Entropy-Based Complexity of Optimization Problems

H. Lydia Deng
Center for Wave Phenomena and Department of Mathematical & Computer Sciences
Colorado School of Mines
Golden, Colorado 80401, USA

John A. Scales
Center for Wave Phenomena and Department of Geophysics
Colorado School of Mines
Golden, Colorado 80401, USA

ABSTRACT

Optimization problems arise in every scientific field, whether as a statement of a fundamental principle, such as the principle of maximum entropy, or as an encapsulation of a given problem, such as maximizing the coherence of the output of some image-processing algorithm. It is often the case in practice that the function to be optimized cannot be specified in closed form in terms of elementary functions, but must be evaluated pointwise via a computer program. It goes without saying that it is desirable to have some way of selecting appropriate optimization tools for a given problem. In order to do so we must have some means of characterizing the structure or complexity of generic optimization problems. Further complicating matters is the high dimensionality of many of the functions encountered. Unable to visualize these high-dimensional functions directly, we propose a measure of complexity based on statistical analysis of samples of models found by some searching algorithms; we believe this measure, to some extent, quantifies how hard the resulting optimization problem is likely to be. Then, we use this measure to analyze some analytic functions. Finally, we analyze the multi-resolution analysis (MRA) of a highly multi-modal function arising in exploration seismology from the standpoint of this new entropy-based complexity measure.

Key words: Optimization, Objective Function, Global Search, Complexity, Topography, Entropy

What Makes an Optimization Hard?

We consider the problem of optimizing a function F (the *objective* or *cost* function) mapping $\mathcal{M} \subset \mathbf{R}^N$ into $\mathcal{Y} \subset \mathbf{R}$. We refer to $\mathcal{M}$ as the *model space*. In most applications, the function F cannot be expressed in closed form in terms of elementary functions, but must be evaluated pointwise. Depending on the application, the problem may be to find the global extremum of F, a single local extremum, or a collection of local extrema. If the structure of F is unknown, then optimization is fundamentally a matter of search in the model space. In order to be able to treat such a broad variety of situations, we begin with an abstract statement of the search algorithm.

Algorithm 1. General Search (GS)
$(\{\mathbf{m}_f\} = GS(F, P, \mathbf{T}, S))$
Let $F : \mathcal{M} \subset \mathbf{R}^N \to \mathcal{Y} \subset \mathbf{R}$, $P = \{\mathbf{m}_0^k\}_{k=1,\cdots,K}$ be an initial population with size $K \geq 1$, $\mathbf{T}$ be a *transition operator*, and S be a stopping criterion.

(i) Iteratively apply the transition operator to generate a new population of model at each iteration, so that $\mathbf{m}_i^k = \mathbf{T}\,\mathbf{m}_{i-1}^k$;
(ii) Repeat (i) until S is satisfied. The final set of models $\{\mathbf{m}_f\}$ are the output of the search.

This is a general statement of searching, which can be considered as an evolution of a population of models on an N-dimensional lattice, where each node $\mathbf{m}$ has a function value $F(\mathbf{m})$ attached to it. The transition operator $\mathbf{T}$ is the rule that determines to which nodes each model walks at each step. For different choices of the initial population P and transition operator $\mathbf{T}$, we have different optimization algorithms.

There is no generally agreed upon characterization of what makes an optimization problem hard. Hardness has to do partly with our goals — do we need a global extremum or will a good local extremum do; partly with the structure of the function — does it have lots of local extrema, how broad is the basin of attraction of the extrema we seek; and partly with the dimensionality of the problem — exhaustive search will be infeasible except for low-dimensional problems, etc. Hardness comprises all of these things and more. Since the functions we are interested in can usually be evaluated only point-wise, some degree of global sampling is essential in order to achieve the characterization we seek.

Unfortunately, there are very few systematic comparisons of global optimization/sampling algorithms, either with other global algorithms, or with repeated application of more elementary methods. For instance, if a simulated annealing algorithm achieves a given result in n iterations, what can we conclude if a straightforward combination of random sampling and hill-climbing can achieve as good a result in fewer function evaluations? Has the sophisticated algorithm really done what we thought it was doing? The sheer size of the problems faced defeats simple answers to such questions. For a combinatorial problem with N unknowns and each of which can take m possible values, the number of possible models is m^N.

Imagine the surface of an objective function being a high-dimensional landscape with hills and basins of different depths and widths scattered on the surface. From our experience, the performance of searching algorithms depends to a large extent on certain topographical features of this landscape, such as the number of basins (local extrema), the widths and depths of these basin of attractions, etc. When the objective function is high-dimensional and can only be sampled point-wise, how best to acquire and analyze this topographical information is not clear.

Some hard combinatorial optimization problems

Hard combinatorial global optimization problems arise in various scientific and engineering fields. Some of the most widely studied include the *spin-glass* problem in statistical physics, the *traveling-salesman* problem (TSP) in computer science, and the *residual statics* problem in exploration seismology. Here we briefly describe these three problems. In this paper, we focus on the last problem; however, we believe that our results can be used to classify other hard optimization problems as well.

Spin-glasses

Spin-glasses (Edwards & Anderson, 1975; Fischer & Hertz, 1991) are disordered magnetic materials in which the orientation of nearby magnetic dipoles may be either parallel or anti-parallel. Models of spin-glasses typically consist of lattices of spins with each spin pointing either up or down. Suppose the entire system has N spins, each up or down; there are a total 2^N possible configurations. Each configuration has a total energy given by a Hamiltonian (Hertz *et al.*, 1991),

$$H = -\sum J_{ij}\ (s_i \times s_j), \quad s_i, s_j = \pm 1, \tag{1}$$

where s_i and s_j are the orientations (up or down) of the two spins, J_{ij} is the energy weighting factor. The goal is to find a configuration, $\{s_i\}_{i=1,\cdots,N}$ that minimizes the Hamiltonian.

Traveling salesman problem (TSP)

The TSP is a well-known NP-complete optimization problem*. Given N cities with distance (or cost) C_{ij} between them. The task is to find the minimum-length (or cost) closed tour that visits each city exactly once and returns to its starting point. The objective function can be formulated as

$$F = \sum_{i=1}^{N} \sum_{j=i}^{N} C_{ij} x_{ij}, \quad x_{ij} = 0, 1, \tag{2}$$

where the unknowns x_{ij} is 1 if the path connecting i and j is traveled and 0 if otherwise. For this problem, both the number of unknowns and computation time grows exponentially with the number of cities N. For a situation with $N = 100$ cities, there are approximately 10^{155} possible tours.

Residual statics

In exploration seismology, statics are the time shifts in seismic reflection data caused by heterogeneous material properties in the near surface. Figure 1 shows synthetic seismic traces (i.e., seismogram) that differ only

* NP-complete is a class of optimization problems for which there are no known polynomial-time algorithms. Empirically, the computation time for solving an NP-complete problem grows exponentially with the size N.

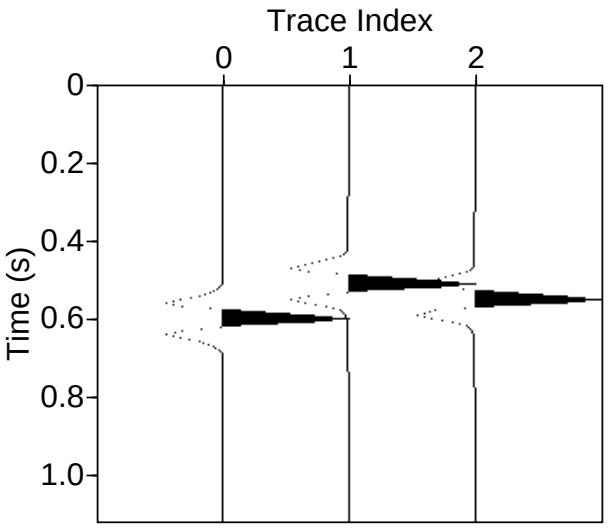

Figure 1. Three synthetic seismic traces that are shifted by random statics.

