The Annals of Applied Probability
2002, Vol. 12, No. 1, 248–271

GLOBAL OPTIMIZATION WITH EXPLORATION/SELECTION ALGORITHMS AND SIMULATED ANNEALING

BY OLIVIER FRANÇOIS

Institut National Polytechnique de Grenoble

This article studies a stochastic model of an evolutionary algorithm that evolves a "population" of potential solutions to a minimization problem. The minimization process is based on two operators. First, each solution is regarded as an individual that attempts a random search on a graph, involving a probabilistic operator called *exploration*. The second operator is called *selection*. This deterministic operator creates interaction between individuals. The convergence of the evolutionary process is described within the framework of simulated annealing. It can be quantified by means of two quantities called the *critical height* and the *optimal convergence exponent*, which both measure the difficulty of the algorithm to deal with the minimization problem. This work describes the critical height for large enough population sizes. Explicit bounds are given for the optimal convergence exponent, using a few geometric quantities. As an application, this work allows comparisons of the evolutionary strategy with independent parallel runs of the simulated annealing algorithm, and it helps deciding when one method should be preferred to the other.

1. Introduction. Evolutionary algorithms are general purpose optimization methods that have demonstrated a capability to yield good approximate solutions, even in case of complicated multi-modal, bumpy or discontinuous problems. These methods involve probabilistic operators based on the model of natural evolution. The resulting algorithms rely on competitive behavior within a *population* of interacting individuals, each of which represents a potential solution to the optimization problem.

The population is initialized arbitrarily, and improved iteratively by means of the randomized operators of *mutation*, *selection* and *recombination* (which is omitted in some algorithmic realizations). The interaction arises from the selection process, which is the main optimizing operator. Mutation is an exploration operator, whose purpose is to introduce new potential solutions in the competition. The new solutions created by this operator are often called the *offspring* resulting from mutation, and the offspring are likely to replace their parents. Although the merit of recombination is often asserted [Goldberg (1989), Cerf (1996)], this operator is less amenable to a quantitative mathematical analysis, and will not be discussed further [see Rabinovitch and Widgerson (1999)].

Received June 2000; revised January 2001.
AMS 2000 subject classifications. Primary 60J10, 92D15.
Key words and phrases. Evolutionary algorithms, generalized simulated annealing.

Three main streams of evolutionary techniques have been identified: genetic algorithms, evolutionary programming and evolution strategies (the later methods sharing many similarities). For an overview of these different instances [see Fogel (1995) and Bäck (1996)]. From this paper point of view, a fourth technique should be added: parallel simulated annealing [Aarts (1988), Trouvé (1996)]. In such case, selection is defined as a rejection/acceptance criterion called the *Metropolis* dynamics. Many authors have proposed a simulated annealing-like approach to genetic algorithms. In Davis and Principle (1991), the "mutation probability" is assumed to converge to zero. This parameterization has been used by Cerf (1994) but with selection (roulette wheel) reinforced as well. Continuous-time versions of genetic algorithms have also been analyzed following this approach by Del Moral and Miclo (1999). The link between these algorithms and simulated annealing has been established through a theory called *generalized simulated annealing* (GSA), based on large deviations results [Trouvé (1996), Catoni (1997)]. On the other hand, few efforts have been devoted to evolution strategies [Bäck (1996)] within the framework of GSA. This paper studies methods called *exploration/selection* (E/S), for which the algorithmic principle is closer to evolution strategies than genetic algorithms. Important differences will appear, however, motivated by the use of GSA. Both methods are based upon a deterministic selection process. The main difference is that the number of offspring obtained by mutation at each generation will be random in E/S, while this number is deterministic in evolution strategies. Furthermore, the fraction of offspring goes to zero, and plays the role of a temperature.

The analysis of the algorithm faithfully follows Cerf's approach of genetic algorithms [Cerf (1996a, b)]. The merit of E/S algorithms is that less intricate arguments are involved in their analysis. As a consequence, some of the critical geometric constants that help to understand the algorithm's behavior can be described explicitly. Let us summarize the results obtained by Cerf for the genetic algorithm. Cerf's algorithm can be modeled as a nonhomogeneous Markov chain controlled by a positive temperature parameter in the spirit of simulated annealing. To prove the convergence of the algorithm toward optimal solutions, a first step was to warrant the concentration of the stationary distribution. Under mild assumptions, there exists a critical population size above which the stationary distribution at low temperature concentrates on copies of such solutions. This critical size is sensitive to the "diameter" of the search space and other constants depending on the objective function [see Cerf (1996a) for definitions]. Choosing the temperature schedule was the second step. Optimal choices actually rely on two geometric quantities called the *critical height* and the *optimal convergence exponent*, which both measure the difficulty of the algorithm to deal with the optimization problem. The definition of these quantities comes from GSA theory. The critical height is related to an infinite horizon perspective while the convergence exponent is related to finite horizon. Unfortunately, no explicit values (or easily computable bounds) are available for these crucial quantities in genetic algorithms.

Our paper is organized as follows. Section 2 gives a definition of the basic algorithm. It uses a coupling of mutation and selection which is different from genetic algorithms. The main results are stated in Section 3. Section 4 gives basic notations and the main tools from the large deviations formalism. Auxiliary results and proofs are deferred to the Sections 5 and 6. The paper concludes with computer implementation issues and comparisons with parallel simulated annealing algorithms.

2. The exploration/selection algorithm. Let E be a finite set and f a nonnegative (non-constant) function defined on E. Let A^* denote the subset of minimal points of f:

$$A^* = \arg\min_{a \in E} f(a). \tag{2.1}$$

Without loss of generality, minimization problems are considered instead of maximization problems. The objective function f is often called the *fitness* function.

The set E is endowed with a graph structure $(E, \mathcal{G})$ called the *exploration graph*. The Exploration/selection strategy uses a vector of potential solutions of the minimization problem. Each solution is regarded as an individual that attempts a random search on the exploration graph. The exploration process acts (almost) independently on each individual, and consists of choosing a random neighbor in the graph. By analogy with genetic algorithms, the exploration process is also termed *mutation process*, and the neighbor resulting from a step of the random walk is said to be the *offspring* resulting from mutation.

The strategy relies upon two parameters: The first parameter is the *population size*

$$n \geq 2, \tag{2.2}$$

and the second parameter is a *mutation probability*

$$0 < p < 1, \tag{2.3}$$

which represents the fraction of offspring by mutation at each generation. The set of all populations is defined as

$$X = E^n, \tag{2.4}$$

which consists of replicas of the search space. For a given population $x = (x_1, \ldots, x_n) \in X$, let

$$A_x = \{x_1, \ldots, x_n\} \subset E \tag{2.5}$$

denote the subset of "types" contained in x. We denote by

$$\hat{x} = \arg\min_{x_i \in A_x} f(x_i) \tag{2.6}$$

the minimal point in A_x with the lowest label.

The exploration graph is assumed to be symmetric (non-oriented) and connected. We denote by $\deg(a)$ the degree of the vertex $a \in E$, and by $N(a) \subset E$ its neighborhood. The algorithm can be described informally as follows.

- Choose n initial individuals in E.
- Repeat
  1. Select $\hat{x}$ from the population;
  2. For each $i = 1, \ldots, n$, create an offspring of individual i and replace the parent by its offspring with probability p; otherwise replace individual i by the individual selected in Step 1 [with probability $(1 - p)$];

  until some stopping criterion is met.

