Artificial Intelligence for Engineering Design, Analysis and Manufacturing (2001), *15*, 37–49. Printed in the USA.

Kriging as a surrogate fitness landscape in evolutionary optimization

ALAIN RATLE

Département de génie mécanique, Université de Sherbrooke, Sherbrooke, Québec, J1K 2R1 Canada

(Received October 9, 1998; Accepted July 25, 2000)

Abstract

The problem of finding optimal values in complex parameter optimization problems has often been solved with success by evolutionary algorithms (EAs). In many cases, these algorithms are employed as black-box methods over imprecisely known domains. Such problems arise frequently in engineering design. The principal barrier to the general use of EAs for those problems is the huge number of function evaluations that is often required. This makes EAs an impractical approach when the function evaluation depends on numerically heavy design analysis tools, for example, finite elements methods. This paper presents the use of kriging interpolation as a function approximation method for the construction of an internal model of the fitness landscape. This model is intended to guide the search process with a reduced number of fitness function evaluations.

Keywords: Evolutionary optimization; Fitness landscape; Function approximation

1. INTRODUCTION

Evolutionary algorithms (EAs) are recognized as a general approach for solving difficult multidimensional parameter optimization problems. These algorithms have, in the last years, been used with success in many branches of engineering design, such as vibration isolation (Keane, 1994, 1995), structural acoustics (Ratle & Berry, 1998), and active noise and vibration control (Baek & Elliott, 1995). However, a problem frequently faced with evolutionary optimization methods is the large number of fitness function evaluations required, since the computational complexity of this function is often a more than significant factor (Ratle & Berry, 1998). An optimization method making a more parsimonious use of fitness function evaluations is clearly preferable.

It is somewhat surprising that EAs make a rather limited use of the information obtained in optimization runs. Although these algorithms use function values to guide further search steps, no attention is paid to emergent structures between the sample points. One exception is found in evolution strategies and evolutionary programming, where individuals may contain information relative to the local variation structure of the fitness landscape (Bäck, 1996). Function approximations have long been used in optimal structural design for dealing with computationally expensive problems. An interesting review of approximation methods is given by Barthelemy and Haftka (1993). Another approach known as *meta-modeling* was proposed by Sartori and Smith (1997) for sensitivity analysis in capital valuation problems using polynomial regressions. It has also been proposed to use neural networks for learning a function equivalent to a complex problem, and then to optimize the network's response instead of the problem itself (Zimmerman, 1999). There is, however, no known optimization approach that makes an explicit use of a global model for extracting information structures beyond fitness values at discrete sampled points, and updating this information on–line as knowledge is gained during the run.

The main idea developed in this paper is an optimization methodology for building a statistical model of the fitness landscape from a small number of samples of the fitness function, and using this model to guide further search steps. Although it may seem to be an odd approach, the interest in using statistical tools for modeling deterministic computer experiments has been shown by Sacks et al. (1989). The samples required to build up the model are obtained during one generation of a simple EA. This statistical model is then exploited as an auxiliary fitness function to get the maximum amount of information out of the initial data

Reprint requests to: Alain Ratle, Institut Supérieur de L'Automobile et des Transports, 58027 Nevers, France. E-mail: Alain.Ratle_ISAT@U-Bourgogne.fr

points. Once a convergence criteria is satisfied, the algorithm turns back to the true fitness landscape for updating the model with fresher samples, expected to be closer to the global optimum. Great care must be taken in the choice of a suitable modeling technique. As a matter of efficiency for numerical optimization, the selected model must have some characteristics:

1. A small computational complexity compared to the real function;
2. An adequate representation of the global trends;
3. Considerations for local fluctuations around the data points, in order to detect emerging local optima.
4. A minimal number of initial hypotheses on the morphology of the landscape.

From the various existing techniques, kriging interpolation, a general tool developed in the framework of linear geostatistics, is chosen since it retains the required features. The paper is organized as follows. Section 2 is a brief introduction to EAs for function optimization. Section 3 presents the theoretical foundations of kriging. Section 4 details the implementation of kriging for evolutionary optimization, and, finally, Section 5 presents a series of computational experiments over a test suite of four problems.

2. EVOLUTIONARY ALGORITHMS

Evolutionary algorithms are a class of optimization methods inspired from the natural evolution of species. Using this paradigm, solutions are represented by a *population* of *individuals*. Each individual is coded by a string, the *genome*, noted **x**, which contains all the information describing a solution. To every genome corresponds a *fitness value* figuring the quality of this solution. Fitness is usually either the function to be optimized, or a scaled version of it.

The algorithm initially creates a population of μ individuals by assigning random values to the elements of the genomes. These individuals constitute the first generation. Subsequent generations are created by applying evolutionary operators: *selection* of parents, *crossover* or *recombination* of the genomes to create an offspring, and *mutation* of the offsprings. New generations are created until a stopping criteria is satisfied. In most cases, this criteria is a fixed number of generations or function evaluations. The basic scheme is as follows:

Algorithm 1. An elementary evolutionary algorithm.

Begin:
  Initialize the population
  while: stopping criterion not satisfied
    Select individuals according to fitness
    Recombine them
    Mutate these individuals
  end: *generation loop*
End: *evolutionary algorithm*

The selection is usually a biased random process, giving a higher probability of reproduction to better individuals. A simple and robust procedure is the *tournament*, where a small number of individuals are randomly picked out, and the best one among them is kept. Many other selection operators exist, as described in Hancock (1994) or Bäck (1996). The crossover operator is intended to perform an *exploitation* of promising regions by the creation of new points based on a recombination of existing genomes. For real-valued variables, crossover operators usually take a weighted sum of the values from two parents, as proposed by Michalewicz (1994). Mutation of real-valued variables consists of adding a random noise of a known distribution to all the variables. The mutation operator actually performs a random *exploration* of the search space, allowing the population to escape from local optima.

3. THEORY OF KRIGING

Kriging is a general tool for modeling experimental data in multidimensional spaces. The method emerged in mining geostatistics for the estimation of geophysical resources using as few samples as possible, because of the cost of the samples. Delineating regions of significant economical interest from waste requires identifying not only the average values of ore content, but also its structures of variation (Matheron, 1965). The analogy with function optimization appears at this point, since a valuable model is expected to guide the optimization toward optima with few or no needs for further samples. The original theory of kriging was formulated for one-, two-, or three-dimensional problems, reflecting physical phenomena. In the function optimization context, it is generalized to an L-dimensional problem. Under the theory of kriging, a phenomena $Z(\mathbf{x})$ is represented on a region $\mathcal{S}$ of the space by a linear combination $U(\mathbf{x})$ of N nonuniformly distributed samples $\mathbf{x}^i$, $i = 1 \ldots N$ (Matheron, 1973):

$$U(\mathbf{x}) = \sum_{i=1}^{N} \lambda_i Z(\mathbf{x}^i), \quad \mathbf{x}^i = x_1^i, \ldots, x_L^i. \tag{1}$$

The best linear unbiased estimator (BLUE) theory provides optimal weights λ_i in such a way that first, the expected values of $U(\mathbf{x})$ and $Z(\mathbf{x})$ are the same for all $\mathbf{x}$ in $\mathcal{S}$, and second, the estimation error is minimal.

3.1. Construction of the BLUE

Before proceeding, the underlying hypothesis should be stated. The general case considers the function as a stationary phenomena: all its statistical moments are assumed to be constant over $\mathcal{S}$. Under this hypothesis, the expected value of Z is a constant, and only stationary fluctuations are allowed. For many cases, this hypothesis is too restrictive, and the order-2 stationarity is employed, where all the statistical moments of order 2 and above are assumed constant. Under order-2 stationarity, the expected value is

https://doi.org/10.1017/S0890060401151024 Published online by Cambridge University Press

represented by a *drift function* noted $a(\mathbf{x})$. In the next section, kriging is presented under the stationarity hypothesis, and then order-2 stationarity is introduced.

