# An Evolutionary Strategy for Global Minimization and Its Markov Chain Analysis

Olivier François

***Abstract*— The mutation-or-selection evolutionary strategy (MOSES) is presented. The goal of this strategy is to solve complex discrete optimization problems. MOSES evolves a constant sized population of labeled solutions. The dynamics employ mechanisms of mutation and selection. At each generation, the best solution is selected from the current population. A random binomial variable $N$ which represents the number of offspring by mutation is sampled. Therefore the $N$ first solutions are replaced by the offspring, and the other solutions are replaced by replicas of the best solution. The relationships between convergence, the parameters of the strategy, and the geometry of the optimization problem are theoretically studied. As a result, explicit parameterizations of MOSES are proposed.**

***Index Terms*— Convergence, evolutionary strategy, genetic algorithms, Markov chains, large deviations, simulated annealing.**

## I. Introduction

EVOLUTIONARY algorithms are global search procedures based upon the evolution of a vector of solutions viewed as a population of interacting individuals. These strategies include simulated annealing, genetic algorithms, evolutionary programming, and simulated evolution [1], [2], [9], [12]. In applying evolutionary strategies to solve large scale and complex optimization problems, one of the most frequently encountered difficulties is convergence toward an undesired attractor. This phenomenon occurs when the population get trapped in a suboptimal state such that the variation operators cannot produce an offspring which outperforms its parents. The relationships between convergence to a global minimum, the parameters of the strategy such as population size or mutation probabilities, and the geometry of the minimization problem are crucial to understand. Many previous studies have investigated such issues for the simulated annealing process [13] and parallel versions [19], the genetic algorithm [4], [5], [12], [15], [17], and the evolution strategies [2]. This work theoretically investigates a simple evolutionary strategy called MOSES (for mutation-or-selection evolutionary strategy) whose purpose is to minimize an arbitrary function on a finite set. MOSES was introduced in [10]. The dynamics employ mechanisms of mutation and selection. A graph structure on the search space defines the allowed mutations. Mutation actually acts as a random walk on this graph. The geometry of the minimization problem is therefore characterized by the values of the objective function on the vertices of the mutation graph.

The strategy relies on two parameters. The first is the size of the population, and the second is a positive temperature that can be imagined as decreasing to zero (as in simulated annealing). The temperature controls the mean number of offspring by mutation. This paper focuses on the choice of the population size and the "cooling" schedules which ensure that the strategy produces an optimal solution. This work demonstrates that the parameters of the algorithm can be chosen so that the dependence on the geometry is weak. In fact, the parameters can be configured with constants depending only on the mutation graph (which is generally predefined). Things are less favorable as far as simulated annealing or genetic algorithms are concerned [4], [13]. In these procedures, the crucial constants strongly depend on the global structure of the minimization problem and are thus unavailable in most practical cases. The theory developed in this paper is analogous to the theory of simulated annealing and is based upon the results of [3], [5], [13], and [19]. The statement of the main result is composed of two parts. First, the population sizes for which the evolution "concentrates" on the optimal solution are determined (Theorem 4.1). Once the population size is fixed, the choice of reasonable cooling schedules is investigated in the spirit of simulated annealing (Theorem 4.3). Section II overviews some results known for simulated annealing and mutation-selection genetic algorithms. Section III presents the strategy in a formal way and gives basic results and notations. Section IV introduces two geometrical indexes which help to determine whether convergence may hold. Section IV also emphasizes the main results and introduces the mathematical formalism which is necessary to prove them. Section V completes the theoretical results with simulation evidence.

## II. State of the Art and Presentation of the Algorithm

The problem to solve in this paper is to find the minimal point of an arbitrary function $f$ which is defined on a finite but generally large set $E$. The set $E$ is endowed with a graph structure. This structure defines a neighborhood for each vertex in $E$. The values of the function on the vertices of the graph describe the *geometry* of the minimization problem. An evolutionary strategy uses a vector (population) of solutions to the minimization problem. Each solution is regarded as an individual which performs a local search on the graph. The premise of evolutionary computation is that cooperative searchers are more efficient than isolated ones. The basic framework for an evolutionary strategy is as follows [9].

Manuscript received December 5, 1997; revised March 10, 1998, May 12, 1998, and July 15, 1998. This work was supported by the MAI-IMAG project.

The author is with the Laboratoire de Modélisation et Calcul, 38041 Grenoble Cedex 9, France.

Publisher Item Identifier S 1089-778X(98)08934-6.

1) Initialize a "population" of solutions in $E$.
2) Evaluate each solution in the population.
3) Propose a number of random changes in the population.
4) Use a rejection criterion to validate each change and evaluate the new solutions.
5) If a stopping criterion is satisfied, return the best solution; if not go to step 3).

Numerous algorithms correspond to this description including simulated annealing, genetic algorithms, and evolutionary programming [1], [9], [12]. Evolution strategies introduced by Rechenberg and Schwefel [2] also fit well to the previous framework, although these techniques are rather devoted to continuous optimization. Since randomness arises at each generation, all these algorithms are Markovian. It is natural to use the formalism of Markov chains to analyze their behavior [13], [15]–[18].

### A. Simulated Annealing and Genetic Algorithm

In the simulated annealing algorithm, the population reduces to a single individual. A potentially new solution is generated by sampling over the neighbors of the current one [step 3)]. The rejection criterion [step 4)] uses Metropolis dynamics with a decreasing temperature schedule. The Markov chain analysis of the algorithm shows that the probability for obtaining the minimal solution converges to 1.0 iff the temperature schedule $(T(t))_{t\geq 1}$ satisfies Hajek's condition [13]

$$\sum_{t\geq 1} \exp(-h_*/T(t)) = \infty \tag{1}$$

where $h_*$ is a constant called critical height. The critical height is a geometrical index which expresses the difficulty for the simulated annealing to find the global minima of the objective function. To interpret $h_*$, it is worth regarding the objective function as an energy. Then this constant is the smallest variation of energy which is necessary to exit from any suboptimal solution in the energy landscape. Unfortunately, this value remains incomputable in practice, because it depends strongly on the geometry of the minimization problem, and assumes a complete knowledge of the energy landscape.

Genetic algorithms proceed by sequentially applying mutation, crossover [step 3)], and selection [step 4)] operators [12]. The links between the geometry of the minimization problem and the convergence properties of the algorithms are not well understood yet. Nevertheless, the results of [5], [7], [12], [15], [17], and [21] can be summarized as follows. Mutation is a crucial step to warrant that the population does not get trapped into a suboptimal state. The mutation parameters, however, must be tuned in a subtle way depending on the problem to minimize. No rule of thumb actually exists to choose these mutation parameters properly in general. On the other hand, crossover is not a necessary feature for convergence to the minimal solution. Although the importance of this operator is often asserted [4], [12], it is relevant to study algorithms which are based on mutation and selection solely. Furthermore, the genetic algorithm without crossover is not limited to code binary strings. As far as mutation is well defined (according to a mutation graph), the algorithm can proceed with arbitrary discrete variables. Many authors have proposed a simulated annealing-like approach to genetic algorithms. In [7], the mutation probability is assumed to converge to zero. A natural way to parameterize this mutation probability is

$$p_T = \exp(-\alpha/T) \tag{2}$$

where $\alpha > 0$ is the intensity of mutation and $T$ is a positive temperature. This parameterization has also been used in [5] but with selection (roulette wheel) reinforced as well. In this setting, the selection probabilities are parameterized with the same temperature $T$. If the current population is equal to $x = (x_1, \ldots, x_n)$, then the probability that the individual $x_i$ is selected in the future generation is

$$\frac{\exp(\beta f(x_i)/T)}{\sum_{j=1}^{n} \exp(\beta f(x_j)/T)} \tag{3}$$

where $\beta > 0$ is the selection intensity. The genetic algorithm has been studied by using the formalism of large deviations [5], [11]. The results obtained in [5] can be restated as follows.

i) Under mild assumptions, there exists a critical population size below which the population gets definitively trapped into a suboptimal solution. If the graph is connected, the critical size is finite. More precisely, a sufficient condition on the population size $n$ for the concentration of the population on a global minima is

$$n > \frac{\alpha D + \beta(D-1)\Delta}{\min(\alpha, \beta\delta)} \tag{4}$$

where $D$ is the diameter of the graph used for mutation (to be redefined later)

$$\Delta = \max\{|f(a) - f(b)|; a, b \in E\} \tag{5}$$

and

$$\delta = \min\{|f(a) - f(b)|; a \neq b \in E\}. \tag{6}$$

ii) The probability for obtaining the minimal solution converges to 1.0 under a condition which is analogous to (1). Unfortunately, the critical constant corresponding to $h_*$ remains unknown.

Equation (4) emphasizes the role of the diameter of the mutation graph. Because (4) is linear in $D$, the bound is sensitive to large diameters. This fact means that large populations are necessary to deal with wide search spaces. The dependence on $\Delta$ and $\delta$ means that the bound is also sensitive to rescaling the objective function. The values of $\Delta$ and $\delta$ are seldom available, and the relevance of such a result to practical situations is weak.

