J Glob Optim (2007) 37:159–175
DOI 10.1007/s10898-006-9042-z

ORIGINAL PAPER

Generating functions and the performance of backtracking adaptive search

W. Baritompa · D. W. Bulger · G. R. Wood

Received: 20 March 2004 / Accepted: 5 May 2006 /
Published online: 7 September 2006

Abstract Backtracking adaptive search is a simplified stochastic optimisation procedure which permits the acceptance of worsening objective function values. Key properties of backtracking adaptive search are defined and obtained using generating functions. Examples are given to illustrate the use of this methodology.

Keywords Adaptive search · Backtracking · Global optimization · Hesitant adaptive search · Markov process

AMS 1991 Subject Classification 90C65 · 90C30 · 65K05

1 Introduction

Backtracking adaptive search (BAS) has been introduced in [7] and used in [6] as an initial model for the study of the convergence behaviour of stochastic global optimisation algorithms. In this paper, we develop tools to examine the performance of BAS. Specifically, we consider and link three performance measures: the distribution of objective function values at a given iteration, the probability that the algorithm has reached a pre-set level (the success rate) at a given iteration, and the distribution of the number of iterations until a preset level is reached.

In [7] a first principles approach (of calculating the expectation of the exponential function a random variable) was used to produce the factorial moment generating function for the number of iterations to first reach a preset level. This paper, com-

W. Baritompa (✉)
Department of Mathematics and Statistics, University of Canterbury, Christchurch, New Zealand
E-mail: b.baritompa@math.canterbury.ac.nz

D. W. Bulger
Department of Statistics, Macquarie University, NSW 2109, Australia

G. R. Wood
Department of Statistics, Macquarie University, NSW 2109, Australia

putes a number of generating functions using a formal power series approach and thus complements that work by relating the generating function found there to other generating functions of interest. These generating functions are used to examine the performance of BAS, and provide a tool to prove a surprising performance result.

Backtracking adaptive search is reviewed in Sect. 2 and generating functions introduced and analysed in Sect. 3. In Sect. 4, we illustrate the results using simple but edifying examples. Section 5, provides a general performance result. The paper concludes with a summary.

2 Backtracking adaptive search and its relevance to global optimization

Our interest centres on the global optimisation problem

$$\text{minimise} f(x), \quad \text{subject to } x \in S,$$

where S is a measurable space and $f\colon S \to [y_*, y^*]$ is a measurable function. An algorithm will be considered to terminate upon sampling a function value sufficiently small. We begin with some background to the current study.

This work has its origins with Zabinsky and Smith [8], where attention was first drawn to the remarkable global optimization performance, on the above problem, of the algorithm termed pure adaptive search (PAS). At the kth iteration, PAS selects a point X_k from the subset of strictly improving points in the domain, according to the restriction of a probability measure δ on S. Under appropriate conditions it was shown to have complexity, which is linear in the dimension of the domain. The key to the analysis was the study of the records of a related non-homogeneous Poisson process.

Subsequently there have been two generalisations of PAS, movements towards more realistic models for objective function values arising from stochastic optimisation algorithms. Hesitant adaptive search (HAS) [1] allows the algorithm to pause at the current level, while backtracking adaptive search (BAS) [7] allows acceptance of worsening values.

We are the first to acknowledge that BAS is still an inadequate model. But it is the best, we have to date; it models the important observed behaviour of backtracking and is capable of full analysis. Better models, more accurately capturing the behaviour of real algorithms and yielding to analysis, are needed. It is hoped that this paper, presenting a second and quite different approach to the analysis of BAS to that given in [7], will assist in stimulating such developments.

2.1 Importance to theory of global optimization

The analysis here is part of an approach, first exposited in [6], to understanding the convergence of global optimisation algorithms. The approach is a familiar one: a parametric model for variable behaviour (here the sequence of objective function values) is established, data (objective function values from a run of a stochastic algorithm) is used to estimate the model parameters, then the fitted model is used to approximate the behaviour of the real algorithm. Such a model allows us to give an answer the question "How long must, we run an algorithm to be within a given distance of the global optimum?", a key question in global optimisation. The method relies on extrapolation, so is dependent on early patterns in objective function values, used

to estimate model parameters, being continued later in the run. The performance of BAS presented in this paper provides a realistic benchmark.