In Step 2 of the algorithm, we impose the additional condition that the offspring cannot be the individual selected in Step 1. (This technical assumption will be useful in defining the large deviations functionals associated with the algorithm.) All choices are made independently. Our stopping criterion is a finite (large) number of iterations (generations).

Modeling genetic algorithms with Markov chains is a standard topic in evolutionary computation [see Nix and Vose (1992), Rudolph (1994), Cerf (1994) and Chakrabordy, Kalyanmoy and Chakraborty (1996)]. The state of the population at generation t is denoted by $X(t)$, and the process $(X(t))$ is actually described by a Markov chain model, for which the transition probabilities are given by

$$\forall x, y \in X, \quad \operatorname{Prob}\big(X(t+1) = y \mid X(t) = x\big) = q(x, y) \tag{2.7}$$

with

$$q(x, y) = \prod_{i=1}^{n} \Big(p\, a(x, y_i) + (1 - p)\delta_{\hat{x}, y_i} \Big) \tag{2.8}$$

and

$$\forall y_i \neq x_i, \qquad a(x, y_i) = \begin{cases} \dfrac{1}{\deg(x_i)}, & \text{if } y_i \in N(x_i) \backslash \{\hat{x}\}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.9}$$

As usual, $\delta_{a,b}$ denotes the Kroneker symbol, equal to 1 if $a = b$, and 0 otherwise. The exploration operator a is well defined as we set

$$a(x, x_i) = 1 - \sum_{y_i \neq x_i} a(x, y_i). \tag{2.10}$$

Let us give now an alternative description of the transition probabilities. For two populations x, y, consider the subset of labels $i \in \{1, \ldots, n\}$ defined as

$$I(x, y) = \{1 \leq i \leq n;\ y_i \neq \hat{x}\}. \tag{2.11}$$

The number of elements in this subset is denoted by

$$V_1(x, y) = \#I(x, y). \tag{2.12}$$

Then, we have

$$\forall x, y \in X, \quad q(x, y) = \pi(x, y) p^{V_1(x,y)} (1 - p)^{n - V_1(x,y)}, \tag{2.13}$$

with

$$\pi(x, y) = \prod_{i \in I(x,y)} a(x, y_i). \tag{2.14}$$

To emphasize the relationship to the simulated annealing algorithm, a positive parameter T is introduced. The mutation probability is therefore represented as

$$p = e^{-1/T}. \tag{2.15}$$

The parameter T is called a *temperature*. As T goes to zero, the transition probabilities $q(x, y)$ satisfy

$$\forall x, y \in X, \qquad q(x, y) \sim \exp\bigl(-V_1(x, y)/T\bigr), \tag{2.16}$$

where the symbol $\sim$ means that

$$-T \log q(x, y) \to V_1(x, y) \qquad \text{as } T \to 0 \tag{2.17}$$

(log is the natural logarithm), and the definition of $V_1(x, y)$ has been extended so that

$$V_1(x, y) = \infty \qquad \text{if } \pi(x, y) = 0. \tag{2.18}$$

NOTE. To compare, let us recall the construction of the Mutation+selection genetic algorithm studied by Cerf (1996a). Starting from $x = x_0 \in X$, the following steps are repeated. For each individual x^i, $i = 1, \ldots, n$, one offspring ζ^i is created by mutation with probability p. Otherwise, ζ^i is equal to x^i. Therefore, a sample $(y^i)_{i=1,\ldots,n}$, $y^i \in \{\zeta^1, \ldots, \zeta^n\}$ is created according to the Boltzmann probability distribution

$$\mathrm{P}(Y = \zeta^i) \propto \exp\bigl(-\theta f(\zeta^i)/T\bigr), \qquad \theta > 0. \tag{2.19}$$

Then, x^i is replaced by y^i for all $i = 1, \ldots, n$.

This description actually corresponds to the genetic algorithm using *roulette wheel* selection. The vector ζ represents the intermediate population vector obtained after the mutation step has been applied. The parameter θ represents the selection intensity, and allows the respective weights of mutation and selection to be balanced. The Boltzmann distribution makes the population converge on the best offspring ζ^* (as T goes to zero), and therefore warrants that some

minimization process actually takes place. A Markov chain model can be associated to the cost

$$V_{GA}(x, y) = \min_{\zeta} \left\{ \sum_{i=1}^{n} \big(1 - \delta(x^i, \zeta^i)\big) + \theta \sum_{i=1}^{n} f(y^i) - f(\zeta^*) \right\}, \tag{2.20}$$

when the transition $x \to y$ is admissible. The contribution $\sum_{i=1}^{n}(1 - \delta(x^i, \zeta^i))$ merely counts the number of offspring by mutation. These cost functionals are by far more complicated than those described in equation (2.12).

NOTE. Similar ideas have been introduced by Del Moral and Miclo (1999). In their work, a coin is tossed to decide which operator of Mutation or Selection should be applied. When the result is "Head," mutation is applied to each individual with probability p. Otherwise, the population is resampled using Boltzmann selection as in Cerf's algorithm. Thus, global decisions are taken in each generation. In contrast, the decision of which individual should mutate or be replaced by a better individual holds locally in the E/S algorithm.

3. Main results.

3.1. *Some definitions.* To start with, we recall some necessary definitions about the exploration graph $(E, \mathcal{G})$. A path is merely a sequence of vertices

$$\gamma\colon a^0 \to a^1 \to \cdots \to a^r, \qquad a^i \in E,\ i = 0, \ldots, r, \tag{3.1}$$

where the symbol $a^i \to a^{i+1}$ denotes an edge between two consecutive vertices. The length of the path is the number r of edges in the path. The distance on the graph, that is, the minimal length of a path between two vertices a and b in $(E, \mathcal{G})$, is denoted by $d(a, b)$. [The diameter of the graph is the maximum distance between two arbitrary vertices in $(E, \mathcal{G})$.]

Next, we define two geometric quantities of crucial importance with regard to the convergence issue. The first is defined as

$$n_* = \max_{a \notin A^*} d(a, A^*), \tag{3.2}$$

where $d(a, A^*)$ is the distance from the subset A^*. We shall call *equifitness* a subset for which the fitness function is constant. The second quantity is defined as

$$D_* = \max_{A: A \cap A^* = \varnothing} \min_{a \in A, b \notin A: f(b) \le f(a)} d(a, b), \tag{3.3}$$

where the maximum runs over one-point subsets and all equifitness subsets A such that a path $\gamma_{aa'}$ having the property

$$f(a^i) \ge f(a), \qquad a^i \in \gamma_{aa'}, \tag{P}$$

exists for all pairs of vertices aa' in A. When f is one-to-one, D_* takes a much simpler form

$$D_* = \max_{a \neq a^*} \min_{b:\, f(b) < f(a)} d(a,b). \tag{3.4}$$

Actually, this quantity measures the greatest distance between a local minimum of the fitness function and a solution which outperforms this minimum (if f has a local minimum). The constant D_* can be regarded as a measure of the "chance" of escape from local minima during the local search [Suzuki (1993)].