### B. Presentation of the Algorithm

This article studies a new procedure called MOSES. This procedure shares similarities with the genetic algorithm without crossover. It can eventually be introduced into a genetic code instead of the mutation/selection operations. The main difference between MOSES and genetic algorithms is that MOSES proceeds by coupling mutation or selection into a single operation while genetic algorithms use mutation and

then selection. An informal description of the algorithm is presented now. The formal description will be given in Section III. The size of the population is equal to $n$. A parameter $p$ is introduced which controls the number of offspring of mutation at each generation. This parameter may depend on the generation. Because $p$ is taken in the interval (0, 1), this parameter is viewed as a mutation probability. The algorithm is as follows.

1) Initialize a population of $n$ labeled individuals in $E$.
2) Repeat
   a) Draw a random number $N$ from the binomial distribution $bin(n,p)$.
   b) Select the optimal individual $x_*$ from the population.
   c) Replace the $N$ first individuals by mutation and the $n - N$ other individuals by $x_*$.
   d) Update $p$.

The main feature of MOSES is that the search is hierarchical. Individuals perform different "degrees" of search according to their position in the population. The individuals with the first labels are allowed to make long random walks in the search space, and some of them may travel along the search space with very weak selection. On the other hand, the individuals with high labels perform a very local search around the best individual, and the selection pressure is strong. This hierarchy is obtained from the use of a random number of offspring by mutation at each generation. The probability that $k$ individuals mutate is given by the binomial distribution $bin(n,p)$. For all $0 \leq k \leq n$, we have

$$\text{Prob}(N = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}. \tag{7}$$

The most expensive step in this algorithm is selection which requires $N$ calls to $f$. On the average, the number of calls is

$$E[N] = np. \tag{8}$$

Nevertheless, this number is smaller than the number of calls in many other strategies (e.g., genetic algorithms or evolution strategies). Besides, a proper choice of the mutation graph can drastically reduce the complexity of computing. Since only comparisons to a fixed value (the best value at the last generation) are needed, neighbors can be chosen so that the difference between the old value and the value of the offspring by mutation is easy to compute.

Many criteria may be used to stop the evolution. The expiration of the computing resource is an expensive criterion. In this paper, different criteria are used. The theoretical analysis assumes that the mutation parameter $p$ decreases to zero. In regard to this method, a natural criterion is to stop the algorithm when $p$ is below a given threshold. Another criterion will be used in Section V where test functions are introduced. The algorithm can be stopped when the value of the objective function is below a given threshold. The drawback of this method is the randomness of the hitting time. Due to this randomness, the control on the length of computation is lost.

MOSES also shares similarities with evolution strategies. In MOSES, $N$ parents (the $N$ first labels in the population) create offspring. Parents and offspring are merged to build the population at the future generation. Again the mutation/selection steps of an evolution strategy require two operations whereas these steps are coupled in a single operation in MOSES.

Overall, the great advantage of MOSES is that the mathematical analysis of the procedure can be described in details. The construction of the algorithm, and especially the choice of the binomial distribution, is motivated by the application of the formalism of large deviations. As in the simulated annealing procedure, a temperature is introduced. This temperature acts on the mean number of offspring of mutation $E[N]$ which decreases to zero. If the population size and the temperature are properly chosen, the evolution concentrates on a global minimum. It has been demonstrated many times that simulated annealing-like theories are relevant to study genetic algorithms [4], [7], [14]. More precisely, the theoretical framework which is used to analyze MOSES is generalized simulated annealing (GSA) [19] which provides a unified formalism for dealing with a large class of evolutionary procedures.

## III. Mathematical Description

### A. Formal Description and Hypotheses

In this section, MOSES is formally described. The necessary assumptions for conducting the mathematical analysis are also given. The objective function $f$ is defined on a set $E$ of finite cardinality. For sake of simplicity, it is assumed that $f$ is one-to-one on $f(E)$ ($f(a) = f(b)$ implies $a = b$). The unique minimal point is denoted by $a^*$. If only a solution among the best ones is wished, this restriction is a little loss of generality. Indeed, a small perturbation of $f$ does not change the minimization problem and can transform $f$ into a one-to-one function. MOSES evolves a population of randomly searching individuals which interact through selection. The size of the population is an integer

$$n \geq 2. \tag{9}$$

The set $X$ of all populations of size $n$ consists of the entire state space $E$ replicated to the $n$th degree

$$X = E^n. \tag{10}$$

For a given population $x = (x_1, \ldots, x_n) \in X$, let

$$E_x = \{x_1, \ldots, x_n\} \subset E \tag{11}$$

denote the subset of individuals contained in $x$. The individual extracted from the population during the selection step of MOSES is the minimal point in $E_x$

$$x_* = \underset{x_i \in E_x}{\arg\min}\, f(x_i). \tag{12}$$

Throughout the entire paper, the uniform population $(a, \ldots, a)$ and the element $a \in E$ are identified by denoting

$$(a) = (a, \ldots, a). \tag{13}$$

A graph structure on $E$ defines the mutations by associating to each $x_i \in E$ a neighborhood $N(x_i) \subset E$. The graph

structure is called mutation graph and is denoted by $(E, \mathcal{G})$. The mutation graph is assumed to be the following:

a) symmetric:

$$x_i \in N(x_j) \quad \text{iff } x_j \in N(x_i); \tag{14}$$

b) connected: there exists a path between two arbitrary points $a$ and $b$ in $(E, \mathcal{G})$;

c) and

$$\forall x \in X, \quad 1 \le i \le n, \quad N(x_i) \cap (E \backslash \{x_*\}) \neq \emptyset \tag{15}$$

where the notation $E\backslash\{x_*\}$ stands for the subset $E$ minus the individual $x_*$.

Symmetry is assumed for sake of convenience. The purpose of this assumption is to alleviate the notations used in the mathematical description of the algorithm. Assumption c) is technical. It is often checked in practice (e.g., 1-bit mutation in binary genetic algorithms or 2-opt in traveling salesman problems). This assumption guarantees that mutations are always possible. From the convergence viewpoint, the fundamental assumption is connectivity, which ensures that the process of evolution is ergodic.

The algorithm is parameterized with a finite temperature $T > 0$. This parameter controls the number of offspring by mutation. Let $X_t^T$ be the state of the population at time $t \ge 1$, and $X_t^T = x \in X$. The population $X_{t+1}^T$ is obtained as follows.

- A subset of offspring by mutation is created by drawing a random number $N$ according to the binomial distribution $bin(n, p_T)$ where

$$p_T = \exp(-1/T). \tag{16}$$

  Formally, this subset is

$$\begin{aligned} &I = \emptyset, \quad \text{if } N = 0 \\ &I = \{1, \ldots, N\}, \quad \text{otherwise.} \end{aligned} \tag{17}$$

- Mutation and replacement are introduced. For $i \in I$, $y_i$ is chosen in $N(x_i) \cap (E\backslash\{x_*\})$ with uniform probability. If $i \notin I$ then $y_i$ is set equal to $x_*$.
- The population at time $t+1$ is $X_{t+1}^T = y$.

The evolution is initialized with an arbitrary population in $X$.

### B. Basic Markov Chain Properties

Such dynamics correspond to a Markov chain on the set $X$ for which the transition probabilities can be formulated explicitly. Let $q_T$ be the Markov transition matrix associated to the chain $(X_t^T)_{t \ge 1}$. We have

$$q_T(x, y) = \text{Prob}\big(X_{t+1}^T = y \,\big|\, X_t^T = x\big). \tag{18}$$

For $x, y$ two populations, consider the subset of successive integers $i \in \{1, \ldots, n\}$ defined as

$$I(x, y) = \{1 \le i \le n; y_i \neq x_*\}. \tag{19}$$

The number of elements in this subset is denoted by

$$C(x, y) = |I(x, y)|. \tag{20}$$

Denote by $1_A$ the indicator function of the subset $A \subset E$. A transition between $x$ and $y$ is possible iff

$$\pi(x, y) = \prod_{i \in I(x,y)} \frac{1_{N(x_i) \cap (E\backslash\{x_*\})}(y_i)}{|N(x_i) \cap (E\backslash\{x_*\})|} \prod_{i \notin I(x,y)} 1_{\{x_*\}}(y_i) \neq 0. \tag{21}$$

In such a case, the probability of a transition from population $x$ to $y$ is given by

$$q_T(x, y) = \text{P}(N = C(x, y))\pi(x, y). \tag{22}$$

In addition, for all populations $y$ that are reachable from $x$ in a single step of the procedure, the transition probabilities satisfy the following inequalities:

$$k_T e^{-C(x,y)/T} \le q_T(x, y) \le K_T e^{-C(x,y)/T} \tag{23}$$

where

$$\begin{aligned} k_T &= (1 - p_T)^n \prod_{i \in I(x,y)} \frac{1}{|N(x_i) \cap (E\backslash\{x_*\})|} \\ K_T &= 2^n \prod_{i \in I(x,y)} \frac{1}{|N(x_i) \cap (E\backslash\{x_*\})|}. \end{aligned} \tag{24}$$

