Length Scale for Characterising Continuous Optimization Problems

Rachael Morgan and Marcus Gallagher

School of Information Technology and Electrical Engineering,
University of Queensland, Brisbane 4072, Australia
r.morgan4@uq.edu.au, marcusg@itee.uq.edu.au

Abstract. In metaheuristic optimization, understanding the relationship between problems and algorithms is important but non-trivial. There has been a growing interest in the literature on techniques for analysing problems, however previous work has mainly been developed for discrete problems. In this paper, we develop a novel framework for characterising continuous optimization problems based on the concept of *length scale*. We argue that length scale is an important property for the characterisation of continuous problems that is not captured by existing techniques. Intuitively, length scale measures the ratio of changes in the objective function value to steps between points in the search space. The concept is simple, makes few assumptions and can be calculated or estimated based only on the information available in black-box optimization (objective function values and search points). Some fundamental properties of length scale and its distribution are described. Experimental results show the potential use of length scale and directions to develop the framework further are discussed.

Keywords: Continuous optimization, Problem properties, Problem characterisation, Fitness landscape analysis.

1 Introduction

A continuous optimization problem with a simple symmetric boundary constraint is to find a solution vector $\mathbf{x}^*$ such that:

$$f(\mathbf{x}^*) \leq f(\mathbf{x}), \forall \mathbf{x} \in \mathcal{S} \tag{1}$$

where $\mathcal{S} = [b_l, b_u]^n \subseteq \mathbb{R}^n$. Given a metaheuristic algorithm, a standard question is how well will the algorithm perform at solving a given problem (in other words, how well-suited is the algorithm to the problem)?. For some types of problems, it is possible to answer these questions rigorously (e.g. if f is convex, smooth and differentiable, then Newton-based algorithms converge rapidly towards the optimum). However if little can be assumed (e.g. f is a 'black-box' problem), then the questions are much more difficult to answer. Metaheuristics utilise multiple heuristics, complex models and randomness, while problems may be high-dimensional, noisy, or have features such as many local optima or

C.A. Coello Coello et al. (Eds.): PPSN 2012, Part I, LNCS 7491, pp. 407–416, 2012.

other complex structures. The relationship between problems and algorithms in practice is the result of interactions between these factors.

Metaheuristics research has been dominated by the development of algorithms, but a recent focus has been to better understand both the relationship between algorithms and problems, and the nature of the problems themselves. For example, fitness landscape analysis has produced a theoretical framework and techniques for studying problems. However, this work has mainly been developed for discrete or combinatorial problems.

In this paper, we develop a framework for characterising continuous optimization problems based on the concept of *length scale*. While previous techniques from fitness landscape analysis can be applied, we argue that length scale is a critical concept for continuous problems that is not captured by these techniques. Sec. 2 reviews previous work on problem analysis with a focus on applicability in the continuous case. In Secs. 3 and 4 we define and develop the notion of problem length scale and its distribution. Illustrative experiments are provided in Sec. 5, and discussion is given in Sec. 6.

2 Discrete and Continuous Fitness Landscape Analysis

The notion of f as an (n-dimensional) 'fitness' landscape defined over $\mathcal{S}$ has been widely used as a model in evolutionary biology and computation. A fitness landscape is defined using f and a graph, G representing $\mathcal{S}$ (i.e $\mathcal{S}$ is discrete). Edges in G can be defined by a move operator and induce a neighbourhood in $\mathcal{S}$. Properties of the landscape can be defined in this framework, e.g. a (strict) local optimum is a point $\mathbf{x}'$ where all neighbours have a fitness worse than $f(\mathbf{x}')$. For a discrete $\mathcal{S}$, it is possible to determine whether or not $\mathbf{x}'$ is a local optimum by exhaustive evaluation of its neighbours. For all but very small problem sizes, enumeration of the landscape is impractical. Fitness landscape analysis typically uses random, statistical or other sampling methods to obtain points of interest (and/or their fitness values) from a landscape. Examples include the distribution of f values (density of states), fitness distance correlation (FDC) between the sample and a point (typically the global optimum), autocorrelation and correlation length statistics of random walks in $\mathcal{S}$. Information content aims at quantifying landscape ruggedness based on transitions observed in f values [15,14]. Previous research has also focussed on problem-specific techniques to characterise properties of combinatorial problems such as the travelling salesman problem. Comprehensive reviews of fitness landscape and problem analysis techniques can be found in [9,10,13].