3.2. *Statements.* This section presents three statements. Theorem 1 describes the behavior of the mean hitting time of the optimal solution. Theorem 2 gives necessary and sufficient conditions for convergence using decreasing cooling schedules in the spirit of simulated annealing. Theorem 3 describes the optimal success probability after a large number of steps.

THEOREM 1 (Hitting time of A^*). *Let $n > n_*$ and*

$$\tau = \inf\{t \geq 1,\ \hat{X}(t) \in A^*\}.$$

Then, we have

$$T \log\Big(\max_{x \in X,\ A_x \cap A^* = \varnothing} E[\tau \mid X(0) = x] \Big) \to D_* \qquad \text{as } T \to 0.$$

Now, let the notation $X(t)$ stand for $X^{T(t)}(t)$ for all $t \geq 1$, meaning that the temperature T can be changed at each generation. This notation includes dependence on $T(t)$, and $(X(t))$ is henceforth a nonhomogeneous Markov chain.

THEOREM 2 (Optimal cooling schedules). *Let $n > n_*$. Consider a non-increasing sequence of temperatures $(T(t))_{t \geq 1}$ that converges to zero. Then we have*

$$\forall x \in X, \qquad \operatorname{Prob}\big(\hat{X}(t) \in A^* \mid X(0) = x\big) \to 1 \qquad \text{as } t \to \infty, \tag{3.5}$$

if and only if

$$\sum_{t=1}^{\infty} e^{-D_*/T(t)} = \infty. \tag{3.6}$$

THEOREM 3 (Optimal convergence exponent). *Let $n > n_*$. There exist two constants $R_1 > 0$ and $R_2 > 0$, and a constant α_* such that, for all $t \geq 1$,*

$$\frac{R_1}{t^{\alpha_*}} \leq \inf_{0 \leq T(t) \leq \cdots \leq T(1)} \max_{x \in X} P\big(\hat{X}(t) \notin A^* \mid X(0) = x\big) \leq \frac{R_2}{t^{\alpha_*}}.$$

Moreover, the constant α_* satisfies

$$\frac{n - n_*}{D_*} \leq \alpha_* \leq \frac{n(n_* + 1 - D_*) + n_* - 1 - D_*}{D_*}.$$

This constant is the convergence exponent of the Markov chain $(X(t))$ (*to be defined in Section* 4).

REMARK. Similar results have been obtained in [Cerf (1996a)] regarding the Mutation+Selection genetic algorithm. Cerf's estimates are however less accurate. Regarding the population size for instance, concentration on absolute minima holds when

$$n > \frac{D + \theta(D-1)\Delta}{\min(1, \theta\varepsilon)}, \tag{3.7}$$

where D is the diameter of the exploration graph,

$$\Delta = \max\{|f(a) - f(b)|;\ a, b \in E\} \tag{3.8}$$

and

$$\varepsilon = \min\{|f(a) - f(b)|;\ a \neq b \in E\}. \tag{3.9}$$

In Del Moral and Miclo (1999), the critical height of the studied Mutation-Selection algorithm remains unknown. Concentration on absolute minima holds (taking $\theta = 1$) if

$$n > \frac{D}{\min(1, \varepsilon)}. \tag{3.10}$$

In conclusion, the upper bounds on critical sizes obtained by Cerf (1996a) and Del Moral and Miclo (1999) can be significantly larger than n_*, whose value is always lower than D.

4. Notation and recalls. The cornerstone of this work is that our model fits in with the formalism of generalized simulated annealing (GSA) presented in Trouvé (1996). This framework has been developed to study Metropolis-like algorithms. A Markov transition kernel q_T defines a generalized Metropolis algorithm, or generalized simulated annealing if there exists $\kappa > 0$ such that

$$\frac{1}{\kappa}\pi(x, y)e^{-V_1(x,y)/T} \leq q_T(x, y) \leq \kappa\pi(x, y)e^{-V_1(x,y)/T} \tag{4.1}$$

where the family V_1 (the communication cost) satisfies $V_1(x, y) \geq 0$ and $V_1(x, y) = +\infty$ iff $\pi(x, y) = 0$.

The *communication cost in many steps* from x to y in X is defined as

$$V(x, y) = \inf\left\{\sum_{k=0}^{r-1} V_1(x_k, x_{k+1}),\ x_0 = x,\ x_k \in X,\ x_r = y,\ r \geq 1\right\}. \tag{4.2}$$

Virtual energy. Specific oriented subgraphs of X (with the same set of vertices) are needed to proceed with the definition. Recall that an x-graph ends at $x \in X$

(no edge starts from x), and contains no cycle (each $y \neq x$ is the starting point of exactly one oriented edge). The set of all x-graphs is denoted by $G(x)$. The *virtual energy* is defined on the set X by

$$\forall x \in X, \qquad W(x) = \min_{g \in G(x)} V(g) \tag{4.3}$$

with

$$V(g) = \sum_{(y \to z) \in g} V(y, z). \tag{4.4}$$

In formula (4.3), the minimum is taken over the set of all x-graphs on X and the sum (4.4) runs over the edges of these graphs. The virtual energy W describes the asymptotic behavior of the chain $X(t)$ as T goes to 0. In Freidlin and Wentzel (1984), a logarithmic equivalent for the stationary probability distribution $\mu_T(x)$ is given by

$$\forall x \in X, \qquad \lim_{T \to 0} -T \log \mu_T(x) = W(x) - W_*, \tag{4.5}$$

where W_* is the minimal value of W over the set X. Let $\mathcal{W}^*$ be the set of all populations in X for which the minimum W_* is attained. Equation (4.5) states that the distribution μ_T concentrates on $\mathcal{W}^*$ as T goes to 0.

Elevations. For $x, y \in X$, $x \neq y$ and each trajectory γ_{xy} of the chain $(X(t))$ between x and y

$$\gamma_{xy} = (x_0 = x \to x_1 \to \cdots \to x_r = y), \tag{4.6}$$

define the *elevation* as

$$H(\gamma_{xy}) = \max_{0 \leq k < r} \{W(x_k) + V(x_k, x_{k+1})\}, \tag{4.7}$$

where the maximum is taken over all vertices in γ_{xy}. Let $H(x, y)$ be the lowest possible value of $H(\gamma_{xy})$ over all self-avoiding trajectories γ_{xy} from x to y. The quantity $H(x, y)$ is called the *communication altitude* between x and y, and is symmetrical in x and y [see Trouvé (1996)].

Definition of the cycles. Let $\lambda \geq 0$ and

$$W_\lambda = \{x \in X;\ W(x) \leq \lambda\}. \tag{4.8}$$

Consider the equivalence classes $\mathcal{C}_\lambda$ of the relation $\mathcal{R}_\lambda$ defined on W_λ as

$$\forall x \neq y \in W_\lambda, \qquad x \mathcal{R}_\lambda y \quad \text{iff } H(x, y) \leq \lambda, \tag{4.9}$$

and $x \mathcal{R}_\lambda x$. A subset $\pi \subset X$ is a cycle if either $\pi = \mathcal{C}_\lambda$ for some $\lambda \geq 0$, or π reduces to a one-point subset.

