Theoretical Analysis of Evolutionary Algorithms With an Infinite Population Size in Continuous Space Part I: Basic Properties of Selection and Mutation

Xiaofeng Qi, *Member, IEEE,* and Francesco Palmieri, *Member, IEEE*

***Abstract*—This paper aims at establishing fundamental theoretical properties for a class of "genetic algorithms" in continuous space (GACS). The algorithms employ operators such as selection, crossover, and mutation in the framework of a multi-dimensional Euclidean space. The paper is divided into two parts. The first part concentrates on the basic properties associated with the selection and mutation operators. Recursive formulae for the GACS in the general infinite population case are derived and their validity is rigorously proven. A convergence analysis is presented for the classical case of a quadratic cost function. It is shown how the increment of the population mean is driven by its own diversity and follows a modified Newton's search. Sufficient conditions for monotonic increase of the population mean fitness are derived for a more general class of fitness functions satisfying a Lipschitz condition. The diversification role of the crossover operator is analyzed in Part II [1]. The treatment adds much light to the understanding of the underlying mechanism of evolution-like algorithms.**

I. Introduction

THE field of global optimization has been rapidly expanding over recent decades, aiming at finding general search algorithms for cost functions with many local minima. The functions being optimized are generally defined on a metric space, which includes multi-dimensional Euclidean space or various discrete spaces. The functions can be nondifferentiable and/or discontinuous, and their domain may be constrained. Combinatorial optimization problems, defined in a discrete setting, and usually NP-complete, represent typical examples. It is now widely recognized that gradient-based nonlinear programming techniques would typically fail in the above situations, and drastically different approaches need to be constructed [2]–[4]. Among the many global search methods proposed over the last two decades, global random search techniques have been considered to be a viable and promising direction of exploration [2], [4]–[12]. A comprehensive survey can be found in [4] and convergence properties for general stochastic search algorithms are discussed in [7] and [8]. Included in this class is the well-known simulated annealing method [13], [14] which has been intensively studied and has found successful application in various combinatorial optimization problems.

Manuscript received May 5, 1993; revised August 13, 1993. This work was partially funded through NSF Grants MIP-9009696 and CTS 9008596.

The authors are with the Department of Electrical and Systems Engineering, The University of Connecticut, Storrs, CT 06269.

IEEE Log Number 9213877.

The focus of this work is on simulated evolution, which is revealing to be a very rich class of stochastic search methods. The genetic algorithm (GA) described in [15] and [16], which has become very popular lately for discrete optimization problems, is part of a class that encompasses the more general evolution strategies [17]–[19], often formulated in a continuous-space framework, and also evolutionary programming [20], [21]. These evolutionary techniques are population-oriented: Successive populations of feasible solutions are generated in a stochastic manner following laws similar to that of natural selection. This is in contrast to standard programming techniques that normally follow just one trajectory (deterministic or stochastic), perhaps repeated many times until a satisfactory solution is reached. In the evolutionary approach, multiple stochastic solution trajectories proceed simultaneously, allowing various interactions among them toward one or more regions of the search space. These approaches can be justified by the fact that a population-oriented algorithm automatically stores in time a sampled replica of the profile of the function being optimized, providing important clues for the global structure of the function. Applications of evolutionary computation can be found in [22]–[25].

In the canonical genetic algorithm, each member of the subsequent generation is selected from the current one with a probability proportional to its function value, or "fitness." After selection, various "genetic operators," such as recombination (which extracts common feature shared by two "good" members within the population in order to explore new regions) and mutation (providing perturbations for selected members in the solution space to extend the dynamic range of selection) are used. Eventually the process is likely to converge to a population dominated by the global maximum (maxima) of the fitness function. Compared with single-trajectory methods, such as simulated annealing, a genetic algorithm is intrinsically parallel and global. Local "fitness" information from different members is mixed through various genetic operators, especially the crossover mechanisms, and probabilistic soft decisions are made concerning removal and reproduction of existing members. Furthermore, GA's require only simple computations, such as additions, random number generations, and logical comparisons, with the only major burden being a large number of fitness function evaluations. Only limited knowledge of the problem is generally necessary to use this strategy. The simulated annealing algorithm, instead, is

intrinsically sequential and needs a long and carefully chosen "cooling" schedule. Parallelism must be explicitly added to the canonical approach with increased computational complexity.

The analysis of the canonical GA's defined in [15] and [16] presents the following two major difficulties: 1) It is formulated after the gene recombination mechanisms, and the solution space is discrete in nature. Therefore it is difficult to efficiently apply a discrete GA as-is to a multi-dimensional optimization problem in a continuous multi-dimensional space, which is the framework frequently needed for practical optimization problems. Suitable encoding of the real space has to be adopted and the validity of the transformation may be a nontrivial issue. Many engineering problems require a real-space formulation and have been approached with random search techniques, including various continuous-space versions of evolutionary computation [26]–[30], when traditional nonlinear programming techniques have proved to be inappropriate. Applications of variants of GA's to the optimal design of artificial neural networks (including weights and connectivity) have also been reported in [31]–[33] and others. 2) A solid analysis for a canonical discrete GA is difficult. It is especially difficult to establish the relationship between the various discrete variables and the fitness function (meaningful solution encoding scheme). Convergence analysis is also difficult to conduct. Limited attempts have been made toward a rigorous analysis of the discrete genetic algorithm [34]–[42], but various restrictive conditions have to be imposed on the algorithm. A complete convergence analysis is still not available, especially for the general case involving the crossover operator.

Attempts have been made to construct and analyze variants of genetic algorithms in continuous space. Ermakov and Zhigljavsky [45] [46] were the first ones to define a class of population-oriented algorithms that bears direct analogy to the canonical discrete GA's defined in [15] and [16], and it is referred to as *Algorithm E* [2]. This class, however, does not contain any recombination operators. Successful attempts in solving a high-dimensional problem ($m = 128$) using Algorithm E are reported in [6], with a rigorous proof of convergence in distribution to a Dirac function located at the global optimum for infinite population size, presented in [46]. The convergence behavior, however, to our knowledge has never been carefully analyzed. Related results concerning Algorithm E can be found in [47] and [48]. In the discrete GA literature, a limited amount of qualitative results are available for the "real-coded genetic algorithms" [49], [50]. Applications of heuristic versions of GA's to constrained global optimization problems in the continuous space have also been reported in [26].

Mathematical analyses for the convergence behavior of continuous-space evolutionary programming can follow different directions. Aside from what has been reported in the discrete GA literature, there are scattered results in other academic disciplines that are related to this subject. We briefly describe these results below, pointing out their limitations, and hoping that they may help to understand the evolutionary programming paradigms from a unified computational point of view.

1) Studies in physics: Problems similar to GACS have been attacked by physicists while studying the behavior of electrons moving in a multi-modal potential field [51]–[54]. The results are intended to guide the formulation of stochastic optimization algorithms that combine two of the universal laws of nature—namely, the law of thermodynamics and that of natural selection. The evolution process with a mutation noise assumed to be very small is modeled by a partial differential equation in space and time, which leads to a Schrödinger-type eigenvalue problem. Initial and asymptotic behavior of the algorithm are analyzed in terms of eigenvalues and eigenfunctions of the Schrödinger operator associated with a given fitness function (potential field) [51]. The analysis of evolutionary programming can thus be carried out in the framework of stochastic differential equations. The assumption of infinitesimally small mutation may be a limiting factor, and it is not clear to us how the important recombination operator (crossover) could be included in this framework. Closed-form results, even for specific fitness functions, to our knowledge, are still unavailable.

2) Studies in population genetics: A large amount of literature exists in the area of population genetics that deals with the asymptotic behavior of a population of chromosomes under various combinations of genetic pressure: selection based on viability (fitness), recombination, mutation, environmental variation, and migration between neighboring populations. Comprehensive coverages can be found in [55]–[59]. Some of the studies concentrate on multi-locus populations, which are much similar to the binary string framework of the canonical GA's. Karlin [60]–[65] has conducted an extensive investigation on the equilibrium behavior of multi-locus systems, given various epistatic structures of the fitness function [60], [61], [63], [64]. We found most interesting his analysis of various phenotypical mating schemes, which are the counterparts of crossover in multi-dimensional Euclidean space [62], [65]. Specifically, Karlin has analyzed convergence behavior of so-called random selective mating and nonrandom mating schemes. The former describes a mating behavior that dictates that a similar pair of vectors (with respect to the Euclidean distance) gets a higher chance of recombination. Various versions of such models were studied in [62]. Nonrandom mating schemes can involve various nonrandom combination mechanisms (nonlinear, linear and/or convex, etc.) for pairs of solution vectors, with conditions provided for convergence of the population toward the global optimum under Gaussian assumptions [65]. It seems that some of the heuristic crossover schemes currently adopted for the continuous space version of the GA's [26] are special cases of the nonrandom mating schemes proposed in [62], [65]. Other results address the roles that crossover plays in the overall evolutionary process [66], [67], which may help answer some of the questions surrounding the computational consequences of crossover in a genetic algorithm [68], [69].

3) Outline of this paper: The purpose of our work is to study evolutionary algorithms in the unified framework of stochastic processes in continuous space. We want to explain quantitatively how the simple idea of selection, mutation and recombination is equivalent to an efficient and robust search.

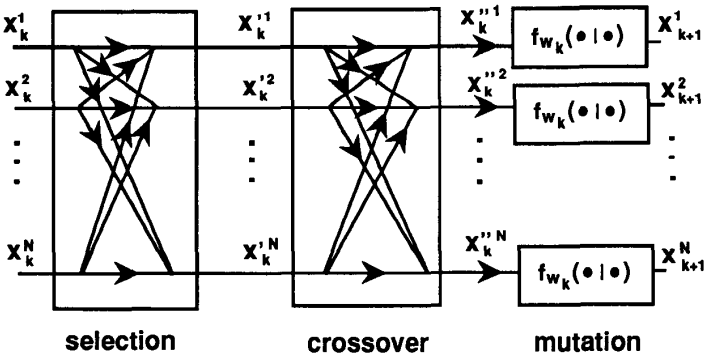

Fig. 1. A schematic illustration of the steps involved in a genetic algorithm with selection, crossover, and generalized mutation scheme. Various notations involved in the analysis are illustrated.

Specifically, we analyze a general class of Genetic Algorithms in Continuous Space (GACS) and formulate it as a discrete-time stochastic process. The states lay in a multi-dimensional Euclidean space and we study the large sample behavior of the process in a rigorous fashion. It is important to point out (and will be shown in the paper) that 1) a large sample analysis is more tractable mathematically, and reveals important patterns of the collective behavior of the population as a whole, and (2) the results of GACS may help us to understand canonical genetic algorithms in discrete space. Furthermore, due to the specific structures of Euclidean space, GACS possesses many unique features not shared by canonical discrete GA's. It should also be emphasized that this paper, being theoretical in nature, may help to provide meaningful guidance for the practical design of other evolutionary procedures.

The approach adopted in our investigation of GACS in this paper has been to formulate the evolutionary process as a discrete-time process with states defined over a multi-dimensional Euclidean space. In the case of a large population, this formulation leads to the study of a sequence of probability density functions characterizing the distribution of the entire population (i.e., the frequency of occurrence of various solution vectors). It needs to be pointed out that the large population assumption is obviously unrealistic in the computational framework. It is obvious that if the whole sample space were filled with population elements, the search should just choose the best individual(s).

However, we maintain that a detailed understanding of the large-population case should constitute the theoretical substrate for meaningful use of genetic algorithm strategies. The study of the more realistic finite population could be seen as an approximation to the large-sample case. A detailed quantitative analysis for the finite-population case seems to be rather difficult, if we want to maintain a good degree of mathematical rigor. We have made a few attempts, that we will not report here, that show how by properly defining the various operators to ensure certain monotonicity conditions on the average fitness of the population, the process qualifies as a *submartingale*.

We have divided the body of this work into two papers. In this first part we formulate the basic algorithm strategy in general terms, and restrict ourselves to the more tractable scenario of populations that undergo selection and mutation only. This particularly elucidates the role of the selection operator in driving the population toward the regions with higher fitness. In the second part that follows [1], we single out the crossover operator, pointing out its peculiar nature and proving a number of results. An attempt to form the global picture is also included in the second part, emphasizing the interaction between crossover and selection. Conclusions and comments follow. We have tried to maintain the paper in a readable format by deferring most of the proofs to the appendices. Some details of the derivation had to be left out to keep this paper within acceptable limits. They can be found in our two technical reports [75] and [74].

II. Basic Algorithm Formulation

We want to solve the following optimization problem:/belowdisplayskip6pt

$$\arg\max_{\mathbf{x}} g(\mathbf{x}), \tag{1}$$

where $\mathbf{x} \in \mathcal{F}$ is the real parameter vector belonging to the feasible region[1] $\mathcal{F}$, $g(\mathbf{x})$ is the so-called fitness function, which measures the goodness of the solution $\mathbf{x}$ and is assumed to satisfy the following conditions:

a) $g(\mathbf{x})$ has only finitely many global maxima;
b) $0 < g(\mathbf{x}) < \infty, \forall \mathbf{x} \in \mathcal{F} \subseteq \mathcal{R}^m$;
c) $g(\mathbf{x})$ has finitely many discontinuous points.

In the following we list the canonical steps of the Genetic Algorithms (see Fig. 1) for a schematic illustration):

S1: Start at time $k = 0$ with N random vectors $\mathbf{x}_0^1, \cdots, \mathbf{x}_0^N$ drawn from an initial probability density function $f_0(\mathbf{x})$.

S2: (Selection) Given $\mathbf{x}_k^1, \cdots, \mathbf{x}_k^N$, select $\mathbf{x}'^1_k, \cdots, \mathbf{x}'^N_k$, such that /belowdisplayskip6pt

$$\Pr\left\{\mathbf{x}'^i_k = \mathbf{x}_k^j | \mathbf{x}_k^1, \cdots, \mathbf{x}_k^N\right\} = \frac{g(\mathbf{x}_k^j)}{\sum_{l=1}^N g(\mathbf{x}_k^l)}, \tag{2}$$

$i, j = 1, \cdots, N$.

[1] Generally $\mathcal{F} \subseteq \mathcal{R}^m$. If a general mutation and/or scheme is invoked so that population may go beyond the feasible region, we assume $\mathcal{F}$ to be $\mathcal{R}^m$ itself for technical convenience.