3.1.1. Conditions on the linear estimator

The unbiasedness condition is verified as long as the expected values of Z and U are equal:

$$E\{U(\mathbf{x}^i)\} = E\{Z(\mathbf{x}^i)\} = m. \tag{2}$$

A linear relation between the λ_i's follows from Eq. (2):

$$\begin{aligned} E\{U(\mathbf{x}^i)\} &= E\left\{\sum_{i=1}^{N} \lambda_i Z(\mathbf{x}^i)\right\} \\ &= \sum_{i=1}^{N} \lambda_i E\{Z(\mathbf{x}^i)\} \\ &\Leftrightarrow \sum_{i=1}^{N} \lambda_i = 1. \end{aligned} \tag{3}$$

The minimization of estimation error cannot be enforced exactly, since the actual error is unknown, but an estimate of this error, the *estimation variance* $\sigma^2(\mathbf{x})$ is computed as follows:

$$\begin{aligned} \sigma^2(\mathbf{x}) &= E\left\{\left[Z(\mathbf{x}) - \sum_{i=1}^{N} \lambda_i Z(\mathbf{x}^i)\right]^2\right\} \\ &= E\{[Z(\mathbf{x})]^2\} - \sum_{i=1}^{N} 2\lambda_i E\{[Z(\mathbf{x})Z(\mathbf{x}^i)]\} \\ &\quad + \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j E\{Z(\mathbf{x}^i)Z(\mathbf{x}^j)\}. \end{aligned} \tag{4}$$

The minimum of $\sigma^2(\mathbf{x})$ is found where its first derivative with respect to each of the λ_i's is zero. A set of equations follows:

$$m - E\{[Z(\mathbf{x})Z(\mathbf{x}^i)]\} + \sum_{j=1}^{N} \lambda_j E\{Z(\mathbf{x}^i)Z(\mathbf{x}^j)\} = 0, \quad i = 1 \ldots N, \tag{5}$$

where m is the Lagrange multiplier associated with the unbiasedness constraint (Trochu, 1993).

3.1.2. Kriging as the BLUE

The N equations defined by Eq. (5) depend on $N + 1$ unknowns: the λ_j's and the term m. Introducing the no-bias condition, a linear system of $N + 1$ equation is defined. Solving the system gives the optimal λ_i's. Whenever experimental data show a global trend, the phenomenon is better modeled by the sum of a drift and a stationary fluctuation:

$$U(\mathbf{x}) = a(\mathbf{x}) + b(\mathbf{x}) \cong Z(\mathbf{x}). \tag{6}$$

The term $b(\mathbf{x})$ represents the stationary fluctuation, and $a(\mathbf{x})$ is the drift built up from a basis of M arbitrary functions, $f_j(\mathbf{x})$:

$$a(\mathbf{x}) = E\{Z(\mathbf{x})\} = \sum_{j=1}^{M} a_j f_j(\mathbf{x}). \tag{7}$$

In the presence of a drift, the no-bias conditions are stated as follows:

$$\sum_{i=1}^{N} \lambda_i f_j(\mathbf{x}^i) = f_j(\mathbf{x}), \quad j = 1 \ldots M, \quad \forall \mathbf{x} \in S. \tag{8}$$

Introducing these conditions, the linear kriging system is defined by

$$\left[\begin{array}{ccc|ccc} K_{11} & \ldots & K_{1N} & f_1(\mathbf{x}^1) & \ldots & f_M(\mathbf{x}^1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ K_{N1} & \ldots & K_{NN} & f_1(\mathbf{x}^N) & \ldots & f_M(\mathbf{x}^N) \\ \hline f_1(\mathbf{x}^1) & \ldots & f_1(\mathbf{x}^N) & 0 & \ldots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ f_M(\mathbf{x}^1) & \ldots & f_M(\mathbf{x}^N) & 0 & \ldots & 0 \end{array}\right] \left\{\begin{array}{c} \lambda_1 \\ \vdots \\ \lambda_N \\ \hline a_1 \\ \vdots \\ a_M \end{array}\right\} = \left\{\begin{array}{c} K_1 \\ \vdots \\ K_N \\ \hline f_1(\mathbf{x}) \\ \vdots \\ f_M(\mathbf{x}) \end{array}\right\}. \tag{9}$$

K_{ij} is the covariance between the sample points $\mathbf{x}^i$ and $\mathbf{x}^j$, that is, $E\{Z(\mathbf{x}^i)Z(\mathbf{x}^j)\}$, and K_i is the covariance between the sample point $\mathbf{x}^i$ and any point $\mathbf{x}$. Solving the system gives the optimal λ_i's at the point $\mathbf{x}$.

3.2. Dual formulation of kriging

The primal formulation given by Eq. (9) depends on the covariance between the samples $\mathbf{x}^i$ and the point $\mathbf{x}$ where the estimation is sought. By the way, the λ_i's also depend on that point. Independent coefficients are obtained with the dual formulation, which is calculated as follows. Since the matrix in Eq. (9) is symmetric, the inverse system has the form

$$\left\{\begin{array}{c} \boldsymbol{\lambda} \\ \hline \mathbf{a} \end{array}\right\} = \left[\begin{array}{c|c} \mathbf{B} & \mathbf{A} \\ \hline \mathbf{A}^T & \mathbf{C} \end{array}\right] \left\{\begin{array}{c} \mathbf{K} \\ \hline \mathbf{f} \end{array}\right\}. \tag{10}$$

Expressing the basic linear estimator [Eq. (1)] as a function of the inverse kriging system we have

$$\begin{aligned} U(\mathbf{x}) &= \{Z(\mathbf{x}^1) \ldots Z(\mathbf{x}^N)\}\mathbf{B}\left\{\begin{array}{c} K_1 \\ \vdots \\ K_N \end{array}\right\} \\ &\quad + \{Z(\mathbf{x}^1) \ldots Z(\mathbf{x}^N)\}\mathbf{A}\left\{\begin{array}{c} f_1(\mathbf{x}) \\ \vdots \\ f_M(\mathbf{x}) \end{array}\right\}. \end{aligned} \tag{11}$$

https://doi.org/10.1017/S0890060401151024 Published online by Cambridge University Press

By the symmetry of **A** and **B**, a new set of coefficients a_i and b_j is defined such that

$$U(\mathbf{x}) = \{b_1 \dots b_N\} \begin{Bmatrix} K_1 \\ \vdots \\ K_N \end{Bmatrix} + \{a_1 \dots a_M\} \begin{Bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_M(\mathbf{x}) \end{Bmatrix}. \quad (12)$$

These coefficients are found by substituting $\{Z(\mathbf{x}^1) \dots Z(\mathbf{x}^N), 0 \dots 0\}^T$ for the right-hand side of Eq. (10). Since the right-hand half of the matrix in Eq. (10) is not explicitly used, the submatrices **A** and **C** are left untouched, and the reinversion gives the dual formulation

$$\left[\begin{array}{ccc|ccc} K_{11} & \dots & K_{1N} & f_1(\mathbf{x}^1) & \dots & f_M(\mathbf{x}^1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ K_{N1} & \dots & K_{NN} & f_1(\mathbf{x}^N) & \dots & f_M(\mathbf{x}^N) \\ \hline f_1(\mathbf{x}^1) & \dots & f_1(\mathbf{x}^N) & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ f_M(\mathbf{x}^1) & \dots & f_M(\mathbf{x}^N) & 0 & \dots & 0 \end{array}\right] \left\{\begin{array}{c} b_1 \\ \vdots \\ b_N \\ \hline a_1 \\ \vdots \\ a_M \end{array}\right\}$$
$$= \left\{\begin{array}{c} Z(\mathbf{x}^1) \\ \vdots \\ Z(\mathbf{x}^N) \\ \hline 0 \\ \vdots \\ 0 \end{array}\right\}. \quad (13)$$

3.3. Covariance and variograms

Equation 13 shows that kriging requires the choice of a covariance function and a drift basis. The former is often the most difficult step, and requires knowledge of the underlying physics. For that purpose, two approaches are amenable. The first is the use of an arbitrarily defined theoretical function. As long as the matrix in Eq. (13) remains positive definite, any function is suitable. In this case the function is a *shape function*, since it has no more relationship with an actual covariance. Kriging in this case is considered an *interpolator*. The other approach is the estimation of an experimental covariance from the samples. Kriging is then considered an *estimator*, and the model reproduces the variance structure of $Z(\mathbf{x})$. Unfortunately, a covariance cannot usually be inferred from experimental data, and does not even exist in some cases (Journel & Huijbregts, 1978). For this purpose, the *variogram* is defined:

$$2\gamma(\mathbf{x}^i, \mathbf{x}^j) = E\{[Z(\mathbf{x}^i) - Z(\mathbf{x}^j)]^2\}. \quad (14)$$