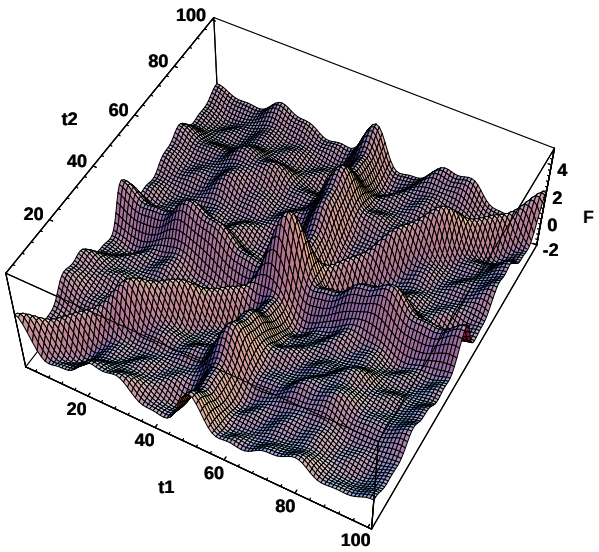

Figure 2. Landscape of a 2-D residual statics objective function.

by random shifts simulating the contamination of such heterogeneity. Essentially, the goal of residual statics estimation is to look for the time shift of each trace by maximizing the alignment of the traces. Consider an example where we need to align three identical traces. Fixing the first trace, we look for time-shifts for the second and third traces, t_1, t_2, so that the sum of squares of the stacked traces (stacking-power) is maximized. Figure 2 shows an example of such a two-dimensional objective function, which has multiple hills and basins of attractions scattered on the landscape. In practice, however, the stacking-power objective function is high-dimensional and highly multi-modal. Sophisticated global searches have generally been thought to be necessary for such problems (Rothman, 1985; Rothman, 1986).

Global search strategies

Among the searching methods defined via Algorithm 1, there are two extreme strategies, *hill-climbing* (HC) and *uniform Monte Carlo* (UMC). HC search is a local search applied to a single model (population size $K = 1$). An initial population $P = \mathbf{m}_0$ is selected (possibly at random) and the transition operators $\mathbf{T}$ are deterministic operators, such as conjugate gradient or downhill simplex, which follow a path downhill (or uphill for maximizing) as far as possible. For objective functions containing more than one local extrema (*multi-modal*), the final optimal model $\mathbf{m}$ using HC strongly depends on the choice of the initial model $\mathbf{m}_0$. UMC, other the other hand, selects points with uniform probability in the model space. The transition operation $\mathbf{T}$ is simply the selection of new points at random and therefore makes no use of information from previous generations. Thus, if there are M parameters and each of them can take n possible values, the probability of finding a particular model is proportional to n^{-M} for each function evaluation.

Many global search strategies have been developed that yield a compromise between two extremes; almost all of these use some stochastic operators, especially in the construction of transition operators. It is important for the success of global searches that the transition operators can make the best use of information provided by the current samples while avoiding being trapped in local extrema. Among all these strategies, the most widely used are *Simulated Annealing* (SA), (Kirkpatrick *et al.*, 1983) and *Genetic Algorithms* (GA) (Holland, 1975) and random hill-climbing (RHC), to be defined shortly.

SA and GA searching strategies use stochastic transition operators $\mathbf{T}$ that are biased towards good samples from the previous generations. Such schemes can therefore be considered as biased random walks on objective-function surfaces. Many variations of SA and GA can be found in the literature (Aarts & Korst, 1989; van Laarhoven & Aarts, 1987; Goldberg, 1989). Although the asymptotic convergence results are known for both SA and GA, these results are hardly useful in practice.

RHC searches, on the other hand, apply deterministic transition operations $\mathbf{T}$ to a randomly chosen population of initial models P. RHC explores locally in multiple areas of objective functions, and the resulting samples are a set of local/global extrema. This search algorithm can be described as

Algorithm 2. Random Hill Climbing

$(\{\mathbf{m}^k\}_{k=0,\cdots,M} = RHC(F, K, \epsilon, cmax))$
Let the initial population size be K. Let the stopping criterion S be that either gradients of all samples are reduced to ϵ or the number of iterations reaches $cmax$. Let $\mathbf{T}_{local}$ be the local hill-climbing transition operator.

(i) Choose initial models $P = \{\mathbf{m}_0^k\}_{k=1,\cdots,K} \in \mathcal{M}$ uniformly at random, where $K \gg 1$;
(ii) Apply Algorithm 1, $\{\mathbf{m}\} = GS(F, P, \mathbf{T}_{local}, S)$.

The final population contains M distinctive models, $\{\mathbf{m}^k\}_{k=0,\cdots,M}$.

In this paper, all RHC numerical results use the nonlinear Conjugate Gradient as transition operators (Deng *et al.*, 1995).

Why complexity?

Highly multi-modal problems such as residual statics are usually treated by Monte Carlo methods. However, it has been observed by Whitley *et al.* (1995b) that for this problem, a RHC scheme can be more effective than a sophisticated GA. But it is not clear whether this is a general feature of statics or of the particular problem studied.

Clearly, the effectiveness of SA and GAs is somehow related to the surface landscape of the objective functions to be optimized. This situation is summarized by Kaufmann (1993):

> Annealing works well only in landscapes in which deep energy wells also drain wide basins. It does not work well on either a random landscape or a "golf course" potential, which is flat everywhere save for a unique "hole". In the latter case, the landscape offers no clue to guide search.

> Recombination (in GAs) is useless on uncorrelated landscapes but useful under two conditions (1) when the high peaks are near one another and hence carry mutual information about their joint locations in genotype space and (2) when parts of the evolving system are quasi-independent of one another and hence can be interchanged with modest chances that the recombined system had the advantage of both parents.

In addition, as RHC searches use hill-climbing transition operators, it is easy to see that the performance of RHC is also largely affected by topographical features of objective functions described above, such as "wide basins", "deep wells", and "random landscape", etc.

In order to study the performance of global optimization algorithms, it is important to quantify these qualitative descriptions of high-dimensional function surfaces. Characterizing the topographical features of high-dimensional functions is the main goal of this research. We believe that a quantitative measure of these features will be useful in designing optimal transition operators.

In this paper, we propose a measure of complexity of high-dimensional functions in the context of information. We first review some related work and discuss the criteria used to define our measure. We then show examples of applying this criterion to some analytical functions. Finally, we use this measure to analyze the behavior of a multi-resolution analysis (MRA) of the seismic statics problem.

Measures of Complexity

Previous work on complexity

Chavent (1991) developed sufficient conditions for an objective function to be locally convex. These conditions are based on the distance × curvature induced by the objective function on trajectories. This local convexity criterion could be generalized to global samples of an objective function, to provide a global measure of complexity.

A measure of the complexity can also be defined by analyzing the distribution of function values for a population of samples found by a given search algorithm (Wolpert & Magready, 1995; Magready & Wolpert, 1995). Such a measure therefore evaluates the performance of a searching algorithm for a given objective function. Based on this, the authors draw two conclusions on complexity of general combinatorial optimization problems:

(i) No algorithm has a better (or worse) average performance than do other algorithms for all possible objective functions. This result is referred as the *no free lunch* (NFL) theorem (Wolpert & Magready, 1995).
(ii) No optimization problem is intrinsically harder than other optimization problems when averaged over all possible search algorithms. However, there do exist the optimal search algorithms for a specific optimization problem (Magready & Wolpert, 1995).