S3: (Crossover) Select two vectors from $\mathbf{x'}_k^1, \cdots, \mathbf{x'}_k^N$ independently, each with equal probabilities among the N members, and perform a crossover operation with probability p between the two vectors. Randomly discard either one of the resulting two vectors and keep the other. Independently repeat this process N times to form a new population $\mathbf{x''}_k^1, \cdots, \mathbf{x''}_k^N$.

S4: (Mutation) Perturb $\mathbf{x''}_k^1, \cdots, \mathbf{x''}_k^N$ to form the next generation $\mathbf{x}_{k+1}^1, \cdots, \mathbf{x}_{k+1}^N$ according to a common conditional probability density function

$$f_{\mathbf{x}_{k+1}^i|\mathbf{x''}_k^i}(\mathbf{y}|\mathbf{x}) = f_{\mathbf{w}_k}(\mathbf{y}|\mathbf{x}), i = 1, \cdots, N, \qquad (3)$$

where $f_{\mathbf{w}_k}(\mathbf{y}|\mathbf{x})$, symmetrical in $\mathbf{y}$ around $\mathbf{x}$, is the conditional probability density function that characterizes the mutation operator at time k.

S5: Check stopping criteria. If not satisfied, update the parameters of crossover and/or mutation according to a specified set of rules and repeat **S2–S4**. Set $k \leftarrow k+1$. If stopping criteria are satisfied, exit.

After the initial settings of a population of randomly chosen solutions, a stochastic selection operator chooses with higher probability the solutions with large fitness values without necessarily discarding the "bad" ones. This has the consequence of concentrating the population in the regions with higher fitness. The resulting population of solutions is now processed through two new stochastic operators, crossover and mutation, that work in the direction opposite to selection. They force the population to increase its diversity, with the obvious purpose of exploring new regions where better solutions may lay. The crossover process performs a sort of random coordinate swapping, and the mutation a random scattering of solutions. In the framework of the discrete genetic algorithms where the solutions are coded as binary strings [15], [16], the crossover corresponds to exchanging part of the strings between the two chosen parents. In our continuous-space framework the crossover operation can be generally thought of as coordinate swapping. The idea (inspired by biological gene crossover) consists of random shuffling of coordinates between two parent vectors with the fundamental consequence of keeping the population within the feasible region while exploring new regions of the solution space. Various heuristic schemes for crossover have been proposed in the literature, particularly in the discrete framework. Variations of the basic idea go from bit-wise to coordinate-wise swapping [50], referred to as "discrete recombination" in [70], and combinations of them. It is rather difficult within a generic discrete setting to visualize the convergence behavior of the crossover operation, although various papers have pointed out the essential features of such operators. In Part II [1] we will restrict our analysis to a specific coordinate-wise crossover scheme. Under this approach, each pair of corresponding coordinates between the two parent vectors is allowed to swap its values independently, with a given probability p. We will discuss in greater detail the convergence of the operation in terms of the population distributions.

We chose to study here the simplest selection scheme—the fitness proportional selection—as defined above. More sophisticated schemes are also possible [82] with the intent of overcoming some of the drawbacks associated with the fitness-proportional one, such as nonuniform selection pressure throughout the process. However, our analysis concentrates on the simplest one, as it possesses the most essential feature associated with any selection operator, namely that of "survival of the fittest."

The crossover operator in **S3** can take various forms, as has been described in the literature for canonical GA's. In Part II of this paper [1] we choose to analyse a simple uniform crossover in much detail, revealing unique diversification properties associated with the crossover mechanism. It should be pointed out that our uniform crossover operation differs slightly from the ones proposed for canonical GA's, in that only one offspring is kept in the new generation, while the other is discarded. However, it possesses properties common to most crossover operators and is chosen for ease of analysis.

The mutation equation in **S4** corresponds to an increase in the diversity of the population induced through a stochastic operation performed on each member independently. This is in contrast to crossover where couples of solutions are combined to form a member of the new population. The formulation in **S4**, in terms of a conditional density, is rather general since it allows member-wise operation dependent on the member value. The result is a generalized convolution of the distribution of each member with the conditional density of the mutation noise. Special cases include: 1) identically independently distributed (IID) additive zero mean noise vectors across the whole population, 2) member-wise zero mean noise with a distribution dependent on the value of the member vector to which it is applied, and 3) noise with a distribution dependent on some global statistics of the entire population, etc.

The free parameters of crossover and/or mutation operators may be adapted as well [74]; for example, when the stopping criteria are not satisfied. This is a rather complicated issue and will not be discussed in this paper. We prefer to concentrate here on the consequence of the basic operations on the behavior of the search algorithm.

Our first approach to the mathematical description of the GACS will be on a simplified version that omits completely crossover (i.e., Step **S3** in the above formulation). The simplified version is interesting because it allows a visualization of how the selection process plays its role toward convergence against the diversity induced by the mutation process. The crossover operator will be treated separately in Part II [1]. The analysis concerning crossover can be combined with those of selection and mutation to form a general picture of the GACS.

In the following we will derive time-recursive relationships for the distribution of the population under the assumption that the population is large. This is done by letting the population size N go to infinity and deriving the consequent limiting behavior of selection, mutation, and crossover on the population distribution. Clearly, as the population size gets large, member points tend to cover the entire solution space continuously; thus, the behavior of the algorithm can be summarized by how dense the points are in the solution space. Our primary goal is to show that after sufficiently long time, the probability density function (PDF) of the population

will be narrowly concentrated around the global maximum (maxima) of the fitness function.

III. Basic Recursive Formulas and Convergence Results

In the theorem that follows, bridging the gap between a finite population GACS and its infinite sample version ($N \to \infty$), provides a recursive equation governing successive population densities under repeated alternations of selection and mutation. The proof of the theorem is similar to that by Ermakov and Zhiglyavsky [46], but includes many more details and is considerably more readable. It was assumed in [46] that the population size can vary with time, provided that it asymptotically approaches infinity. [46] also allows bounded measurement noise on the sampled fitness values. In our statement of the theorem, the population size is fixed over time, although considered in its asymptotic behavior, and the observed fitness values of the member solutions are noiseless. We state the theorem below and defer the proof to Appendix A.

Theorem 1: Let the algorithm be formulated according to (2) and (3), and let the mutation operator act on each of the N members of the population independently with the same *conditional probability law $f_{\mathbf{x}^i_{k+1}|\mathbf{x}'^i_k}(\cdot|\cdot)$ for $i = 1, \cdots, N$, hereafter denoted as $f_{\mathbf{w}_k}(\cdot|\cdot)$. Assume the fitness function $g(\mathbf{x})$ and the mutation conditional density $f_{\mathbf{w}_k}$ satisfy the following conditions:*

$$(\mathbf{A}) \quad 0 < g_{min} \le g(\mathbf{x}) \le g_{max} < \infty, \forall \mathbf{x} \in \mathcal{F},$$

$$(\mathbf{B}) \quad \sup_{\mathbf{x},\mathbf{z} \in \mathcal{F}} f_{\mathbf{w}_k}(\mathbf{x}|\mathbf{z}) \le M < \infty.$$

Then as $N \to \infty$, the time history of the simplified GACS (without crossover) can be described by a sequence of random vectors $\{\mathbf{x}_k\}_{k=0}^{\infty}$, $\mathbf{x}_k \in \mathcal{F}$, with densities:

$$f_{\mathbf{x}_{k+1}}(\mathbf{x}) = \frac{\int_{\mathcal{F}} f_{\mathbf{x}_k}(\mathbf{y}) g(\mathbf{y}) f_{\mathbf{w}_k}(\mathbf{x}|\mathbf{y}) d\mathbf{y}}{\int_{\mathcal{F}} f_{\mathbf{x}_k}(\mathbf{y}) g(\mathbf{y}) d\mathbf{y}}. \tag{4}$$

Similar recursions were also reported by Karlin [62], that studied the behavior of biological population evolution under various diversification operators. The results demonstrates that selection on a large population leads to a normalized multiplication of the current density with the fitness function. This is a sort of modulation of the population that emphasizes the part of the population with higher fitness. The mutation operator just results in a generalized convolution of the population density function with the mutation conditional density.

An immediate consequence of the above theorem, in the typical case of additive mutation noise:

$$\mathbf{x}^i_{k+1} = \mathbf{x}'^i_k + \mathbf{w}^i_k, \quad i = 1, 2, \cdots, N, \tag{5}$$

where $\mathbf{w}^i_k$, $i = 1, 2, \cdots, N$ are independent and identically distributed m-dimensional random vectors with zero mean and a common density $f_{\mathbf{w}_k}(\cdot)$, is that as $N \to \infty$, the sequence of population densities $\{f_{\mathbf{x}_k}\}_{k=0}^{\infty}$ satisfies the following recursion:

$$f_{\mathbf{x}_{k+1}}(\mathbf{x}) = \frac{[f_{\mathbf{x}_k}(\mathbf{x}) g(\mathbf{x})] * f_{\mathbf{w}_k}(\mathbf{x})}{\int_{\mathcal{F}} f_{\mathbf{x}_k}(\mathbf{y}) g(\mathbf{y}) d\mathbf{y}}, \tag{6}$$

where $*$ denotes the m-dimensional linear convolution.

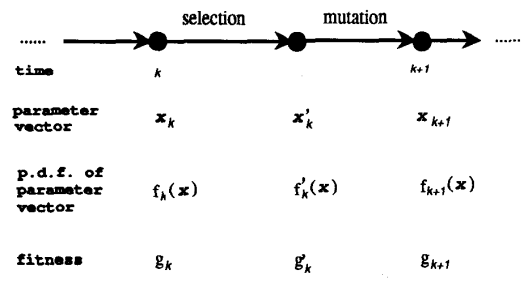

Fig. 2. Schematic illustration of selection and mutation steps for the GACS with corresponding notations.

For notational convenience of the analysis that follows, let us adopt the following simplified notations: denote $\mathbf{x}'_k$ the intermediate population *after* selection at time k, but *before* mutation. Also define:

$$f_k(\mathbf{x}) \triangleq f_{\mathbf{x}_k}(\mathbf{x}),$$
$$f'_k(\mathbf{x}) \triangleq f_{\mathbf{x}'_k}(\mathbf{x}),$$
$$g_k \triangleq g(\mathbf{x}_k),$$
$$g'_k \triangleq g(\mathbf{x}'_k).$$

Fig. 2 depicts the corresponding schematic diagram. Now (6) corresponding to additive independent mutation can be rewritten as:

$$f_{k+1}(\mathbf{x}) = \frac{[f_k(\mathbf{x}) g(\mathbf{x})] * f_{\mathbf{w}_k}(\mathbf{x})}{\int_{\mathcal{F}} f_k(\mathbf{y}) g(\mathbf{y}) d\mathbf{y}}. \tag{7}$$

To emphasize the individual behavior of the two random operators let us divide (7) into two parts. The effect of selection is expressed by the formula

$$f'_k(\mathbf{x}) = \frac{f_k(\mathbf{x}) g(\mathbf{x})}{E[g_k]}, \tag{8}$$

where $E[\cdot]$ denotes the mathematical expectation operator, and that of additive mutation by

$$f_{k+1}(\mathbf{x}) = f'_k(\mathbf{x}) * f_{\mathbf{w}_k}(\mathbf{x}). \tag{9}$$

The above expressions clearly show how the evolution process consists of an alternation of multiplication with the fitness function (selection) and convolution with the mutation density (mutation). The former tends to "squeeze" the density of $\mathbf{x}$ around the global maximum of the fitness function, whereas the latter "spreads" the resulting distribution. If the latter effect gradually diminishes (with a decreasing noise over time), the population will narrowly concentrate around the global maximum. Two pictorial examples are shown in Fig. 3 of a one-dimensional density that undergoes only repeated selection with two types of fitness functions (one unimodal and the other trimodal). The example shows how the population tends to concentrate itself on a point corresponding to the maximum fitness. The following theorem establishes the result rigorously:

Theorem 2 (Repeated Selection Alone): Let us consider a nonnegative function $g(\mathbf{x})$ defined over a bounded subset $\mathcal{F} \in \mathcal{R}^m$. Assume that $g(\mathbf{x})$ possesses a unique global maximum at $\mathbf{x} = \mathbf{x}^* \in \mathcal{F}$ with $g^* \triangleq g(\mathbf{x}^*)$. Also assume that there exists a

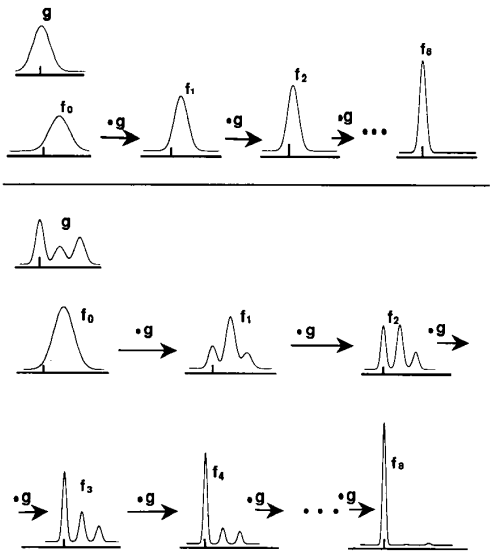

Fig. 3. Evolutions of a one-dimensional density under repeated selection (no mutation) for a unimodal fitness function (above) and a trimodal fitness function (below). The global maximum of the fitness function is marked on the horizontal axis.

finite neighborhood around $\mathbf{x}^*$ *that is simply connected*[2] *and that* $g(\mathbf{x})$ *is continuous within this neighborhood. If the initial density* f_0 *is nonzero at the optimal point* $\mathbf{x}^*$*, then the sequence*

$$f_{k+1}(\mathbf{x}) = \frac{f_k(\mathbf{x})g(\mathbf{x})}{\int_{\mathcal{F}} f_k(\mathbf{x})g(\mathbf{x})d\mathbf{x}}, k = 0, 1, \cdots, \qquad (10)$$

converges to $\delta(\mathbf{x} - \mathbf{x}^*)$[3] as $k \to \infty$.

The proof is deferred to Appendix B. The conclusion is intuitively clear, since raising a function to a power many times, while maintaining a unit volume underneath it, results in sharp peaks at the maximal points of the function. Note that the above arguments also apply to functions with multiple global maxima. In this case the sequence of densities for function values $\{f_{g(\mathbf{x}_k)}(\mathbf{y})\}_{k=0}^{\infty}$ would converge to $\delta(\mathbf{y} - g^*)$, where g^* is the common function values at the multiple global maxima.