The variogram is estimated from a set of data points. Under strict stationarity, it depends only on the scalar distance h between two points:

$$2\gamma(h) = E\{[Z(\mathbf{x}) - Z(\mathbf{x} + h)]^2\}, \quad \forall \mathbf{x} \in \mathcal{S}. \quad (15)$$

Adding a scalar h to the vectorial quantity **x** means that the direction of h is not significant. In this case, the variogram is said to be *isotropic*. However, real-world phenomena often show different behaviors with respect to each of the variables. In these cases, an oriented variogram $\gamma(\mathbf{h}) = \gamma(h_1, h_2, \dots, h_N)$ is defined, depending on the vectorial distance **h**. Such a variogram is said to be *orthotropic*. In presence of a drift (order-2 stationarity), the variogram is computed from the stationary fluctuation:

$$2\gamma(h) = E\{[Z(\mathbf{x}) - a(\mathbf{x}) - Z(\mathbf{x} + h) + a(\mathbf{x} + h)]^2\}, \quad \forall \mathbf{x} \in S. \quad (16)$$

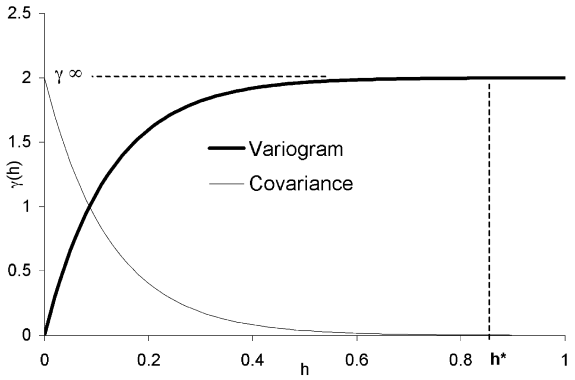

Fig. 1. Typical behavior of a variogram and a covariance.

Typical shapes of a variogram and a covariance are shown on Figure 1. The variogram is usually strictly increasing, although exceptions are found for oscillatory phenomena. Whenever the function is stationary, it presents a ceiling value γ^∞ at a distance h^*, the *distance of influence*. Absence of a ceiling tells that the stationarity assumption is wrong. It can be shown (Journel & Huijbregts, 1978) that a covariance may be calculated from an experimental variogram using the relation

$$K(h) = \gamma^\infty - \gamma(h). \quad (17)$$

3.4. Nested structures

Samplings from a physical phenomena often present different properties at different scales. The typical case is in gold ore estimation, where a point sample always has either 100% or 0% gold content, depending on whether it was taken from a nugget or from the surrounding material. A mapping from point samples would be totally useless, but samples averaged over a nonzero volume always have something between 0 and 100% ore content, and give a more useful mapping. Mathematically speaking, although the covariance between two very close points is high, it drops to nearly zero when the distance is increased by a small amount. The variogram $\gamma_0(h)$ of the point samples has a sharp slope at $h = 0$ and reaches a ceiling rapidly, whereas the variogram from nonzero volume samples is smoother. As long as order-2 statistics are concerned, the theory of nested struc-

https://doi.org/10.1017/S0890060401151024 Published online by Cambridge University Press

tures (Journel & Huijbregts, 1978) states that a general variogram can be built up as the sum of variograms observed at different scales:

$$\gamma(h) = \gamma_0(h) + \gamma_1(h) + \ldots + \gamma_i(h). \tag{18}$$

The limit case is a variogram with a discontinuity at $h = 0$, the *nugget effect* in reference to its physical signification:

$$\gamma(h) = \begin{cases} 0 & \text{if } h = 0 \\ C_0 + \gamma_1(h) & \text{else} \end{cases} \tag{19}$$

The constant C_0 is the nugget effect parameter, the amplitude of the discontinuity, and $\gamma_1(h)$ is a continuous function with a null value at the origin. Kriging, as presented up to this point, is an exact interpolator passing through all the data points. Using a nugget effect, the model does not pass through the points anymore, but tends toward the average behavior. The nugget effect is enforced by adding a constant to the N first diagonal elements of Eq. (13). In the case $C_0 \to \infty$, kriging reduces to a least-square regression on the drift basis. This phenomenon can be used by an optimization algorithm, by focusing on the global behavior during the first generations and taking into account the local fluctuations only in later stages.

4. IMPLEMENTATION

This section presents practical details of useful forms of kriging concerning theoretical covariance functions, experimental variograms, and models of drift basis. An optimization algorithm using fitness landscape approximation is then introduced.

4.1. Theoretical covariance functions

Four models are proposed. Two of them rely on the *distance of influence* introduced by Trochu (1993) for functional interpolation. It reflects the fact that the actual covariance between two distant points is often small enough to be considered null. The general covariance $K(h)$ might be designed in such a way that $K(h) = 0$ if $h > h^*$. An interesting feature of using a finite influence is that the kriging matrix becomes sparse, making possible the use of sparse system solvers.

Pure nugget effect: This model is the limit case where local fluctuations are considered insignificant. The model is useful for noisy data and when a global estimator is required:

$$K(h) = \begin{cases} 1 & \text{if } h = 0 \\ 0 & \text{else} \end{cases} \tag{20}$$

Linear model: This model vanishes linearly between $h = 0$ and $h = h^*$. Combined with a linear drift, it is strictly equivalent to a linear interpolation of the data points:

$$K(h) = \begin{cases} 1 - h/h^* & \text{if } h < h^* \\ 0 & \text{else} \end{cases} \tag{21}$$

Cubic covariance: This model ensures the continuity of $U(\mathbf{x})$ and its first derivatives by enforcing the nullity of $\partial K/\partial h$ at $h = 0$ and $h = h^*$. Two other conditions are $K(0) = 1$ and $K(h^*) = 0$. Solving for these four conditions over a cubic polynomial gives

$$K(h) = \begin{cases} 1 - 3(h/h^*)^2 + 2(h/h^*)^3 & \text{if } h < h^* \\ 0 & \text{else} \end{cases} \tag{22}$$

Gaussian covariance: This model has an infinite influence and is infinitely differentiable, which make it interesting for its robustness:

$$K(h) = e^{(-h^2/2\sigma^2)}. \tag{23}$$

The influence is controlled by the parameter σ. This model is strictly equivalent to the Radial Basis Functions found in the Support Vector Machines literature (Schölkopf et al., 1998).

4.2. Estimation of an experimental variogram

In this section, the estimation of a variogram function from the $\frac{1}{2}N(N-1)$ variogram samples available by taking pairwise N samples of $Z(\mathbf{x})$ is discussed.

4.2.1. Extracting the stationary fluctuation: Variogram of the residuals

In the general case where $Z(\mathbf{x})$ is nonstationary, a transformed function must be found before computing $\gamma(h)$. Should the drift function be known *a priori*, the stationary part would be deduced directly. A drift function $a(\mathbf{x})$ can be estimated by solving the kriging system under the hypothesis of pure nugget covariance. Then, the residuals $Z(\mathbf{x}) - a(\mathbf{x})$ give an approximation of the stationary part of $Z(\mathbf{x})$, as long as the drift basis contains enough degrees of freedom. The main drawback of this approach is that the system is solved twice, but it allows the use of arbitrary samples.