The critical height and the convergence exponent. The *critical height* is a geometric quantity defined as

$$H_1 = \max\{H_e(\pi);\ \pi \text{ cycle not intersecting } \mathcal{W}^*\}. \tag{4.10}$$

The exit height $H_e(\pi)$ can be viewed as the limit of $T\log(E[\tau_\pi \mid X_0 = x])$ as T goes to zero, where τ_π is the exit time of π and $x \in \pi$. (The definition is independent of the starting point x.) An algorithmic definition of exit heights can be found in Trouvé (1996) (because of its length, this definition cannot be reproduced here). Following Theorem 7, we shall be able to give a simpler definition of H_1 for $n > n_*$, which will be more amenable to computations. The *convergence exponent* is given by

$$\alpha_* = \min\left\{\frac{W(\pi) - W_*}{H_e(\pi)};\ \pi \text{ cycle not intersecting } \mathcal{W}^*\right\}, \tag{4.11}$$

where $W(\pi)$ is the minimum of W over the subset π. The convergence exponent describes the minimum error of the algorithm after a large number of generations.

Trouvé's Theorem. We recall here Trouvé (1996), Theorem 2.22, page 981.

THEOREM 4. *For all decreasing cooling schedules $(T(t))_{t\geq 1}$ converging to* 0, *we have*

$$\sup_{x\in X} \operatorname{Prob}\big(X(t) \notin \mathcal{W}^* \mid X_0 = x\big) \to 0 \qquad \text{as } t \to \infty \tag{4.12}$$

if and only if

$$\sum_{t=1}^{\infty} e^{-H_1/T(t)} = \infty. \tag{4.13}$$

We assume that $\alpha_* < \infty$. There exist two constants $R_1 > 0$ and $R_2 > 0$ such that, for all $t \geq 1$

$$\frac{R_1}{t^{\alpha_*}} \leq \inf_{0\leq T(t)\leq\cdots\leq T(1)} \max_{x\in X} P(X(t) \notin \mathcal{W}^* \mid X_0 = x) \leq \frac{R_2}{t^{\alpha_*}}.$$

5. Concentration and uniform populations. Throughout the whole paper, the uniform population $(a, \ldots, a)$ and the element $a \in E$ are identified by denoting

$$(a) = (a, \ldots, a). \tag{5.1}$$

The subset of all uniform populations is denoted by U. From the definition of V_1 [equation (2.12)], we have,

$$\forall x \in X, \qquad V_1\big(x, (\hat{x})\big) = 0. \tag{5.2}$$

Moreover, for all $x \in X$ and $y \neq (\hat{x})$,

$$V_1(x, y) > 0. \tag{5.3}$$

Define a new functional on the subset of uniform populations as follows. For all $x, y \in U$, put

$$V_{1,U}(x, y) = \inf \left\{ \sum_{k=0}^{r-1} V_1(x_k, x_{k+1}),\ x_0 = x,\ x_k \notin U\ (1 \le k < r),\ x_r = y,\ r \ge 1 \right\}.$$

Therefore, define V_U from $V_{1,U}$ in the same way as V is defined from V_1. For uniform populations (a) and (b), $V_U(a, b)$ and $W(a)$ will stand for $V_U((a),(b))$ and W((a)) (parentheses are omitted when dealing with uniform populations).

LEMMA 1. *For two uniform populations* (a) *and* (b), *we have*

$$W(a) = \min_{g \in G_U(a)} \sum_{(b \to c) \in g} V_U(b, c)$$

where the minimum runs over the a-graphs defined over the subset U (*or* E). *Moreover*, $H(a, b)$ *can be computed from* V_U *instead of* V.

PROOF. It follows from Theorem 5.8 of Cerf (1996a) applied with $H = U$.

NOTE. This lemma can be interpreted as follows. The chain $(X(t))$ can be identified with the chain of successive visits to U. This induced chain is defined over the subset U. It satisfies the large deviation property associated with the functional V_U. The functionals W and H coincide on U with their analog definitions built from V_U instead of V.

LEMMA 2. *Let* $a^* \in A^*$ *and consider a vertex* $a \in E$ *such that* $a \neq a^*$. *Then we have*

$$V_U(a, a^*) = d(a, a^*), \tag{5.4}$$

where d *is the distance on the graph* $(E, \mathcal{G})$.

PROOF. Obviously, we have $V_U(a, a^*) \ge d(a, a^*)$. Consider a path that realizes the min in the definition of $d(a, a^*)$:

$$a_0 = a \to a_1 \to \cdots \to a_r = a^*,$$

and the trajectory defined as follows:

$$
\begin{array}{c}
x_0 = (a) \\
\downarrow \\
x_1 = (a_1, a, \ldots, a) \\
\cdots \\
\downarrow \\
x_k = (a_k, \tilde{x}_k, \ldots, \tilde{x}_k) \\
\downarrow \\
\cdots \\
\downarrow \\
x_r = (a^*, \tilde{x}_r, \ldots, \tilde{x}_r) \\
\downarrow \\
x_{r+1} = (a^*)
\end{array}
\tag{5.5}
$$

where we put

$$\tilde{x}_1 = a \tag{5.6}$$

and

$$\forall k = 2, \ldots, r, \qquad \tilde{x}_k = \begin{cases} a_{k-1}, & \text{if } f(a_{k-1}) \leq f(\tilde{x}_{k-1}), \\ \tilde{x}_{k-1}, & \text{otherwise.} \end{cases} \tag{5.7}$$

Therefore, we have

$$\forall k = 0, \ldots, r-1, \qquad V_1(x_k, x_{k+1}) = 1, \tag{5.8}$$

and then $V_U(a, a^*) \leq d(a, a^*)$.

The following result warrants that the chain converges on a uniform population consisting of n copies of a global minimum $a^* \in A^*$.

THEOREM 5 (Concentration on the subset of global minima). *Let $n > n_*$ where*

$$n_* = \max_{a \notin A^*} d(a, A^*). \tag{5.9}$$

Then, we have $\mathcal{W}^* = A^*$, and for all $x \in X$,

$$\lim_{T \to 0} \lim_{t \to \infty} \operatorname{Prob}\big(X(t) \textit{contains an element of } A^* \mid X(0) = x\big) = 1. \tag{5.10}$$

PROOF. See the Appendix.

6. Proofs.

6.1. *Hitting times.* Again, specific subgraphs of X are needed to proceed with the computation of the hitting time of A^* by the chain $(X(t))$. Let A be a subset

of vertices in E, and $G(A)$ be the set of all graphs that end in A (no edge starts from A), and such that for $g \in G(A)$, each $a \notin A$ is the starting point of exactly one edge (g contains no cycle). Let

$$\tau = \inf\{t \geq 1,\ \hat{X}(t) \in A^*\}. \tag{6.1}$$

Put

$$H_0 = \lim_{T \to 0} T \log\Big(\max_{x \in X,\, A_x \cap A^* = \varnothing} E[\tau \mid X(0) = x] \Big). \tag{6.2}$$

LEMMA 3. *Let $n > n_*$. Then we have*

$$H_0 = \min_{g \in G(A^*)} V_U(g) - \min_{a \notin A^*} \min_{g \in G(A^* \cup a)} V_U(g).$$

PROOF. It follows from Theorem 5 and Freidlin and Wentzell (1984), Chapter 6.

LEMMA 4. *Let $a, b \in E$ be any pair of vertices such that $a \notin A^*$, and there exists a path having the property (P). Then the cost $V_U(a, b)$ is equal to the length of a minimal path having this property.*

PROOF. It is similar to Lemma 2.