Similar relationships have been derived for various forms of the selection operator in discrete space (see [82]), where exact solutions are obtained for difference equations describing a number of interesting selection schemes.

Note that the above theorem no longer holds in its present form when the population size is finite. Random sampling error during selection can lead a finite population of size N towards one of the N absorbing states, corresponding to each of the N initial members [41]. Theorem 2 essentially asserts that the population will be dominated by the member having the highest fitness if selection does not introduce sampling errors. The infinite population algorithm certainly represents such a case. There are other (say, deterministic) selection schemes on a finite population that also converge to the member of the largest fitness value in the initial generation. In addition, the more important process of selection combined with mutation no longer has well-defined absorbing states in the finite population case.

Next we study the combined effect of selection and mutation. The combination is illustrated in Fig. 4 for two one-dimensional examples with Gaussian additive mutation and two different fitness functions (one unimodal and the other trimodal). Note how the two operators play their opposite roles: selection emphasizes the regions with higher fitness, and mutation spreads the distribution. The competition between these two operators results in a final population concentrated in the neighborhood of the optimal point(s). Note also that Fig. 4 remains valid even if the initial population does not include the global maximum, since repeated mutation will eventually cover that point.

One may ask at this point the natural question: *Why should there be mutation at all if selection alone leads to convergence toward the optimal point?* To answer this we have to consider the large-population assumption. Essentially, we assume that the whole space is covered with nonzero probability by elements of the population. Therefore the optimal solution would already be contained in the sample and selection alone would suffice to find the optimal point(s). The global process of alternation between selection *and* mutation needs to be visualized in the context of a finite population where the mutation is responsible for allowing members to explore regions not previously covered, and that may correspond to larger fitness values. The purpose of the large sample assumption is to facilitate a quantitative description of the average behavior of the algorithm to which a finite sample algorithm approximates. In fact, the mutation process in the large sample assumption is certainly disruptive because it may lead to excessive spreading of the distribution without a guarantee of convergence. The mutation density has to be "narrow" enough to ensure convergence to a neighborhood of the optimal point. We first state the following lemma showing how selection alone increases the concentration of points in regions with current fitness above population average.

Lemma 1: Define the set of above-average vectors at time k *as*

$$B_k \triangleq \{\mathbf{x} \in \mathcal{F} : g(\mathbf{x}) \geq E[g(\mathbf{x}_k)]\}. \qquad (11)$$

Then the probability after selection of a member of the population to be in the above-average set is non-decreasing over time. Namely, $\Pr\{\mathbf{x}'_k \in B_k\} \geq \Pr\{\mathbf{x}_k \in B_k\}, \forall k \geq 0$.

The proof is straightforward since

$$\Pr\{\mathbf{x}'_k \in B_k\} = \int_{B_k} f'_k d\mathbf{x} = \int_{B_k} \frac{f_k g}{E[g(\mathbf{x}_k)]} d\mathbf{x} \geq \int_{B_k} f_k d\mathbf{x} = \Pr\{\mathbf{x}_k \in B_k\}.$$

The relationship holds also pointwise, since from (8), $f_{\mathbf{x}'_k} \geq f_{\mathbf{x}_k}$, $\forall \mathbf{x} \in B_k$. In fact, aside from the trivial case in which the fitness function is a constant and/or the population has converged to a single fitness level, points within the above-average set generally have fitness values strictly greater than the mean fitness; therefore the inequality in the above lemma is in general a strict one. This result, which can be considered

[2] Roughly speaking, simple connectedness implies that there is(are) no hole(s) within the neighborhood.

[3] $\delta(\mathbf{x} - \mathbf{x}^*)$ denotes Dirac's function, located at $\mathbf{x}^*$, namely an infinitely narrow peak at $\mathbf{x}^*$.

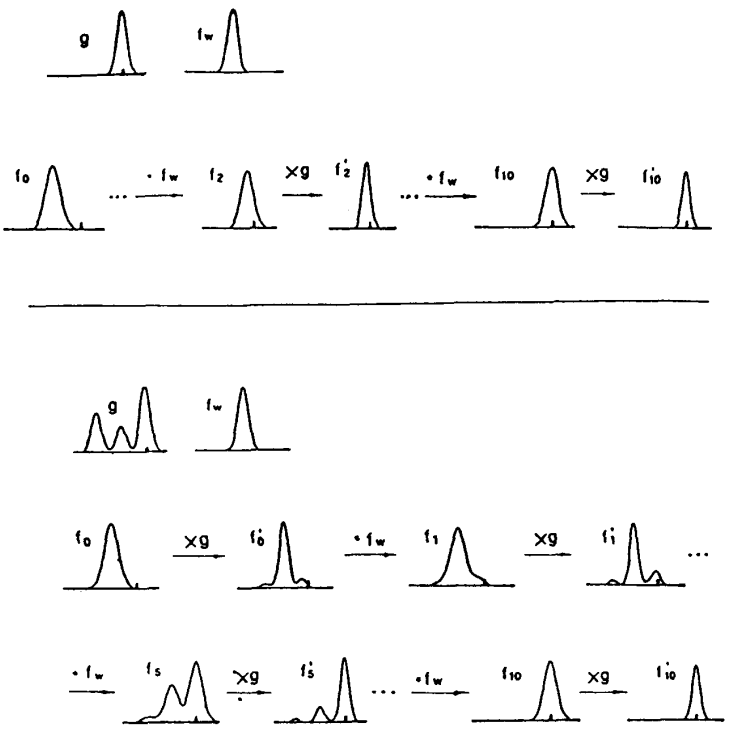

Fig. 4. Evolutions of a one-dimensional density function under selection and mutation. The additive mutation follows a Gaussian zero-mean distribution. The upper portion shows the case of a unimodal fitness function, and the lower portion shows a trimodal fitness function. "$*$" denotes linear convolution. The maximum of the fitness function is marked on the horizontal axis.

to be the continuous-space analog of Holland's Fundamental Theorem of GA [15], shows how the selection process rewards population members with above-average fitness. The mutation process, or more generally a diversification operation (which will include crossover, as we will discuss later), has to move the population probability mass outside the above-average-fitness regions; or in other words, "smooth" the distribution. Ermakov and Zhiglyavsky [47] have proved that under fairly non-restrictive condition(s), $f_k(\mathbf{x})$ will always converge to a δ-function concentrated on the unique maximum if σ_w^2, the variance of mutation noise, tends to zero. We have proved a similar theorem that establishes the existence of a sequence of mutation densities leading to monotone convergence of the population towards the global maximum (maxima).

Theorem 3 (Selection and Mutation): Consider A Function $g(\mathbf{x}) \geq 0$, $\forall \mathbf{x} \in \mathcal{F}$, *with a global maximum value equal to* g^* *(at possibly many locations). Suppose that continuity and simple-connectedness are satisfied within the neighborhood of each global optimal point, and the initial probability mass is nonzero in at least one of such neighborhoods; then there exist a sequence of* m*-dimensional mutation densities* $\{f_{\mathbf{w}_k}\}_{k=0}^{\infty}$ *with covariance matrices* $\{\sigma_k^2 \mathbf{I}\}_{k=0}^{\infty}$, *such that the sequence of densities defined as:*

$$f_{k+1} = \frac{[f_k g] * f_{\mathbf{w}_k}}{\int_{\mathcal{F}} f_k g d\mathbf{x}},$$

yields increasing average fitness converging to the global maximum value. Namely, $E[g_{k+1}] \geq E[g_k]$ *and* $\lim_{k-\infty} E[g_k] = g^*$.

The proof can be found in Appendix C. Note that no assumption on the shape of the noise distribution is made other than that it is not impulsive (it does not contain any Dirac functions). The theorem basically says that selection coupled with mutation can find the global maximum (maxima) of the fitness function with a sequence of populations having increasing average fitness, as long as the mutation noise is small enough. The result may appear to be useless since it is obvious from the above arguments that small perturbation will lead to convergence if the mutation density is sufficiently dense. However, if we want to apply the results of the infinite sample analysis to the finite population algorithms we have to use, in general, a mutation noise as large as possible to guarantee sufficient coverage of the feasible region. In a practical finite sample case we would want to use the sequence of mutation densities with the largest variance among those that guarantee convergence. The next section addresses this issue.

IV. Sufficient Conditions For Monotonic Increase of the Average Fitness (Selection and Mutation)

Practical use of theoretical results generally requires (as for almost any optimization algorithms) more specific assumptions on the feature of the fitness function. Let us remember that the evolutionary algorithm aims at the solution of difficult problems with possibly multi-modal discontinuous fitness functions. This poses the difficult problem of iden-

tifying global features of a fitness function and their roles in the convergence of the corresponding search. We derive here sufficient conditions for a simplified GACS to have monotonically increasing average fitness, given that the fitness function under consideration satisfies a Lipschitz condition. This class obviously excludes the discontinuous functions for which the Lipschitz number would be infinity, but it represents at least a first step toward a self-tuning GACS that can adjust its own parameters (probability of mutation, population size, etc.) according to the statistics of the population.

Suppose that in addition to conditions **a**)–**c**) of Section II, $g(\mathbf{x})$ is sufficiently smooth and satisfies the Lipschitz condition:

d)

$$|g(\mathbf{x}) - g(\mathbf{y})| \leq L\|\mathbf{x} - \mathbf{y}\|, \forall \mathbf{x}, \mathbf{y} \in \mathcal{F} \subseteq \mathcal{R}^m, \quad (12)$$

where $0 < L < \infty$ is the Lipschitz number.[4] We also assume the mutation noise to be zero mean and additive. We are looking for the "largest" mutation (the exact meaning of "largest" will be made clear below) that still guarantees monotonic increase of the average fitness

$$E[g_{k+1}] \geq E[g_k], \forall k. \quad (13)$$

Since

$$E[g_{k+1}] = \int_{\mathcal{F}} g(\mathbf{y}) f_{k+1}(\mathbf{y}) d\mathbf{y}, \quad (14)$$

substituting (9) into (14), we have

$$E[g_{k+1}] = \int_{\mathcal{F}} [f'_k(\mathbf{x}) \int_{\mathcal{F}} g(\mathbf{y}) f_{\mathbf{w}_k}(\mathbf{y} - \mathbf{x}) d\mathbf{y}] d\mathbf{x}. \quad (15)$$

The Lipschitz condition on $g(\cdot)$ implies that[5]

$$g(\mathbf{y}) \geq g(\mathbf{x}) - L\|\mathbf{y} - \mathbf{x}\|, \forall \mathbf{x}, \mathbf{y} \in \mathcal{F}, \quad (16)$$

and lower bounding (15) with (16), we have

$$\begin{aligned} E[g_{k+1}] &\geq \int_{\mathcal{F}} [f'_k(\mathbf{x}) g(\mathbf{x}) \int_{\mathcal{F}} f_{\mathbf{w}_k}(\mathbf{y} - \mathbf{x}) d\mathbf{y}] d\mathbf{x} \\ &\quad - L \int_{\mathcal{F}} [f'_k(\mathbf{x}) \int_{\mathcal{F}} \|\mathbf{y} - \mathbf{x}\| f_{\mathbf{w}_k}(\mathbf{y} - \mathbf{x}) d\mathbf{y}] d\mathbf{x} \\ &= E[g'_k] - L \int_{\mathcal{F}} [f'_k(\mathbf{x}) \int_{\mathcal{F}} \|\mathbf{y} - \mathbf{x}\| f_{\mathbf{w}_k} \\ &\quad \times (\mathbf{y} - \mathbf{x}) d\mathbf{y}] d\mathbf{x}. \end{aligned} \quad (17)$$

Where the second line in (17) holds since $\int_{\mathcal{F}} f_{\mathbf{w}_k}(\mathbf{y} - \mathbf{x}) d\mathbf{y} = 1, \forall \mathbf{x} \in \mathcal{F}$ according to the definition of conditional probability densities. Let us restrict ourselves to mutation densities that are spherically symmetrical, with an average radius of mutation defined as

$$\bar{r}(k, \mathbf{x}) \triangleq \int_{\mathcal{F}} \|\mathbf{y} - \mathbf{x}\| f_{\mathbf{w}_k}(\mathbf{y} - \mathbf{x}) d\mathbf{y}, \quad (18)$$

that in general can depend also on the current state $\mathbf{x}'_k$ (the point at which the mutation operates) and measures the average "spread" of the mutation. Substituting (18) into (17), we have

$$E[g_{k+1}] \geq E[g'_k] - L \int_{\mathcal{F}} f'_k(\mathbf{x}) \bar{r}(k, \mathbf{x}) d\mathbf{x}. \quad (19)$$

Substituting (8) in (19) and using the result to bound $E[g_{k+1}]$ from below, the monotonicity condition is satisfied if

$$E[g'_k] - \frac{L}{E[g_k]} \int_{\mathcal{F}} g(\mathbf{x}) \bar{r}(k, \mathbf{x}) f_k(\mathbf{x}) d\mathbf{x} \geq E[g_k], \quad (20)$$

or

$$\frac{L}{E[g_k]} \int_{\mathcal{F}} g(\mathbf{x}) \bar{r}(k, \mathbf{x}) f_k(\mathbf{x}) d\mathbf{x} \leq E[g'_k] - E[g_k]. \quad (21)$$

The right side of (21) represents the increase of average fitness due to selection alone. From (8) we have immediately that (the proof is given later in Section V)

$$E[g'_k] - E[g_k] = \frac{Var[g_k]}{E[g_k]},$$

therefore (21) becomes

$$\int_{\mathcal{F}} g(\mathbf{x}) \bar{r}(k, \mathbf{x}) f_k(\mathbf{x}) d\mathbf{x} \leq \frac{1}{L} \text{ Var } [g_k]. \quad (22)$$

Equation (22) is a sufficient condition for monotonic increase of average fitness. This condition can be further simplified by assuming that the mutation density depends only on the statistics associated with the entire population, instead of each individual member of the population; namely, $f_{\mathbf{w}_k}(\cdot)$ is independent of $\mathbf{x}'_k$ and depends only on the statistics (mean, variance, etc.) of the population. Consequently, the average radius is only a function of the time index k. The sufficient condition for monotonic increase of the average fitness (22) now becomes:

$$\bar{r}(k) \leq \frac{\text{Var}[g_k]}{LE[g_k]}. \quad (23)$$

This expression, although it is just a sufficient condition, leads to the following observations:

a) A non-smooth fitness functions (large Lipschitz number L) may require mutation with small radius for monotonic convergence.
b) A large current average fitness requires small noise to guarantee monotonicity. A large average expresses the fact that the population has already concentrated itself on regions with large fitness.
c) A large fitness variance corresponds to a population that is still rather spread out and can tolerate large mutation effects.
d) At convergence $\text{Var}[g_k] \rightarrow 0$ the mutation can be reduced accordingly.