If $\mathcal{S}$ is continuous, landscape features conceptually similar to the discrete case can be defined mathematically (as suggested in [9]), but evaluating them in practice is problematic. Each solution has an infinite number of neighbours in theory, yet a finite but extremely large number in practice due to finite-precision floating-point representation. Another significant difference between discrete and continuous landscapes is tied to the *distance* between points in $\mathcal{S}$ (using some metric). For a discrete landscape, the minimum possible distance will occur between a point and one of its neighbours, with a finite set of possible distance values between all points in G. For a continuous landscape, the minimum distance between points can be made arbitrarily small (in practice until the limit of precision is reached) and the number of possible distance values is infinite.

The reason for these difficulties lies in the difference between continuous and discrete problems. Consider a combinatorial problem with binary representation, $S = [0, 1]^n$. To solve the problem is to determine whether each variable x^i in the solution vector should take the value 0 or 1. A metric (e.g. Hamming distance) can be defined, but there is no notion of the *scale* of x^i. For a continuous problem however, finding an appropriate scale for each x^i is critical (e.g. does the objective function vary in a significant way with changes in x^i of order 10^3? 10^{-3}? 10^{-30}?). Fitness landscape techniques originating from the assumption of a discrete $\mathcal{S}$ do not capture such information because it is not relevant for the discrete case.

Despite these issues, there have been some adaptations of landscape analysis to continuous problems. Gallagher calculated FDC for the training problem in multi-layer perceptron neural networks [4]. For the learning tasks considered (student-teacher model), the global optimum is known, however this would not normally be the case for such a problem. Points were sampled from within a specified range around the global optimum. Wang and Li calculate FDC in the context of a continuous NK-landscape model and on some standard test functions [16]. Müller and Sbalzarini [7] analyse the CEC 2005 benchmark function set using FDC on points uniformly sampled from $\mathcal{S}$. While these results show interesting structure and differences between problems, the limitations of FDC noted for discrete problems remain (e.g. [7] concludes that FDC alone is not sufficient for problem design or measuring difficulty).

Dispersion is a recently-proposed problem metric [6] which measures the average distance between pairs of high quality solutions. Quality is determined by sampling points and retaining a percentage with the best fitnesses (according to a specified threshold). Dispersion is shown to be a useful metric in studying the performance of CMA-ES on a number of functions. Dispersion makes only limited use of the f values of points via the threshold used to produce the sample. Pairwise distances between solutions have also been analysed in samples of apparent local minima for multi-layer perceptron training [4].

In summary, there are some important limitations of existing techniques for the analysis of continuous problems, stemming from the adaptation of techniques developed for discrete problems and/or a limited use of the available information from sampling solution and their fitness values.

3 Length Scale in Optimization

We aim to develop a framework to study the topological/structural characteristics of a problem landscape independent of any particular algorithm. Importantly, the framework should utilise all information available in the black-box optimization setting, be estimated easily from data and be amenable to statistical and information theoretic analysis.

Definition 1. *Let* $\mathbf{x}^i$ *and* $\mathbf{x}^j$ *be two distinct solutions in the search space* ($\mathbf{x}^i \neq \mathbf{x}^j$) *with corresponding objective function values* $f(\mathbf{x}^i)$ *and* $f(\mathbf{x}^j)$. *The* ***length scale****,* r*, is defined as:*

$$r : [0, \infty) = \frac{|f(\mathbf{x}^i) - f(\mathbf{x}^j)|}{\|\mathbf{x}^i - \mathbf{x}^j\|} \quad (2)$$

The length scale intuitively measures how much the objective function value changes with respect to a step between two points in the search space. In this paper, we use Euclidean distance, however any appropriate metric can be used. Length scale is defined simply as a magnitude over a finite interval in the search space: directional information about a step from $\mathbf{x}^i$ to/from $\mathbf{x}^j$ is not considered.