These inequalities are immediate from the definition of the binomial distribution. Indeed, we have

$$(1 - p_T)^n \le \binom{n}{k}(1 - p_T)^{n-k} \le \max_k \binom{n}{k} \le 2^n. \tag{25}$$

Besides, if $p_T \le 1/2$, then $(1 - p_T)^n \ge (1/2)^n$, and

$$\frac{1}{\kappa}\pi(x, y)e^{-C(x,y)/T} \le q_T(x, y) \le \kappa\pi(x, y)e^{-C(x,y)/T} \tag{26}$$

with $\kappa = 2^n$. These estimates are the starting point to develop the analysis of the algorithm and apply the formalism of large deviations. By convention, set $C(x, y) = +\infty$ when $q_T(x, y) = 0$. The constants $k_T$ and $K_T$ satisfy

$$\lim_{T \to 0} -T \ln k_T = \lim_{T \to 0} -T \ln K_T = 0. \tag{27}$$

The transition probabilities $q_T(x, y)$ are therefore logarithmically equivalent to $e^{-C(x,y)/T}$ as $T$ goes to zero

$$\lim_{T \to 0} -T \ln q_T(x, y) = C(x, y). \tag{28}$$

The exact transition model which is given by (22) is hardly tractable, and the analysis will rely on the approximation given in (23). The quantity $C(x, y)$ represents the number of individuals in the population $y$ which are different from the best individual in $x$. During the analysis, this quantity is viewed as a one-step communication cost between populations $x$ and $y$. It expresses the difficulty for the chain $(X_t^T)$ to move from $x$ to $y$ in one step. This cost satisfies the following properties:

i) for all $x \in X$

$$C(x, (x_*)) = 0; \tag{29}$$

ii) for all $x \in X$, $y \neq (x_*)$

$$C(x, y) > 0; \tag{30}$$

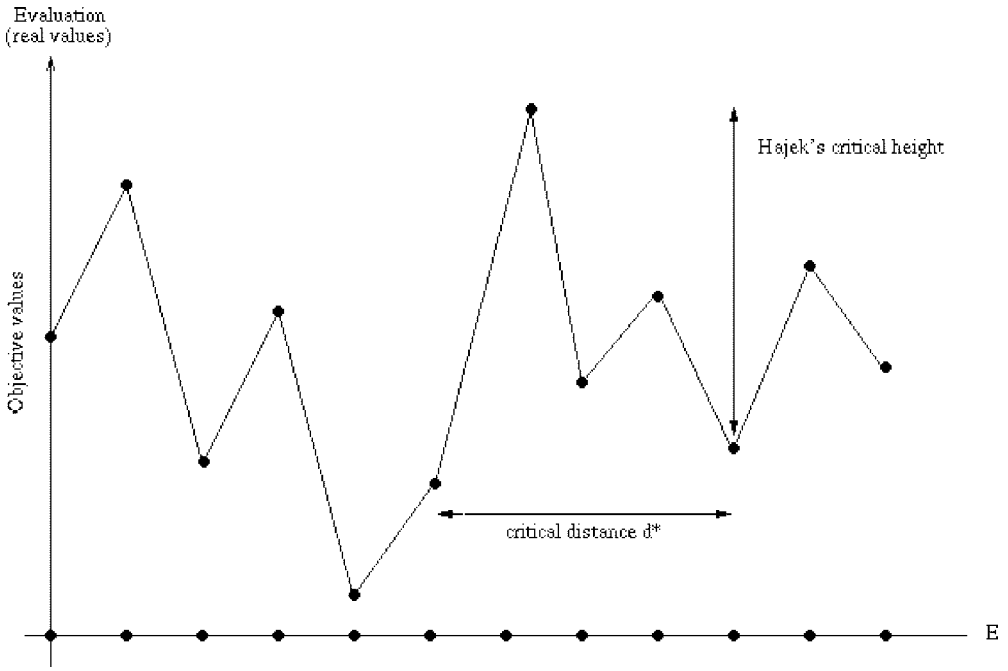


Fig. 1. A very simple example of minimization problem. The state space $E$ is depicted as the set of dots, and the mutation graph is determined by the nearest neighbors on the line. In this example, $n_* = 7$ and $d_* = 4$. The value of Hajek's critical height is also shown.

iii) by condition (15), for all $(x, y) \in X \times X$, there exists a sequence $(x_0 \to x_1 \to \cdots \to x_r)$ such that

$$x_0 = x, \quad x_k \in X, \quad x_r = y \tag{31}$$

and

$$\sum_{k=0}^{r-1} C(x_k, x_{k+1}) < \infty. \tag{32}$$

Such a sequence is referred to as a path in the sequel.

## IV. Convergence to the Minimal Population

### A. Geometrical Indexes

This section introduces two geometrical indexes which are useful to quantify the convergence of MOSES toward the minimal solution. A geometrical index is a quantity expressed in terms of the objective function $f$, and in terms of variables depending on the mutation graph.

*Notation:* The distance on the graph, i.e., the minimal length of a path between two arbitrary points $a$ and $b$ in $(E, \mathcal{G})$, is denoted by $d(a, b)$.

The first index is an index of population size. It is defined as

$$n_* = \max_{a \neq a^*} d(a, a^*) \tag{33}$$

where $a^*$ is the minimal point of $f$. This index corresponds to the maximal distance between the minimal point of $f$ and an arbitrary point in the graph $(E, \mathcal{G})$. Initializing the algorithm with a state which is very far from the minimum plays in disfavor of convergence. Intuitively, large population sizes can attenuate this matter. If the population size is greater than $n_*$, it is possible to create a chain of individuals from any solution to the best solution, avoiding mutation to suboptimal states.

The second index is a convergence index. Again, it quantifies the difficulty of the algorithm to deal with the geometry of the minimization problem. Formally, it is defined as

$$d_* = \max_{a \neq a^*} \min_{b: f(b) < f(a)} d(a, b). \tag{34}$$

Clearly, the maximum in (34) is attained at a vertex $a$ which presents a local minimum of the function $f$. Actually, this index measures the greatest distance between a local minimum of $f$ and a solution which outperforms this minimum. The constant $d_*$ can be regarded as a measure of the "chance" of escape from local minima during the local search. Such measure has been called mutation order by Suzuki [18], who introduced a similar quantity in the context of modified elitist genetic algorithms (a less general setting).

An important feature is that the two indexes $n_*$ and $d_*$ are not directly dependent from the values of the function. They actually only depend on the ordering of these values. In particular, $n_*$ and $d_*$ are not sensitive to affine rescaling of $f$ (for instance, they remain unchanged if $f$ is dilated or contracted). On the other hand, the critical height $h_*$ introduced in the theory of simulated annealing is strongly dependent from such rescaling. Fig. 1 gives a simple example for which the indexes $n_*$ and $d_*$ are explicitly computed.

Another important remark is that $n_*$ and $d_*$ are bounded by the diameter $D$ of the mutation graph ($D$ is the maximal distance between two arbitrary vertices in the graph). This bound does not depend on the function to minimize and is useful to describe universal parameterizations of MOSES.

In brief, it is worthwhile to imagine that MOSES creates its own representation of the energy landscape. In contrast, the classical statistical mechanics formalism introduced to deal with simulated annealing relies on a direct representation of the objective function as an energy. The virtual landscape is

related to the geometry of the optimization problem in complex way. This landscape is built upon distances from local minima to solutions with better evaluation.

### B. Main Results

This section states the main results concerning the convergence of MOSES. Proofs are deferred to the Appendix. This paper presents two results. Theorem 4.1 gives population sizes which ensure that the process "concentrates" on the best solution. Theorem 4.3 gives cooling schedules for the mutation parameter $p_T$ which ensure that the algorithm converges to the best solution.

According to condition (32), the Markov transition matrix $q_T$ which is associated with the algorithm satisfies the classical convergence conditions of the Perron–Frobenius theorem [8]

$$\forall (x,y) \in X \times X, \quad \exists r \geq 1 \quad q_T^{(r)}(x,y) > 0 \tag{35}$$

(and aperiodicity obviously holds). Thus, the Markov chain $(X_t^T)$ converges to a unique stationary distribution as $t$ goes to infinity. We shall denote by $\mu_T$ the stationary distribution. As in simulated annealing, the convergence of the algorithm relies upon the concentration of $\mu_T$ on $(a^*) = (a^*, \ldots, a^*) \in X$ as $T$ goes to zero.

*Theorem 4.1:* Let

$$n > n_*. \tag{36}$$

The invariant distribution of the chain $(X_t^T)$ concentrates on the uniform population $(a^*)$ as $T$ goes to zero, where $a^*$ is the minimal point of $f$.

*Note:* This result justifies the introduction of $n_*$ as an index of population size. A consequence of this result is that the critical size is finite. The connectivity of the mutation graph has a major role in this matter of fact. Depending on the context, however, it may be difficult to compute the exact value of $n_*$. In such a case, the best uniform bound that can be achieved by this method is the diameter $D$ (by uniform, we mean that no particular knowledge is required on the location of $a^*$).

To implement a practical search of $a^*$, the temperature $T$ must be decreased to zero. As in the annealing procedure, it is desirable to decrease this parameter at each step of the algorithm. Let the notation $X_t$ stand for $X_t^{T(t)}$ for all $t \geq 1$. Similarly, let the notation $p_t$ stand for $p_{T(t)}$ for all $t \geq 1$.