Although the Wolpert-Magready measure and conclusions are informative, they do not characterize the topography of the objective function.

Another proposal is that functions can be characterized by their spatial correlation properties (Weinberger, 1990; Stadler, 1992a). Several typical combinatorial optimization problems were investigated by studying the correlation in landscapes: the TSP (Stadler, 1992c), graph-bipartitioning problem (Stadler, 1992b), and the NK model problems, a spin-glass like problem in biology (Kauffman & Weinberger, 1989). Using correlation features of the objective function's landscape as a criterion, they study the effectiveness of some global algorithms for certain types of landscapes.

In addition, analyzing the topography of high-dimensional energy functions is important in phys-

ics. Berry and Breitengraser-Kunz (1995) studied topography and dynamics of multidimensional inter-atomic potential surfaces by analyzing a population of local minima, each of which has two saddle points connected to them. By connecting these samples in a certain order, the high-dimensional function surface is represented by a series of one-dimensional lines. By looking at these one-dimensional plots, the topography information is represented by the width and depth of the primary, secondary or tertiary basins of attractions (Berry & Breitengraser-Kunz, 1995).

Furthermore, the complexity of high-dimensional Hamiltonians can also be studied by means of entropy (Falcioni *et al.*, 1995). For an N-dimensional Hamiltonian, some local extrema are first found by some local search algorithm. Contributions of these local-minima to the complexity is represented by a probability distribution $\{P^{(k)}(N)\}_{k=1,\cdots,M}$ where

$$P^{(k)}(N) \propto \Delta^{(k)}(N).$$

Here, $\Delta^{(k)}(N)$ is the estimated width of the kth basin of attraction. The complexity of the N-dimensional surface can be characterized by the following entropy,

$$S(N) = -\sum_k P^{(k)}(N) \ln P^{(k)}(N) = \left\langle \ln\left(\frac{1}{\Delta(N)}\right)\right\rangle . \quad (3)$$

Information measure of the complexity

From our experience and previous studies, the hardness of optimization problems largely depends on topography of the objective functions. Therefore, the complexity of an optimization procedure can be considered equivalent as the topographical complexity of the objective function. The topographic features of a landscape are mostly attributed to the distribution of local extrema, such as,

(i) number of basin of attractions (local extrema);
(ii) width of each basin of attractions;
(iii) relative depths of the basins (function values of the extrema).

As shown in Algorithm 2, RHC explores various regions of the model space and takes initial samples downhill to the bottom of basins on the surface of functions. Therefore, the results of systematic RHC searches can be used to characterize the complexity of objective functions. However, the entropy-based complexity, as defined in equation (3), represents global features of the objective function landscape only by the number of basins of attractions and the widths of these basins. It does not take into account the relative depths of each basins. Figure 3 shows two functions with the same number of local minima and widths of basins of attractions. Using the criterion in equation (3), they would have the same complexity. However, the difficulty of minimizing these functions is different: the left function has identical basins of attractions, while the one on the right has a dominant global minimum and decreasingly important local minima away from the center.

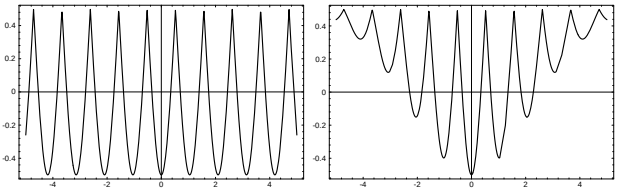

Figure 3. Two functions with the same number of local minima and widths of basins of attractions. The basins of the function on the left have the same depths, while those of the function on the right have depths decreasing with distance from the center with a dominate global minima at $x = 0$.

We define our entropy-based complexity measure as follows:

Definition 1. Entropy-Based Complexity
Let $F : \mathcal{M} \subset \mathbf{R}^N \rightarrow \mathcal{Y} \subset \mathbf{R}$. Let $\{\mathbf{m}_i\}_{i=1,\cdots,M} = RHC(F, K, \epsilon, cmax)$ be the distinct converged models of an RHC search, $\{h_i\}_{i=1,\cdots,M}$ be the histogram of the final population, and $\{y_i = F(\mathbf{m}_i)\}_{i=1,\cdots,M}$ be their corresponding function values. Let $y_m = \min\{y_i\}$, and $\sigma = \langle (y_i - y_m)^2 \rangle$. Define the entropy-based complexity C_e as

$$\begin{aligned} C_e &= -\sum_{i=1}^{M} P(\mathbf{m}_i) \ln(P(\mathbf{m}_i)) \\ &= \left\langle \ln\left(\frac{1}{P(\mathbf{m}_i)}\right)\right\rangle , \end{aligned} \quad (4)$$

where $P(\mathbf{m}_i) \propto h_i\, v(y_i)$, in which

$$v(y_i) \equiv \begin{cases} 1, & \text{if } \sigma = 0; \\ \exp(-\frac{|y_i - y_m|}{\sqrt{\sigma}}), & \text{otherwise,} \end{cases}$$

and $P(\mathbf{m}_i)$ is normalized to $\sum_{i=0}^{M-1} P(\mathbf{m}_i) = 1$.

C_e defined in Definition 1 measures the amount of information gained about the global features of the objective function after an RHC. $P(\mathbf{m}_i)$ is a probability representing contributions to the global information by the ith converged model found by the RHC. This probability is proportional to the width of basin of attractions (h_i), weighted by the relative depths of the basins ($v(y_i)$). This complexity measure is similar to the the entropy measure used in characterizing Hamiltonians (Falcioni *et al.*, 1995), except that we take into account the distribution of function values at each local extrema.

Let us consider some examples of the application of this measure to a family of two-dimensional functions.

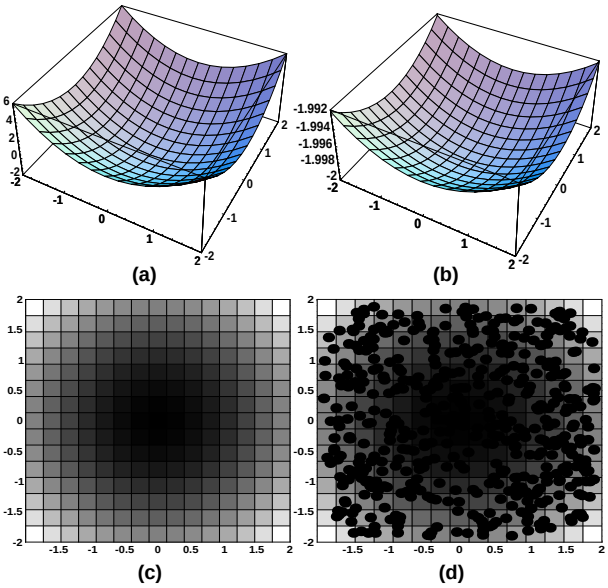

Figure 4. Random hill-climbing with population of $K = 500$. (a) and (b) show the function surface defined in equation(6) when $f = 0$, $b = 1$ and $a = 1,\ 0.001$ respectively. (c) and (d) show the convergence of 500 initial models in the 2-D model space on functions (a) and (b). In the first case $C_e = 0$, in the second $C_e = 4.11$.