Our definition of r is related to the *difference quotient* (also known as *Newton's quotient* and is a generalisation of *finite difference* techniques) from calculus and numerical analysis. The difference quotient is defined as $\frac{f(x+h)-f(x)}{h}$, and can be used to estimate the gradient at a point x, as $h \to 0$ [8]. Implementations of gradient-based algorithms utilise approximations of this form if the gradient of f is not available. Finite difference methods are widely used in the solution of differential equations, but are not directly related to this paper. Length scale is also related to the *Lipschitz constant*, defined as a constant, $L \geq 0$, where $|f(\mathbf{x}^i) - f(\mathbf{x}^j)| \leq L\|\mathbf{x}^i - \mathbf{x}^j\|, \forall \mathbf{x}^i, \mathbf{x}^j$ [17]. However, r does not assume that f is continuous and captures information about *all* rates of change of f over $\mathbf{x}$.

In some cases, it is possible to derive a simple expression for the length scale of a problem, as illustrated by the following examples.

Example 1. *1-D linear objective function*

Given $f = ax$ *where* $(x, a \in I\!R)$*, the length scale between* $\mathbf{x}^i$ *and* $\mathbf{x}^j$ *is:*

$$\begin{aligned} r &= \frac{|f(\mathbf{x}^i) - f(\mathbf{x}^j)|}{\|\mathbf{x}^i - \mathbf{x}^j\|} \\ &= \frac{|ax^i - ax^j|}{|x^i - x^j|} \\ &= |a| \end{aligned}$$

For this function, r captures the intuition that any step in $\mathcal{S}$ will be accompanied by a proportional change in f. The length scale of any finite set of samples from the search space (e.g. the points visited by an optimization algorithm) is invariant to the location(s) in $\mathcal{S}$ or the order in which the points were taken. The length scale of a (n-D) neutral or flat landscape is also a special case of this.

For most continuous problems, r will not be a constant over $\mathcal{S}$. In different regions of the space, the length scale value will depend on the local topology of the fitness landscape (varying slope, basins of attraction, ridges, saddle points, etc.).

Example 2. *1-D quadratic objective function*

Given $f = ax^2$ *where* $(x, a \in I\!R)$*, the length scale between* $\mathbf{x}^i$ *and* $\mathbf{x}^j$ *is:*

$$\begin{aligned} r &= \frac{|f(\mathbf{x}^i) - f(\mathbf{x}^j)|}{\|\mathbf{x}^i - \mathbf{x}^j\|} \\ &= \frac{\|ax^{i2} - ax^{j2}\|}{\|x^i - x^j\|} \\ &= \frac{|a|\|(x^i - x^j)(x^i + x^j)\|}{\|x^i - x^j\|} \\ &= |a|\|x^i + x^j\| \end{aligned}$$

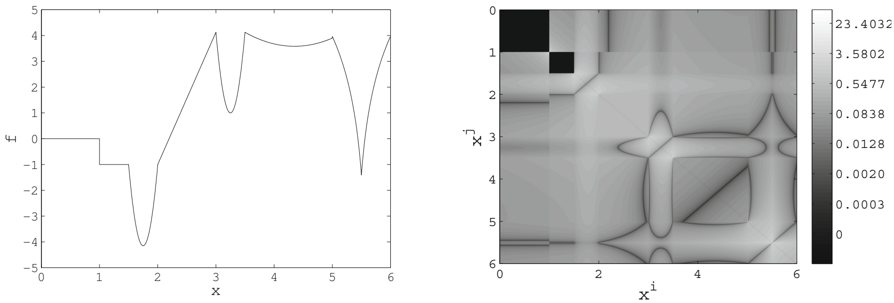
Here, steps between points that are relatively close to the optimum result in relatively small length scales compared to the same-sized steps further from the optimum. This suggests that an algorithm needs to reduce the size of the steps it makes to successfully approach the optimum of this function (e.g. gradient descent). To illustrate the richness of length scale information we construct an artificial 1-D function with a variety of different topological features.