THEOREM 6. *Let $n > n_*$. Then we have $H_0 = D_*$.*

PROOF. Let $g_* \in G(A^*)$ be a graph such that

$$V_U(g_*) = \min_{g \in G(A^*)} V_U(g). \tag{6.3}$$

In g_*, an edge which starts from the uniform population (a) ends at a uniform population (b) such that $f(b) \leq f(a)$. Since g_* contains no cycle, the maximal edge cost is necessarily equal to D_*. Now, consider an edge $a_* \to b_*$ for which

$$V_U(a_*, b_*) = D_*. \tag{6.4}$$

Then, build a minimal graph

$$g_{**} \in G(A^* \cup a_*) \tag{6.5}$$

by deleting the edge $a_* \to b_*$ in g_*. Hence, we have

$$\min_{a \notin A^*} \min_{g \in G(A^* \cup a)} V_U(g) = V_U(g_{**}) = \min_{g \in G(A^*)} V_U(g) - D_*. \tag{6.6}$$

This proves that

$$H_0 = D_*. \tag{6.7}$$

6.2. *Cycles.* A subset $U_f \subset U$ is called *equifitness* if $f(a) = f$ for all $(a) \in U_f$, and some constant $f \geq 0$.

LEMMA 5. *Let π be a cycle over U which is not an equifitness subset of U. Let f_π be the minimal value of $f(a)$ over π and $a \in \pi$ such that $f(a) = f_\pi$. Then, there exists $b \in \pi$, $f(b) > f_\pi$, such that*

$$H(a, b) = W(a) + V_U(a, b). \tag{6.8}$$

PROOF. The proof is by induction on the number of elements in π. First, assume that $\#\pi = 2$, that is, $\pi = \{a, b\}$. By definition, $H(a, b) = W(c) + V_U(c, d)$, for some $c \neq d \in U$. Since $H(a, b) \leq \lambda$ for some $\lambda \geq 0$, we have $H(c, d) \leq \lambda$ and $c, d \in \pi$.

Now, take a cycle π of arbitrary size and consider the minimal λ such that $H(a, b) \leq \lambda$ for all $a, b \in \pi$. Let

$$U_{f_\pi} = \{a \in \pi;\ f(a) = f_\pi\}, \tag{6.9}$$

and consider all subcycles in π. The following dichotomy holds:

- Either there exists a strict subcycle which intersects U_{f_π} and which is not equifitness. Then, the induction argument applies.
- Or U_{f_π} is a subcycle for all $\mu < \lambda$.

In the second case, the chain exits optimally from U_{f_π} following an edge $a \to b$ where $a \in U_{f_\pi}$, $b \notin U_{f_\pi}$ and

$$H(a, b) = W(a) + V_U(a, b) = \lambda. \tag{6.10}$$

□

THEOREM 7. *Let $n > n_*$. Over U, a cycle either intersects A^* or consists of an equifitness subset of U.*

PROOF. Let π be a cycle over U which is not an equifitness subset of U. Then, by lemma 5, there exists two vertices a and b in π such that

$$V_U(a, b) \geq n \tag{6.11}$$

[at least n simultaneous mutations are necessary to go from (a) to (b)], and

$$H(a, b) = V_U(a, b) + W(a) \leq \lambda \tag{6.12}$$

for some $\lambda > 0$. Now, take $a^* \in A^*$ such that $d(a, a^*) = d(a, A^*)$. By lemma 2, we have $V_U(a, a^*) = d(a, a^*)$. Hence

$$H(a, a^*) \leq W(a) + V_U(a, a^*) = W(a) + d(a, a^*) \leq W(a) + n_*. \tag{6.13}$$

Moreover,

$$W(a) + n_* \leq W(a) + V_U(a, b) \leq \lambda. \tag{6.14}$$

Putting together equations (6.13) and (6.14) shows that $(a^*) \in \pi$. □

6.3. *Optimal cooling schedules.*

THEOREM 8. *Let $n > n_*$. We have $H_1 = D_*$.*

The proof of this result requires a preliminary lemma.

LEMMA 6. *Let π be a cycle and an equifitness subset of U. Then, the virtual energy is constant over π.*

PROOF. Take a cycle π such that $\#\pi \geq 2$. Assume that ℓ populations have the same virtual energy $\ell \leq \#\pi$, and let π_ℓ denote this subset in π. Now, let $(a) \in \pi \setminus \pi_\ell$ and $(b) \in \pi_\ell$, such that $d(a, b)$ is minimal. Then, by Lemma 4,

$$H(a, b) = W(a) + d(a, b) = H(b, a) = W(b) + d(b, a). \tag{6.15}$$

Then, we have $W(a) = W(b)$ and π is equifitness. □

PROOF OF THEOREM 8. According to Theorem 7 and the definition of H_1, we have

$$H_1 = \max_{a \notin A^*} \min_{a^* \in A^*} \{H(a, a^*) - W(a)\}. \tag{6.16}$$

(In this definition, we have used Lemma 6.) In order to prove that $H_1 \leq D_*$, let $a \notin A^*$, $a^* \in A^*$ and consider all self-avoiding trajectories in U defined as

$$\gamma\colon (a^0) = (a) \to (a^1) \to \cdots \to (a^r) = (a^*), \tag{6.17}$$

where each edge $a^i \to a^{i+1}$ has been taken so that $f(a^i) \geq f(a^{i+1})$ and there exists a path $\gamma_{a^i a^{i+1}}$ having the property (P). In addition, assume that $\gamma_{a^i a^{i+1}}$ denotes a minimal path. Then for every trajectory, we have

$$H(a, a^*) \leq \max_{i=0,\ldots,r-1} \{W(a^i) + \text{length}(\gamma_{a^i a^{i+1}})\} \tag{6.18}$$

and

$$H_1 \leq D_*. \tag{6.19}$$

To prove the reverse inequality, Proposition 14 of Catoni (1997) can be used. According to this (classical) result, one has

$$H_0 \leq \max_{a \notin A^*} \min_{a^* \in A^*} \{H(a, a^*) - W(a)\}. \tag{6.20}$$

The left-hand side equals D_* by Theorem 6.

The proof of Theorem 2 follows from Theorem 4.