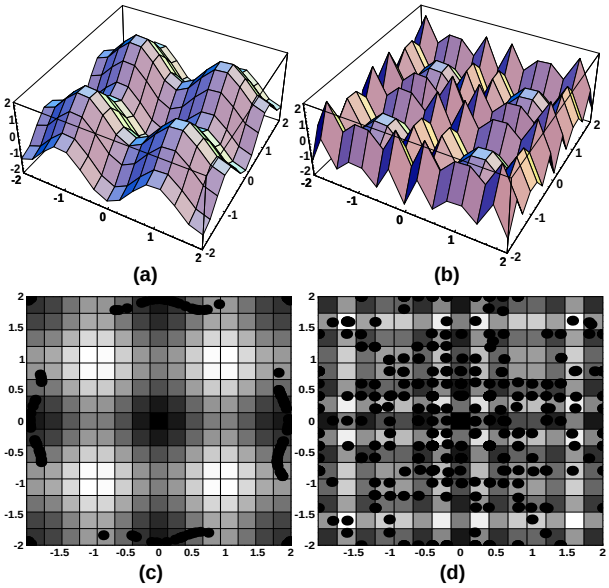

Figure 5. Random hill-climbing with population of $K = 500$. (a) and (b) show the function surface defined in equation (6) when $a = 0.1$, $b = 1$ and $f = 1,\ 10$ respectively. (c) and (d) show the convergence of 500 initial samples in the 2-D model space on functions (a) and (b). In the first case $C_e = 2.14$ and second case $C_e = 3.93$.

$$F(\mathbf{m}) = \mathbf{m}^T A\mathbf{m} + \mathbf{b}^T\,\mathbf{w}(\mathbf{f}, \mathbf{m}), \tag{5}$$

where A is a square-matrix, $\mathbf{b}$ is a constant vector and $\mathbf{w}(\mathbf{f}, \mathbf{m})$ is an oscillatory vector function

$$w_i = -\cos(f_i\, m_i), \quad i = 0, 1$$

where f_i is the frequency in each direction. The function $F(\mathbf{m})$ in equation (5) is multi-modal, and the local minima are caused by the oscillating term $\mathbf{w}(\mathbf{m})$. Hill-climbing searches of this function may converge to local extrema for arbitrary initial models. The difficulties of finding the global extremum, however, vary with the coefficients. For simplicity, we only consider equation (5) when

$$A = a\, I_2, \quad \mathbf{b} = b,\, I_1 \quad \text{and} \quad \mathbf{f} = f\, I_1,$$

where I_2 is a 2×2 identity matrix and $I_1 = (1, 1)^T$. In this case, $\mathbf{m} = (m_0, m_1)$ and equation (5) becomes

$$F(m_0, m_1) = a\,(m_0^2 + m_1^2) - b\,\cos(f\, m_0) - b\,\cos(f\, m_1). \tag{6}$$

We perform RHC on equation (6) with a population of 500 models. The searches are stopped when the residuals are reduced to $\epsilon = 10^{-5}$ or when the number of non-linear Conjugate Gradient iterations reaches $camx = 200$.

When the frequency of the oscillation is zero, $f = 0$, and $F(m_0, m_1)$ becomes quadratic with a unique global minimum. Figures 4(a) and (b) show the function surfaces when $f = 0,\ b = 1$ and $a = 1,\ 0.001$ respectively. Both functions are quadratic. Noticing the scale of the two plots, however, Figure 4(b) is much flatter than (a). This difference in the curvature of surfaces influences the result of optimization significantly. Figures 4(c) and (d) show the converged models for the RHC search when the stopping criterion are met. In Figure 4(c), we see that all 500 models converge to the global minimum when the curvature of the quadratic function is large enough. As a result, $C_e = 0$. However, when the curvature is small, as in Figure 4(b) the maximum number of iterations is exceeded and the final models are scattered about the domain. In this case $C_e = 4.1$. This demonstrates that even for unimodal objective functions, the hardness of global search could also depend on the curvature of the landscape.

When the frequency f is not zero, function F is multi-modal. Figures 5(a) and (b) show the function surface when $a = 0.1,\ b = 1$ and $f = 1,\ 10$ respectively; both functions have the same quadratic term, but have different number of local minima. Figures 5(c) and (d) show the converged models using the RHC search when stopping criterion are met. The converged models are more clustered at bottoms of basins on the low-frequency function surface than on the high-frequency one, and their entropy-based complexities C_e are 2.14 and 3.93, respectively. Figure 6 shows the complexity C_e as a function of the spatial frequency f when the curvature a is

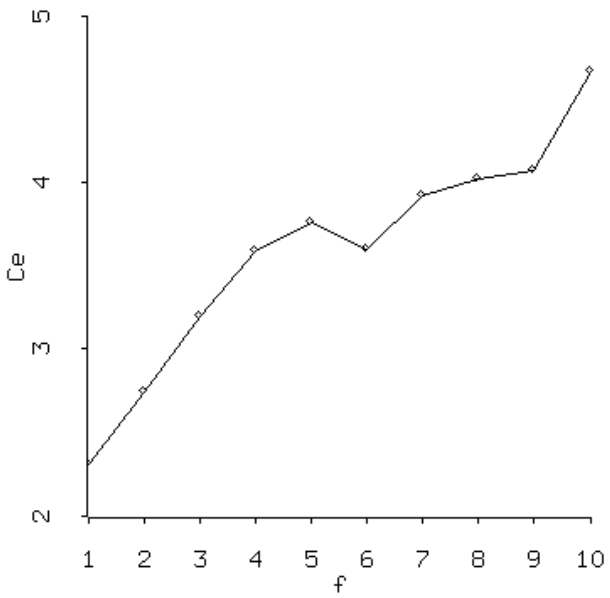

Figure 6. Plot of C_e as a function of the spatial frequency f for the function defined in equation (6) when $a = 0.1$ and $b = 1$.

fixed. We see that C_e increases as the number of local extrema increases.

Fixing the spatial frequency, the complexity of a multi-modal function is also influenced by the spatial curvature. Figure 7 shows the function surfaces of equation (6) when $f = 10$, $b = 1$ and $a = 1,\ 10$ in (a) and (b) respectively; both functions have the same number of local minima as does in Figure 5(b). Figures 7(c) and (d) show the converged samples when the stopping criterion are met. Compared with Figure 5(d), the converged models are more clustered with the increase of spatial curvature despite of the fact that all three functions have the same number of local minima in the model space. The entropy-based complexity C_e of these two functions are 3.17 and 1.02, respectively. Figure 8 shows the complexity C_e as a function of the spatial curvature a for the same function when $f = 10$ and $b = 1$.

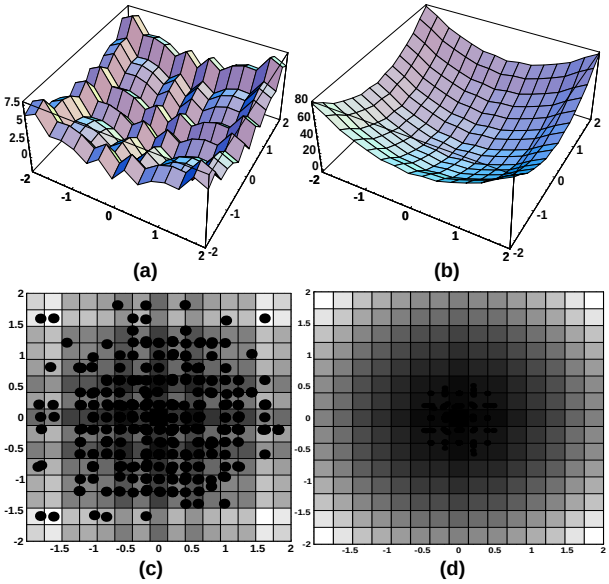

Figure 7. Random hill-climbing with population of $K = 500$. (a) and (b) show the function surface defined in equation (6) when $f = 10$, $b = 1$ and $a = 1,\ 10$ respectively. Both functions have the same number of local minima, though those in (b) are too small to be noticed. (c) and (d) show the convergence of 500 initial models in the 2-D model space on functions (a) and (b). In the first case, $C_e = 3.17$, in the second $C_e = 1.02$.