Example 3. *1-D 'mixed-structure' function defined as follows and shown in Fig. 1(a).*

$$f(x) = \begin{cases} -1 \text{ if } 1 \leq x < 1.5 \\ 50(x-1.75)^2 - 4.15 \text{ if } 1.5 \leq x < 2 \\ 5.125x - 11.25 \text{ if } 2 \leq x < 3 \\ 50(x-3.25)^2 + 1 \text{ if } 3 \leq x < 3.5 \\ 0.75(x-4.35)^2 + 3.583 \text{ if } 3.5 \leq x < 5 \\ 3\log(|x-5.6|) + 5.5 \text{ if } 5 \leq x < 5.5 \\ 3\log(|x-5.4|) + 5.5 \text{ if } 5.5 \leq x < 6 \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

Since f is a 1-D problem ($x \in [0, 6]$) it is possible to enumerate length scales over the entire search space to a certain level of numerical precision. Fig. 1(b) shows the length scales calculated between pairs of points, x^i, x^j, at increments of 10^{-3} across $\mathcal{S}$. We have coloured the values using a logarithmic scale to better visualise magnitudes of change. The plot is symmetric across the diagonal, which follows from the definition of r. The thin black line along the diagonal is approximately a zero-length step ($x^i = x^j$). The two flat regions of the function produce black squares where $r = 0$. The dark lines and curves in the plot show steps in the space where $f(x^i) \approx f(x^j)$, e.g. moving from a point on one side of a basin or funnel to a point on the other side of the minimum at the same height. Within the plot, it can be seen that components of the function combine to produce different patterns and gradients of r values.

Overall, it is clear that r reflects the structure of f: if f has complex structure then this will also be captured in r. In addition, Fig. 1(b) gives an indication of how much variety is contained in the search points and f values that an algorithm encounters as it attempts to search a landscape effectively.



(a) (b)

Fig. 1. (a) 1-D 'mixed-structure' function. (b) Enumeration of length scales in the 1-D 'mixed-structure' function.

4 Length Scale Distribution

Length scale values produce information about problem structure. While it is possible to enumerate r over a large set of values for a 1-D problem, this is clearly infeasible for higher dimensions. One possibility is to summarise the values of r that occur over a given landscape. A good summary of r may be usable to predict the values of r we will see if further exploration of the landscape is conducted (particularly if the sampling technique is the same).

Definition 2. *Consider r as a continuous random variable. Then, let the* ***length scale distribution*** *be defined as the probability density function $p(r)$.*

Consider again Example 1. Since $r = |a|$, $p(r)$ is a Dirac delta function:

$$p(r) = \begin{cases} 1 \text{ if } r = |a| \\ 0 \text{ otherwise} \end{cases}$$

It follows that the n-D flat function also results in a Dirac delta function with a spike at $r = 0$.

Now reconsider Example 2. Let Z be the sum of two independent, continuous uniform random variables bounded by $[b_l, b_u]$. This produces a triangular distribution [5]:

$$p_Z(z) = \begin{cases} \frac{z-2b_l}{(b_u-b_l)^2} \text{ if } 2b_l \leq z \leq (b_l + b_u) \\ \\ \frac{2b_u-z}{(b_u-b_l)^2} \text{ if } (b_l + b_u) < z \leq 2b_u \\ \\ 0 \text{ otherwise} \end{cases}$$

The length scale distribution for Example 2 is the absolute value of $p_Z(z)$:

$$p(r) = |p_Z(r)| = p_Z(r) + p_Z(-r), \forall r \geq 0$$

Therefore, $p(r)$ of the 1-D quadratic function is a 'folded' triangular distribution.