4.2.2. Polynomial regression of variogram data

A simple method for estimating a model from scattered samples is the least square regression over a function basis. An estimate $\gamma'(h)$ of $\gamma(h)$ can be expressed by a sum of k functions $g_j(h)$, $j = 1 \ldots k$, depending on k coefficients p_j:

$$\gamma'(h) = \sum_{j=0}^{k} p_j g_j(h). \tag{24}$$

A least-square optimal set of p_j's is found by minimizing the expected sum of squared error over the N samples:

$$E\{(\gamma'(h) - \gamma(h))^2\} = \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=0}^{k} p_j g_j(h_i) - \gamma(h_i) \right)^2. \tag{25}$$

https://doi.org/10.1017/S0890060401151024 Published online by Cambridge University Press

Equating to zero the first derivatives of Eq. (25) with respect to the p_j's gives a linear system whose solution corresponds to the optimal set of weights. Using $g_j(h) = h^j$ as a function basis gives the normal least-squares system, which is severely ill conditioned even for low orders, since it contains small and large terms (Gerald & Whitley, 1990). A better approach is the use of an orthogonal set of constant order polynomials, such as the Bézier-Bernstein polynomials, well known in computer-aided design for their flexibility and low computing cost (Farin, 1990). Using this basis, all the terms have the same order of magnitude, giving a better-conditioned matrix. Bézier-Bernstein polynomials are primarily intended for drawing parametric curves according to a variable number of *control points*. These points define the Bézier polygon, for which the curve is considered an approximation. For estimating a variogram, a one-dimensional curve is drawn with the distance h as a parameter normalized between 0 and 1. Least-square optimal coefficients are substituted to control points. The kth order polynomial $P(h)$ is a weighted sum of the Bernstein polynomials $B_{j,k}(h)$:

$$P(h) = \sum_{j=0}^{k} p_j B_{j,k}(h) \tag{26}$$

$$B_{j,k}(h) = \binom{k}{j}(1-h)^{k-j}h^j. \tag{27}$$

The expected squared error is obtained by introducing Eq. (26) into Eq. (25), and optimal weights are found by minimizing this error function analytically.

4.3. Drift function basis

Polynomial drift bases are common in linear geostatistics (Journel & Huijbregts, 1978). In the function optimization framework, the choice of basis for multidimensional spaces is restricted by the explosion in the number of required samples. A complete basis of order k comprises terms depending on all of the possible subsets of variables of size 1 to k:

$$a(\mathbf{x}) = a_0 + \sum_{i=1}^{L} a_i x_i + \sum_{i=1}^{L}\sum_{j=i}^{L} a_{ij} x_i x_j + \sum_{i=1}^{L}\sum_{j=i}^{L}\sum_{k=j}^{L} a_{ijk} x_i x_j x_k + \ldots. \tag{28}$$

The sum of linear terms contains L terms or, equivalently speaking, $\binom{L}{1}$ terms, the quadratic sum, $\binom{L+1}{2}$ terms, and the kth order sum shall be made of $\binom{L+k-1}{k}$ terms. From the additive properties of binomial coefficients (Graham, 1994), the sum of all the terms up to order k is equal to

$$\sum_{m=0}^{k}\binom{L+m-1}{m} = \binom{L+k}{k}. \tag{29}$$

Table 1. *Number of undetermined coefficients as a function of the dimensionality and polynomial order, for a diagonal and complete basis*

	Order, diagonal basis				Order, complete basis			
Dimensions	1	2	3	4	1	2	3	4
2	3	5	7	9	3	6	10	15
10	11	21	31	41	11	66	286	1001
20	21	41	61	81	21	231	1771	10,626

A simple alternative is the *diagonal* basis, which keeps the terms depending on only one dimension:

$$a(\mathbf{x}) = a_0 + \sum_{i=1}^{L}\sum_{j=1}^{k} a_{ij}(x_i)^j. \tag{30}$$

The diagonal basis needs $1 + kL$ coefficients, a quantity linear in L for any k, and the size of the complete basis is proportional to a polynomial of order k. Table 1 presents the number of coefficients of diagonal and complete basis of order 1 to 4 in 2-, 10-, and 20-dimensional spaces. Since a sample size at least equivalent to the number of drift coefficients is required, the use of a complete basis rapidly becomes impractical.

4.4. Optimization algorithm

The proposed optimization + fitness modeling algorithm is based on the real-coded EA presented in Section 2. The algorithm maintains two populations, a genetic population **G**, subject to genetic evolution, and a model population **M**, the long term memory. In the first generation, **G** is initialized similarly to the basic EA. Fitness is evaluated using the true function, **M** is initialized as a copy of **G**, and a first kriging model is built up from **M**. This model is exploited as a surrogate fitness function for several generations until a stopping criterion is satisfied. Then, in the next generation, some points in **G** are evaluated and added to **M** according to a given sampling scheme, and the kriging model is updated using these new points. The process is repeated as long as the global stopping criterion is not satisfied. To simplify the analysis, the model update criterion will be simply a fixed number of generations. The complete procedure is as follows:

Algorithm 2. Evolutionary optimization algorithm with fitness landscape approximation.

Begin:
Initialize the population **G**
Evaluate fitness
Copy **G** into **M** and build a fitness landscape model
fitness function ← model

https://doi.org/10.1017/S0890060401151024 Published online by Cambridge University Press

while: Global stopping criterion not satisfied
  if: Criterion for model update is satisfied
    fitness function ← original function
    Evaluate some points in **G** according to the sampling scheme
    Add these points to **M** and update the fitness model
    fitness function ← new model
  end: *if*
  Evaluate the population in **G** with the model
  Select individuals in **G** according to fitness
  Recombine them
  Mutate these individuals
end: *generation loop*
End: *Algorithm*

4.5. Sampling schemes

Using this algorithm, the model is updated after g generations have been evaluated with a given model. An important issue is how new samples should be taken from the true fitness function at every update, and how these new samples should replace the old ones. This task might be handled in many different ways, but two approaches are considered in this paper. The first strategy, called *rebuild from scratch*, consists of evaluating μ new individuals, that is, the complete generation, and replacing all the old samples with them. The model is therefore completely rebuilt from scratch every time. It is expected that the first model is at best a global approximation, while the subsequent ones will be more precise in some limited regions. The other approach, called *incremental growth*, consists of evaluating with the true fitness function the n_{best} best points in the current population, and possibly n_{rand} random points. These points are added into the model. However, to keep the model relatively cheap to evaluate, its size is limited. If the size exceeds the limit value, some points are randomly deleted before the new points are included.

5. COMPUTATIONAL EXPERIMENTS

The fitness landscape approximation algorithm has been tested over a set of four analytic functions with various properties. Numerous benchmark problems have been devised in the past; interesting cases are found in Michalewicz (1994), Bäck (1996), or Floudas and Pardalos (1996). The four chosen functions are presented in Figure 2. Computa-