*Theorem 4.2:* Let $n > n_*$. Assume that

$$\sum_{t=1}^{\infty} e^{-d_*/T(t)} = \infty \tag{37}$$

then we have

$$\mathrm{Prob}(X_t = (a^*) \mid X_0 = x) \to 1.0 \tag{38}$$

as $t$ tends to infinity (MOSES converges to the minimal solution).

*Note:* The conclusion of the proof of this result (see the Appendix) is a little more general than the previous statement. Under condition (37), the probability that the vector of solutions contains at least one coordinate equal to $a^*$ also converges to 1.0 as $t$ tends to infinity.

Equivalently, the mutation parameter $p_t$ at generation $t$ can be chosen so that

$$\sum_{t=1}^{\infty} p_t^{d_*} = \infty. \tag{39}$$

Finally, the algorithm can be parameterized in the following way.

*Theorem 4.3:* Let $n > n_*$ ($n > D$ is sufficient). Assume that

$$p_t = t^{-1/d_*} \quad \text{for all } t \geq 1 \tag{40}$$

or

$$p_t = t^{-1/D} \quad \text{for all } t \geq 1. \tag{41}$$

Then, we have

$$\mathrm{Prob}(X_t = (a^*) \mid X_0 = x) \to 1.0 \tag{42}$$

as $t$ tends to infinity (MOSES converges to the minimal solution).

*Note:* Once the graph is fixed, the parameterization which uses (41) is universal since it only depends on the graph and no longer on $f$. Of course, optimal mutation strategies are dependent on the function to minimize. If the optimal diameter $D$ is desired, then this constant must depend on the objective function.

This subsection is concluded with a comparison between the bounds obtained for MOSES and those obtained for simulated annealing and genetic algorithms. Regarding to the standard annealing procedure, it is easy to see that $d_*$ is analogous to the critical height $h_*$ of Hajek [13]. Hajek's constant, however, is of minor practical interest, because it remains incomputable unless the geometry of the optimization problem is entirely known. The use of the classical Metropolis algorithm requires a series of tedious trial-and-error tests which contribute to slow down the overall process of optimization and make the quality of the response uncertain. In contrast, the parameterization given in (41) only relies on the diameter $D$ of the mutation graph.

In [5], a mutation-selection genetic algorithm is described which is similar to MOSES. The emphasis is that MOSES proceeds by applying mutation *or* selection whereas a genetic algorithm sequentially applies mutation, *and* thereafter selection. In MOSES, a single operator is involved at each step and this makes the analysis easier (see the Appendix). Furthermore, the results stated in [5] give few reasons to implement the mutation-selection algorithm. Concentration is proved to hold above a value which may be huge [see (4)] depending on the ratio $\Delta/\delta$. Theorem 4.3 holds for mutation-selection genetic algorithms as well. Nevertheless, the geometrical indexes which determine the behavior of the genetic algorithm are different from $n_*$ and $d_*$. Actually, the constant which corresponds to $d_*$ is unknown, and all the available bounds strongly depend on the objective function.

### C. Large Deviations Arguments

A unifying formalism called GSA [19] has been developed to study simulated annealing and genetic algorithms. In [19], Markov transition kernel $q_T$ are studied with the assumption that there exists $\kappa$ such that

$$\frac{1}{\kappa}\pi(x,y)e^{-C(x,y)/T} \leq q_T(x,y) \leq \kappa\pi(x,y)e^{-C(x,y)/T} \quad (43)$$

where the family $C$ (the communication cost) satisfies $C(x,y) \geq 0$ and $C(x,y) = +\infty$ iff $\pi(x,y) = 0$. Actually, GSA theory applies under the slightly more general conditions (23) and (27) and yields the same results. This formalism generalizes the results of Hajek [13] concerning the standard simulated annealing algorithm. The basic fact underlying GSA is that a deterministic mechanism is perturbed at each step of the algorithm. GSA relies itself on Freidlin and Wentzell [11] in which a theoretical framework is developed for dealing with the Markovian perturbations of dynamical systems. The idea is to replace the classical energy function in simulated annealing by a virtual energy which is expressed in terms of costs on the trajectories of the system. In MOSES, the deterministic mechanism is easy to identify. It consists of assigning to each pool of solutions $x$ the uniform element $(x_*)$ (as in [4] and [5]).

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

$$V_1(x,y) = \inf\left\{\sum_{k=0}^{r-1} C(x_k, x_{k+1}),\ x_0 = x,\right.$$
$$\left. x_k \in X,\ x_r = y,\ r \geq 2\right\}. \quad (44)$$

Specific subgraphs of $X$ are needed to proceed with the definition of virtual energy. Let $G(x)$ be the set of all spanning trees on $X$ rooted at $x$. Recall that an $x$-graph ($g \in G(x)$) ends at $x$ and contains no cycle (each $y \neq x$ is the starting point of exactly one oriented edge). The *virtual energy* is defined on the set $X$ by

$$\forall x \in X, \quad W(x) = \min_{g \in G(x)} \sum_{(y \to z) \in g} V_1(y,z). \quad (45)$$

In formula (45), the minimum is taken over the set of all $x$-graphs on $X$ and the sum runs over the edges of these graphs. The virtual energy $W$ describes the asymptotic behavior of the chain $(X_t^T)$ as $T$ goes to zero. In [11], a logarithmic equivalent for the stationary probability distribution $\mu_T(x)$ is given

$$\forall x \in X, \quad \lim_{T \to 0} -T \ln \mu_T(x) = W(x) - W_{\min} \quad (46)$$

where $W_{\min}$ is the minimal value of $W$. Let $\mathcal{W}^*$ be the set of all populations in $X$ for which $W_{\min}$ is attained. In reference to statistical mechanics, the distribution $\mu_T$ can be regarded as a Gibbsian distribution

$$\mu_T(x) \approx \frac{e^{-W(x)/T}}{Z_T} \quad (47)$$

associated with the virtual energy $W(x)$. Equation (46) states that the distribution $\mu_T$ concentrates on $\mathcal{W}^*$. Clearly, $\mathcal{W}^*$ is contained in the subset $U \subset X$ of uniform populations. Theorem 4.1 shows that for sufficiently large population sizes, the subset $\mathcal{W}^*$ reduces to the singleton $\{(a^*)\}$ which corresponds to the optimal population.

TABLE I
FINAL VALUE FOUND BY MOSES AFTER $10^5$ ITERATIONS. $D$ IS THE DIAMETER OF THE EXPLORATION GRAPH AND $n$ THE POPULATION SIZE

| $D$ | 2 | 3 | 4 | 5 | 6 | 10 |
|---|---|---|---|---|---|---|
| $n = 10$ | .74203 | .04631 | .00663 | .00446 | .00294 | .00279 |
| $n = 50$ | .05453 | .00625 | .00385 | .00279 | .00279 | .00279 |
| $n = 100$ | .00735 | .00279 | .00279 | .00279 | .00279 | .00279 |

Turn now to the results concerning the choice of the cooling schedules. The work of [3] and [19] regarding GSA is used extensively. A chain satisfying condition (23) and (27) is viewed as a generalization of the Metropolis algorithm, and the optimal cooling schedules can be given as in [13] and [20]. Trouvé's paper [19, Thm. 2.22] has established this result in a formal way. This can be restated as follows.

*Theorem 4.4:* There exists a nonnegative constant $H_1$ such that for all decreasing cooling schedules $(T(t))_{t \geq 1}$ converging to zero we have

$$\sup_{x \in X} \text{Prob}(X_t \notin \mathcal{W}^* \mid X_0 = x) \to 0 \quad \text{as } t \to \infty \quad (48)$$

if and only if

$$\sum_{t=1}^{\infty} e^{-H_1/T(t)} = \infty. \quad (49)$$

In [19], an explicit description of $H_1$ is given in terms of the decomposition of $X$ into cycles. The definition of $H_1$, however, is quite intricate and intractable. In this work, an alternative description of $H_1$ is preferred [3]. A characterization of $H_1$ in terms of paths of $(X_t)$ is proposed. Recall some definitions. For each path

$$\gamma_{xy} = (x_0 = x \to x_1 \to \cdots \to x_r = y) \quad (50)$$

between $x$ and $y$ in $X$, define

$$H(\gamma_{xy}) = \max_{0 \leq k < r} \{W(x_k) + C(x_k, x_{k+1})\} \quad (51)$$

where the maximum is taken over all vertices in $\gamma_{xy}$. Let $H(x,y)$ be the lowest possible value of $H(\gamma_{xy})$ over all self-avoiding paths $\gamma_{xy}$ from $x$ to $y$. The quantity $H(x,y)$ is the communication altitude between $x$ and $y$. Then, following the results of [3], $H_1$ is given by

$$H_1 = \max_{x \neq (a^*)} H(x, (a^*)) - W(x). \quad (52)$$

For MOSES, an upper bound on $H_1$ can be easily obtained (see the Appendix). This bound is

$$H_1 \leq d_*. \quad (53)$$

Theorem 4.3 follows from Theorem 4.4 and the previous inequality.