We would like to emphasize that the condition derived is simply a condition for monotonic convergence of the average fitness value. Although Theorem 3 ensures the existence of a mutation sequence that guarantees monotonicity and convergence to the global maximum, condition (23) does not guarantee convergence to the global optimal point. The algorithm may still experience premature convergence to a local maximum. We have not been able to establish a sufficient

[4] $\|\mathbf{x}\|$ represents the usual Euclidean norm $\sqrt{\mathbf{x}^T\mathbf{x}}$.

[5] Note that the Lipschitz condition implies that $-L\|\mathbf{y} - \mathbf{x}\| \leq g(\mathbf{x}) - g(\mathbf{y}) \leq L\|\mathbf{y} - \mathbf{x}\|, \forall \mathbf{x}, \mathbf{y} \in \mathcal{F} \subseteq \mathcal{R}^m$. Using the right-hand side of it we have (16).

condition for global convergence, but have observed from our simulations on benchmark problems that conditions similar to (23) lead to satisfactory solutions in many cases. We regret not being able to include the results here due to space restriction. In this paper we focus on the above results for common classes of radially symmetric mutation densities.

A. The Generalized Gaussian Density

The Gaussian density can be generalized as follows:

$$f^{(l)}(\mathbf{x}) = \left[\sigma^m C_m \frac{m}{l}\Gamma\left(\frac{m}{l}\right)\right]^{-1} \exp\left[-\left(\frac{\|\mathbf{x}\|}{\sigma}\right)^l\right], \quad l > 0, \sigma > 0. \tag{24}$$

Its average radius is

$$\bar{r} = \frac{\Gamma\left(\frac{m+1}{l}\right)}{\Gamma\left(\frac{m}{l}\right)}\sigma, \tag{25}$$

where $\Gamma(\cdot)$ denotes the Gamma function. The calculations for the radius are carried out in Appendix D. Popular special cases are as follows:

1. For l=1 we have the radially symmetrical exponential density function that from (25) has $\bar{r} = m\sigma$, with the adaptive mutation law (23) becoming
$$\sigma(k) \leq \frac{\mathrm{Var}[g_k]}{m\mathrm{L}\ \mathrm{E}[g_k]}. \tag{26}$$
2. For l =2 we have the m-dimensional Gaussian density with zero mean and covariance matrix $\frac{\sigma^2}{2}\mathbf{I}$. Thus, condition (23) now becomes:
$$\sigma(k) \leq \frac{\Gamma\left(\frac{m}{2}\right)}{\Gamma\left(\frac{m+1}{2}\right)}\frac{\mathrm{Var}[g_k]}{\mathrm{L}\ \mathrm{E}[g_k]}. \tag{27}$$
Specifically, as m gets large, we can use Stirling's formula to approximate the factorials involved in the Gamma function obtaining
$$\frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \approx \sqrt{m},$$
and
$$\sigma(k) \leq \frac{\mathrm{Var}[g_k]}{L\sqrt{m}E[g_k]}. \tag{28}$$

It is interesting to see that in all the above cases as the dimensionality of the solution space grows, the mutation variance should become smaller in order for the sufficient condition of monotonicity to be satisfied.

B. The Gaussian/Gaussian Mixture

The above result is easily extended to the case of the mutation density being a Gaussian/Gaussian mixture (in fact, any convex mixture of densities belonging to the class (24)):

$$f_{\mathbf{w}_k} \sim (1-\epsilon)\mathcal{N}\left(\mathbf{0}, \frac{\sigma_1^2(k)}{2}\mathbf{I}\right) + \epsilon\mathcal{N}\left(\mathbf{0}, \frac{\sigma_2^2(k)}{2}\mathbf{I}\right), 0 \leq \epsilon \leq 1. \tag{29}$$

It is easy to verify that the condition for monotonicity (23) becomes:

$$\epsilon\sigma_2(k) + (1-\epsilon)\sigma_1(k) \leq \frac{\Gamma\left(\frac{m}{2}\right)}{\Gamma\left(\frac{m+1}{2}\right)}\frac{Var[g_k]}{LE[g_k]}. \tag{30}$$

Similar results can be obtained if the covariance matrices of the two constituent densities are kept constant over time, while the combination coefficients, ϵ and $1-\epsilon$, change with time. This corresponds to the case where the mutation noise is of "impulsive" type, with a large probability of making small jumps, and a small probability of making large jumps (i.e., $1-\epsilon(k) \gg \epsilon(k)$, and $\sigma_1 \ll \sigma_2$). The coefficient ϵ can also be made adaptive to the state of the population.

Finite population versions of (23) are derived in [74] where all the ensemble averages are replaced by their corresponding sample averages over the finite population:

$$\bar{r}(k) \leq \frac{\overline{Var}[g_k]}{L \cdot \bar{g}_k}. \tag{31}$$

where $\overline{Var}[g_k]$ and $\bar{g}_k$ are the sample variance and sample mean of the fitness, averaged over the finite population at time k.

V. Evolution of the Moments

In this section we analyze numerous consequences of the basic recursion (7). Specifically, we analyze the evolutions of the first and second moments of the random vectors $\{\mathbf{x}_k\}_{k=0}^{\infty}$ representing the population over time, as well as those of $\{g(\mathbf{x}_k)\}_{k=0}^{\infty}$. These quantities represent important statistics (mean and covariance) of the population, and convey interesting qualitative global behavior of the search. These recursions are derived under the general assumption of selection-mutation combination with the mutation density assumed to be symmetrical and invariant over time.

A. Evolution of the Mean of Fitness

Let us look at the mean increment of the fitness function $g(\mathbf{x})$ to be maximized, due to both selection and mutation. For simplicity, the arguments of the functions are dropped where no confusion can occur. (7) is rewritten as

$$f_{k+1} = \frac{(f_k \cdot g) * f_w}{E[g_k]}, \tag{32}$$

and the mean fitness at time $k+1$ is

$$E[g_{k+1}] = \frac{1}{E[g_k]}\int [(f_k g) * f_w] g\, d\mathbf{x} = \frac{1}{E[g_k]}\int\int f_k(\mathbf{s}) g(\mathbf{s}) f_w(\mathbf{x}-\mathbf{s}) d\mathbf{s}\, g(\mathbf{x}) d\mathbf{x} \tag{33}$$

Since f_w is symmetrical, $f_w(\mathbf{x}-\mathbf{s}) = f_w(\mathbf{s}-\mathbf{x})$; therefore the integration with respect to $\mathbf{x}$ becomes a convolution and we have

$$E[g_{k+1}] = \frac{1}{E[g_k]}\int f_k g[g * f_w] d\mathbf{s} = \frac{E[g_k(g * f_w)(k)]}{E[g_k]}, \tag{34}$$

where $(g * f_w)(k)$ denotes the function value of $(g * f_w)$ at $\mathbf{x}_k$. If mutation is not present, then $f_w = \delta$, where δ is the m-dimensional Dirac function and

$$E[g_{k+1}] - E[g_k] = \frac{\mathrm{Var}[g_k]}{E[g_k]} > 0. \tag{35}$$

Therefore, the mean increment of the fitness function due tos election alone at any time instant is always positive and proportional to the variance of the fitness, and inversely

proportional to its mean value. This leads to an interesting interpretation: 1) diversity within the population leads to greater improvement of the mean solution, and 2) improvements at relatively high mean fitness values are smaller.

B. Evolution of the Population Mean

Since the mutation process has zero mean, it has no effect on the evolution of the mean of $\mathbf{x}_k$ (Note that in a finite population, zero-mean mutation does affects the actual course taken by the algorithm. The result here should be considered to represent the average trajectory of a large population); therefore from (8):

$$E[\mathbf{x}_{k+1}] = E[\mathbf{x'}_k] = \frac{1}{E[g_k]} \int f_k g\, \mathbf{x}\, d\mathbf{x} = \frac{E[g_k \mathbf{x}_k]}{E[g_k]}, \quad (36)$$

hence,

$$E[\mathbf{x}_{k+1}] - E[\mathbf{x}_k] = \frac{E[g_k \mathbf{x}_k] - E[g_k]E[\mathbf{x}_k]}{E[g_k]} = \frac{E[[g_k - E[g_k]][\mathbf{x}_k - E[\mathbf{x}_k]]]}{E[g_k]}. \quad (37)$$

The numerator is the cross-covariance between the fitness function and the parameter vector at time k. We can conclude that the mean of the population evolves in the direction of the space mostly determined by the components that are highly correlated with the fitness value. At convergence, the random vector representing the population and its fitness will be uncorrelated. This is much similar to the zero-gradient condition for equilibrium points in conventional nonlinear programming techniques where no improvement occurs due to movements of the solution vector [43]. However the equilibrium condition here is a

global one, as opposed to the local one in nonlinear programming.

C. Evolution of the Fitness Variance

With a derivation similar to that used for the average fitness, we have that

$$E[g_{k+1}^2] = \frac{E[g_k (g^2 * f_w)(k)]}{E[g_k]}. \quad (38)$$

If no mutation is present, $f_w = \delta$. Hence

$$\mathrm{Var}[g_{k+1}] = \frac{E[g_k]E[g_k^3] - E^2[g_k^2]}{E^2[g_k]}. \quad (39)$$

Unfortunately the above expressions do not seem to lend themselves to an immediate interpretation.

D. Evolution of the Population Correlation Matrix

We can easily verify that

$$E[\mathbf{x}_{k+1}\mathbf{x}_{k+1}^T] = \sigma_w^2 \mathbf{I}_m + \frac{E[g_k \mathbf{x}_k \mathbf{x}_k^T]}{E[g_k]}, \quad (40)$$

and observe that the correlation among the coordinates of the solution space is weighted by the fitness function during selection.

VI. Evolution with a Quadratic Cost Function Behaves as a Newton Search

In this section we analyze the large-sample behavior of the simplified GACS with a quadratic cost function. This classical case has been the foundation for analyzing the convergence behavior of virtually all nonlinear programming algorithms [43], and it is perhaps one of the few tractable cases for detailed analysis. Although the evolutionary approach aims at the solution of complicated non-convex problems, the analysis of this case lends important clues to the near-convergence behavior of the search. Due to the nature of the Darwinian strategy, we transform the minimization of the quadratic cost function into the maximization of a Gaussian fitness function. The analysis leads to closed-form results regarding the convergence speed for the population distribution and fitness [75]. The results parallel standard analysis carried out in the optimization literature for gradient-decent algorithms [43].

A. The Fitness Function

Let the optimization problem be

$$arg\min_{\mathbf{x}} \frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T \mathbf{Q} (\mathbf{x} - \mathbf{x}^*), \quad (41)$$

where[6] $\mathbf{x}, \mathbf{x}^* \in \mathcal{F}$, and $\mathbf{Q} = \mathbf{Q}^T > 0$.

The fitness function could be any monotonically decreasing function of the cost function. For mathematical convenience (which will be justified in the sequel) we choose the negative exponential

$$g(\mathbf{x}) \triangleq \exp[-\frac{1}{2}(\mathbf{x} - \mathbf{x}^*)^T \mathbf{Q} (\mathbf{x} - \mathbf{x}^*)]. \quad (42)$$

Equation (42) above is a multi-dimensional Gaussian function (except for a positive scaling factor). As it will be seen later, this conversion greatly facilitates the analysis. Now the minimization problem is converted to the maximization of $g(\mathbf{x})$. This conversion technique is rather standard and can be applied to translate any unconstrained optimization problem into the general GACS framework, with the proper scaling of the original function.

B. Iteration of the Population Mean and Covariance Matrix

Let us assume that the algorithm starts with the Gaussian distribution $f_0(\mathbf{x}) = \mathcal{N}(\mu_0, \mathbf{\Sigma}_0)$. According to (7), all the densities in the sequence $\{f_k\}$ will be Gaussian, since $g(\mathbf{x})$ is Gaussian, and multiplication (as in (8)), followed by convolution (as in (9)), of two Gaussian functions still results in a Gaussian. Hence

$$\mathbf{x}(k) \sim \mathcal{N}(\mu_k, \mathbf{\Sigma}_k)$$
$$\mathbf{x}'(k) \sim \mathcal{N}(\mu_k', \mathbf{\Sigma}_k'), \forall k. \quad (43)$$

We can state the following theorem.

Theorem 4: For the GACS algorithm (selection and mutation), if

a) *the fitness function to be maximized is Gaussian;*
b) *the mutation process consists in addition of an m-dimensional Gaussian independent process with zero mean and covariance matrix* $\sigma_w^2 \mathbf{I}_m$*;*

[6] $\mathbf{Q} > 0$ indicates that $\mathbf{Q}$ is a positive definite matrix.

c) *the initial distribution is Gaussian with an arbitrary mean* μ_0 *and a positive definite covariance matrix* Σ_0;

then the population distribution is Gaussian at any time instant, with the mean and the covariance matrix sequences satisfying the following recursive equations:

$$\Sigma_{k+1} = \sigma_w^2 \mathbf{I}_m + (\mathbf{Q} + \Sigma_k^{-1})^{-1} \tag{44}$$

$$\mu_{k+1} = \mu_k - (\mathbf{Q} + \Sigma_k^{-1})^{-1} \nabla c(\mu_k). \tag{45}$$

Where $c(\cdot)$ denotes the quadratic cost function to be minimized (see equation (41)), and the term $\nabla c(\mu_k)$ denotes the gradient of the cost function evaluated at the population mean μ_k. The proof can be found in Appendix E. The results of the above theorem are quite striking since they show how the increments for the population mean are of the Newton type. The Hessian of the cost function (41) is $\mathbf{Q}$ and it is easy to verify that if the eigenvalues of $\Sigma_k \mathbf{Q}$ are large, the algorithm is approximately a Newton algorithm with increment $\mathbf{Q}^{-1} \nabla c(\mu_k)$.[7]

Due to the model that involves products and convolution, all the results presented here hold in a weak sense for rather arbitrary initial distribution and mutation. A weak form of the central limit theorem could be used to justify this claim: A sufficiently large number of convolutions will produce a near-Gaussian distribution; consequently, the iterations for the population mean and covariance matrix approximate those given for the Gaussian case. However, we have not yet been able to produce results that rigorously quantify this conjecture. In particular, the relationships derived here can be extended to the case of a uniform initial population; note that Theorem 4 holds trivially if the initial population concentrates on a single point in the solution space. In the case of a uniform initial distribution, the population density distribution can be thought of as the superposition of infinitely many Gaussian densities with equal strength, each resulting from a single starting point in the initial distribution. Therefore, the population density at any time instant takes the form of an integral of the Gaussian density over the uniform initial distribution, and similar equations can be derived for the iteration of the population mean and covariance matrix. Obviously, (44) remains the same under the assumption of a uniform initial population, since it does not depend on the starting point of the population. The exact form of the other equations under the assumption of a uniform initial distribution is currently under investigation.