6.4. *Optimal convergence exponent.*

THEOREM 9. *Let $n > n_*$. We have*

$$\frac{n - n_*}{D_*} \leq \alpha_* \leq \frac{n(n_* + 1 - D_*) + n_* - 1 - D_*}{D_*}.$$

PROOF. In light of Theorem 7 and Lemma 6, we have

$$\alpha_* = \min_{b \notin A^*} \left\{ \frac{W(b) - W_*}{H_e(b)} \right\}. \tag{6.21}$$

Hence, by the definition of H_1, we have

$$\alpha_* \geq \frac{\min_{b \notin A^*} W(b) - W_*}{H_1} \tag{6.22}$$

Let $b_* \notin A^*$ be such that $W(b_*) = \min_{b \notin A^*} W(b)$, and consider a b_*-graph that realizes the minimum in the former definition. From this graph, create an a^*-graph ($a^* \in A^*$), say g, by deleting the edge $a^* \to a$ and adding the edge $b_* \to a^*$. By Lemma 2, we obtain

$$V_U(g) - W(b_*) = d(b_*, a^*) - V_U(a^*, a) \geq W_* - W(b_*) \tag{6.23}$$

and hence

$$W(b_*) - W_* \geq V_U(a^*, a) - d(b_*, a^*) \geq n - n_*. \tag{6.24}$$

This establishes that

$$\alpha_* \geq \frac{n - n_*}{D_*}. \tag{6.25}$$

To compute the upper bound on α_*, notice that

$$\alpha_* \leq \frac{W(b^*) - W_*}{H_1} \tag{6.26}$$

where b^* is such that $H_e(b^*) = H_1$. Now create a b^*-graph, say g', by deleting the edge $b^* \to b$ and adding the edge $a^* \to b^*$ in an a^*-graph which realizes the minimum in the definition of W_*. We obtain

$$V(g') - W_* = V(a^*, b^*) - D_* \geq W(b^*) - W_* \tag{6.27}$$

In order to obtain an upper bound on $V_U(a^*, b^*)$, assume that $D_* > 1$. Consider the path that realizes the minimum in the definition of this quantity

$$\gamma\colon (a^0) = (a^*) \to (a^1) \to \cdots \to (a^r) = (b^*). \tag{6.28}$$

We have $f(a^{r-2}) < f(b^*)$ and $f(a^{r-1}) \geq f(b^*)$. Now less than $(n_* + 1 - D_*)$ edges are needed to connect (a^*) and (a^{r-1}). In the worst case, n simultaneous mutations are need at each step. Moreover, we have

$$V(a^{r-1}, b^*) \leq n_* - 1.$$

Therefore the cost is such that

$$V(a^*, b^*) \leq n(n_* + 1 - D_*) + n_* - 1. \tag{6.29}$$

This proves that

$$\alpha_* \leq \frac{n(n_* + 1 - D_*) + n_* - D_* - 1}{D_*}. \tag{6.30}$$

If $D_* = 1$, this inequality can obviously be improved as follows

$$\alpha_* \leq nn_*. \tag{6.31}$$

Again, the proof of Theorem 3 follows from Theorem 4.

7. Computer implementation and examples.

7.1. *Computer implementation.* A slightly different version of the algorithm proposed in Section 2 has been implemented. This version relies upon the binomial sampling of the number of offspring by mutation. Mutation is applied to the first labelled individuals in the population. Informally, the algorithm is as follows.

- Initialize a population x of n labelled individuals in E
- Repeat
  1. Draw a random number N from the binomial distribution $bin(n, p)$
  2. Select the best individual $\hat{x}$ from the population
  3. Create offspring $x'_1, \ldots, x'_N$ of the N first individuals $x_1, \ldots, x_N$, and replace the parents by their offspring. Replace the $n - N$ remaining individuals by $\hat{x}$
  4. Decrease the mutation parameter p

  until some stopping criterion is met.

The probability of a transition $q_M(x; y)$ from population x to $y \in X$ is given by

$$q_M(x, y) = \mathrm{P}\big(N = V_1(x, y)\big)\pi_M(x, y). \tag{7.1}$$

where $\pi_M(x, y) = 0$ if $y_i \notin N(x_i) \cap (E \backslash \{\hat{x}\})$ for some $i \leq V_1(x, y)$ or $y_i \neq \hat{x}$ for some $i > V_1(x, y)$. Otherwise, $\pi(x, y)$ is a positive number (independent of p) that corresponds to the choices of the neighbors on the exploration graph. The new Markov chain still satisfies large deviations estimates. For some positive constant κ_M, we have

$$\frac{1}{\kappa_M}\pi_M(x, y)e^{-V_1(x,y)/T} \leq q_M(x, y) \leq \kappa_M \pi_M(x, y)e^{-V_1(x,y)/T} \tag{7.2}$$

where the cost function V_1 is the same as for the transition kernel q. As a consequence, all results stated in Section 3 hold for this version as well.

A merit of this way of programming is that the actual population size is $N + 1$. Using this version allows saving memory space, as useless assignments of $\hat{x}$ to high-labeled individuals can be avoided. Furthermore, the main feature of this algorithm is that it provides some kind of hierarchical search. Such ideas have been considered in the past [see, e.g., Dawson (1987)], but do not seem often used within the three streams of evolutionary computation. Individuals perform different 'degrees' of search according to their position in the population. For large population sizes, the first labeled individuals may travel along the search space with very weak selection pressure, due to the standard fluctuations of the binomial distribution. On the other hand, the individuals with labels close to np perform a local search around the best individual, and the selection pressure is strong. This version has been experimented by Francois (1998) and turns to be computationally more efficient than the one proposed in Section 2. It has been preferred for running the numerical simulations presented in the next section.

7.2. *Comparisons with simulated annealing.* Given two algorithms A and B and a class of test functions, a crucial question is to determine whether algorithm A is better than algorithm B or not. By the "no free lunch" theorem of Wolpert and Macready (1997), the decision may be difficult when the test class is too broad. Our approach allows comparisons between algorithms on the basis of optimal convergence exponents, since these quantities give estimates of the success probability for the algorithms. Thanks to this constant, classes of problems can be built for which an algorithm outperforms (at least asymptotically) the other. To carry out this program, explicit estimates on the convergence exponent must be available.

In this section, we compare the E/S algorithm, denoted by A, with n simulated annealing running in parallel, denoted by B, where n is the population size of A. The optimal exponent of algorithm B is lower than $n(\Delta - h_*)/h_*$ where h_* is Hajek's critical height [Hajek (1988)] and

$$\Delta = \max\{|f(a) - f(b)|;\ a, b \in E\}. \tag{7.3}$$

In light of Theorem 3, algorithm A should be preferred to algorithm B if

$$h_* > (\Delta - h_*)D_*, \tag{7.4}$$

and n is chosen so that

$$n\left(1 - \frac{(\Delta - h_*)D_*}{h_*}\right) > n_*. \tag{7.5}$$

To assess the value of such a claim, numerical simulations have been performed on a very simple test problem, for which the critical constants h_*, n_* and D_* are easy to compute. The test problem is defined on the search space $E = \mathbb{Z}/16\mathbb{Z}$, and the exploration operator acts as a random walk on this space (a neighborhood

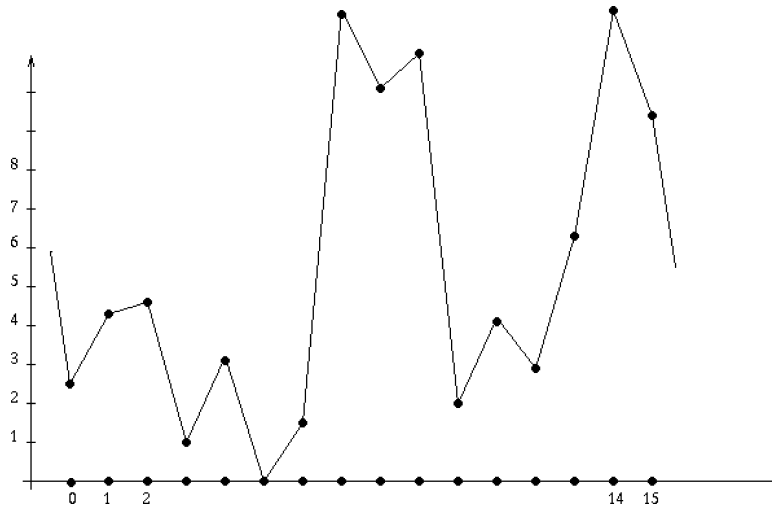

FIG. 1. *Objective values for the first test problem defined on the cyclic group with 16 elements.* $f(0) = 2.5$, $f(1) = 4.3$, $f(2) = 4.6$, $f(3) = 1.0$, $f(4) = 2.9$, $f(5) = 0.0$, $f(6) = 1.5$, $f(7) = 12.0$, $f(8) = 10.2$, $f(9) = 11.0$, $f(10) = 2.0$, $f(11) = 4.1$, $f(12) = 2.8$, $f(13) = 6.3$, $f(14) = 12.0$, $f(15) = 9.4$.