While $p(r)$ can be derived for some functions, in general it can be approximated using probability density estimation based on r values sampled from the landscape (see Sec. 5). The length scale distribution is not unique for a problem but will vary depending on the structure present in that problem.

We can utilise concepts from information theory to compare landscapes via their length scale distributions. Shannon entropy is used as a measure of the uncertainty of a random variable [2]. Entropy measures the expected amount of information needed to describe the random variable. The entropy of $p(r)$ is:

$$h(r) = -\int_{0}^{\infty} p(r)\log_2\big(p(r)\big)\,dr \tag{4}$$

We conjecture that problems with structure of similar complexities should yield a similar $h(r)$, and hence, $h(r)$ is potentially very useful for categorising problems. The Dirac delta function has the smallest entropy of all density functions, meaning the n-D flat and 1-D linear functions minimize $h(r)$. The uniform density function (in a bounded region) has the largest entropy of any other density function bounded within the same region. To obtain a uniform $p(r)$, there must be length scales of uniformly varying size, e.g. random noise functions. Therefore, random noise functions maximize $h(r)$. This results in two extreme values of $h(r)$ that any given landscape is within.

5 Length Scales of the BBOB'10 Test Functions

In this section we examine length scales of the Black-Box Optimization Benchmarking 2010 (BBOB'10) test functions [3]. We aim to investigate whether or not there is a relationship between the 'difficulty' of functions (as measured by the best performing algorithms in BBOB'10) and length scale. Problems with largely varying length scales may contain a richer, more complex structure, and may be more difficult to solve. The methodology used in these experiments is general and can be easily applied to other black-box problems. Source code used is available at http://www.itee.uq.edu.au/∼uqrmorg4/length-scale-bbob.html.

We use a random Levy walk to sample $\mathcal{S}$ and corresponding f values. Levy walks generally yield good coverage of the search space at varying magnitudes of step sizes [12]. The Levy distribution pertaining to step size is parameterised by scale (γ) and location (δ) parameters, here both set to 0.001. This type of walk has frequent small steps (with 0.001 being the minimum), and infrequent large steps.

A Levy walk of 10^5 steps was conducted for each of the BBOB'10 functions. Walks were bounded by $[-5, 5]^{10}$, with proposed steps outside the boundary rejected. Length scales were calculated for each pair of solutions, producing 50005000 values of r. Kernel density estimation was then used to estimate $p(r)$. The kernel bandwidth was calculated using the 'solve-the-equation plug-in'

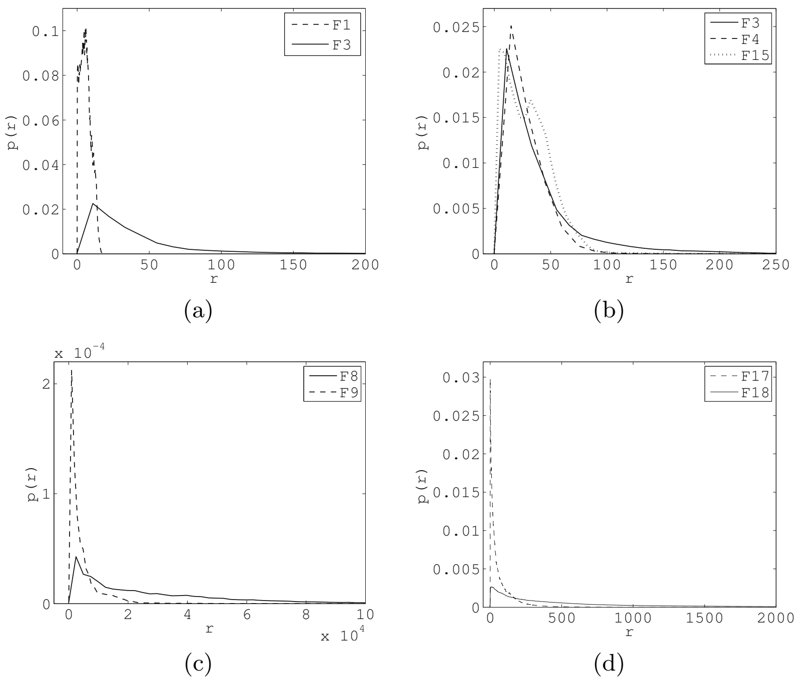


Fig. 2. Examples of similar length scale distributions estimated on BBOB'10 functions

method [11]. The resulting length scale distributions were quite varied across the problems, however there were a few notable similarities, shown in Fig. 2.