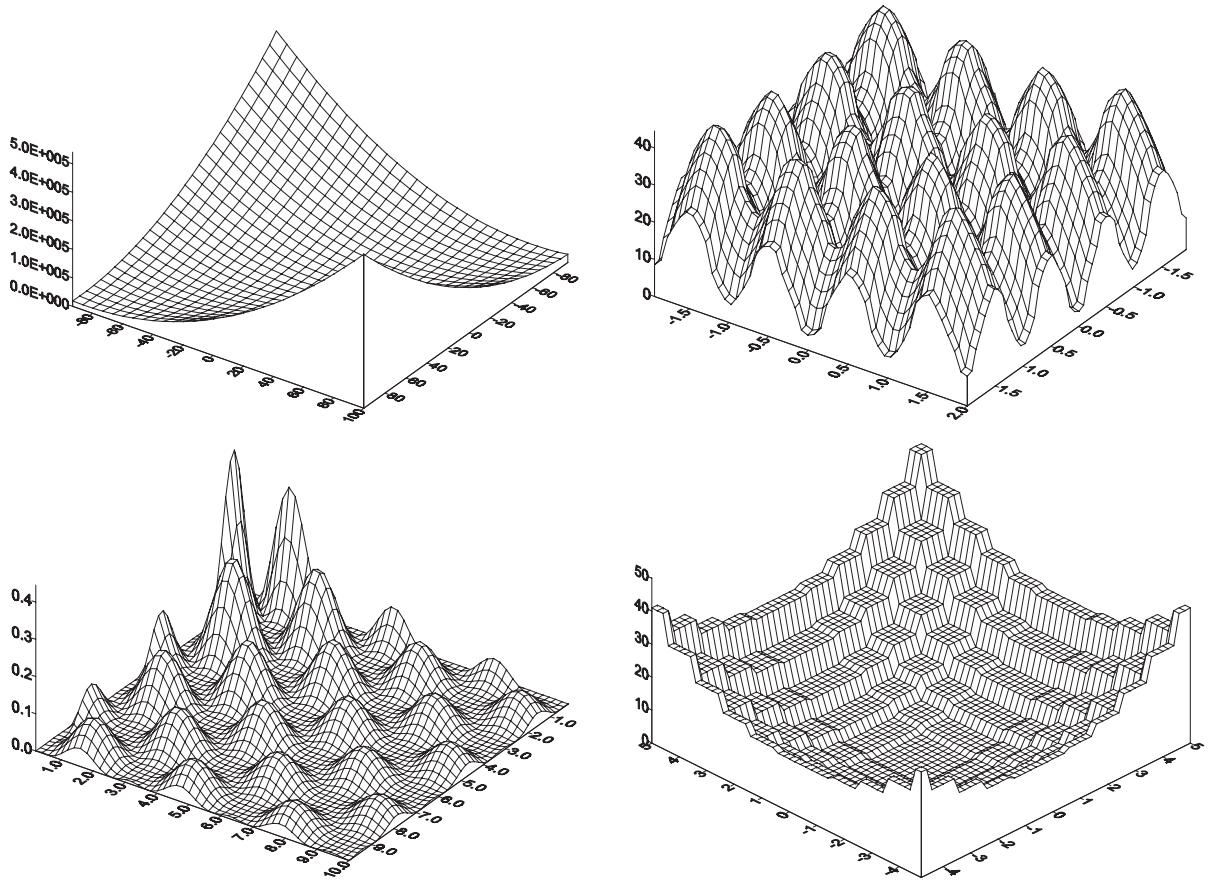

Fig. 2. The four test functions: Quadratic Problem f_1 (top left), Rastrigin's Multimodal Function f_2 (top right), Keane's Constrained Multimodal Function f_3 (bottom left), and the Step Function f_4 (bottom right).

https://doi.org/10.1017/S0890060401151024 Published online by Cambridge University Press

tional experiments for testing the effect of various parameters are presented:

- Population size, with the rebuild-from-scratch strategy;
- Influence of the drift basis order;
- Covariance functions, theoretical *versus* experimental;
- Constraint handling;
- Sampling strategies: the basic approaches, and the parameters: g, n_{best}, and n_{rand}.

Results are always represented by the fitness of the best-so-far individual found averaged over five runs, as a function of the number of real fitness evaluations. Although it is clearly not true for the analytical functions considered in this article, it is assumed that the cost of evaluating the model is insignificant compared to that of the real fitness function. The basic EA taken as a reference makes use of real-valued coding, Gaussian mutation with autoadaptive variance, and tournament selection. To keep this article concise, not all the parameters are tested for all the test functions; only a selection of the most significant results is presented. The four test cases are described as follows.

Function f_1: Quadratic. This first problem is a simple quadratic form:

$$\begin{aligned} &\text{Minimize } f_1(\mathbf{x}) = \tfrac{1}{2}\mathbf{x}^T\mathbf{A}\mathbf{x} + \mathbf{b}\mathbf{x} + c, \\ &\text{subject to } -100 \leq x_i \leq 100. \end{aligned} \tag{31}$$

Although deterministic algorithms can solve this problem faster than an EA, the interest is to show how a simple problem can be solved efficiently if the optimizer is allowed to build up an internal model.

Function f_2: Rastrigin's Multimodal. The Rastrigin's function is considered a difficult case since it consists of a large number of valleys hiding the minima. The problem is scaleable to an L-dimensions space, and is defined by

$$\begin{aligned} &\text{Minimize } f_2(\mathbf{x}) = 10L + \sum_{i=1}^{L} (x_i^2 - 10\cos(2\pi x_i)), \\ &\text{subject to } -2 \leq x_i \leq 2. \end{aligned} \tag{32}$$

It is observed from Figure 2 that the function possesses a strong periodical structure hiding a quadratic shape. The absence of a direct relationship between the local fitness landscape and global behavior makes the problem difficult unless the periodical structure is identified and taken into account in the fitness landscape estimation.

Problem f_3: Keane's Constrained Multimodal. This case was proposed by Keane (1994) for the study of EAs on constrained problems when the solution lies close to the edge of the feasible domain. The problem is as follows:

$$\begin{aligned} &\text{Maximize } f_3(\mathbf{x}) = \left| \frac{\sum_{i=1}^{L} \cos^4(x_i) - 2\prod_{i=1}^{L} \cos^2(x_i)}{\sqrt{\sum_{i=1}^{L} i x_i^2}} \right|, \\ &\text{subject to } 0 \leq x_i \leq 10 \\ &\text{and } \prod_{i=1}^{L} x_i > 0.75, \quad \sum_{i=1}^{L} x_i < 7.5L. \end{aligned} \tag{33}$$

This function also has a variable dimensionality. The constrained global optimum lies on the surface $\prod_{i=1}^{L} x_i = 0.75$. Solutions for the two-dimensional problem are found analytically to be 0.67367 at $(x_1, x_2) = (1.3932, 0)$ for the unconstrained version, and 0.36498 at $(x_1, x_2) = (1.60086, 0.46850)$ with constraints. The two-dimensional case has been successfully solved using a dynamic penalty function (Smith & Coit, 1997):

$$f_3'(\mathbf{x}) = f_3(\mathbf{x}) - \frac{t}{t_{\max}} \text{pen}(\mathbf{x}) \tag{34}$$

$$\text{pen}(\mathbf{x}) = \begin{cases} 1 - \dfrac{4}{3}\displaystyle\prod_{i=1}^{L} x_i & \text{if } \displaystyle\prod_{i=1}^{L} x_i < 0.75 \\ 0 & \text{otherwise.} \end{cases} \tag{35}$$

The time t corresponds to the current generation and $t_{\max}$ is the maximum number of generations. The function f_3' allows a free exploration of the space during the first generations, with a penalty increasing in later stages to push the population back into the feasible domain.

Problem f_4: Step. This last case, adapted from DeJong (1975), consists of an assembly of flat plateaus with a minimum value of zero in the square region $-0.5 \leq x_i \leq 0.5\ \forall i$:

$$\begin{aligned} &\text{Minimize } f_4(\mathbf{x}) = \sum_{i=1}^{L} \lfloor x_i + 0.5 \rfloor^2 \\ &\text{subject to } -5 \leq x_i \leq 5. \end{aligned} \tag{36}$$

The problem cannot be solved by a local search algorithm, since the gradient value is zero everywhere. The only way for an optimizer to solve such a problem is to have some *global view* of the fitness landscape. Since the minimum is a large region rather than a point, the case is well suited to a landscape approximation approach; the object is just to point out this region.

5.1. Population size with the rebuild-from-scratch strategy

The problem of choosing a suitable population size is studied for two cases: a very simple problem, the two-dimensional quadratic surface (f_1), and multidimensional cases: the step function (f_4) with 20 and 100 dimensions. The problems have been solved using the parameters described in Table 2, except for the population size.

The $f_1 - 2D$ problem being exactly defined by six points, the global minima is theoretically obtainable within six fitness evaluations. Since the sampling strategy requires that a final generation with the true fitness be calculated before terminating, in order to make sure the optimal solution really belongs to the fitness landscape, the algorithm can theoretically terminate after 12 evaluations. Results presented in Figure 3 show that the solution is not obtained within the limit of 12 evaluations with a population of 6. However, using a population of 20, the minimum is always reached in

https://doi.org/10.1017/S0890060401151024 Published online by Cambridge University Press

Table 2. *Optimization parameters for all the test problems*

	Problem						
	Quadratic	Rastrigin		Keane		Step	
Dimensions	2	2	10	2	20	20	100
Population size	20	50	200	50	300	50	250
Max. evaluations	400	1000	5000	2500	60,000	600	3000
Tournament size	2	2	3	2	4	2	2
Elitism	1	10	20	1	10	5	15
P_{mut}	1/3	1/4	1/40	1/2	1/20	1/40	1/200
P_{cross}	1	1	1	1	1	1	1
Drift	2C	2D	2D	3C	3D	2D	2D
Covariance	nugget	experimental		cubic		nugget	
Sampling	rebuild (if not specified)						

2C: Quadratic complete drift, 2D: quad. diagonal., 3C: cubic comp., 3D: cubic diag., P_{mut}: probability of mutation per variable, P_{cross}: probability of crossover per individual

40 evaluations, which is twice the population size. This suggests a lower bound on population size for an efficient use of the EA.

For the step function, the quadratic diagonal drift employed requires 41 samples in 20 dimensions, and 201 in 100 dimensions. Results are presented in Figure 4 for the 20-dimensional case with a population of 50. In this case, the kriging algorithm points out the optimal region after two generations with the true fitness landscape, which represents the maximum efficiency. On the 100-dimensional problem, the influence of population size is greater. Results shown in Figure 5 present three cases. The first makes use of an insufficient sampling with 150 points, the second one a critically sized population of 201, and the third one, 250 points. In all cases, the algorithm settles down to a constant fitness value at the second generation and does not improve anymore. However, only the larger sampling ensures the settling to the best fitness value, whereas the insufficient and critical samples make the algorithm converge to a suboptimal value.

