Math. Oper. Res.* **2** 311–329.
[14] HAJEK, B. and SASAKI, G. H. (1989). Simulated annealing—to cool or not. *Systems Control Lett.* **12** 443–447.

[15] HAJEK, B. and SASAKI, G. H. (1988). The time complexity of maximum matching by simulated annealing. *J. Assoc. Comput. Mach.* **35** 387–403.

[16] HOLLAND, J. H. (1975). *Adaptation in Natural and Artificial Systems*. Univ. Michigan Press, Ann Arbor.

[17] HOLLEY, R. A., KUSUOKA, S. and STROOCK, D. W. (1988). Asymptotics of the spectral gap with applications to the theory of simulated annealing. *J. Funct. Anal.* **83** 333–347.

[18] HOLLEY, R. A. and STROOCK, D. W. (1989). Simulated annealing via Sobolev inequalities. *Comm. Math. Phys.* **115** 553–569.

[19] HWANG, C. R. and SHEU, S. J. (1992). Singular perturbed Markov chains and exact behaviours of simulated annealing process. *J. Theoret. Probab.* **5** 223–249.

[20] JERRUM, M. R. (1992). Large cliques elude the Metropolis process. *Random Structures and Algorithms* **3** 347–359.

[21] JERRUM, M. R. and SINCLAIR, A. J. (1993). Polynomial-time approximation algorithms for the Ising model. *SIAM J. Comput.* **22** 1087–1116.

[22] LUNDY, M. and MEES, A. (1986). Convergence of an annealing algorithm. *Math. Programming* **34** 111–124.

[23] SINCLAIR, A. J. (1992). Improved bounds for mixing rates of Markov chains and multicommodity flow. *Combin. Probab. Comput.* **1** 351–370.

[24] SORKIN, G. B. (1991). Efficient simulated annealing on fractal energy landscapes. *Algorithmica* **6** 367–418.

[25] TROUVÉ, A. (1992). Convergence optimale pour les algorithmes de recuits généralisés. *C. R. Acad. Sci. Paris Sér. I Math.* **315** 1197–1202.

[26] TROUVÉ, A. (1993). Parallélisation massive du recuit simulé. Ph.D. dissertation, University Paris XI.

[27] TROUVÉ, A. (1995). Asymptotical behaviour of several interacting annealing processes. *Probab. Theory Related Fields* **102** 123–143.

[28] TROUVÉ, A. (1996). Cycle decompositions and simulated annealing. *SIAM J. Control Optim.* **34** 966–986.

[29] TROUVÉ, A. (1996). Rough large deviation estimates for the optimal convergence speed exponent of generalized simulated annealing algorithms. *Ann. Inst. H. Poincaré.* **32** 299–348.

CNRS URA 743
UNIVERSITÉ PARIS SUD
MATHÉMATIQUES, BÂTIMENT 425
91405 ORSAY CEDEX
FRANCE
E-MAIL: raphael.cerf@math.u-psud.fr