consists of the right and left vertices). The values of the objective function are given in Figure 1. We have

$$h_* = 10.0, \quad n_* = 7, \quad D_* = 4 \tag{7.6}$$

and

$$1 - \frac{(\Delta - h_*)D_*}{h_*} = 0.196. \tag{7.7}$$

Therefore, n has been taken equal to 40, so that equation (7.5) can be checked.

The probability of hitting the best solution $a^* = 5$ has been estimated from a (huge) number of runs. The population has been started from the vertex 10, which corresponds to the less favorable situation. First of all, note that the infinite temperature strategy is efficient, because of the low complexity of the minimization problem. As far as E/S can be modified to become elitist, A and B are equivalent in this infinite parameter setting. With 10 iterations, we have about a 0.75 probability of hitting a^*, and this probability increases to 0.99 with 10 additional iterations.

Results concerning six non-trivial temperature schedules are displayed in Tables 1 and 2. First, the temperature has been kept fixed to the value $T = 0.83$ (a mutation probability of 0.3). The results show that algorithm A has a 0.96 probability of reaching the absolute minima within 25 iterations, while the independent simulated annealing B has almost no chance to do so. With the temperature $T = 0.33$ ($p = 0.05$), algorithm A has a 0.99 probability of

TABLE 1
Success probability for the exploration/selection algorithm (population size $n = 40$), varying with inverse temperature schedules and the number of iterations on the first test problem. The results were averaged over 1,000 runs

	$-\log(0.3)$	$-\log(0.05)$	β_1	β_2	β_3	β_4
20	0.88	–	0.65	0.98	0.66	0.98
25	0.97	0.12	0.92	1	0.78	1
50	1	0.30	0.97	1	0.89	1
100	1	0.61	1	1	0.91	1
500	1	0.99	1	1	1	1

finding a^* while this probability is neglectible for algorithm B. Analogous results are reported when decreasing schedules were experimented. Consider

$$\forall t \geq 1, \qquad \beta(t) = 1/T(t). \tag{7.8}$$

The logarithmic schedules $\beta_1(t) = \log(t)/4 + 1.2$ and $\beta_2(t) = \log(t)/10 + 0.5$ have been tested first. With β_1, the absolute minimum is attained within 50 iterations of algorithm A with probability 0.97, while this probability is 0 for algorithm B even with 500 iterations. The second schedule seems more favorable to B, but A is still more efficient. The same phenomenon occurs with linear schedules $\beta_3(t) = 0.83 + 0.05t$, and $\beta_4(t) = 0.51 + 0.02t$. With β_3, the absolute minimum is attained within 100 iterations of algorithm A with probability 0.91, while this probability is 0.0 for algorithm B even with 500 iterations. With β_4, the probability increases to 0.15 for algorithm B, but A is again better. Note that the number of function evaluations is significantly lower in algorithm A's runs than in B's. As the evaluation of the objective function can be considered as the major source of complexity in solving the minimization problem, then E/S strategies reveal themselves computationally more efficient than n simulated annealing running in parallel.

TABLE 2
Success probability for 40 simulated annealing running in parallel, varying with inverse temperature schedules and the number of iterations on the first test problem. The results were averaged over 1,000 runs

	$-\log(0.3)$	$-\log(0.05)$	β_1	β_2	β_3	β_4
20	–	–	–	–	–	–
25	–	–	–	–	–	–
50	–	–	–	–	–	–
100	0.001	–	0.0	0.19	0.0	0.14
500	0.03	0.0	0.0	0.31	0.0	0.15

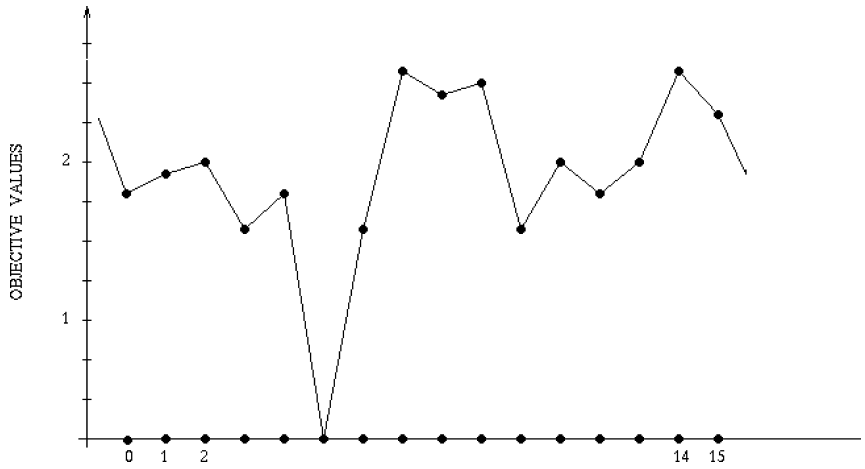

FIG. 2. *Objective values for the second test problem defined on the cyclic group with* 16 *elements.* $f(0) = 1.8$, $f(1) = 1.9$, $f(2) = 2.0$, $f(3) = 1.6$, $f(4) = 1.8$, $f(5) = 0.0$, $f(6) = 1.6$, $f(7) = 2.6$, $f(8) = 2.4$, $f(9) = 2.5$, $f(10) = 1.6$, $f(11) = 2.0$, $f(12) = 1.8$, $f(13) = 2.0$, $f(14) = 2.6$, $f(15) = 2.3$.

Analogous arguments apply to decide when B may outperform A. Algorithm B should be preferred when

$$\frac{n(n_* + 1 - D_*) + n_* - D_* - 1}{D_*} < \frac{n\delta}{h_*} \tag{7.9}$$

with

$$\delta = \min\{f(a) - f(a_*);\ a \neq a_* \in E\}. \tag{7.10}$$

To assess the value of this claim, numerical simulations have been performed for a second test problem defined on the same search space (Figure 2). For this problem, the critical constants h_*, n_* and D_* are

$$h_* = 1.0, \quad n_* = 7, \quad D_* = 4, \tag{7.11}$$

and $n = 8$, so that equation (7.9) can be checked. The same six schedules have been studied for this problem, except for β_2 which has been changed to $\beta_2(t) = 0.5 + \log(t)$, taking into account the rescaling of function f.