5.2. Choosing a suitable drift basis order

The influence of the drift basis is illustrated for two cases: the 2-dimensional quadratic problem, in Figure 6 and the 10-dimensional Rastrigin's problem in Figure 7. Both problems have a quadratic general trend. In the two cases, it is observed that the best results are obtained when the appropriate basis is employed (quadratic). This points out the fact that there is room for *a priori* knowledge of the physics of the problem, in order to obtain the best possible optimization results. Whenever the nature of the problem is ignored, the algorithm takes a longer time, or converges to suboptimal solutions.

5.3. Covariance functions: Theoretical *versus* experimental

Figure 8 presents the effect of various covariance functions for the two-dimensional Rastrigin's problem. A simple EA is compared with reconstruction algorithms using either a

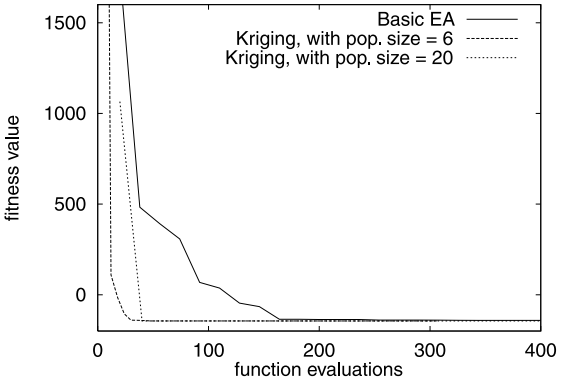

Fig. 3. Influence of population size with the 2-D function f_1.

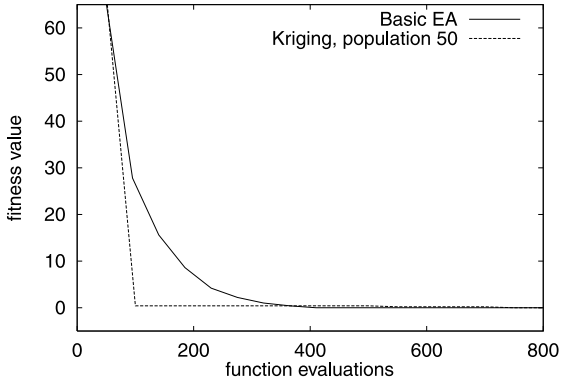

Fig. 4. Influence of population size with the 20-D step function f_4.

https://doi.org/10.1017/S0890060401151024 Published online by Cambridge University Press

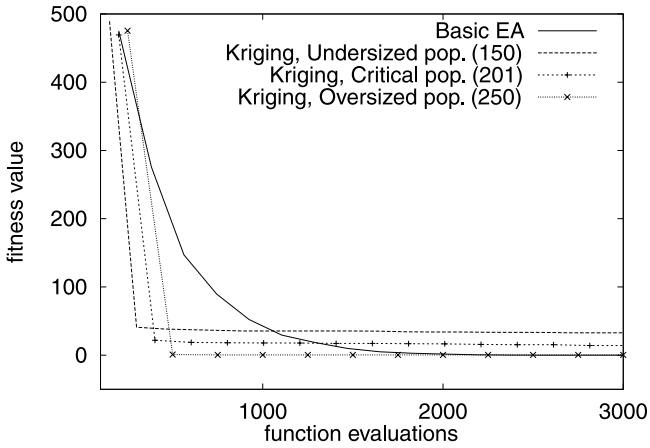

Fig. 5. Influence of population size with the 100-D step function f_4.

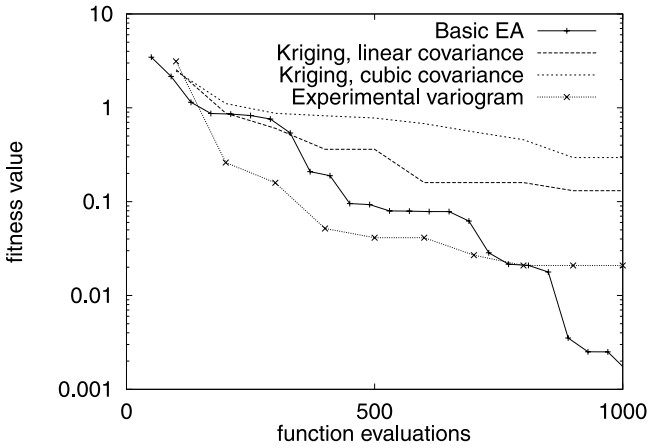

Fig. 8. Covariance functions for the 2-D Rastrigin's function f_2.

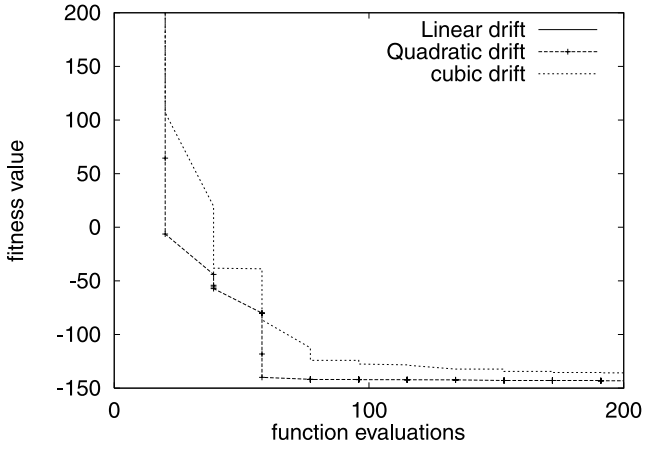

Fig. 6. Drift basis for the 2-D quadratic function f_1.

theoretical covariance or an experimental variogram. Since the difficulty of this case lies in the numerical precision of the solution rather than on the rough identification of the attractor, results are presented on a logarithmic scale of fitness value. As would be expected from the periodic structure of this function, theoretical covariance models, either linear or cubic, give only poor results compared to a simple EA, since these models cannot reproduce the covariance structure of the true landscape. Using an experimental variogram estimated from an order-4 Bézier polynomial allows us to find a fitness value of about 0.04 in roughly 400 fitness function evaluations compared to 800 for the basic EA. However, in the long run, the precision of the model is still insufficient to outperform the solution quality obtained without fitness approximation. For this problem, it is also clear from Figure 9 that the use of an isotropic variogram is not appropriate; better results are obtained when a different variogram is calculated for each of the dimensions.

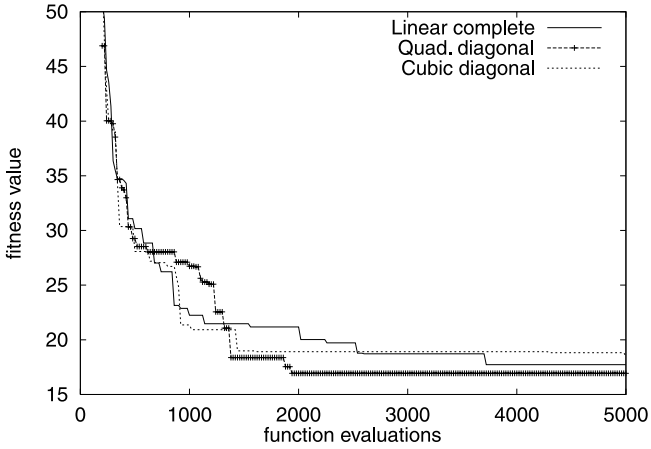

Fig. 7. Drift basis for the 10-D Rastrigin's function f_2.