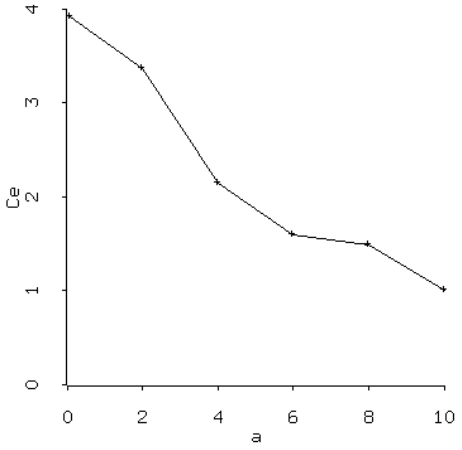

Figure 8. Plot of C_e as function of the spatial curvature a for the function in equation (6) when $f = 10$ and $b = 1$.

Applications to High-dimensional Test Functions

In this section, we use the complexity C_e to study two commonly used optimization test functions, the Rosenbrock function and the Griewangk function.

N-dimensional Rosenbrock function

An N-dimensional Rosenbrock function can be written as

$$R(\mathbf{x}) = \sum_{i=1}^{N-1} \left[100(x_i - x_{i-1}^2)^2 + (1 - x_{i-1})^2\right], \qquad (7)$$

where $\mathbf{x} = (x_0, \ldots, x_N)$. Although unimodal, the long and narrow basin is a challenge for searching algorithms. Figure 9 shows the function surface and its contour when $N = 2$. When $N \geq 2$, the function is still unimodal, but

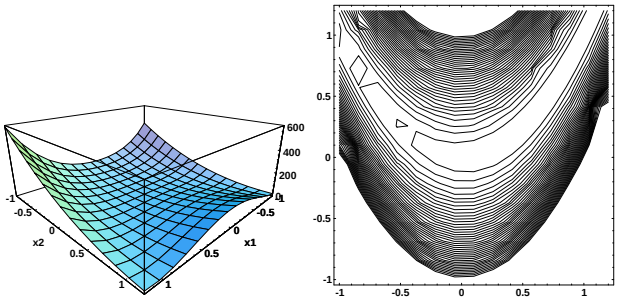

Figure 9. Two-dimensional Rosenbrock function. The figure on the left is a 3-D plot of the function surface, while the one on the right shows the contour plot of the same function.

it is not easy to see how the increase of dimensionality alters the difficulty of optimization.

One way of studying the spatial curvature of functions is by looking at the ratio of largest and smallest eigenvalues (*condition number*) of the Hessian. The Hessian for equation (7) is a tri-diagonal matrix,

$$\begin{pmatrix} a_0 & c_0 & 0 & \cdots & 0 \\ b_1 & a_1 & c_1 & 0 & \cdots \\ 0 & b_2 & a_2 & c_2 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & b_{N-1} & a_{N-1} \end{pmatrix} \tag{8}$$

where

$$\begin{aligned} a_0 &= 2 + 1200\, x_0^2 - 400\, x_1 \\ a_i &= 202 + 1200 x_{i-1}^2 - 400 x_i, \quad 0 < i < N-1 \\ a_{N-1} &= 200 \\ b_i &= -400\, x_{i-1} \\ c_i &= -400\, x_{i+1}. \end{aligned}$$

At the global minimum $(1, 1, \cdots, 1)$, the tri-diagonal matrix equation (8) becomes Toeplitz except for a_0 and a_{N-1}:

$$\begin{aligned} a_i &= 1002, \quad 0 < i < N-1 \\ b_i &= -400, \quad 0 < i \leq N-1 \\ c_i &= -400, \quad 0 \leq i < N-1 \\ a_0 &= 802 \\ a_{N-1} &= 200. \end{aligned}$$

The condition number of the Hessian at the global minimum reaches an asymptote with increasing dimension, as shown in Figure 10. Figure 11 shows the complexity measure C_e as a function of the number of dimensions; it shows the same asymptotic trend as does the condition number. Thus the increasing complexity for low dimensions is the result of increasing ill-conditioning of the Hessian and has nothing to do with local minima.

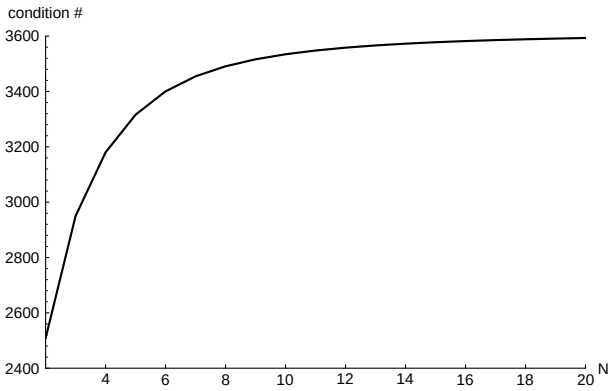

Figure 10. Condition numbers of the Hessian matrix for N-dimensional Rosenbrock functions at the global minimum.

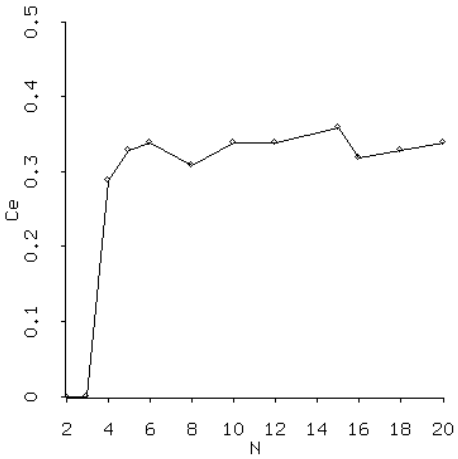

Figure 11. Complexity C_e for N-dimensional Rosenbrock functions as a function of N.

High-dimensional Griewangk functions

The Griewangk function is another test function being used for testing optimization algorithms (Whitley *et al.*, 1995a):

$$g(\mathbf{x}) = 1 + \sum_{i=1}^{N} \frac{x_i^2}{4000} - \prod_{i=1}^{N} (\cos(x_i)/\sqrt{i}) \tag{9}$$

The cosine term makes equation (9) multimodal. Figure 12 shows a one-dimensional slice of the Griewangk function along the diagonal of the hypercube for dimensions $1, 3, 5, 9$. Whitley and Mathias (1995a) observed such slices and concluded that "as the dimensionality increases the local optima induced by the cosine decrease in number and complexity".

However, such pictures can be misleading since they tell us only about low-dimensional projects of the function. Figure 13 shows slices of the same functions when all but one variables are fixed to be 0. The increasing dimen-

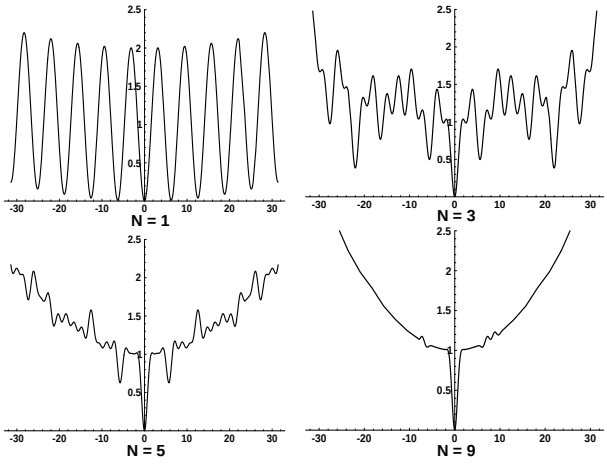

Figure 12. Diagonal slices of N-dimensional Griewangk functions.