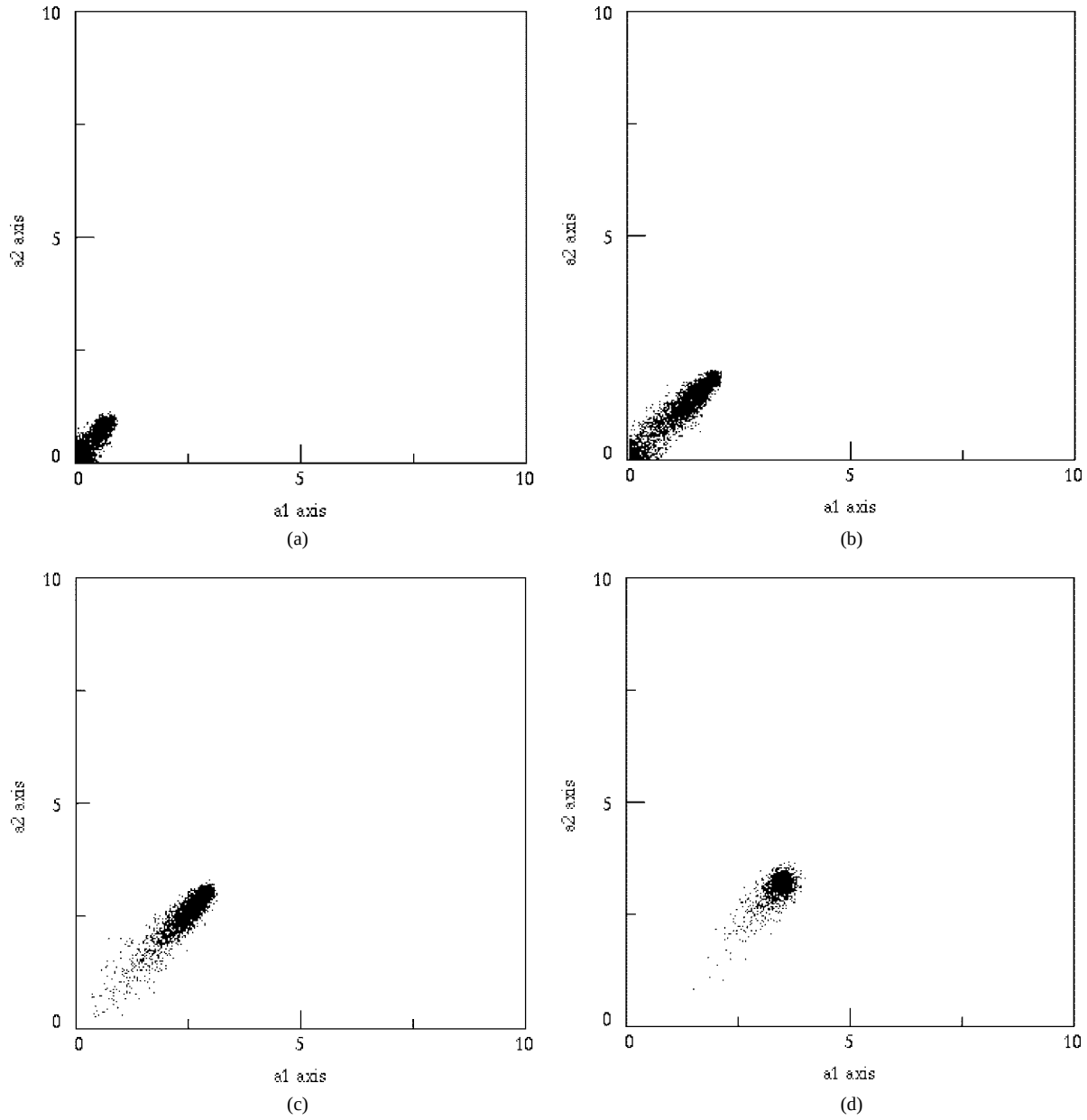


Fig. 2. Two-dimensional plot of the population in the test problem (54). The diameter is $D = 252$, and the population size is $n = 2000$. The individuals were initially located at (0, 0). The cooling schedule is given by (41). The population is displayed each 100 generations $(a - g)$. The optimal solution is located at the center of every image. The best individual is located in front of the cluster observed in every image.

## V. Experimental Results

### A. Role of Diameter and Size

To extend the previous theoretical results, numerical simulations were carried out on a specific minimization problem. We empirically investigated the effects of the variations of the population size and of the diameter of the exploration graph. For a graph with a fixed diameter $D \in \{1, \ldots, 10\}$, the size of the population $n$ was varied from 10 to $n_{\max} = 100$. The cooling schedule defined by (41) was used in all cases. MOSES was stopped after $10^5$ generations. The optimization problem involved the search of the global minimum of the function

$$f(a_1, a_2) = 0.2\big((a_1 - 5)^2 + (a_2 - 5)^2\big) + 2\sin(10(a_1 + a_2 - 10)) + 2 \qquad (54)$$

with $(a_1, a_2) \in [0, 10]^2$. The function $f$ oscillates and admits a large number of barriers which are difficult to cross. The global minimum $a^*$ is located near the point (5, 5) and the minimal value is close to 0.0 (the value at (5, 5) is 2.0, and is actually not minimal but $f(4.92, 4.92) \approx 0$). The square [0, 10]$^2$ was discretized into a grid of mesh $\epsilon = 0.004$. A squared neighborhood has been imposed to each individual, i.e., a square of sidelength $r$ for the Euclidean metric (centered on the individual). A graph of diameter $D$ was obtained by choosing $r = 10.0/D$. The evolution is started from the point (0, 0) where the difficulty is the highest. The results reported in Table I show the optimal value found by the algorithm at the final generation in a typical trial.

With small diameters, only suboptimal solutions are produced. Obtaining the optimal solution is then a very slow

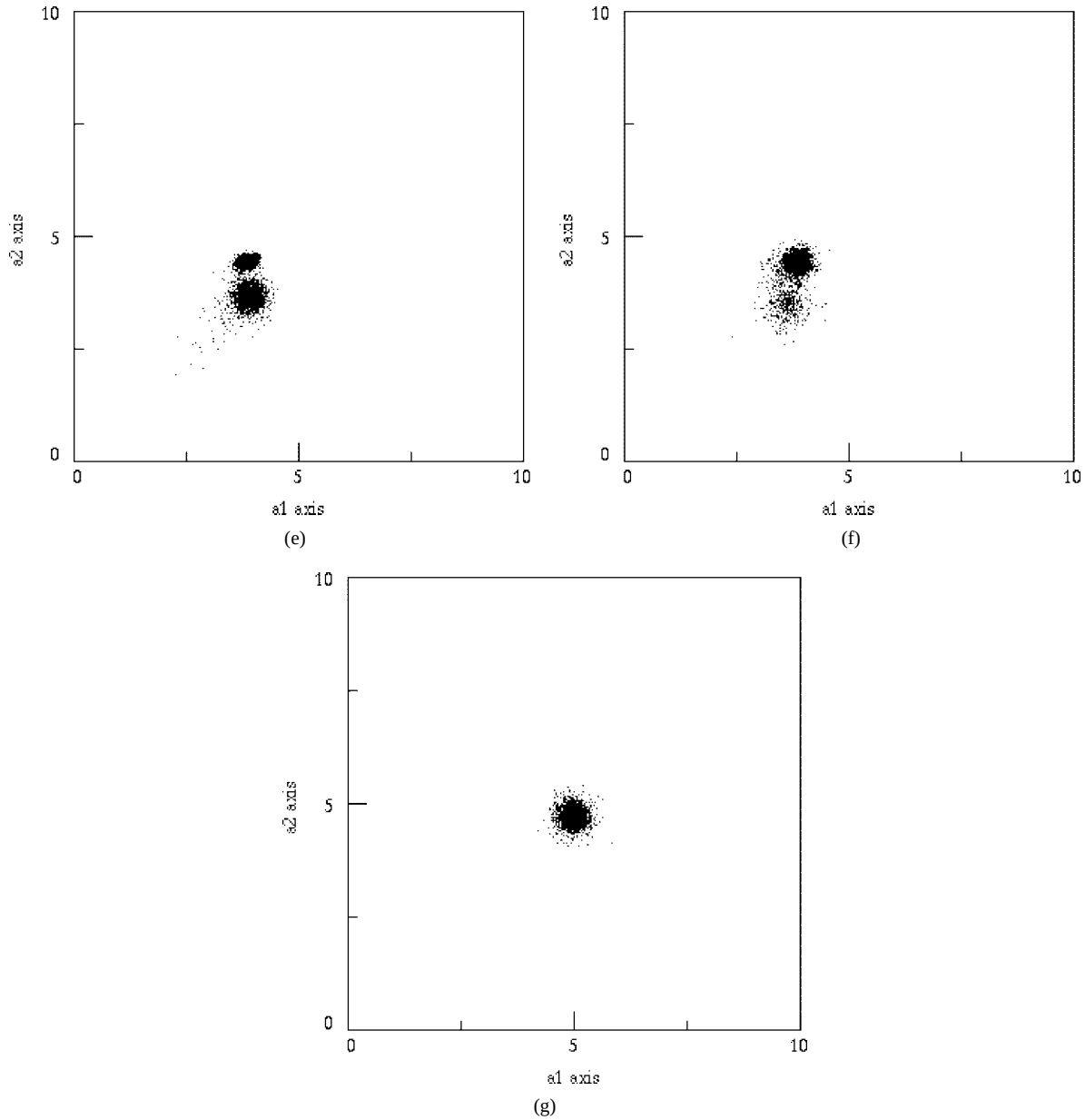


Fig. 2. (*Continued.*) Two-dimensional plot of the population in the test problem (54). The diameter is $D = 252$, and the population size is $n = 2000$. The individuals were initially located at (0, 0). The cooling schedule is given by (41). The population is displayed each 100 generations $(a - g)$. The optimal solution is located at the center of every image. The best individual is located in front of the cluster observed in every image.