We have obtained more results concerning the convergence behavior of the evolutionary algorithm in the quadratic case, that provide a full analysis of this case, including time complexity as a function of the mutation noise level and the eigen-structure of the quadratic form. They are contained in one of our technical reports [75]. We prefer not to include them here due to limited space. Also, a number of simulations for a quadratic cost function are included in the technical report, and show excellent agreement with the theory, even for a relatively small population size.

[7] Actually the form of the increments is a type of modified Newton search that is used when the Hessian matrix is ill-conditioned or near-singular. In those cases another positive definite matrix is added to the Hessian before inversion [43]. In the case of the genetic search the addition of this matrix is implicit in the algorithm.

VII. Conclusion and Discussion

This work aims at establishing a clear conceptual understanding of evolutionary programming by casting it into the unifying framework of stochastic processes in continuous space. This approach should facilitate a quantitative analysis of the dynamic behavior of the algorithm. Our work has been primarily concentrated on large population behavior of the evolutionary algorithms in continuous space. The following results have been achieved:

- **Algorithm Formulation:** The genetic algorithm has been formulated in the general framework of Markov chains, which greatly facilitates the application of mature techniques in stochastic processes theory to the analysis of the algorithms.
- **Evolution of the Population:** Recursive formulae for the distribution of the population under repeated applications of various "genetic operators" have been derived rigorously. The microscopic effect of each operator has been described quantitatively, and attempts have been made to combine the individual analysis showing how the operators interact with each other. Special attention has been paid to the selection-mutation combination showing how the alternation between contraction (selection) and spreading (mutation) of the population distribution drives the members of the population toward the optimal point(s) of the fitness landscape. The analysis also predicts interesting dynamic behavior of the statistic moments associated with the population.
- **Monotonicity condition under time-varying mutation:** A condition for monotonic convergence has been derived for the class of Lipschitz fitness functions. We are still investigating under which condition(s) the convergence corresponds to the global optimal point(s), although it is clear that the monotonicity condition derived indeed guarantees convergence to at least one local optimum. Analysis also shows that: 1) The exploration of the blind mutation operator should be more cautious (i.e., with smaller mutation noise) if the solution space has a higher dimension and the fitness landscape is rough. On the other hand, mutation jumps are more unrestricted for low-dimensional space with a smooth fitness function; and 2) the mutation noise should be small if the process is approaching a good solution when the average fitness of the population is high and the fitness variation over the population is low, as would naturally be expected.
- **Quadratic cost function and near convergence behavior:** Convergence behavior for the classical quadratic cost function has been analyzed revealing important clues of the near-convergence behavior of the algorithms. The asymptotic population mean is the optimal point regardless of the value of the noise power. The process of reaching the optimum follows modified Newton's steps. For large values of the noise power, the step become approximately a Newton's step and the mean reaches the optimum in very few iterations. The eigenvalues of the covariance matrix, however, get very large. It can be seen that a fundamental feature of evolution-

like algorithms is that the convergence of the mean of parameter vector towards the optimum is driven not only by the average function values of the population but also by the diversity of the population. On the other hand, a low degree of diversity is eventually required to achieve the required precision. Therefore, for any evolution-like algorithm there is always a fundamental trade-off between a high convergence rate of the mean and a low variance of the population. The reader is referred to our technical report [75] for more details, where a full account has been given for the convergence rates of the algorithm as well as bounds on them, all in terms of the eigenvalues of the Hessian of the quadratic cost function and the mutation variance. It should be emphasized that although the evolutionary algorithm is primarily targeted at multi-modal fitness functions, the unimodal Gaussian fitness function associated with the quadratic cost function represents a mathematically tractable case for a detailed analysis. This has been done to provide insights into the quantitative behavior of the algorithm. The quadratic case is also the most extensively studied case in the traditional nonlinear programming community [43], opening the way to a comparison of evolutionary programming with nonlinear programming.

In the following we discuss the results in the light of their possible extensions.

- **Extension to discrete-space GA:** The results of large sample analyses obtained in this paper can be extended to the canonical discrete-space GA with some modifications. The selection operator in discrete space can be studied in exactly the same manner as for the continuous-space one, since the selection operation does not depend on the metric norm of the space involved. The mutation operator can generally be formulated as a process of switching among different states of the solution space; therefore it can be characterized by a conditional probability density (or mass) function of a new state, given an old one. If a proper metric is defined over the space under consideration, then the mutation operator associated with the space can take a much simpler form. In the case of the Euclidean space studied in this report, the structure of the Euclidean norm permits us to define additive noise. From a formal point of view, there is no fundamental difference between a canonical discrete-space GA and its continuous-space counterpart. We are in the process of constructing formal stochastic models to extend the results of this paper to the discrete-space GA's.
- **Time-varying mutation:** The condition for monotonic convergence described in Section IV suggests an adaptive mutation scheme, and it represents a first step toward a general framework for the adaptation of various parameters (population size, probabilities of mutation, and crossover, etc.) of the algorithm based on certain statistics of the population. A more elaborate description of a set of adaptive mutation rules with simulation results can be found in [74]. The establishment of this new framework requires a full understanding of how various genetic operators affect the average fitness and the fitness variance of the population, in order to design the adaptation rules that ensure certain monotonicity (submartingale) conditions on the population average fitness. Adaptive GA's can at least partially eliminate the task of choosing the parameters for the algorithms, which is usually carried out heuristically. Actually, an adaptive GA might very well be the only one for which global convergence in probability can be proved. It should be noted that as long as the adaptation rules do not explicitly depend on time, the resulting Markov chain is still a time-homogeneous one; thus the analysis for an adaptive GA might not be more difficult than the non-adaptive ones.
- **Advantages and limitations of the large population assumption:** The large sample assumption greatly facilitates the understanding of the collective dynamic behavior of the entire population, as many asymptotic results in stochastic processes theory can be readily applied in this case. The assumption has been made in most current literature on genetic algorithms as well as in population genetics research. However the information on cross interaction between the members of the population is largely lost during the process of taking limits as the population size goes to infinity, since we are only looking at the marginal distribution of one generic member of the population. The large sample analysis does provide important insights into the average behavior of the algorithm of which the finite sample version is an approximation, and ignores the actual behavior of a single run of the algorithm. In a finite population, random sampling errors at certain point during the process may lead to locally optimal points. It is well conceded that exact convergence behavior for an arbitrarily finite population will be very difficult to obtain. In the finite sample case, since the joint distribution of the entire population does not depend on the particular order in which the members are arranged, the property of stochastic exchangeability of the member vectors constituting the population may be exploited (see [71] and [72] for attempts made in population genetics). The analysis seems to have been just preliminary. A more interesting work related to GACS is concerned with evolution of a population with both infinite size and infinite number of states [73]. The treatment, however, concentrates on a very restrictive class of fitness functions arising from purely biological concerns.
- **How infinite population results relate to a finite population algorithm:** The results of infinite population analysis presented in this paper are closely related to the case of a finite population algorithm: as the population size gets large, the typical behavior of a finite population algorithm approximates that predicted by the infinite sample analysis. We describe below how some of the results obtained in this paper can be directly extended to the finite sample case.

  Results of the evolution of the population densities: These results are essentially all derived from Theorem 1, and are mostly contained in Section III. It is proved in Appendix A that the histogram for the members of

a finite population of size N approaches the population density function predicted by (4), with a difference on the order of $\frac{1}{\sqrt{N}}$. This essentially states that in a finite population of reasonably large size, the percentage of population members within a particular region of the solution space can be approximated by the probability mass calculated from (4) over that region. Therefore, the finite population results can readily predict the typical behavior of a finite sample algorithm with a large population size. This applies particularly to the prediction of the mean behavior of a generic member of the population; e.g., the evolution of its mean position and correlations among its coordinates (loci), as has been shown in Section V. However, the joint stochastic characteristics among members of a finite population can not be captured by the large sample analysis in its present form.

Results of adaptive mutation schemes: Results contained in Section IV can essentially be extended directly to the finite population case, with all the ensemble statistics involved in the equations replaced by their sampled counterparts. A complete analysis of the variable mutation schemes for the finite population algorithms can be found in [81].

Results of the evolution of population moments: Some of the results in Section V, namely, that for the population mean and the average fitness, have been extended to the finite population case with both selection and mutation present [75], [81]. More general extensions, however, seems difficult.

Results of the Gaussian fitness case: These results, contained in Section VI, have been extended to the finite population case in [75] and [81] with some nonessential approximations. It is interesting to observe that the population mean vector follows a path jointly determined by both the negative gradient at the "center of mass" of the current population and those at each individual member positions, with the step size determined by the sample covariance matrix of the population.

In general, the finite population algorithm should be formulated and analyzed in a somewhat different manner from its infinite-sample counterpart. In the finite population case we are concerned about whether the actual configuration of the population is eventually dominated by the global optimal points, or points close to them. Therefore, the process can be characterized in terms of the almost sure convergence and/or convergence in probability of the sequence of populations towards a random population, whose members we expect to be concentrated around the global optimal point(s). For an infinite population algorithm, which is much easier to analyze, we are primarily interested in how the expected population would be dominated by global optimal points, after repeated use of genetic operators; therefore, the process is best described in terms of convergence in distribution of the sequence of populations. The infinite sample results can be related to the finite sample scenario in the following way: As the population size grows, the histogram over the member configurations will become ever closer to the density functions predicted by infinite-sample analysis.

In summary, the on-going investigation of GACS establishes some interesting properties of evolutionary optimization paradigms in general, and GACS in particular. The results shed much light on the understanding of canonical GA's in the discrete domain. They also reveal many unique properties of GACS, due to the special structure of Euclidean space, that are not shared by canonical GA's.

APPENDICES

Appendix A. Proof of Theorem 1

Let the population at time k consist of N random row vectors $\mathbf{x}_k^1, \cdots, \mathbf{x}_k^N \in \mathcal{F}$. Let $f_{\mathbf{x'}_k^i}(\mathbf{x})$ and $f_{\mathbf{x}_{k+1}^i}(\mathbf{x})$ be the marginal probability density functions of the ith member of the population after selection and mutation, respectively, at time k, with $i = 1, \cdots, N$. Refer to Fig. 1 for a schematic diagram. Let us define the $(N \times m)$-dimensional vector $\mathbf{X}_k \triangleq [\mathbf{x}_k^1, \cdots, \mathbf{x}_k^N]$ containing the entire population at time k. Its joint probability density function is denoted as $f_{\mathbf{X}_k}(\mathbf{x}^1, \cdots, \mathbf{x}^N)$. Similarly we denote the entire population after selection at time k, but before mutation by another $N \times m$-dimensional vector $\mathbf{X'}_k \triangleq [\mathbf{x'}_k^1, \cdots, \mathbf{x'}_k^N]$, where each $\mathbf{x'}_k^i, i = 1, \cdots, N$, is selected independently from the N alternatives $\mathbf{x}_k^1, \cdots, \mathbf{x}_k^N$ independently with conditional probabilities:

$$\Pr\{\mathbf{x'}_k^i = \mathbf{x}_k^j | \mathbf{X}_k\} = \frac{g(\mathbf{x}_k^j)}{\sum_{l=1}^N g(\mathbf{x}_k^l)}, \quad i, j = 1, 2, \cdots, N.$$

Therefore, the probability density function of the ith element of the population after selection at time k, conditioned on the whole population at time k, is:

$$f_{\mathbf{x'}_k^i|\mathbf{X}_k}(\mathbf{z}|\mathbf{y}^1, \cdots, \mathbf{y}^N) = \frac{\sum_{j=1}^N \delta(\mathbf{z} - \mathbf{y}^j) g(\mathbf{y}^j)}{\sum_{l=1}^N g(\mathbf{y}^l)}, \forall i, j = 1, 2, \cdots, N, \forall \mathbf{z}, \mathbf{y}^1, \cdots, \mathbf{y}^N \in \mathcal{F}. \tag{A1}$$

where $\delta(\cdot)$ is the m-dimensional Dirac's function. Since the mutation process, described by a single conditional density $f_{\mathbf{x}_{k+1}^i|\mathbf{x'}_k^i}(\mathbf{x}|\mathbf{z}) = f_{\mathbf{w}_k}(\mathbf{x}|\mathbf{z})$, is performed independently on each element of the population, we have:

$$f_{\mathbf{x}_{k+1}^i|\mathbf{X}_k}(\mathbf{x}|\mathbf{y}^1, \cdots, \mathbf{y}^N) = \int_{\mathcal{F}} f_{\mathbf{x}_{k+1}^i|\mathbf{x'}_k^i}(\mathbf{x}|\mathbf{z}) f_{\mathbf{x'}_k^i|\mathbf{X}_k}(\mathbf{z}|\mathbf{y}^1, \cdots, \mathbf{y}^N) d\mathbf{z} \tag{A2}$$

$$= \int_{\mathcal{F}} f_{\mathbf{w}_k}(\mathbf{x}|\mathbf{z}) f_{\mathbf{x'}_k^i|\mathbf{X}_k}(\mathbf{z}|\mathbf{y}^1, \cdots, \mathbf{y}^N) d\mathbf{z} \tag{A3}$$