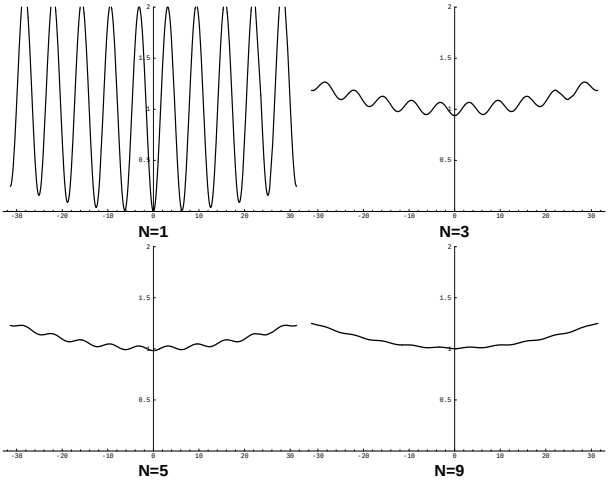

Figure 14. Slices of N-dimensional Griewangk functions. All variables but one are fixed at 2.

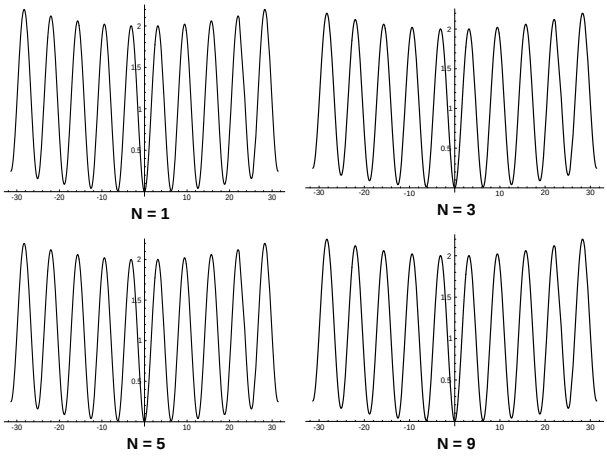

Figure 13. Slices of N-dimensional Griewangk functions. All variables but one are fixed at 0.

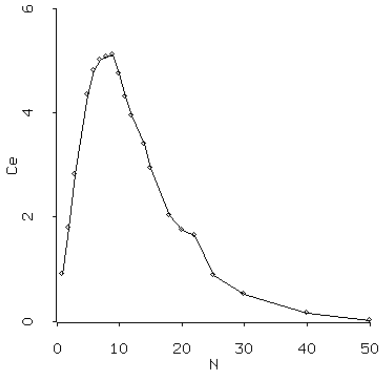

Figure 15. The complexity C_e as a function of dimension N for the Griewangk function.

sionality does not change the shape of the slices. On the other hand, Figure 14 shows slices of this function when all but one variables are fixed to be 2. Oscillation of the function has been reduced with the increasing dimensionality, but the curvature has not been increased as it was in Figure 12. Therefore, studying the overall performance of high dimensional functions could be tricky. We compute C_e for the Griewangk function with a population 500 models in the domain of $-5 \leq x_i \leq 5, \quad i = 0, \cdots, N-1$. Figure 15 shows the resulting complexity C_e for dimensions up to 50. This result gives us more confidence that we really understand the dimensional-dependence of complexity than by simply looking at hyper-planes.

Complexity of Residual Statics Problems

A synthetic residual statics problem

Let us return to the residual statics problem previously described. Figure 2 shows a statics objective function with two unknowns. In practice, however, the time-shifts of the traces are not independent. The statics of each trace are caused by the combined time distortion of near-source and near-receiver heterogeneities (*source-statics* and *receiver-statics*). Figure 16 illustrates the similarity of travel paths near each source and each receiver.

The recorded reflection seismic signals are usually sorted into *midpoints* y (of the source and receiver locations) and *offsets* h (half distance between the source and receivers). Letting $\vec{s}$ and $\vec{r}$ be unknown vectors of source- and receiver-statics, this optimization problem can be formulated as

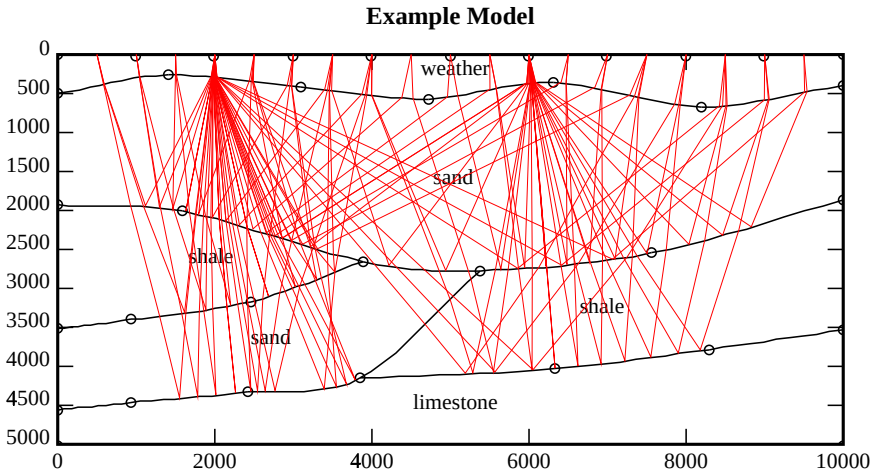

Figure 16. Static-shifts of seismic traces are caused by the combined time-distortions of near-source and near-receiver heterogeneities.

$$\max_{\vec{s},\vec{r}} F(\vec{s},\vec{r}) = \sum_{y} \sum_{h_1 \neq h_2} \Phi^{y}_{h_1,h_2}(\tau(\vec{s},\vec{r})), \tag{10}$$

where $\Phi^{y}_{h_1,h_2}(\tau)$ is the cross-correlation between traces (after nominal correction for normal moveout) of offsets h_1 and h_2 at midpoint y evaluated at

$$\tau = s_{i(y,h_1)} + r_{j(y,h_1)} - s_{i(y,h_2)} - r_{j(y,h_2)},$$

and $i(y,h)$ and $j(y,h)$ are the source and receiver indices for midpoint y and offset h, respectively. Function $F(\vec{s},\vec{r})$ in equation (10) is usually called *stacking-power function*.

Figure 17 shows the recording geometry of one example synthetic data set. This data set has 20 sources, 35 distinct receivers and 320 traces. All traces are identical except for random source and receiver statics. These are generated by repeatedly shifting a single trace of field data. Thus, the objective function of equation (10) has 55 unknowns. When there are no statics in the data, the global maximum of the function is at the origin ($\vec{s}_i = \mathbf{0},\ \vec{r}_j = \mathbf{0}$).

Figure 18 shows 2-D hyper-planes of the stacking-power function from slices in which all parameters are fixed at their correct values except the 10th and 11th source statics (left) and the 10th source and 20th receiver statics (right). For both slices, the function appears to have many local maxima, regularly scattered throughout the slices. On the other hand, Figure 19 shows the stacking-power function as functions of the 10th and 11th source statics (left) and the 10th source statics and the 20 receiver statics, where all other unknowns statics are chosen at random. The topography of the function is much less regular in this case.