process. Better results are observed with larger diameters. Augmenting the size of the population speeds up the convergence of the process. To understand this, the virtual energy must be studied more deeply. Since the communication cost is linear in $n$, such a study must be very similar to [5]. Intuitively, when $n$ is large the energy landscape looks like a large basin and the optimization process is greatly facilitated. Fig. 2 illustrates a possible scenario with $n = 2000$ and $D = 252$ (these unrealistic values are used for the purpose of demonstration).

### *B. Nonasymptotical Experimental Study*

It must be pointed out that the above given result (Theorem 4.3) is an asymptotical result. Although it can be large, the computing resource is always finite. Applying the conditions of Theorem 4.3 ensures the reliability of the procedure in the long run. Better results may be obtained, however, in short computing time without following the (sufficient) conditions of the theorem. In practice, useful information can be gained from MOSES even when the population is far from convergence. For instance, there is no need for the whole population to consist of optimal solutions. A relevant event is the occurrence of the first visit to $a^*$ of a single individual in the population. To assess the occurrence of this event, 50 simulations were run at fixed temperature $T = 1.0 \quad (p_T \approx 0.368)$ and with population size $n = 500$. The initial population was sampled from the uniform distribution over the square $[0, 10]^2$. In Table II, the mean time of the first visit to $a^*$ is reported. The simulations shed light on the fact that the diameter of the search graph plays an important role. The

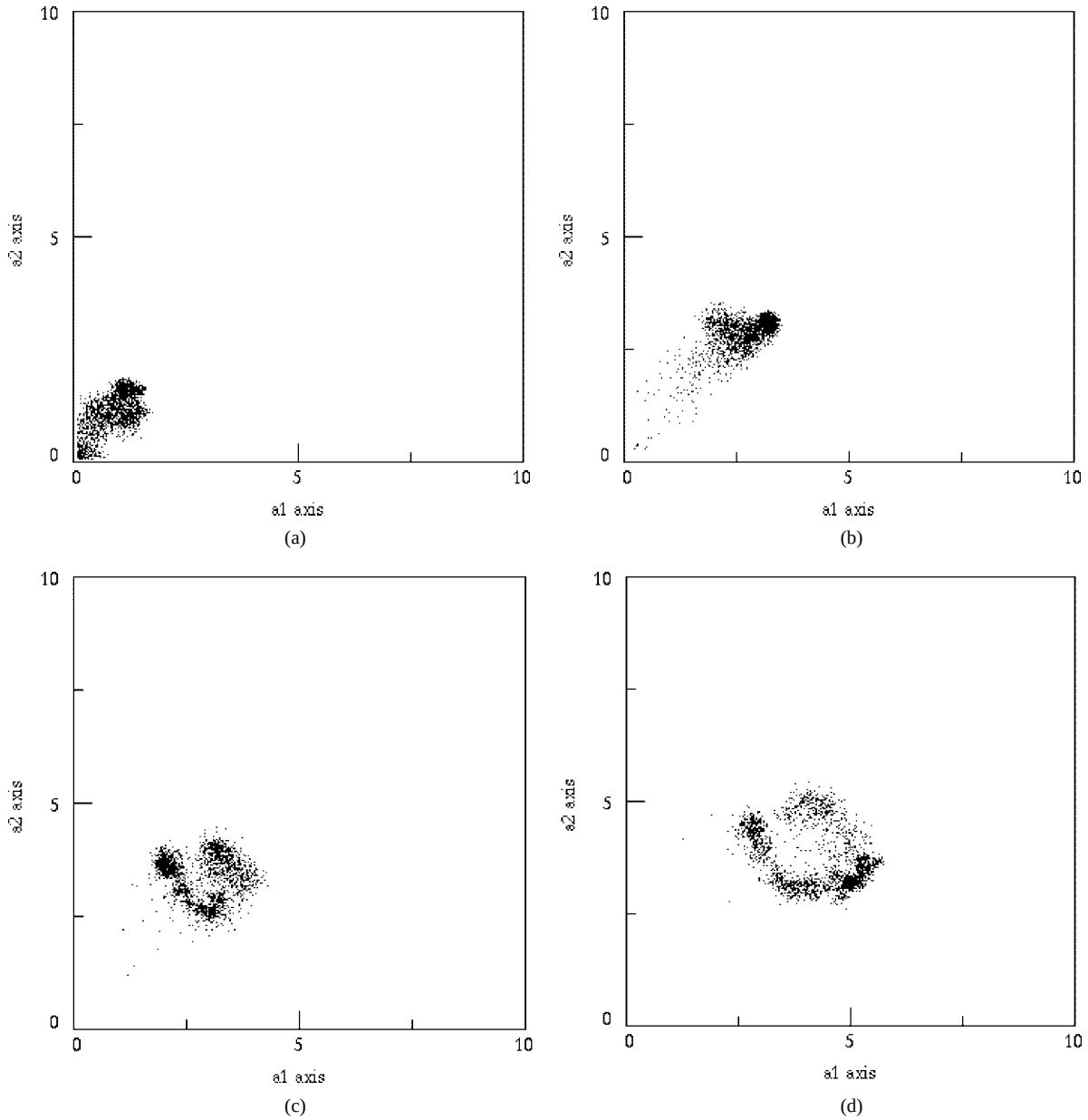


Fig. 3. Two-dimensional plot of the population in the time dependent problem. The diameter is $D = 252$, and the population size is equal to $n = 2000$. The individuals were initially located at (0, 0). The cooling schedule is given by (41). The population is displayed each 100 generations $(a - g)$. The optimal solution moves on circle of radius 2 at the center of every image. The best individual is located in front of the cluster observed in every image. The algorithm uses self organization to find the solution in real time, and individuals are ordered on the trajectory of the minimum.

TABLE II
MEAN TIME OF THE FIRST VISIT TO $a^*$ OVER 50 REPETITIONS. $n = 500$ AND $T = 1.0$. $D$ IS THE DIAMETER OF THE MUTATION GRAPH

| Diameter $D$ | 5 | 10 | 30 | 35 | 40 | 45 | 50 | 55 | 60 | 70 | 80 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Mean time | 1842 | 514 | 136 | 63 | 53 | 50 | 24 | 25 | 36 | 50 | 140 |

best value is obtained for $D = 50$. The average number of evaluations necessary to find the solution is $24np_T \approx 4400$. (The total number of comparisons required to find the minimum by enumeration is approximately 6 500 000.) With small diameters, the whole space is scanned, and local minima are rapidly found. Nevertheless, these solutions may be far from optimal. On the other hand, the process of evolution is slow when very large diameters are used. This is due to the very local search performed by most of the individuals. In practical situations, a balance between the choice of the diameter and the computing resource must be investigated. We have yet no theoretical result in that direction. It seems difficult to tackle this issue since it assumes that the geometry of the problem varies in time.