Substituting (A1) into (A3), we have:

$$f_{\mathbf{x}_{k+1}^i|\mathbf{X}_k}(\mathbf{x}|\mathbf{y}^1, \cdots, \mathbf{y}^N) = \int_{\mathcal{F}} f_{\mathbf{w}_k}(\mathbf{x}|\mathbf{z}) \frac{\sum_{j=1}^N g(\mathbf{y}^j) \delta(\mathbf{z} - \mathbf{y}^j)}{\sum_{l=1}^N g(\mathbf{y}^l)} d\mathbf{z} \tag{A4}$$

$$= \frac{\sum_{j=1}^N g(\mathbf{y}^j) \int_{\mathcal{F}} f_{\mathbf{w}_k}(\mathbf{x}|\mathbf{z}) \delta(\mathbf{z} - \mathbf{y}^j) d\mathbf{z}}{\sum_{l=1}^N g(\mathbf{y}^l)} \tag{A5}$$

$$= \frac{\sum_{j=1}^{N} g(\mathbf{y}^j) f_{\mathbf{w}_k}(\mathbf{x}|\mathbf{y}^j)}{\sum_{l=1}^{N} g(\mathbf{y}^l)} \times \forall i = 1, \cdots, N. \quad \text{(A6)}$$

Using total probability theorem and exchanging the order of summation and integration, we have

$$f_{\mathbf{x}_{k+1}^i}(\mathbf{x}) = \int_{\mathcal{F}^N} f_{\mathbf{x}_{k+1}^i|\mathbf{x}_k}(\mathbf{x}|\mathbf{y}^1, \cdots, \mathbf{y}^N) \times f_{\mathbf{x}_k}(\mathbf{y}^1, \cdots, \mathbf{y}^N) d\mathbf{y}^1 \cdots d\mathbf{y}^N \quad \text{(A7)}$$

$$= \sum_{j=1}^{N} \frac{\int_{\mathcal{F}^N} g(\mathbf{y}^j) f_{\mathbf{w}_k}(\mathbf{x}|\mathbf{y}^j)}{\sum_{l=1}^{N} g(\mathbf{y}^l)} \times f_{\mathbf{x}_k}(\mathbf{y}^1, \cdots, \mathbf{y}^N) d\mathbf{y}^1 \cdots d\mathbf{y}^N \quad \text{(A8)}$$

The elements of the population are exchangable (not necessary independent) since the labeling of the elements resulting from selection is arbitrary and the mutation acts on each element independently with the same conditional probability law. Therefore all the coordinates of $f_{\mathbf{X}_k}(\mathbf{y}^1, \cdots, \mathbf{y}^N)$ can be permutated at will without affecting the functional value. The integrals involved in (A8) will be the same for all $j = 1, \cdots, N$. This yields

$$f_{\mathbf{x}_{k+1}^i}(\mathbf{x}) = N \int_{\mathcal{F}^N} \frac{g(\mathbf{y}^j) f_{\mathbf{w}_k}(\mathbf{x}|\mathbf{y}^j)}{\sum_{l=1}^{N} g(\mathbf{y}^l)} \times f_{\mathbf{x}_k}(\mathbf{y}^1, \cdots, \mathbf{y}^N) d\mathbf{y}^1 \cdots d\mathbf{y}^N \quad \text{(A9)}$$

$\forall i = 1, 2, \cdots, N$, where j is any of the indices $\{1, 2, \cdots, N\}$.

Define two new random variables:

$$\eta_k^N \triangleq \frac{1}{N} \sum_{l=1}^{N} g(\mathbf{x}_k^l) \quad \text{(A10)}$$

$$\xi_k(\mathbf{x}) \triangleq g(\mathbf{x}_k^j) f_{\mathbf{w}_k}(\mathbf{x}|\mathbf{x}_k^j), j \in \{1, 2, \cdots, N\}. \quad \text{(A11)}$$

Note that since $\mathbf{x}_k^j$ is a random vector, $\xi_k(\mathbf{x})$ is a random variable, even though $f_{\mathbf{w}_k}(\cdot|\cdot)$ is a fixed density function. We now have:

$$f_{\mathbf{x}_{k+1}^i}(\mathbf{x}) = E\left[\frac{\xi_k(\mathbf{x})}{\eta_k^N}\right], \quad \text{(A12)}$$

where $E[\cdot]$ is the expectation over the density $f_{\mathbf{X}_k}(\mathbf{y}^1, \cdots, \mathbf{y}^N)$. The law of large numbers for symmetrically dependent random variables [83] implies that

$$\lim_{N \to \infty} \eta_k^N = \eta_k, \quad a.s. \quad \text{(A13)}$$

where η_k is itself a random variable with mean:

$$E[\eta_k] = E[g(\mathbf{x}_k^j)] = \int_{\mathcal{F}} g(\mathbf{y}) f_{\mathbf{x}_k^j}(\mathbf{y}) d\mathbf{y}, \forall k, \text{ and any } j. \quad \text{(A14)}$$

Furthermore, η_k is independent of η_k^N for any finite N. In particular, η_k is independent of $\eta_k^1 = g(\mathbf{x}_k^j)$, for any j. This implies that η_k is independent of $\xi_k(\mathbf{x})$.

Now define

$$\Delta_k(N, \mathbf{x}) \triangleq \left| f_{\mathbf{x}_{k+1}^i}(\mathbf{x}) - \frac{\int_{\mathcal{F}} f_{\mathbf{x}_k^i}(\mathbf{y}) g(\mathbf{y}) f_{\mathbf{w}_k}(\mathbf{x}|\mathbf{y}) d\mathbf{y}}{\int_{\mathcal{F}} f_{\mathbf{x}_k^i}(\mathbf{y}) g(\mathbf{y}) d\mathbf{y}} \right|, \quad \text{(A15)}$$

We want to prove that $\lim_{N \to \infty} \Delta_k(N, \mathbf{x}) = 0, \forall k = 0, 1, 2, \cdots, \forall \mathbf{x} \in \mathcal{F}$. But

$$\Delta_k(N, \mathbf{x}) = \left| E\left[\frac{\xi_k(\mathbf{x})}{\eta_k^N}\right] - \frac{E[\xi_k(\mathbf{x})]}{E[\eta_k]} \right|, \quad \text{(A16)}$$

since η_k is positive and independent of both $\xi_k(\mathbf{x})$ and η_k^N, we have

$$\Delta_k(N, \mathbf{x}) = \frac{1}{E[\eta_k]} \left| E\left[\frac{\xi_k(\mathbf{x})\eta_k}{\eta_k^N}\right] - \xi_k(\mathbf{x}) \right| = \frac{1}{E[\eta_k]} \left| E\left[\frac{\xi_k(\mathbf{x})}{\eta_k^N}(\eta_k - \eta_k^N)\right] \right| \leq \frac{1}{E[\eta_k]} E\left[\frac{\xi_k(\mathbf{x})}{\eta_k^N}|\eta_k - \eta_k^N|\right].$$

Recall that $\eta_k, \eta_k^N \geq g_{\min}$, and $\xi_k(\mathbf{x}) \leq g_{\max} M$, we have:

$$\Delta_k(N, \mathbf{x}) \leq \frac{M g_{\max}}{g_{\min}^2} E[|\eta_k^N - \eta_k|]. \quad \text{(A17)}$$

Since the inequality in [83], p. 157:

$$E[|\eta_k^N - \eta_k|] \leq \frac{1}{\sqrt{N}} + ess\sup |\eta_k^N - \eta_k| \Pr\left\{|\eta_k^N - \eta_k| \geq \frac{1}{\sqrt{N}}\right\}, \quad \text{(A18)}$$

and the central limit theorem for symmetrically distributed random variables [84], it follows that

$$E[|\eta_k^N - \eta_k|] = O(\frac{1}{\sqrt{N}}). \quad \text{(A19)}$$

Therefore,

$$\lim_{N \to \infty} f_{\mathbf{x}_{k+1}^i}(\mathbf{x}) = \frac{\int_{\mathcal{F}} f_{\mathbf{x}_k^i}(\mathbf{y}) g(\mathbf{y}) f_{\mathbf{w}_k}(\mathbf{x}|\mathbf{y}) d\mathbf{y}}{\int_{\mathcal{F}} f_{\mathbf{x}_k^i}(\mathbf{y}) g(\mathbf{y}) d\mathbf{y}}, \quad i = 1, \cdots, N. \quad \text{(A20)}$$

This implies that for a relatively large population, the statistical behavior of the entire population over time can be summarized by a sequence of probability measures over time, which consists of the marginal probability density functions of a single member of the population at successive time steps. In other words, the time history of the entire population can be represented by a discrete time stochastic process $\{\mathbf{x}_k\}_{k=0}^{\infty}$ with $\mathcal{F}$ being the state space. The associated sequence of densities is:

$$f_{\mathbf{x}_{k+1}}(\mathbf{x}) = \frac{\int_{\mathcal{F}} f_{\mathbf{x}_k}(\mathbf{y}) g(\mathbf{y}) f_{\mathbf{w}_k}(\mathbf{x}|\mathbf{y}) d\mathbf{y}}{\int_{\mathcal{F}} f_{\mathbf{x}_k}(\mathbf{y}) g(\mathbf{y}) d\mathbf{y}}. \quad \text{(A21)}$$

Q.E.D.

Appendix B. Proof of Theorem 2

Given (10) through direct substitution, it is easy to verify that

$$f_k(\mathbf{x}) = \frac{f_0(\mathbf{x}) g^k(\mathbf{x})}{\int_{\mathcal{F}} f_0(\mathbf{x}) g^k(\mathbf{x}) d\mathbf{x}}, k = 0, 1, 2, \cdots, \quad \text{(B1)}$$

Given the uniqueness of the global maximum $\mathbf{x}^*$, the continuity and simple connectedness of $g(\mathbf{x})$ within its neighborhood, we can always find an $\epsilon > 0$, and a simply connected neighborhood of $\mathbf{x}^*$ defined as $B_{\mathbf{x}^*}(\epsilon) \triangleq \{\mathbf{x} \in \mathcal{F} : g(\mathbf{x}) \geq g^* - \epsilon\}$, such that for all $\mathbf{x} \in \mathcal{F} \backslash B_{\mathbf{x}^*}(\epsilon)$ we have $g(\mathbf{x}) \leq g^* - \epsilon$. See Fig. B1 for an example in a one-dimensional space. We

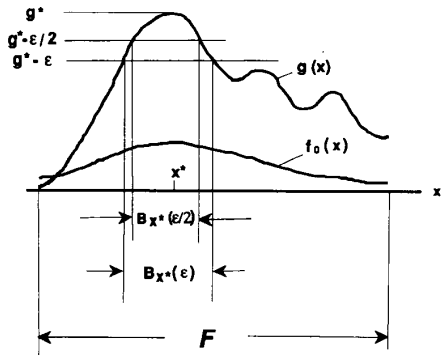

Fig. B1. Illustration of notations involved in the proof of Theorem 2.

want to prove that

$$\lim_{k\to\infty} \frac{\int_{\overline{B_{x^*}(\epsilon)}} g^k(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x}}{\int_{B_{x^*}(\epsilon)} g^k(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x}} = 0, \tag{B2}$$

where $\overline{B_{\mathbf{x}^*}(\epsilon)} \triangleq \mathcal{F} \backslash B_{\mathbf{x}^*}(\epsilon)$. This would guarantee that all the mass of the distribution is asymptotically contained in an arbitrarily small neighborhood around the global optimal point, and would guarantee the convergence of $\{f_k\}_{k=0}^{\infty}$ to $\delta(\mathbf{x}-\mathbf{x}^*)$.

For a sufficiently small ϵ in a simply connected neighborhood of $\mathbf{x}^*$ we have that

$$\frac{\int_{\overline{B_{x^*}(\epsilon)}} g^k(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x}}{\int_{B_{x^*}(\epsilon)} g^k(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x}} \le \frac{\int_{\overline{B_{x^*}(\epsilon)}} g^k(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x}}{\int_{B_{x^*}(\frac{\epsilon}{2})} g^k(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x}}, \tag{B3}$$

where the neighborhood $B_{\mathbf{x}^*}(\frac{\epsilon}{2})$ is an arbitrary set contained in $B_{\mathbf{x}^*}(\epsilon)$. Since

$$\int_{B_{x^*}(\frac{\epsilon}{2})} g^k(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x} \ge (g^* - \frac{\epsilon}{2})^k \int_{B_{x^*}(\frac{\epsilon}{2})} f_0(\mathbf{x}) d\mathbf{x}$$

$$\int_{\overline{B_{x^*}(\epsilon)}} g^k(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x} \le (g^* - \epsilon)^k \int_{\overline{B_{x^*}(\epsilon)}} f_0(\mathbf{x}) d\mathbf{x},$$

we have that

$$\frac{\int_{\overline{B_{x^*}(\epsilon)}} g^k(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x}}{\int_{B_{x^*}(\epsilon)} g^k(\mathbf{x}) f_0(\mathbf{x}) d\mathbf{x}} \le \left(\frac{g^* - \epsilon}{g^* - \frac{\epsilon}{2}}\right)^k \frac{\int_{\overline{B_{x^*}(\epsilon)}} f_0(\mathbf{x}) d\mathbf{x}}{\int_{B_{x^*}(\frac{\epsilon}{2})} f_0(\mathbf{x}) d\mathbf{x}} \le \left(\frac{g^* - \epsilon}{g^* - \frac{\epsilon}{2}}\right)^k \frac{1}{\int_{B_{x^*}(\frac{\epsilon}{2})} f_0(\mathbf{x}) d\mathbf{x}}. \tag{B4}$$

Since $f_0(\mathbf{x}^*) > 0$, we have

$$\int_{B_{x^*}(\frac{\epsilon}{2})} f_0(\mathbf{x}) d\mathbf{x} > 0,$$

and (B2) follows. **Q.E.D.**

Appendix C. Proof of Theorem 3

Construct the set of above-average vectors (see Fig. C1 for notations involved in this proof):

$$B_k \triangleq \{\mathbf{x} \in \mathcal{F} : g(\mathbf{x}) \ge E[g(\mathbf{x}_k)]\}.$$