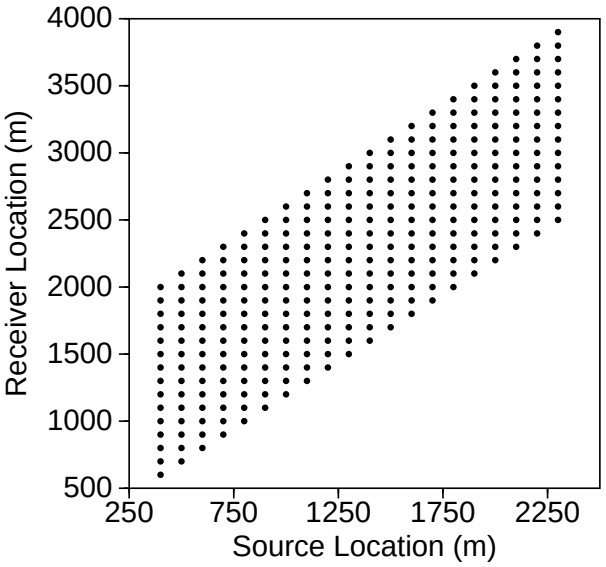

Figure 17. Recording geometry of a synthetic data set. The horizontal axis is source position and vertical axis the receiver position.

Behavior of the multi-resolution analysis

Rather than use a Monte Carlo global optimization method to solve the statics problem, Deng (1995) has proposed simplifying the optimization via a multi-resolution analysis (MRA). The idea is to use a wavelet decomposition to generate successively simpler representations of the seismic data, thereby eliminating progressively more local extrema from the objective function. We now apply the entropy-based measure of complexity to this multi-resolution analysis and see if we can get a deeper un-

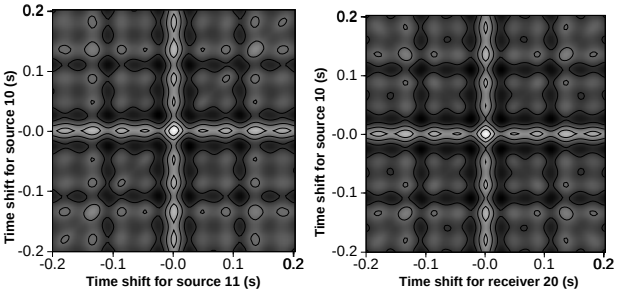

Figure 18. Contour plots of slices of the stacking-power function with all other statics being correct but two components. The left is a slice as a function of the 10th and 11th source statics, and the right is a slice as a function of the 10th source statics and 20th receiver statics.

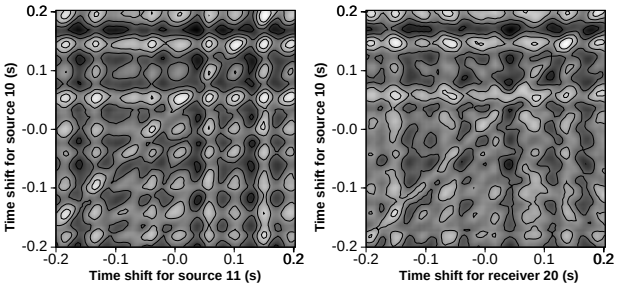

Figure 19. Contour plots of slices of the stacking-power function. The left is a slice as a function of the 10th and 11th source statics, and the right is a slice as a function of the 10th source statics and 20th receiver statics. All other components of $\vec{s}$ and $\vec{r}$ are chosen at random.

derstanding of just what is being accomplished by the MRA.

First let us define a multi-scale RHC algorithm:

Algorithm 3. Multi-scale Random Hill-Climbing $(\{\mathbf{m}\} = MRHC(F, L, \epsilon, cmax))$

Let $\{\mathbf{S}_i\}_{i=L,\cdots,0}$ be a sequence of decreasingly smooth operators to be defined below, and $\mathbf{S}_0$ be an identity.

(i) Let $f_L = \mathbf{S}_L F$; choose an initial population $\{\mathbf{m}_0^k\}_{k=1,\cdots,K}$ with size K at random; apply Algorithm 2, so $\{\mathbf{m}^k\}_{k=1,\cdots,M_L} = RHC(f_L, K, \epsilon, cmax)$, and $i = L-1$.

(ii) Let $f_i = \mathbf{S}_i F$, $(L > i \geq 0)$ and $\{\mathbf{m}_0\} = \{\mathbf{m}^k\}_{k=1,\cdots,M_{i-1}}$; run Algorithm 2, $\{\mathbf{m}^k\}_{k=1,\cdots,M_i} = RHC(f_i, K, \epsilon, cmax)$.

(iii) Decrease i by 1, repeat (ii) until $i = 0$. The final set of models $\{\mathbf{m}^k\}$ is the solution.

The smoothing operators $\{\mathbf{S}_i\}_{i=L,\cdots,0}$ could be a sequence of low-pass filters with increasingly wider passband (Chen, 1994; Bunks *et al.*, 1995), or a sequence of increasingly fine wavelet operators (Deng, 1995). The sequence of smoothing operators should be such that the resulting functions, $\{f_i\}_{i=L,\cdots,0}$, have the same global feature as the objective function F with increasing number of local optima, and $f_0 = F$. Deng (1995) showed that this could be achieved using the shift-invariant basis of Saito and Beylkin (1993).

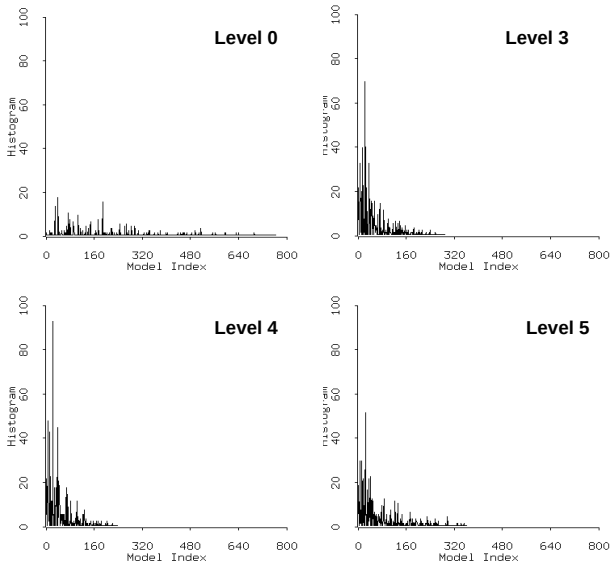

Figure 20. Histogram plots of converged models RHC of population of 1000 for MRA decomposed stack-power functions.

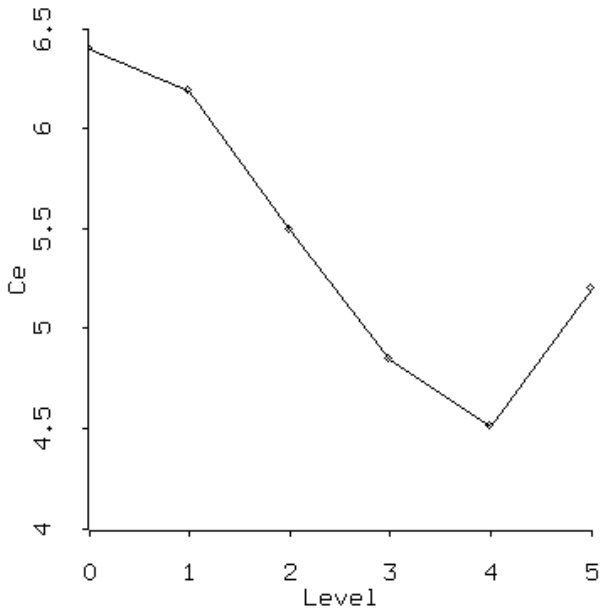

Figure 21. Complexity C_e as a function of the level of wavelet decomposition.