### C. Time-Dependent Problems

Attention is called here that MOSES can be used to solve time-dependent problems. By using a hierarchical population, the procedure is able to produce features of self-organization. Individuals may organize to track the optimal solution in real time. Such a claim is illustrated

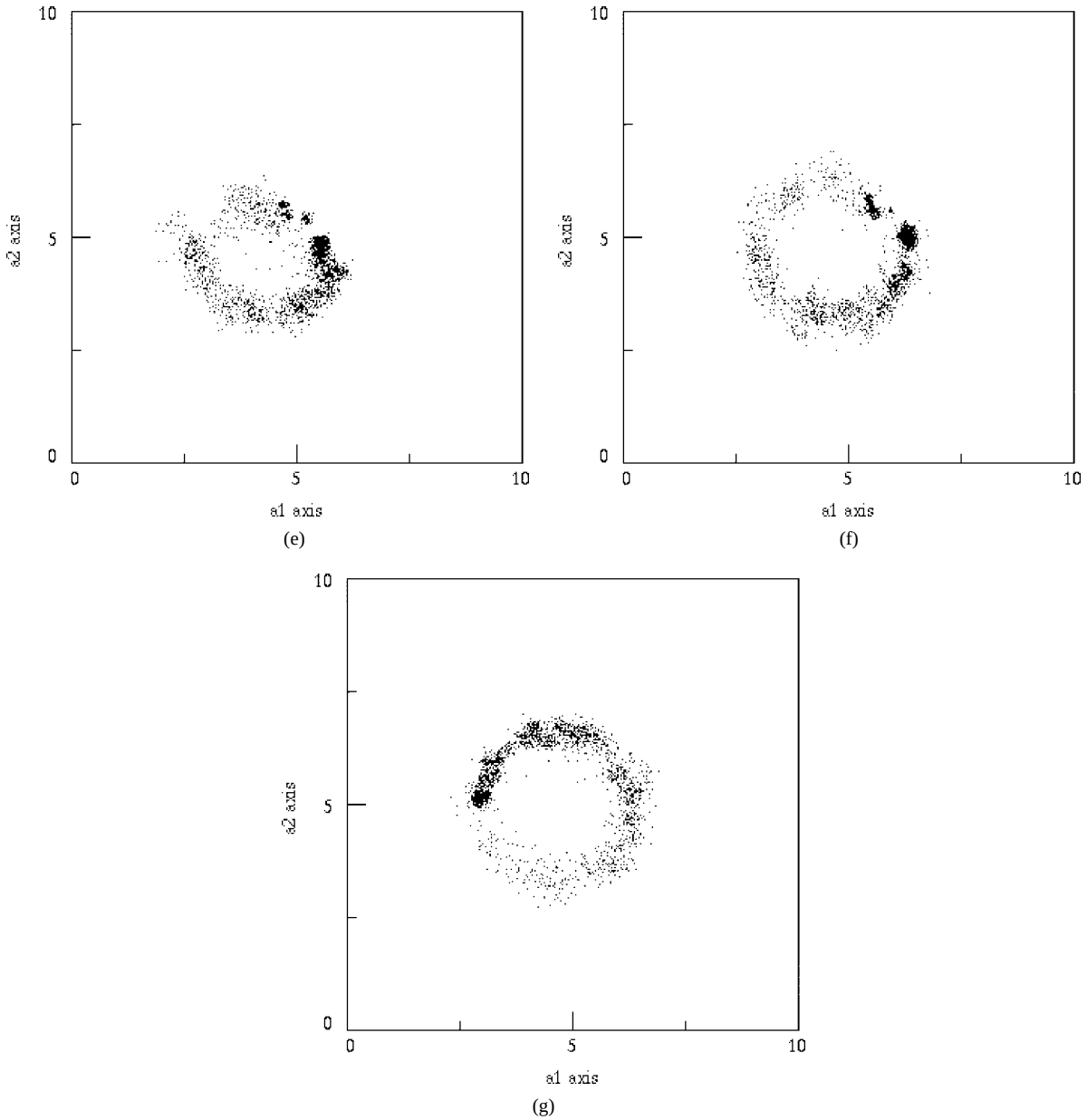


Fig. 3. (*Continued.*) Two-dimensional plot of the population in the time dependent problem. The diameter is $D = 252$, and the population size is equal to $n = 2000$. The individuals were initially located at (0, 0). The cooling schedule is given by (41). The population is displayed each 100 generations $(a - g)$. The optimal solution moves on circle of radius 2 at the center of every image. The best individual is located in front of the cluster observed in every image. The algorithm uses self organization to find the solution in real time, and individuals are ordered on the trajectory of the minimum.

with the following problem. The function to minimize is

$$f_t(a_1, a_2) = 0.2\big((a_1 - \alpha_t)^2 + (a_2 - \beta_t)^2\big) + 4\sin(10(a_1 - \alpha_t + a_2 - \beta_t)) + 4 \quad (55)$$

with $(a_1, a_2)$ in the square $[0, 10]^2$ and

$$\alpha_t = 5 + 2\cos(2\pi t/100) \quad (56)$$

$$\beta_t = 5 + 2\sin(2\pi t/100). \quad (57)$$

For all $t$, there exists a unique global minimum of $f_t$: it is close to $(\alpha_t, \beta_t)$. The trajectory of this minimum is a circle of radius 2 centered at a point close to (5, 5). A typical scenario is shown in Fig. 3. In Fig. 4, the optimal value found by MOSES is plotted as a function of time.

## VI. Conclusion

MOSES shares similarities with both the simulated annealing [1], [13] and the mutation-selection genetic algorithm [5]. In the present analysis, the formalism of GSA introduced in [3] and [19] was used. This approach replaces the standard energy function of the Metropolis algorithm by a virtual energy

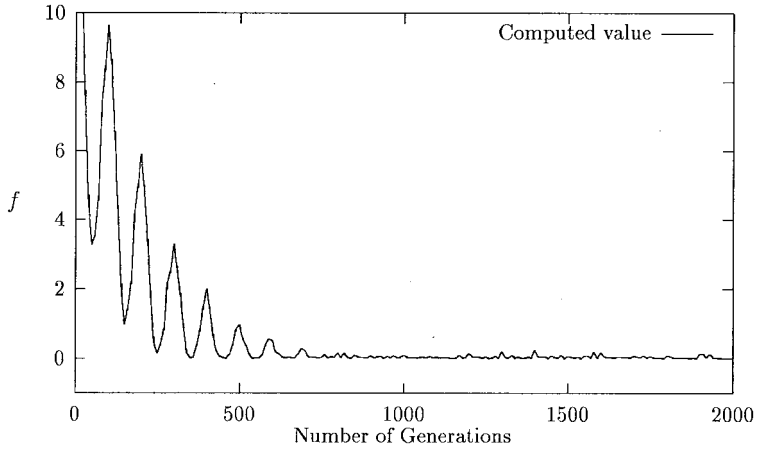


Fig. 4. Typical evolution of the optimal value computed by MOSES in the time-dependent problem. The minimal value is 0.0.

TABLE III
A VARIATION ON MOSES WITH POPULATION SIZE EQUAL TO $n$ AND MUTATION PROBABILITIES EQUAL TO $p_T$. THIS ALGORITHM IS CONVERGENT (NO CRITICAL SIZE) AND THE COOLING SCHEDULE CAN BE CHOSEN AS IN (39)

1. Initialize the population randomly.
2. Repeat
   - Draw a random number $N$ from the binomial distribution $bin(n-1, p_T)$.
   - Select the optimal individual $x_*$ from the population.
   - Replace the $N$ first individuals by mutation and the $n-N$ other individuals by $x_*$.
   - Update $p_T$.

which is expressed in terms of costs on the trajectories of the system.

In this paper, the relationships between convergence, the parameters of the strategy, and the geometry of the optimization problem were investigated, and a universal parameterization of the strategy has been proposed. The conclusions are similar to those of the theory of simulated annealing. Nevertheless, the constants involved in both approaches differ and are easier to estimate in MOSES.

The main result in this paper may be summarized as follows. Convergence of MOSES is ensured when the population size is greater than the diameter of the graph on which mutating individuals explore the search space. Once the size is fixed, the temperature can be decreased according to a logarithmic schedule which is again proportional to the diameter of the graph.

The analysis is mainly asymptotical. A nonasymptotical study would directly rely on (22) which is hardly tractable. Mutation probabilities are assumed to be small and the computing time is assumed to be large. It seems natural that these conditions are broken in realistic situations. The asymptotical study, however, has emphasized the relevant geometrical quantities such as the diameter $D$ and the critical constant $d_*$. As in genetic algorithms, the information on the best solution ever found by the procedure may be lost in course of the evolution. A large size minimizes the influence of this event. An artifact can be introduced, however, to avoid such events. The artifact merely consists in sampling from the binomial distribution $bin(n-1, p_T)$ instead of $bin(n, p_T)$. Such a variant is elitist: it becomes impossible to lose the best individual, and the minimal solution $(a^*)$ becomes an absorbing state of the Markov chain $(X_t^T)$. In this situation the question of convergence is obviously solved. The population converges to $(a^*)$ as soon as the population consists of at least two individuals (see Table III). Concerning the cooling schedules, the conclusions are exactly analogous to those of Section IV-B (except that $n$ may be arbitrary). The behavior of MOSES is therefore asymptotically identical to the variant with population size equal to two. Obviously, this is not true in the nonasymptotic regime. This remark underlines a weakness of the asymptotic approach. Further efforts must be devoted to develop the nonasymptotic approach.

Although MOSES will be outperformed by genetic algorithms or simulated annealing on some problems, applying these techniques requires a difficult preliminary study of the optimization problem whereas the parameters of MOSES can be configured with little (or no) prior knowledge on $f$.

## APPENDIX

*Note:* We shall work with paths in $E$ and with paths in $X$. To avoid confusion, paths in $E$ are paths in the graph $(E, \mathcal{G})$ while paths in $X$ are trajectories of the chain $(X_t^T)$.

Theorem 5.8 of [5] is used to compute the relevant quantities: $W$ and the communication altitude $H$ (defined Section

IV). Theorem 5.8 of [5] says that if for all $x \in X$, we have $(x_*) \in U$ and $V_1(x,(x_*)) = 0$, then $W$ and $H$ can be computed on $U$ with

$$V(x,y) = \inf\left\{\sum_{k=0}^{r-1} C(x_k, x_{k+1}),\ x_0 = x,\right.$$
$$\left. x_k \notin U (1 \le k < r),\ x_r = y,\ r \ge 2\right\} \quad (58)$$

instead of the cost $V_1$ defined in (44). For uniform populations $(a)$ and $(b)$, $V(a,b)$ and $W(a)$ will now stand for $V((a),(b))$ and $W((a))$ (parentheses are omitted when dealing with uniform populations).