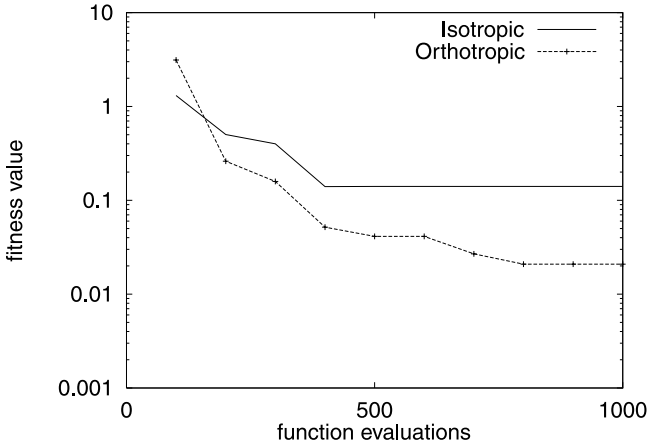

Fig. 9. Influence of an orthotropic variogram, 2-D Rastrigin's function.

https://doi.org/10.1017/S0890060401151024 Published online by Cambridge University Press

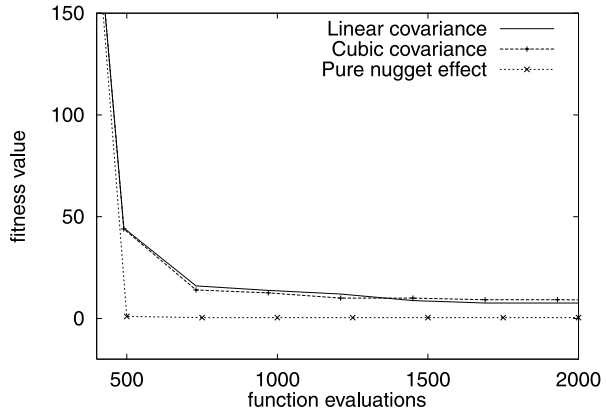

Fig. 10. Covariance models for the 100-D step function.

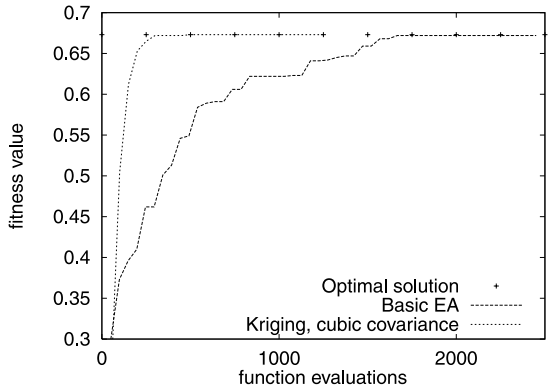

Fig. 12. Optimization results for the unconstrained 2-D Keane's function f_3.

For the 100-dimensional step function, opposite results are obtained. In this case, the best thing to do is to keep the model simple and global, since only a rough estimate of the optimal region is required, due to the discrete nature of this problem. It is seen from Figure 10 that a pure nugget effect covariance outperforms the other models in terms of solution quality. The 20-dimensional Keane's function also presents a different optimal covariance function. In this case (Fig. 11), the best results are obtained with the cubic covariance, while the linear covariance makes the algorithm converge to a suboptimal value.

5.4. Constraint handling

The fitness approximation algorithm has been tested on the Keane's problem to study the effect of a dynamically varying landscape. Results are presented in Figure 12 for the unconstrained problem with the function f_3 and in Figure 13 for the constrained version using f_3'. In both cases, using an approximated landscape built from a cubic covariance (theoretical) allows the solution to be found in about a quarter of the number of function evaluations required by the basic EA, with the parameters of Table 2.

The problem has also been studied in a 20-dimensional space. This case has long been considered a difficult problem for which no standard method gave satisfactory results (Michalewicz & Schoenauer, 1996). Some good solutions have been found by the same authors using specialized evolutionary operators intended to explore only the surface defining the edge of feasible domain. A similar approach is considered together with a fitness landscape reconstruction algorithm, using fitness function samples taken exclusively from the boundary of the feasible domain. This procedure gave the results presented in Figure 14, where a certain gain is obtained with the use of a landscape approximation.

5.5. Sampling strategies

The last important point to be discussed is how the model should be updated. Three parameters are studied, the number of generations between updates, g, the number of best individuals, n_{best}, reevaluated with the true function, and

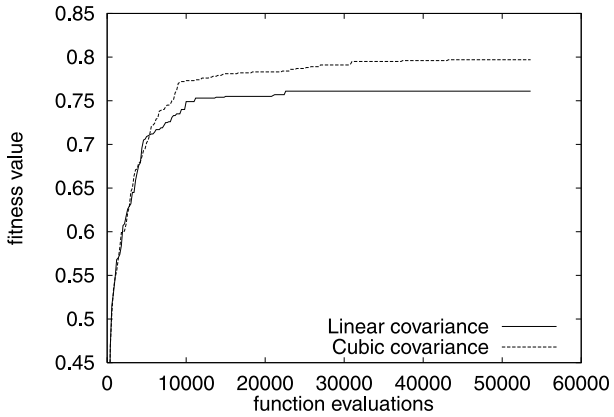

Fig. 11. Covariance models for Keane's 20-D constrained function f_3.

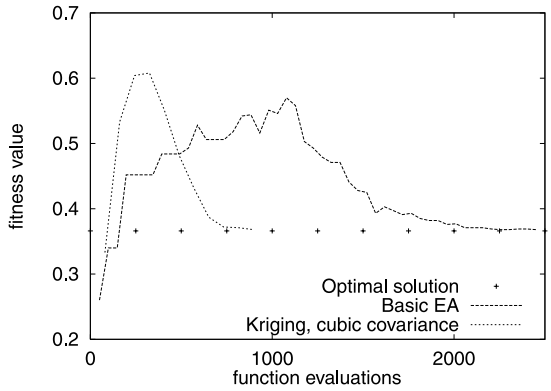

Fig. 13. Constrained 2-D Keane's function f_3 with dynamic penalty function.

https://doi.org/10.1017/S0890060401151024 Published online by Cambridge University Press

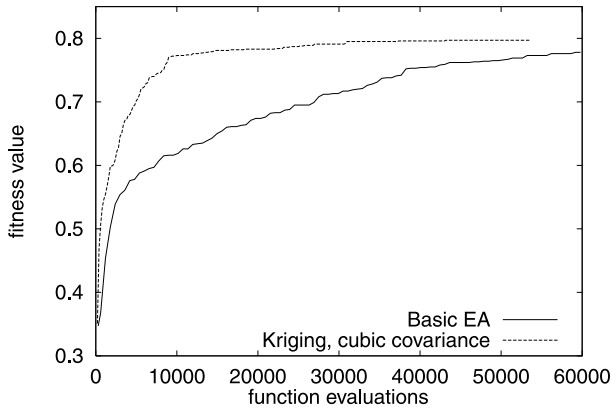

Fig. 14. Constrained 20-D Keane's function f_3 using boundary operators.

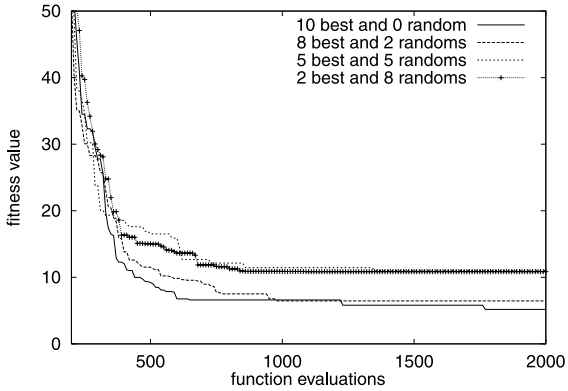

Fig. 16. Influence of the evaluation of random individuals, Rastrigin's 10-D function.

the number of random evaluations, n_{rand}. These parameters have been studied for the 10-dimensional Rastrigin problem. Experiments were conducted with the parameters of Table 2, except, of course, for the sampling strategies. In all the cases, the model size was limited to 500 points. The influence of the number of best individuals evaluated, with no random evaluations, and an update every four generations is illustrated in Figure 15. It is observed that there is a minima of fitness value obtained with the evaluation of 10 individuals, that is, 5% of the population size. From other experiments not presented in this article, this 5% value seems to be a rather robust choice.