We have already proved that

$$\Pr\{\mathbf{x}'_k \in B_k\} > \Pr\{\mathbf{x}_k \in B_k\}, \tag{C1}$$

therefore we can always find a time-dependent constant $K_k > 1$, such that

$$\Pr\{\mathbf{x}'_k \in B_k\} = K_k \Pr\{\mathbf{x}_k \in B_k\}.$$

Because $f'(\mathbf{x}) > f(\mathbf{x}), \forall \mathbf{x} \in B_k$, we can always choose a sufficiently small distance $\epsilon_k > 0$ and construct a set $B'_k \subset B_k$ with its boundary parallel to that of B_k and inside B_k by an equal amount ϵ_k (see Fig. C1), such that

$$\Pr\{\mathbf{x}'_k \in B'_k\} > \Pr\{\mathbf{x}_k \in B_k\}.$$

Thus we can define, for each $k > 0$, a constant $K'_k > 1$ and have

$$\Pr\{\mathbf{x}'_k \in B'_k\} = K'_k \Pr\{\mathbf{x}_k \in B_k\}. \tag{C2}$$

Now consider mutation

$$f_{k+1}(\mathbf{x}) = \int_{\mathcal{F}} f'_k(\mathbf{y}) f_{\mathbf{w}_k}(\mathbf{x}-\mathbf{y}) d\mathbf{y},$$

therefore,

$$\begin{aligned}\Pr\{\mathbf{x}_{k+1} \in B_k\} &= \int_{B_k}\int_{\mathcal{F}} f'_k(\mathbf{y}) f_{\mathbf{w}_k}(\mathbf{x}-\mathbf{y}) d\mathbf{y} d\mathbf{x} \\ &\ge \int_{B_k} [\int_{B'_k} f'_k(\mathbf{y}) f_{\mathbf{w}_k}(\mathbf{x}-\mathbf{y}) d\mathbf{y}] d\mathbf{x} \\ &= \int_{B'_k} f'_k(\mathbf{y}) [\int_{B_k} f_{\mathbf{w}_k}(\mathbf{x}-\mathbf{y}) d\mathbf{x}] d\mathbf{y} \\ &\ge \int_{B'_k} f'_k(\mathbf{y}) [\int_{\|\mathbf{x}-\mathbf{y}\|\le\epsilon_k} f_{\mathbf{w}_k}(\mathbf{x}-\mathbf{y}) d\mathbf{x}] d\mathbf{y}.\end{aligned}$$

But

$$\int_{\|\mathbf{x}-\mathbf{y}\|\le\epsilon_k} f_{\mathbf{w}_k}(\mathbf{x}-\mathbf{y}) d\mathbf{x} = \Pr\{\|\mathbf{w}_k\| \le \epsilon_k\},$$

that using Chebychev's Inequality gives

$$\Pr\{\|\mathbf{w}_k\| \le \epsilon_k\} \ge 1 - \frac{E[\|\mathbf{w}_k\|^2]}{\epsilon_k^2} = 1 - \frac{m\sigma_k^2}{\epsilon_k^2}. \tag{C3}$$

Now $\Pr\{\mathbf{x}_{k+1} \in B_k\}$ can be lower bounded as below:

$$\Pr\{\mathbf{x}_{k+1} \in B_k\} \ge \Pr\{\mathbf{x}'_k \in B'_k\}\left(1 - \frac{m\sigma_k^2}{\epsilon_k^2}\right). \tag{C4}$$

Substituting (C2) into (C4) we have:

$$\Pr\{\mathbf{x}_{k+1} \in B_k\} \ge K'_k \Pr\{\mathbf{x}_k \in B_k\}\left(1 - \frac{m\sigma_k^2}{\epsilon_k^2}\right).$$

In order to have $\Pr\{\mathbf{x}_{k+1} \in B_k\} > \Pr\{\mathbf{x}_k \in B_k\}$ it is sufficient that

$$K'_k\left(1 - \frac{m\sigma_k^2}{\epsilon_k^2}\right) > 1,$$

or

$$\sigma_k^2 < \frac{\epsilon_k^2}{m}\left(1 - \frac{1}{K'_k}\right). \tag{C5}$$

Therefore if (C5) is true for every k, then the probability mass contained in the set of above-average points will strictly increase as time goes on. This implies

$$E[g(\mathbf{x}_{k+1})] > E[g(\mathbf{x}_k)],$$

$$\lim_{k\to\infty} E[g(\mathbf{x}_{k+1})] = g^*.$$

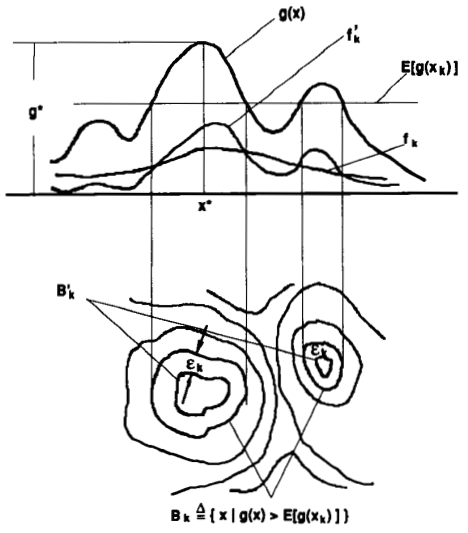

Fig. C1. One- and two-dimensional views of the effect of selection on the population density.

Q.E.D.

Appendix D. Evaluation of the Mean Radius of a Class of Generalized Gaussian Distributions

The average radius $\bar{r}$ of any m-dimensional spherically symmetrical density function $f(\mathbf{x})$ is defined as follows:

$$\bar{r} \triangleq \int_{\mathcal{R}^m} \|\mathbf{x}\| f(\mathbf{x}) d\mathbf{x}. \tag{D1}$$

Let us consider a special class of spherically symmetrical density functions known as generalized Gaussian densities [86].[8]

$$f^{(l)}(\mathbf{x}) \triangleq K(l) \exp\left(-\frac{\|\mathbf{x}\|}{\sigma}\right)^l, l > 0,$$

$$K(l) \triangleq \left[\int_{\mathcal{R}^m} \exp\left(-\frac{\|\mathbf{x}\|}{\sigma}\right)^l d\mathbf{x}\right]^{-1}.$$

Let us define a consant C_m such that the "volume" of a m-dimensional ball of radius r can be expressed as $C_m r^m$. Then

$$\begin{aligned}\int_{\mathcal{R}^m} \exp\left(-\frac{\|\mathbf{x}\|}{\sigma}\right)^l d\mathbf{x} &= \int_0^\infty \exp\left(-\frac{r}{\sigma}\right)^l C_m dr^m \\ &= \sigma^m C_m \int_0^\infty \exp\left(-\frac{r}{\sigma}\right)^l d\left(\frac{r}{\sigma}\right)^m \\ &= \sigma^m C_m \int_0^\infty \exp\left(-\frac{r}{\sigma}\right)^l d\left[\left(\frac{r}{\sigma}\right)^l\right]^{\frac{m}{l}} \\ &= \sigma^m C_m \frac{m}{l} \int_0^\infty x^{\frac{m}{l}-1} \exp(-x) d\mathbf{x} \\ &= \sigma^m C_m \frac{m}{l} \Gamma\left(\frac{m}{l}\right),\end{aligned}$$

[8]This definition is slightly different from Kassam's in [86] since in Kassam's definition the standard deviation is held a constant. In this case the parameter σ does not necessarily represent the standard deviation.

where $\Gamma(\cdot)$ is the standard Gamma function. Therefore

$$K(l) = \left[\sigma^m C_m \frac{m}{l} \Gamma\left(\frac{m}{l}\right)\right]^{-1}, \tag{D2}$$

and

$$f^{(l)}(\mathbf{x}) = \left[\sigma^m C_m \frac{m}{l} \Gamma\left(\frac{m}{l}\right)\right]^{-1} \exp\left(-\frac{\|\mathbf{x}\|}{\sigma}\right)^l. \tag{D3}$$

$\bar{r}$ defined by (D1) is

$$\bar{r} = \left[\sigma^m C_m \frac{m}{l} \Gamma\left(\frac{m}{l}\right)\right]^{-1} \int_{\mathcal{R}^m} \|\mathbf{x}\| \exp\left(-\frac{\|\mathbf{x}\|}{\sigma}\right)^l d\mathbf{x}. \tag{D4}$$

Going through calculations similar to those leading to (D2) we have:

$$\int_{\mathcal{R}^m} \|\mathbf{x}\| \exp\left(-\frac{\|\mathbf{x}\|}{\sigma}\right)^l = \sigma^{m+1} C_m \frac{m}{l} \Gamma\left(\frac{m+1}{l}\right). \tag{D5}$$

Substituting (D5) into (D4) the average radius is

$$\bar{r} = \frac{\Gamma\left(\frac{m+1}{l}\right)}{\Gamma\left(\frac{m}{l}\right)} \sigma. \tag{D6}$$

Appendix E. Proof of Theorem 4

We derive here the iteration formulae for μ_k and $\mathbf{\Sigma}_k$. First let us look at the effects of selection. From (42) and (43) the product of $g(\mathbf{x})$ and $f_k(\mathbf{x})$ has an exponent

$$\begin{aligned}&-\frac{1}{2}(\mathbf{x}-\mathbf{x}^*)^T\mathbf{Q}(\mathbf{x}-\mathbf{x}^*) - \frac{1}{2}(\mathbf{x}-\mu_k)^T\mathbf{\Sigma}_k^{-1}(\mathbf{x}-\mu_k) \\ &\quad = -\frac{1}{2}\mathbf{x}^T(\mathbf{Q}+\mathbf{\Sigma}_k^{-1})\mathbf{x} + (\mathbf{Q}\mathbf{x}^* + \mathbf{\Sigma}_k^{-1}\mu_k)^T\mathbf{x} \\ &\qquad -\frac{1}{2}(\mathbf{x}^{*T}\mathbf{Q}\mathbf{x}^* + \mu_k^T\mathbf{\Sigma}_k^{-1}\mu_k).\end{aligned} \tag{E1}$$

In (E1) only the first two terms involve $\mathbf{x}$ and are:

$$\begin{aligned}&-\frac{1}{2}\mathbf{x}^T(\mathbf{Q}+\mathbf{\Sigma}_k^{-1})\mathbf{x} + (\mathbf{Q}\mathbf{x}^* + \mathbf{\Sigma}_k^{-1}\mu_k)^T\mathbf{x} \\ &= -\frac{1}{2}[\mathbf{x} - (\mathbf{Q}+\mathbf{\Sigma}_k^{-1})^{-1}(\mathbf{Q}\mathbf{x}^* + \mathbf{\Sigma}_k^{-1}\mu_k)]^T(\mathbf{Q}+\mathbf{\Sigma}_k^{-1}) \\ &\quad \cdot[\mathbf{x} - (\mathbf{Q}+\mathbf{\Sigma}_k^{-1})^{-1}(\mathbf{Q}\mathbf{x}^* + \mathbf{\Sigma}_k^{-1}\mu_k)] \\ &\quad +\frac{1}{2}(\mathbf{Q}\mathbf{x}^* + \mathbf{\Sigma}_k^{-1}\mu_k)^T(\mathbf{Q}+\mathbf{\Sigma}_k^{-1})^{-1}(\mathbf{Q}\mathbf{x}^* + \mathbf{\Sigma}_k^{-1}\mu_k).\end{aligned} \tag{E2}$$

Now, only the first term in (E2) contains $\mathbf{x}$ with the second term being a constant. After substituting (E2) into (E1) and normalizing as in (8), the terms in (E1) not containing $\mathbf{x}$ disappear. Hence the resulting normalized product $g(\mathbf{x})f_k(\mathbf{x})/K$ is Gaussian with mean μ_k' and covariance matrix $\mathbf{\Sigma}_k'$ given as follows:

$$(\mathbf{\Sigma}_k')^{-1} = \mathbf{Q} + \mathbf{\Sigma}_k^{-1}, \tag{E3}$$

$$\mu_k' = (\mathbf{Q}+\mathbf{\Sigma}_k^{-1})^{-1}[\mathbf{Q}\mathbf{x}^* + \mathbf{\Sigma}_k^{-1}\mu_k]. \tag{E4}$$

The effects of the mutation is much simpler:

$$\mathbf{\Sigma}_{k+1} = \mathbf{\Sigma}_k' + \Sigma_w^2\mathbf{I}_m, \tag{E5}$$

and

$$\mu_{k+1} = \mu_k'. \tag{E6}$$

since f_w is zero mean.

Combining equations for both selection and mutation, we finally have

$$\mathbf{\Sigma}_{k+1} = \Sigma_w^2\mathbf{I}_m + (\mathbf{Q}+\mathbf{\Sigma}_k^{-1})^{-1}, \tag{E7}$$

and

$$\mu_{k+1} - \mu_k = -(\mathbf{Q} + \mathbf{\Sigma}_k^{-1})^{-1}\mathbf{Q}(\mu_k - \mathbf{x}^*). \quad \text{(E8)}$$

Noticing that

$$\mathbf{Q}(\mu_k - \mathbf{x}^*) = \nabla c(\mu_k), \quad \text{(E9)}$$

where $\nabla c(\mu_k)$ is the gradient of the quadratic cost function $c(\mathbf{x})$ at $\mathbf{x} = \mu_k$, we have

$$\mu_{k+1} - \mu_k = -(\mathbf{Q} + \mathbf{\Sigma}_k^{-1})^{-1}\nabla c(\mu_k). \quad \text{(E10)}$$

Q.E.D.

REFERENCES

[1] X. F. Qi and F. Palmieri, "Theoretical analysis of evolutionary algorithms with infinite population size: Part II. Analysis of the diversification role of crossover," this issue.

[2] A. Torn and A. Zilinskas, *Global Optimization*. Berlin, Heidelberg: Berlin, Heidelberg: Springer-Verlag, 1989.

[3] R. Horst and H. Tuy, *Global optimization—Deterministic Approaches*. Berlin, Heidelberg: Springer-Verlag, 1990.

[4] A. A. Zhigljavsky, *Theory of Global Random Search*. Kluwer Academic Publishers, 1991.

[5] F. Alluffi-Pentini, V. Parisi, and F. Zirilli "A global optimization algorithm using stochastic differential equations," *ACM Trans. on Math. Software*, vol. 14, no. 4, pp. 345–365, 1988.

[6] A. A. Zhigljavsky, "A Monte Carlo method for estimating functionals of suprema," Dr. Sci. Thesis, Leningrad Univ., (in Russian), 1987.

[7] J. Pinter, "Convergence properties of stochastic optimization procedures," *Math. Operat. Stat. Ser. Optimization*, vol. 15, pp. 405–427, 1984.