*Lemma 6.1:* Let $a \neq a^*$ where $a^*$ is the minimal point of $f$. Then we have

$$V(a,a^*) = d(a,a^*) \quad (59)$$

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

*Proof:* Obviously, we have $V(a,a^*) \ge d(a,a^*)$. Now, consider a path in $(E,\mathcal{G})$ which realizes $d(a,a^*)$ : $a_1^0 = a \to a_1^1 \to \cdots \to a_1^r = a^*$ and the path in $X$

$$\begin{array}{c} x_0 = (a) \\ \downarrow \\ x_1 = (a_1^1, \dots) \\ \downarrow \\ \cdots \\ \downarrow \\ x_r = (a^*, \dots) \\ \downarrow \\ x_{r+1} = (a^*) \end{array} \quad (60)$$

such that

$$\forall k = 0, \dots, r-1, \quad C(x_k, x_{k+1}) = 1. \quad (61)$$

(If for instance $f(a_1^1) < f(a)$ then $x_1 = (a_1^1, a, \dots, a)$ and $x_2 = (a_1^2, a_1^1, \dots, a_1^1)$ etc.) Then we have $V(a,a^*) \le d(a,a^*)$.

The concentration on $(a^*)$ is obtained by controlling the cost function $V$ on the subset $U$. □

*Lemma 6.2:* Assume that there exists an $a^* \in E$ such that

$$\forall a,b \in E, \quad a,b \neq a^*, \quad V(a,a^*) < V(a^*,b). \quad (62)$$

Then, for all $a \neq a^*$, $W(a^*) < W(a)$: the chain concentrates on $(a^*)$.

*Proof:* Let $a \in E$ such that $a \neq a^*$ and $g$ an $a$-graph on $U$ for which

$$W(a) = \sum_{(u \to v) \in g} V(u,v). \quad (63)$$

Since $a \neq a_*$ and $g$ is a spanning tree on $U$ rooted at $a$, there must be $a,b \in U$ 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

$$W(a^*) \le W(a) + V(a,a^*) - V(a^*,b) < W(a). \quad (64)$$

□

The problem that is addressed now is to force $\mathcal{W}^*$ to be equal to $\{(a^*)\}$ by correctly determining the size $n$ of the population. Here is the proof of Theorem 4.1.

*Proof:* Let $(a)$ and $(b)$ be uniform populations, $a,b \neq a^*$. We have

$$n > \max_{a \neq a^*} \{\text{length of the shortest path from } a \text{ to } a^* \text{ in } E\} \quad (65)$$

and hence, by Lemma 6.1

$$n > V(a,a^*). \quad (66)$$

Moreover, the path of minimal cost which exits from $(a^*)$ in $U$ necessarily involves $n$ simultaneous mutations

$$\forall b \neq a^*, \quad V(a^*,b) \ge n. \quad (67)$$

We obtain

$$V(a^*,b) > V(a,a^*). \quad (68)$$

Under this condition, Lemma 6.2 applies. □

The last statement to prove is as follows.

*Lemma 6.3:* We have $H_1 \le d_*$.

*Proof:* First, use again Theorem 5.8 of [5] to compute $H$ and $W$ from $V$ instead of $C$. Then we have

$$H_1 = \max_{a \neq a^*} H(a,a^*) - W(a). \quad (69)$$

Let $a \in E$ and consider the path

$$\gamma: \quad (a) \to (b) \to (a^*) \quad (70)$$

in $X$ (or equivalently in $U$) where $b$ realizes

$$\min_{b: f(b) < f(a)} d(a,b). \quad (71)$$

Obviously, we have $V(a,b) \ge d(a,b)$. Now, consider a path in $(E,\mathcal{G})$ which realizes $d(a,b)$

$$a_1^0 = a \to a_1^1 \to \cdots \to a_1^r = b$$

and the path in $X$ defined by

$$\begin{array}{c} x_0 = (a) \\ \downarrow \\ x_1 = (a_1^1, a, \dots, a) \\ \downarrow \\ \cdots \\ \downarrow \\ x_r = (b, a, \dots, a) \\ \downarrow \\ x_{r+1} = (b). \end{array} \quad (72)$$

Then, the cost of this path in $X$ is exactly

$$\sum_{k=0}^{r} C(x_k, x_{k+1}) = d(a,b) \quad (73)$$

and $V(a,b) = d(a,b)$. Then, we have

$$H(\gamma) \le W(a) + V(a,b) = W(a) + d(a,b). \quad (74)$$

Obviously, this implies that

$$H(a, a^*) - W(a) \leq d(a, b). \tag{75}$$

## ACKNOWLEDGMENT

The author is grateful to R. Cerf for corrections and helpful suggestions. He thanks D. Zaharie and M. Beguin for various conversations and comments.

## REFERENCES

[1] E. H. L. Aarts and J. H. M. Korst, *Simulated Annealing and Boltzmann Machines*. New York: Wiley, 1988.
[2] T. Bäck, *Evolutionary Algorithms in Theory and Practice*. New York: Oxford Univ. Press, 1996.
[3] O. Catoni, "Simulated annealing algorithms and Markov chains with rare transitions," lecture notes, DEA Univ., Paris XI, 1997.
[4] R. Cerf, "Une théorie asymptotique des algorithmes génétiques," Ph.D. dissertation, Montpellier II, 1994.
[5] ———, "The dynamics of mutation-selection algorithms with large population sizes," *Ann. Inst. H. Poincaré Probab. Statist.*, vol. 32, pp. 455–508, 1996.
[6] U. K. Chakraborty, D. Kalyanmoy, and M. Chakraborty, "Analysis of selection algorithms: A Markov chain approach," *Evol. Comput.*, vol. 4, no. 2, pp. 133–167, 1996.
[7] T. E. Davis and J. C. Principe, "A simulated annealing like convergence theory for the simple genetic algorithm," in *Proc. Fourth Int. Conf. Genetic Algorithms*, R. K. Belew and L. B. Booker, Eds. San Mateo, CA: Morgan Kauffman, 1991, pp. 174–181.
[8] W. Feller, *An Introduction to Probability Theory and Its Applications*. New York: Wiley, 1971, vol. II.
[9] D. B. Fogel, *Evolutionary Computation: Toward a New Philosophy of Machine Intelligence*. Piscataway, NJ: IEEE Press, 1995.
[10] O. François, "Convergence in simulated evolution algorithms," *Complex Syst.*, vol. 10, pp. 311–319, 1996.
[11] M. I. Freidlin and A. D. Wentzell, *Random Perturbations of Dynamical Systems*. New York: Springer-Verlag, 1984.
[12] D. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*. Reading, MA: Addison-Wesley, 1989.
[13] B. Hajek, "Cooling schedules for optimal annealing," *Math. Oper. Res.*, vol. 13, pp. 311–329, 1988.
[14] L. Ingber and B. Rosen, Genetic algorithm and very fast simulated annealing—A comparison," *Math. Comput. Model.*, vol. 16, no. 11, pp. 87–100, 1992.
[15] Y. Leung, Y. Gao, and Z. B. Xu, "Degree of population diversity—A perspective on premature convergence in genetic algorithms and its Markov chain analysis," *IEEE Trans. Neural Networks*, vol. 8, pp. 1165–1175, 1997.
[16] A. Nix and M. Vose, "Modeling genetic algorithms with Markov chains," *Ann. Math. Art. Intell.*, vol. 5, no. 1, pp. 79–88, 1992.
[17] G. Rudolph, "Convergence analysis of canonical genetic algorithms," *IEEE Trans. Neural Networks*, vol. 5, pp. 96–101, 1994.
[18] J. Suzuki, "A Markov chain analysis on a genetic algorithm," in *Proc. Fifth Int. Conf. on Genetic Algorithms*, S. Forrest, Ed., Urbana-Champaign, IL, 1993, pp. 146–153.
[19] A. Trouvé, "Cycle decomposition and simulated annealing," *SIAM J. Control Optim.*, vol. 34, no. 3, pp. 966–986, 1996.
[20] J. N. Tsitsiklis, "Markov chains with rare transitions and simulated annealing," *Math. Oper. Res.*, vol. 14, pp. 70–90, 1989.
[21] M. Vose, "A closer look at mutation in genetic algorithms," *Ann. Math. Art. Intell.*, vol. 10, no. 4, pp. 423–435, 1994.

**Olivier François** was born in Saint Vallier, France, in August 1966. He received the Magistère de Mathématiques et Applications in 1989 and the Ph.D. degree in mathematics from the University of Grenoble, France, in 1992.

Since 1993, he has been an Associate Professor with the Ecole Nationale Supérieure d'Informatique et de Mathématiques Appliqués de Grenoble. His research interests include applied probability, Markov chains, neural networks, simulated evolution, statistical mechanics, and complex systems.