The results reported in Tables 3 and 4 show that 8 simulated annealing performs better than the E/S algorithm with size $n = 8$. For constant schedules, the success probability (of hitting $a^* = 5$) is close to 1 within 100 iterations of algorithm B. For algorithm A, this probability is smaller (almost 0 for $p = 0.05$). Algorithm B significantly outperforms algorithm A for the other schedules. This results show that condition (7.9) have some value although it is based on rough estimates.

TABLE 3
Success probability for 8 *simulated annealing running in parallel, varying with inverse temperature schedules and the number of iterations on the second test problem. The results were averaged over* 1,000 *runs*

	$-\log(0.3)$	$-\log(0.05)$	β_1	β_2	β_3	β_4
20	0.68	0.26	0.60	0.49	0.66	0.78
25	0.82	0.41	0.73	0.61	0.81	0.88
50	0.99	0.84	0.99	0.87	0.97	1
100	1	0.99	1	0.97	0.99	1

APPENDIX

Proof of Theorem 5. The proof generalizes a previous result stated in Francois (1998). It is given here for sake of completeness. According to Lemma 2, we have

$$V_U(a, a^*) = d(a, a^*) \tag{A.1}$$

for all $a^* \in A^*$, and $a \neq a^*$.

The Markov transition matrix q which is associated to the algorithm satisfies the classical irreducibility condition

$$\forall (x, y) \in X \times X, \qquad \exists r \geq 1, \quad q^{(r)}(x, y) > 0, \tag{A.2}$$

and, the Markov chain $(X(t))$ converges to the stationary distribution μ_T, as t goes to infinity.

To prove that $A^* \subset \mathcal{W}^*$, consider $a^* \in \mathcal{W}^*$ and the a^*-graph, say g_*, which realizes

$$V_U(g_*) = W_*. \tag{A.3}$$

Since the graph is minimal, an edge which starts from the uniform population (a) ends at a uniform population (b) such that

$$V_U(a, b) = \min_{b' :\, f(b') \leq f(a)} d(a, b'). \tag{A.4}$$

TABLE 4
Success probability for the exploration/selection algorithm (population size $n = 8$*), varying with inverse temperature schedules and the number of iterations on the second test problem. The results were averaged over* 1,000 *runs*

	$-\log(0.3)$	$-\log(0.05)$	β_1	β_2	β_3	β_4
20	–	–	–	–	–	0.36
25	0.24	–	0.05	–	0.17	0.49
50	0.54	–	0.13	–	0.19	0.73
100	0.82	0.01	0.19	0.02	0.23	0.79

Since g_* contains no cycle, there exists an edge $a_* \to b_*$ such that $a_* \in A^*$ and $b_* \in \mathcal{W}^*$. Now, create a a_*-graph, say g, by reversing the path from a_* to a^* in g_*. We have

$$V(g) = W_* - V(a_*, a^*) + V(a^*, a_*) = W_*, \tag{A.5}$$

and $a_* \in \mathcal{W}^*$.

Now, let (a) be a uniform population such that $a \notin A^*$. We have

$$n > \max_{a \notin A^*} \{d(a, A^*)\}. \tag{A.6}$$

Hence, by Lemma 2, there exists an $a^* \in A^*$ such that

$$n > V_U(a, a^*). \tag{A.7}$$

On the other hand, at least n simultaneous mutations are required to exit from the subset of minimal populations

$$\forall b \notin A^*, \quad V_U(a^*, b) \geq n. \tag{A.8}$$

Let g be a a-graph on U for which

$$V_U(g) = W(a) = \sum_{(u \to v) \in g} V_U(u, v). \tag{A.9}$$

Since $a \notin A^*$ and g is a spanning tree on U rooted at a, there must be a population $b \in U$ such that $(a^* \to b)$ is in g. We build an a^*-graph by deleting the edge $(a^* \to b)$ in g and introducing the edge $(a \to a^*)$. Thus, we have

$$W(a^*) \leq W(a) + V_U(a, a^*) - V_U(a^*, b). \tag{A.10}$$

Thus, we have

$$\forall a \notin A^*, \quad W(a) > W(a^*), \tag{A.11}$$

and $\mathcal{W}^* \subset A^*$.

REFERENCES

AARTS, E. H. L and KORST, J. H. M. (1988). *Simulated Annealing and Boltzmann Machines.* Wiley, New York.

BÄCK, T. (1996). *Evolutionary Algorithms in Theory and Practice.* Oxford Univ. Press.

CATONI, O. (1997). Simulated annealing algorithms and Markov chains with rare transitions. Lectures notes, Univ. Paris XI.

CERF, R. (1994). Une théorie asymptotique des algorithmes génétiques. Ph.D. thesis, Montpellier II.

CERF, R. (1996a). The dynamics of mutation-selection algorithms with large population sizes. *Ann. Inst. H. Poincaré Probab. Statist.* **32** 455–508.

CERF, R. (1996b). A new genetic algorithm. *Ann. Appl. Probab.* **6** 778–817.

CHAKRABORTY, U. K., KALYANMOY, D. and CHAKRABORTY, M. (1996). Analysis of selection algorithms: a Markov chain approach. *Evol. Comput.* **4** 133–167.

DAVIS, T. E. and PRINCIPE, J. C. (1991). A simulated annealing like convergence theory for the simple genetic algorithm. In *Proceedings of the Fourth International Conference on Genetic Algorithms* (R. K. Belew and L. B. Booker, eds.) 174–181. Morgan Kauffman, San Mateo, CA.

DAWSON, D. A. (1987). Stochastic models of parallel systems for global optimization. *IMA Vol. Math. Appl.* **9** 25–44.

DEL MORAL, P. and MICLO, L. (1999). On the convergence and the applications of the generalized simulated annealing. *SIAM J. Control Optim.* **37** 1222–1250.

FOGEL, D. B. (1995). *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence.* IEEE Press, Piscataway, NJ.

FRANÇOIS, O. (1998). An evolutionary strategy for global minimization and its Markov chain analysis. *IEEE Trans. Evolutionary Comput.* **2** 77–90.

FREIDLIN, M. I. and WENTZELL, A. D. (1984). *Random Perturbations of Dynamical Systems.* Springer, New York.

GOLDBERG, D. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, MA.

HAJEK, B. (1988). Cooling schedules for optimal annealing. *Math. Oper. Res.* **13** 311–329.

NIX, A. and VOSE, M. (1992). Modeling genetic algorithms with Markov chains. *Ann. Math. Artificial Intelligence* **5** 79–88.

RABINOVITCH, Y. and WIGDERSON, A. (1999). Techniques for bounding the convergence rate of genetic algorithms. *Random Structures Algorithms* **14** 111–137.

RUDOLPH, G. (1994). Convergence analysis of canonical genetic algorithms. *IEEE Trans. Neural Networks* **5** 96–101.

SUZUKI, J. (1993). A Markov chain analysis on a genetic algorithm. In *Proceedings of the Fifth International Conference on Genetic Algorithms* (S. Forrest, ed.) 146–153. Morgan Kauffman, San Mateo, CA.

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

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

WOLPERT, D. H. and MACREADY, W. G. (1997). No free lunch theorems for optimization. *IEEE Trans. Evolutionary Comput.* **1** 67–82.

TIMC
FACULTÉ DE MÉDECINE
F38706 LA TRONCHE CEDEX
FRANCE
E-MAIL: Olivier.Francois@imag.fr