Fig. 2(a) shows $p(r)$ for Sphere (F1) and Rastrigin (F3). We do not expect these distributions to be identical, however since the global structure of F3 is F1, we observe some similarity. In Fig. 2(b), we see almost identical length scale distributions, resulting from sampling Rastrigin-like functions. The length scale distributions for Rosenbrock (F8) and Rosenbrock Rotated (F9) can be seen in Fig. 2(c). The larger peak in the F9 distribution indicates that there are more low-valued length scales. Both functions are variations of Rosenbrock and we observe similar changes in fitness, and hence similar ranges of length scales. This observation is also true for the Schaffer F7 (F17) and Shaffer F7 Moderately Ill-Conditioned (F18) distributions (Fig. 2(d)). It is clear that problems with similar structure have similar length scale distributions, while problems with vastly different structure have different length scale distributions.

To examine the relationship between r and problem difficulty, we use the expected running time (ERT) for the best-performing algorithm in the BBOB'10 results [1] as a proxy for problem difficulty. We use the results within a precision of 10^{-8} of the global optimum (e.g. the BFGS algorithm performs best on F1 with an ERT of 23, and so we use '23' to indicate problem difficulty). Given the kernel density estimate of each $p(r)$, we can estimate $h(r)$. Fig. 3 shows $h(r)$ vs ERT for the BBOB'10 functions. There is an interesting relationship

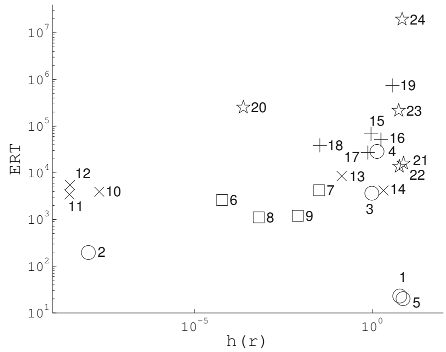


Fig. 3. Length scale distribution entropy vs ERT for BBOB'10 problems. F1 to F5 are ◦, F6 to F9 are □, F10 to F14 are ×, F15 to F19 are + and F20 to F24 are ⋆.

between $h(r)$, ERT and the function type. Separable functions are denoted by ◦; low to moderately conditioned functions by □; high and uni-modal functions by ×; multi-modal with adequate global structure by +; and multi-modal with weak global structure by ⋆. Fig. 3 clearly shows clustering of problems within categories, e.g. F10-12, F6-7 and F15-19. In fact, for problems F6-12, ERT is incapable of distinguishing the categories, while $h(p(r))$ can. This illustrates that length scales capture valuable information about problem structure.

In Fig. 3, there is a great distinction between the uni-modal (◦, □ and ×) and multi-modal functions (+ and ⋆). In general, multi-modal functions are more difficult, and so we expect ERT to separate uni-modal functions from multi-modal functions. This is observed, however we can additionally see that $h(r)$ is capable of characterising uni-modal and multi-modal functions.

6 Summary and Conclusions

We have proposed a framework for characterising continuous optimization problems using the notion of length scale and its distribution. The framework is based on utilising all available information in black-box optimization and is readily calculated using points from $\mathcal{S}$ and their f values. This paper has discussed some properties of length scale, presented examples and experimental results using the BBOB'10 competition results.

We believe that there is considerable scope for future work. It should be possible to explore the relationship between features such as landscape modality or ruggedness and the shape of $p(r)$. Entropy was used to summarise the distribution, but other ideas from statistics and information theory deserve investigation. Our experimental results assume that the sampling methodology used produces a representative sample of the search space. This requires quantification. It would be interesting to analyse different real-world and benchmark continuous problems using length scale. The set of points that an algorithm evaluates during a run could also be analysed to examine the length scales visited by the algorithm.