[8] G. Rappl, "Konvergenzraten von Random Search—Verfahren zur Globalen Optimierung," Ph.D. Thesis, Hochschule der Bundeswehr, Munchen, 1984.

[9] F. J. Solis and R. J-B. Wets, "Minimization by random search techniques," *Math. Operations Res.*, vol. 6, pp. 19–30, 1981.

[10] L. A. Rastrigin, "Random search, theory and application," *Systematic Bibliography, Inst. of Electronics and Computer Science of the Latvian SSR Acad. of Sci.*, Riga, p. 80 (in Russian), 1973.

[11] R. Z. Khas'minskij, "Random search applications to the problems of optimization and recognition," *Problems of Information Transmission*, vol. 65, no. 3, pp. 113–117 (in Russian), 1965.

[12] S. H. Brooks, "A discussion of random methods for seeking minima," *Operations Research* vol. 6, no. 2, pp. 244–251, 1958.

[13] E. H. L. Aarts and J. Korst, *Simulated Annealing and Boltzmann Machines*. New York: John Wiley and Sons, Ltd., 1990.

[14] H. Haario and E. Saksman, "Simulated annealing process in general state space," *Adv. Appl. Prob.*, vol. 23, pp. 866–893, 1991.

[15] J. H. Holland, *Adaptation in Natural and Artificial Systems*, Ann Arbor, MI: The Univ. of Michigan Press, 1975.

[16] D. E. Goldberg, *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Publishing Company, 1988.

[17] T. Bäck, F. Hoffmeister, and H.-P. Schwefel, "A survey of evolution strategies," in R. K. Belew, and L. B. Booker, Eds. *Proc. 4th International Conference on Genetic Algorithms*, pp. 2–9, San Mateo, CA: Morgan Kaufmann, 1991.

[18] I. Rechenberg, *Evolutionsstrategie—Optimierung Technischer Systeme nach Prinzipien der Biologische Information*. Stuttgart: Frommann Verlag, 1973.

[19] H. P. Schwefel, *Numerische Optimierung von Computer-Modellen der Evolutionsstrategie*. Basel: Birkhauser, 1977.

[20] L. J. Fogel, A. J. Owens, and M. J. Walsh, *Artificial Intelligence through Simulated Evolution*, New York: John Wiley, 1966.

[21] D. B. Fogel, *System Identification through Simulated Evolution: A Machine Learning Approach to Modeling*. Needham Heights, MA: Ginn Press, 1991.

[22] T. Boseniuk, V. Ebeling, and A. Engel, "Boltzmann and Darwin strategies in complex optimization," *Physics Letters A*, vol. 125, no. 6/7, pp. 307–310, 1987.

[23] R. M. Brady, "Optimization strategies gleaned from biological evolution," *Nature*, vol. 317, pp. 804–806, 1985.

[24] S. H. Bukatova, "Global search in evolutionary simulation," in *Methods of Global Optimization*, V. V. Fiodorov, Ed. (In Russian), pp. 60–80 1985.

[25] H. Mühlenbein, M. Gorges-Schleuter and O. Krämer, "Evolution algorithms in combinatorial optimization," *Parallel Computing*, vol. 7, pp. 65–88, 1988.

[26] Z. Michalewicz and C. Z. Janikow, "Handling constraints in genetic algorithms," in *Proceedings of the Fourth International Conference on Genetic Algorithms*, 1991, pp. 151.

[27] N. Baba, "Global optimization of functions by the random optimization method," *Int. J. Control*, vol. 30, pp. 1061–1065, 1977.

[28] R. A. Jarvis, "Adaptive global search by the process of competitive evolution," *IEEE Trans. on Systems, Man, and Cybernetics*, vol. 5, no. 3, pp. 297–311, 1975.

[29] C. Karnopp, "Random search techniques for optimization problems," *Automatica*, vol. 1, no. 213, pp. 111–121, 1963.

[30] G. J. McMurtry and K. S. Fu, "A variable structure automaton used as a multimodal searching technique," *IEEE Trans. on Automatic Control*, vol. 11, pp. 379–387, July 1966.

[31] D. Whitley, T. Starkweather, and D. Shaner, "The traveling salesman and sequence scheduling: Quality solutions using genetic edge recombination," in *The Genetic Algorithms Handbook*, L. Davis, Ed. 1991.

[32] D. Whitley, S. Dominic, and R. Das, "Genetic reinforcement learning for multilayered neural networks," Technical Report CS-91-107, Department of Computer Science, Colorado State University, May 1991.

[33] S. A. Harp and T. Samad, "Genetic synthesis of neural network architecture," in The Genetic Algorithm Handbook, L. Davis, Ed. Van Nostrand Reinhold, Chap. 15, 1990.

[34] T. E. Davis, "Toward an extrapolation of the simulated annealing convergence theory onto the simple genetic algorithm," Ph.D. thesis, University of Florida, Gainesville, FL., 1991.

[35] T. E. Davis and J. C. Principe, "A simulated annealing like convergence theory for the simple genetic algorithm," *Proceedings of the Fourth International Conference on Genetic Algorithms*, p. 174, 1991.

[36] M. D. Vose and G. E. Liepins, "Punctuated equilibria in genetic search," *Complex Systems*, vol. 3, pp. 31–44, 1991.

[37] G. E. Liepins and M. D. Vose, "Polynomials, basis sets, and the deceptiveness in genetic algorithms," *Complex Systems*, vol. 5, pp. 45–61, 1991.

[38] A. E. Eiben, E. H. L. Aarts, and K. M. Van Hee, "Global convergence of genetic algorithms: a Markov chain analysis," in *Proc. First Workshop on Problem Solving from Nature*, H.-P. Schwefel and R. Männer, Eds. Berlin, Heidelberg: Springer-Verlag, pp. 4–12, 1990.

[39] D. E. Goldberg, "Genetic algorithms and Walsh functions: Part I, A gentle introduction," *Complex Systems*, vol. 3, pp. 129–152, 1989.

[40] D. E. Goldberg, "Genetic algorithms and Walsh functions: Part II, Deception and its analysis," *Complex Systems*, vol. 3, pp. 153–171, 1989.

[41] D. B. Fogel, "Evolving artificial intelligence," Ph.D. thesis, University of California, San Diego, 1992.

[42] G. Rudolph, "Convergence analysis of canonical genetic algorithms," *IEEE Transactions on Neural Networks*, Special Issue on Evolutionary Programming, 1993.

[43] D. G. Luenberger, *Linear and Nonlinear Programming*, Addison-Wesley Publishing Company, 2nd edition, 1989.

[44] W. Feller, *An Introduction to Probability Theory and its Applications*. Vols. I, II, New York: John Wiley, 1957.

[45] S. M. Ermakov, Ed. *Mathematical Theory of the Design of Experiments*. Moscow: Nauka, pp. 392 (in Russian), 1983.

[46] S. M. Ermakov and A. A. Zhiglyavskij, "On random search of global extremum," *Probability Theory and Applications*, vol. 83, no. 1, pp. 129–136, 1983.

[47] S. M. Ermakov, A. A. Zhiglyavskij, and V. Solncev, "On a general scheme of random search for the extremum of a function," in *Monte Carlo Methods in Computational Mathematics and Mathematical Physics 1*, (in Russian), Novosibirsk, pp. 17–24, 1979.

[48] S. M. Ermakov and L. V. Mitiugova, "On one method of search of the extremum of a function based on estimation of a covariance matrix," *Antomatika i Vychistel'naya Tekhnika*, (in Russian), vol. 77, no. 5, pp. 38–41, 1977.

[49] D. E. Goldberg, "Real-coded genetic algorithms, virtual alphabets, and blocking," *Complex Systems*, vol. 5, pp. 139–167, 1991.

[50] A. H. Wright, "Genetic algorithms for real parameter optimization," in *Foundations of Genetic Algorithms*, G. J. E. Rawlins, Ed. San Mateo, CA: Morgan Kaufmann, pp. 205–218, 1991.

[51] W. Ebeling, "Pattern dynamics and optimization by reaction diffusion systems," *Journal of Statistical Physics*, vol. 45, nos. 5/6, pp. 841–903, 1986.

[52] W. Ebeling and A. Engel, *Syst. Anal. Model. Simul.*, vol. 3, pp. 377, 1986.

[53] W. Ebeling, A. Engel, and R. Feistel, "Diffusion and models of evolution processes," *Journal of Statistical Physics*, vol. 37, nos. 3/4, pp. 369–384, 1984.

[54] W. Ebeling and R. Feistel, "Stochastic theory of molecular replication processes with selection character," *Ann. Physik.*, 7. Folge, bd. 34, pp. 81–90, 1977.
[55] Yu. M. Svirezhev and V. P. Passekov, *Fundamentals of Mathematical Evolutionary Genetics*, The Netherlands: Kluwer Academic Publishers, 1990.
[56] W. J. Ewens, *Mathematical Population Genetics*. Berlin, Heidelgberg: Springer-Verlag, 1979.
[57] J. S. Gale, *Theoretical Population Genetics*, London: Unwin Hyman, 1990.
[58] M. Kimura and T. Ohta, *Theoretical Aspects of Population Genetics*. Princeton, NJ: Princeton University Press, 1971.
[59] T. Maruyama, *Stochastic Problems in Population Genetics*, in *Lecture Notes in Biomathematics*, vol. 17, S. Levin, Ed. Berlin: Springer-Verlag, 1977.
[60] S. Karlin and U. Liberman, "The reduction property for central polymorphisms in nonepistatic systems," *Theoretical Population Biology*, vol. 22, pp. 69–95, 1982.
[61] S. Karlin and H. Avni, "Analysis of central equilibria in multilocus systems: a generalized symmetric viability regime," Theoretical Population Biology, vol. 20, pp. 241–280, 1981.
[62] S. Karlin, "Models of multifactorial inheritance: I–VI," *Theoretical Population Biology*, vol. 15, pp. 308–439, 1979, and vol. 17, pp. 255–297, 1980.
[63] S. Karlin, "Central equilibria in multilocus systems: I–II," *Genetics*, vol. 91, pp. 777–816, 1979.
[64] S. Karlin, "Principles of polymorphism and epistasis for multilocus systems," *Proc. Natl. Acad. Sci. USA*, vol. 76, no. 1, pp. 541–545, 1979.
[65] S. Karlin, "Equilibrium behavior of population genetic models with nonrandom mating," Part I–Part II, *J. App. Prob.*, vol. 5, pp. 231–313 and pp. 487–566, 1968.
[66] L. R. Ginzburg and C. A. Braumann, "Multilocus population genetics: relative importance of selection and recombination," *Theoretical Population Biology*, vol. 17, pp. 298–320, 1980.
[67] I. Eshel and M. W. Feldman, "On the evolutionary effect of recombination," *Theoretical Population Biology*, vol. 1, pp. 88–100, 1970.
[68] J. D. Schaffer and L. J. Eshelman, "On crossover as an evolutionarily viable strategy," in *Proc. 4th International Conference on Genetic Algorithms*, R. K. Belew and L. B. Booker, Eds. San Mateo, CA: Morgan Kaufmann, pp. 61–68, 1991.
[69] M. Lipsitch, "Adaptation on rugged landscapes generated by iterated local interactions of neighboring genes," *Proceedings of the Fourth International Conference on Genetic Algorithms*, pp. 128, 1991.
[70] H.-P. Schwefel, *Numerical Optimization of Computer Models*. Chichester, U.K.: John Wiley, 1981.
[71] J. F. C. Kingman, "Uses of exchangeability," *Ann. of Probability*, vol. 6, no. 2, pp. 183–197, 1978.
[72] J. F. C. Kingman, "Coherent random walks arising in some genetical models," *Proc. R. Soc. Lond. A.*, vol. 351, pp. 19–31, 1976.
[73] I. Eshel, "On evolution in a population with an infinite number of types," *Theoretical Population Biology*, vol. 2, pp. 209–236, 1971.
[74] X. F. Qi and F. Palmieri, "General properties of genetic algorithms in the euclidean space with adaptive mutation and crossover," Technical Report EE-92-04, May, 1992.
[75] F. Palmieri and X. F. Qi, "Analysis of Darwinian algorithms in the continuous space," Technical Report EE-92-01, May, 1991.
[76] X. F. Qi and F. Palmieri, "Analyses of the genetic algorithms in the continuous space," *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing, San Francisco, CA*, March, 1992.
[77] X. F. Qi and F. Palmieri, "Large sample analyses of Darwinian algorithms," *Proc. 26th Annual Conference on Information Science and Systems*, Princeton, NJ: Princeton University, March 18–20, 1992.
[78] X. F. Qi and F. Palmieri, "Analyses of the genetic algorithms in the continuous space," *Proc. IEEE International Joint Conference on Neural Networks*, Baltimore, MD, June 7–11, 1992.
[79] X. F. Qi and F. Palmieri, "Adaptive Mutation in the Genetic Algorithm," *Proc. 2nd Annual Conference on Evolutionary Programming*, San Diego, CA, Feb. 25–26, 1993.
[80] X. F. Qi and F. Palmieri, "The Diversification Role of Crossover in Genetic Algorithms," *Proc. of the Fifth International Conference on Genetic Algorithms*, University of Illinois at Urbana-Champaign, July 1993.
[81] X. F. Qi, "Analysis and applications of Darwinian optimization algorithms in multi-dimensional spaces," Ph.D. dissertation, the University of Connecticut, April 1993.
[82] D. E. Goldberg and K. Deb, "A comparative analysis of selection schemes used in genetic algorithms," in *Foundations of Genetic Algorithms*, G. J. E. Rawlins, Ed. San Mateo, CA: Morgan Kaufmann, pp. 69–93, 1991.
[83] M. Loève, *Probability Theory*. Princeton, NJ: Van Nostrand, 3rd ed., 1963.
[84] J. R. Blum, H. Chernoff, M. Rosenblatt, and H. Teicher, "Central limit theorems for interchangeable process," *Canadian J. Math.*, vol. 10, pp. 222–229, 1958.
[85] L. C. W. Dixon and G. P. Szego, "The optimization problem: An introduction," *Towards Global Optimization 2*, L. C. W. Dixon and G. P. Szego, Eds. Amsterdam: North-Holland, 1978.
[86] S. A. Kassam, *Signal Detection in Non-Gaussian Noise*, Berlin Heidelberg: Springer-Verlag, 1988.

For photographs and biographies of **Xiaofeng Qi** and **Franceso Palmieri,** see Part II of this article.