It would make sense *a priori*, for multimodal functions, to reevaluate some random individuals to prevent convergence to a local optima. However, the results of Figure 16 show that there is no benefit from doing so: The best results are obtained when only the 10 best and no random individuals are evaluated with the true function. The interval between model update is studied in Figure 17. These results show that there is no point in letting the algorithm run on a given model for more than one generation: optimal results are obtained when the model is evaluated with 5% of the current population every generation. Finally, the rebuild-from-scratch sampling strategy has been compared with the incremental model growth on the 10-dimensional Rastrigin function. In this case, the results of Figure 18 indicate that the progressive update of the model population gives the best results. It should be recognized, however, that this latter strategy relies on many parameters; a setting with the wrong parameters can severely worsen the results, as shown in Figures 15, 16, and 17.

6. CONCLUSION

This article presented the use of kriging interpolation for improving the utilization of available information by evolutionary optimization methods. The estimation of an approximated fitness landscape is an efficient way to reduce the computational cost of an optimization problem when the original fitness function is at the same time too difficult for local search methods and numerically too heavy for standard heuristic methods like EAs. The results presented

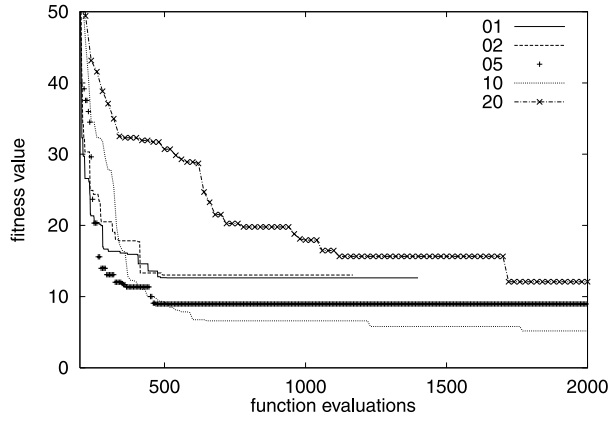

Fig. 15. Number of best individuals evaluated, 10-D Rastrigin's function.

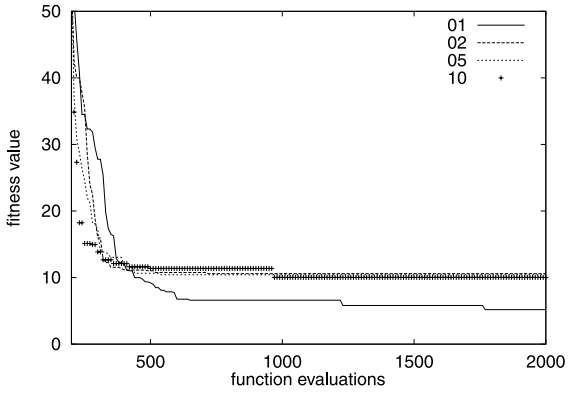

Fig. 17. Number of generations between model updates, Rastrigin's 10-D function.

https://doi.org/10.1017/S0890060401151024 Published online by Cambridge University Press

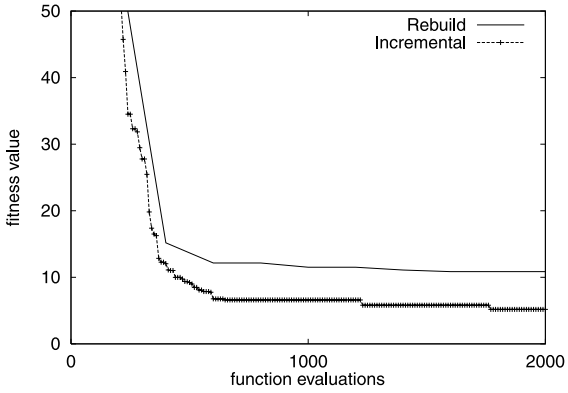

Fig. 18. Comparison rebuild/incremental, 10-D Rastrigin function.

for various classes of problems suggest that building an internal model of fitness landscape is a very efficient approach for problems where only a rough estimate of the best solution is required. In the cases where a high numerical precision is expected, things get a bit more complex. On the one hand, for very simple problems, the gain obtained through the estimation of a fitness model is marginal, since the solution will be found quickly by lighter methods. On the other hand, for highly rugged, multimodal, and multidimensional landscapes, it becomes impractical to build a global model of the whole space. This suggests that there is a specific niche for global function approximation algorithms over the spectra of problem complexity, whereas a local or midrange approximation might work better in other cases.

The quality of optimization results were in many cases somehow sensitive to modeling parameters. A promising approach seems to be the use of an experimentally estimated correlation structure between the samples. Real-world problems seldom present a completely random structure. For complex multimodal problems where some structure is present in the fitness landscape, variogram estimation methods are likely to give interesting results. The problem of scaling up to high-dimensional spaces is one of the points that is still unresolved.

Some aspects have not been addressed in this article. All the problems considered were defined in a real-valued search space. Since EAs are well known to be robust for many different data representations (binary strings, integers, trees, etc.) the generalization of the fitness landscape reconstruction approach to other space metrics is an important issue. The comparison with other estimation paradigms, such as neural networks or support vector machines, is also an important research issue that is left for further investigations.

ACKNOWLEDGMENTS

The author addresses a special acknowledgment to François Trochu for all the discussions that helped improve this research. This work was made possible by the financial support of the IRSST.

REFERENCES

Bäck, T. (1996). *Evolutionary Algorithms in Theory and Practice*. Oxford University Press, New York.

Baek, K.H. & Elliott, S.J. (1995). Natural algorithms for choosing source locations in active control systems. *Journal of Sound and Vibration 186(2)*, 245–267.

Barthelemy, J.-F.M. & Haftka, R.T. (1993). Approximation concepts for optimum structural design—A review. *Structural Optimization 5*, 129–144.

DeJong, K.A. (1975). *An analysis of the behaviour of a class of genetic adaptive systems*. PhD thesis. University of Michigan, Ann Arbor, Michigan.

Farin, G. (1990). *Curves and Surfaces for Computer Aided Geometric Design*. Academic Press, New York.

Floudas, C.A. & Pardalos, P.M. (1996). *A Collection of Test Problems for Constrained Global Optimization Algorithms* (Vol. 455). Springer-Verlag, Berlin.

Gerald, C.F. & Whitley, P.O. (1990). *Applied Numerical Analysis*, 4th ed. Addison Wesley, Reading, United Kingdom.

Graham, R.L., Knuth, D.E. & Patashnik, O. (1994). *Concrete Mathematics*. Addison-Wesley, Reading, United Kingdom.

Hancock, P.J.B. (1994). An empirical comparison of selection methods in evolutionary algorithms. In *Evolutionary Computing, AISB Workshop*, (Fogarty, Terence C., Ed.) pp. 80–94.

Journel, A.G. & Huijbregts, Ch.J. (1978). *Mining Geostatistics*. Academic Press, New York.

Keane, A.J. (1994). Experience with optimizers in structural design. *Adaptive Computing in Engineering Design and Control*, 14–27.

Keane, A.J. (1995). Passive vibration control via unusual geometries: The application of genetic algorithm optimization to structural design. *Journal of Sound and Vibration 185(3)*, 441–453.

Matheron, G. (1965). *Les variables régionalisées et leur estimation*. Masson et Cie, Paris.

Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability 5*, 439–468.

Michalewicz, Z. (1994). *Genetic Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, Berlin.

Michalewicz, Z. & Schoenauer, M. (1996). Evolutionary algorithms for constrained parameter optimization problems. *Evolutionary Computation 4(1)*, 1–32.

Ratle, A. & Berry, A. (1998). Use of genetic algorithms for the vibroacoustic optimization of a plate carrying point-masses. *Journal of the Acoustical Society of America 104*, 3385–3397.

Sacks, J., Mitchell, T.J., Welch, W.J., & Wynn, H.P. (1989). Design and analysis of computer experiments. *Statistical Science 4(4)*, 409–435.

Sartori, D.E. & Smith, A.E. (1997). A metamodel approach to sensitivity analysis of capital project valuation. *The Engineering Economist 43(1)*, 1–24.

Schölkopf, B., Burgess, C., & Smola, A. (1998). *Advances in Kernel Methods*. MIT Press, Cambridge, MA.

Smith, A.E. & Coit, D. (1997). Penalty functions. In *Handbook of Evolutionary Computation*, section C5.2. Oxford University Press, New York.

Trochu, F. (1993). A contouring program based on dual kriging interpolation. *Engineering with Computers 9*, 160–177.

Zimmerman, D.C. (1999). Navigating expensive and complex design spaces using genetic algorithms. *ASME Design Engineering Technical Conf.*