Acknowledgments. We thank Mike Preuss and Olaf Mersmann for making available pre-processed results data from BBOB'10.

References

1. Auger, A., Finck, S., Hansen, N., Ros, R.: BBOB 2010: Comparison Tables of All Algorithms on All Noiseless Functions. Technical Report RT-388, INRIA (2010), http://hal.inria.fr/inria-00516689
2. Cover, T.M., Thomas, J.A.: Elements of information theory. Wiley Interscience, New York (1991)
3. Finck, S., Hansen, N., Ros, R., Auger, A.: Real-parameter black-box optimization benchmarking 2009: Noiseless functions definitions. Tech. Rep. 2009/20, Research Center PPE (2009)
4. Gallagher, M.: Multi-layer perceptron error surfaces: visualization, structure and modelling. Ph.D. thesis, Dept. Computer Science and Electrical Engineering, University of Queensland (2000)
5. Killmann, F., von Collani, E.: A note on the convolution of the uniform and related distributions and their use in quality control. Economic Quality Control 16(1), 17–41 (2001)
6. Lunacek, M., Whitley, D.: The dispersion metric and the CMA evolution strategy. In: GECCO 2006, pp. 477–484. ACM, New York (2006)
7. Müller, C.L., Sbalzarini, I.F.: Global Characterization of the CEC 2005 Fitness Landscapes Using Fitness-Distance Analysis. In: Di Chio, C., Cagnoni, S., Cotta, C., Ebner, M., Ekárt, A., Esparcia-Alcázar, A.I., Merelo, J.J., Neri, F., Preuss, M., Richter, H., Togelius, J., Yannakakis, G.N. (eds.) EvoApplications 2011, Part I. LNCS, vol. 6624, pp. 294–303. Springer, Heidelberg (2011)
8. Overton, M.: Numerical Computing with IEEE Floating Point Arithmetic. Cambridge University Press (2001)
9. Pitzer, E., Affenzeller, M.: A Comprehensive Survey on Fitness Landscape Analysis. In: Fodor, J., et al. (eds.) Recent Advances in Intelligent Engineering Systems. SCI, vol. 378, pp. 161–191. Springer, Heidelberg (2012)
10. Reidys, C., Stadler, P.: Combinatorial landscapes. SIAM Rev. 44(1), 3–54 (2002)
11. Sheather, S., Jones, M.: A reliable data-based bandwidth selection method for kernel density estimation. Journal of the Royal Statistical Society. Series B (Methodological) 683–690 (1991)
12. Shlesinger, M.F., West, B.J., Klafter, J.: Lévy dynamics of enhanced diffusion: Application to turbulence. Physical Review Letters 58, 1100–1103 (1987)
13. Smith-Miles, K., Lopes, L.: Measuring instance difficulty for combinatorial optimization problems. Computers and Operations Research 39(5), 875–889 (2011)
14. Steer, K., Wirth, A., Halgamuge, S.: Information Theoretic Classification of Problems for Metaheuristics. In: Li, X., Kirley, M., Zhang, M., Green, D., Ciesielski, V., Abbass, H.A., Michalewicz, Z., Hendtlass, T., Deb, K., Tan, K.C., Branke, J., Shi, Y. (eds.) SEAL 2008. LNCS, vol. 5361, pp. 319–328. Springer, Heidelberg (2008)
15. Vassilev, V.K., Fogarty, T.C., Miller, J.F.: Information characteristics and the structure of landscapes. Evol. Comput. 8, 31–60 (2000)
16. Wang, Y., Li, B.: Understand behavior and performance of real coded optimization algorithms via NK-linkage model. In: CEC 2008, pp. 801–808. IEEE (2008)
17. Wood, G.R., Zhang, B.P.: Estimation of the Lipschitz constant of a function. Journal of Global Optimization 8, 91–103 (1996)