To achieve a more comprehensive picture of the performance of the multi-scale algorithm, we applied the Multi-scale Random Hill-Climbing algorithm to the 55 parameter statics problem using a population of 1000

models with $0-5$ levels of wavelet decomposition. Figure 20 shows histograms of converged models achieved with a non-linear conjugate gradient approach. It can be seen that several levels of decomposition result in fewer distinct converged models. However, too much decomposition (level 5) may result in a function that is too flat for optimization. Figure 21 shows C_e calculated from one RHC for decomposition levels varying from 0 to 5, where level 0 represents using the original data. Repeated experiments on different initial populations and stopping criteria display the same shape of the C_e curve, though each C_e may vary according to population size and stopping criterion. These results indicate that complexity of the optimization is reduced via MRA, up to a point (level 4, in this example).

Conclusions & Future Work

We have developed an entropy-based measure of the complexity of function surfaces and used it to study the difficulty associated with generic optimization problems. We have shown the application of this measure to various analytic test functions as well as a practical problem in exploration seismology. In addition, we have used this measure to gain insights into the behavior of multiresolution analysis.

Our criterion is based on a statistical analysis of the results of random hill-climbing, and measures the amount of information the hill-climbing can achieve about the topography of the function. The front-end for this calculation is the specification of an initial population of points in the domain of the function. At present we choose these with uniform probability on the interior of this domain. It is likely that there are more efficient ways of specifying these points in high dimensional spaces, for example, by quasi-Monte Carlo methods, or by subdividing the domain into small sub-domains and extrapolating the results.

The factors contributing to the information encapsulated by this measure are: the number of basins of attractions on the function landscape and the widths and depths of these basins. The number of distinct converged models from the random hill-climbing can estimate number of local optima, the histogram of the converged models is used to approximate widths of the basin of attractions, and the distribution of function values of each basin of attractions is the weighting factor of each probability. When $C_e = 0$, the function is unimodal with only one extremum. On the other hand, when the function is a constant, the complexity C_e is maximum: $C_e = \ln K$, where K is the population size of the random hill-climbing.

Ideally, this complexity, as a characterization of the topography of surfaces, should be independent of the searching algorithm and computing time. In practice, however, we can estimate statistically the topography of functions from only limited samples. Further, as the dimensionality increases, volumes become increasingly compressed near their boundaries. So it may be difficult to choose the initial population of models in such a way as to ensure efficient sampling of a function's domain.

At first glance it seems disturbing that a measure of complexity would be influenced by numerical issues such as ill-conditioning. But as a practical matter it is often difficult to untangle the results of ill-conditioning (flat landscape) from those of multi-modality, and so while it remains to investigate the sensitivity of our measure to the stopping criteria and search algorithms used, we believe that these are important aspects of what makes an optimization problem hard.

Acknowledgment

This work was partially supported by the sponsors of the Consortium Project on Seismic Inverse Methods for Complex Structures at the Center for Wave Phenomena, Colorado School of Mines, the Shell Foundation and Oak Ridge National Laboratory.

References

Aarts, E.H.L., & Korst, Jan. 1989. *Simulated Annealing and Boltzman Machines.* N.Y: Wiley.

Berry, R. S., & Breitengraser-Kunz, R. 1995. Topography and Dynamics of Multidimensional Interatomic Potential Surfaces. *Physical Review Letters*, **74**, 3951–3954.

Bunks, C., Saleck, F. M., Zaleski, S., & Chavent, G. 1995. Multiscale seismic waveform inversion. *Geophysics*, **60**(5), 1457–1473.

Chavent, G. 1991. On the theory and practice of nonlinear least-squares. *Adv. Water Resources*, **14**, 55–63.

Chen, Tong. 1994. *Multilevel Differential Semblance Optimization for Waveform Inversion.* Tech. rept. CWP-153. Center for Wave Phenomena, Colorado School of Mines.

Deng, H. L. 1995. *Using Multi-Resolution Analysis to Study the Complexity of Inverse Calculations.* Tech. rept. CWP-183. Center for Wave Phenomena, Colorado School of Mines.

Deng, H. L., Gouveia, W., & Scales, J. A. 1995. The CWP Object-Oriented Optimization Library. *submitted to The Leading Edge.*

Edwards, S. F., & Anderson, P. W. 1975. The theory of spin glasses. *Journal of Physics, F*, **5**, 965.

Falcioni, M., Marconi, U. M. B., Gianneschi, P. M., & Vulpiani, A. 1995. Complexity of the Minimum Energy Configurations. *Physical Review Letters*, **75**, 637–640.

Fischer, K. H., & Hertz, J. A. 1991. *Spin Glasses.* Cambridge Studies in Magnetism. Cambridge University Press.

Goldberg, D. E. 1989. *Genetic Algorithms in Search, Optimization, & Machine Learning.* Addison-Wesley.

Hertz, J., Krogh, A., & Palmer, R. G. 1991. *Introduction to the theory of neural computation.* Computation and neural systems series. Addison-Wesley.

Holland, J. H. 1975. *Adaptation in Natural and Artificial Systems.* University of Michigan Press.

Kauffman, S. A., & Weinberger, E. D. 1989. The NK model of rugged fitness landscapes and its application to maturation of the immune response. *Journal of Theoretical Biology*, **141**(2), 211.

Kaufmann, S. 1993. *The Origins of Order: Self-Organization and Selection in Evolution.* Oxford University Press. Chap. 2-3, pages 33–117.

Kirkpatrick, S., Gelatt, C.D., & Vecchi, M.P. 1983. Optimization by simulated annealing. *Science*, **220**, 671–680.

Magready, W. G., & Wolpert, D. H. 1995. *What Makes an Optimization Problem Hard?* Tech. rept. SFI-TR-95-05-046. the Santa Fe Institute.

Rothman, D. H. 1985. Nonlinear inversion, statistical mechanics, and residual statics estimation. *Geophysics*, **50**, 2797–2807.

Rothman, D. H. 1986. Automatic estimation of large residual statics corrections. *Geophysics*, **51**, 332–346.

Saito, N., & Beylkin, G. 1993. Multiresolution representations using the auto-correlation functions of compactly supported wavelets. *IEEE Transactions on Signal Processing*, **41**, 3585–3590.

Stadler, P. F. 1992a. Correlation in landscapes of combinatorial optimization problems. *Europhysics Letters*, **20**(6), 479–482.

Stadler, P. F. 1992b. Correlation structure of the landscape of the graph-bipartitioning problem. *Journal of Physics. A, Mathematical and General*, **25**(11), 3103–3110.

Stadler, P. F. 1992c. The landscape of the traveling salesman problem. *Physics Letters A*, **161**, 337–344.

van Laarhoven, P.J.M., & Aarts, E.H.L. 1987. *Simulated Annealing: Theory and Practice.* Dordrecht: Reidel.

Weinberger, E. D. 1990. Correlated and uncorrelated fitness landscapes and how to tell the difference. *Biological Cybernetics*, **63**, 325.

Whitley, D., Mathias, K., Rana, S., & Dzubera, J. 1995a. *Evaluating Evolutionary Algorithms: The Perils of Poor Empiricism.* Department of Computer Science, Colorado State University.

Whitley, D., Mathias, K., & Stork, C. 1995b. *Global Optimization for Geophysical Applications using Genetic Algorithms.* Tech. rept. CASI-TR-95-09. Department of Computer Science, Colorado State University.

Wolpert, D. H., & Magready, W. G. 1995. *No Free Lunch Theorems for Search.* Tech. rept. SFI-TR-95-02-010. The Santa Fe Institute.