Before, we launch into the paper proper, we offer a philosophical comment. We consider it important to work at all levels in global optimisation, from the very practical to the very theoretical. This paper verges towards the latter end, unapologetically. Theory development generally follows application; in this case, analysis of adaptive search methods are preceding application. It is interesting to note, however, that recently they are showing scope for practical application in [2, 4].

2.2 A description of backtracking adaptive search

We now present an informal description of BAS, followed by a formal description. In BAS, the first sample point is chosen according to δ. Each following iteration X_{k+1} either worsens (with probability $w(X_k)$), hesitates (with probability $h(X_k)$) or betters (with probability $b(X_k)$). Once, the decision has been made to either worsen or better, X_{k+1} is chosen according to the restriction of δ to the current worsening or bettering set in S, respectively. Note that $w + h + b \equiv 1$. The PAS is the special case of BAS occurring when $h \equiv 0$ and $w \equiv 0$, while HAS is the special case of BAS occurring when only $w \equiv 0$. We now present BAS formally.

Let w, h and b be non-negative real valued functions on $[y_*, y^*]$ with $w+h+b \equiv 1$. The worsening function w is assumed to be integrable, while the bettering function b is assumed to be integrable and for all y, bounded away from zero on $(y, y^*]$.

2.3 Backtracking adaptive search

Step 1 Set $k = 0$. Generate X_0 according to δ. Set $Y_0 = f(X_0)$.

Step 2 Generate X_{k+1} according to the normalised restriction of δ to

$$\begin{array}{lll} \{x \in S : f(x) > Y_k\} & \text{with probability} & w(Y_k), \\ \{x \in S : f(x) = Y_k\} & \text{with probability} & h(Y_k), \\ \{x \in S : f(x) < Y_k\} & \text{with probability} & b(Y_k). \end{array}$$

Set $Y_{k+1} = f(X_{k+1})$.

Step 3 If a stopping criterion is met, stop. Otherwise, increment k and return to Step 2.

The BAS is considered to have converged when it first passes below some value y with $y_* < y \leq y*$, that is, when $Y_k \leq y$ for the first time. We denote the termination region $[y_*, y]$ by T and let $N(y)$ be the number of iterations before landing in the termination region. Thus $N(y) = \min\{k : Y_k \leq y\}$, since, the first iteration is denoted Y_0.

An extremely convenient summary measure is ρ, the range probability measure defined by $\rho(A) = \delta(f^{-1}[A])$ for each measurable subset A of $[y_*, y^*]$ (for brevity, we write $\rho(\{t\})$ as $\rho(t)$). We assume throughout that $\rho(T) > 0$. The range cumulative distribution function (CDF) associated with ρ we denote p, so for $t \in \mathbb{R}$,

$$p(t) = \rho((-\infty, t]) = \delta\left(f^{-1}(-\infty, t]\right).$$

In this paper, we assume that p is continuous and that the probability density function f_k of Y_k exists. Thus, without loss of generality we can deal with a standardised BAS

range process. That is, $Y_0, Y_1, \ldots$, is a Markov chain on $[0,1]$ with uniform initial distribution (so $Y_0 \sim \lambda$, where λ is Lebesgue measure on $[0,1]$) and kernel

$$\beta(x,A) = b(x)\frac{\lambda(A \cap [0,x))}{x} + h(x)\delta_x(A) + w(x)\frac{\lambda(A \cap (x,1])}{1-x}.$$

We let $f_n(y)$ and $F_n(y)$ be the probability distribution function (PDF) and CDF, respectively, for Y_n. The Markov kernel for BAS gives the recursion $f_{n+1}(y) = \int_{x=0}^{1} f_n(x)\, d\beta(x,y)$, which in this case gives $f_0(y) = 1$ and

$$f_{n+1}(y) = \int_y^1 \frac{b(x)}{x} f_n(x)\mathrm{d}x + f_n(y)h(y) + \int_0^y \frac{w(x)}{1-x} f_n(x)\mathrm{d}x. \tag{1}$$

Intuitively, the new PDF $f_{n+1}(y)$[1] at y is found by averaging over all possible states x at the previous iteration. Here, $x \in [0,y)$ can lead to y if worsening occurs, $x \in (y,1]$ can lead to y if bettering occurs, and y leads to itself if hesitation happens.

3 Main results

Two measures are of interest when, we run a stochastic global minimisation algorithm. First, we want to know the distribution of the minimum value achieved at the kth iteration, so we compute the cumulative probabilities $P(\min\{Y_0, Y_1, \ldots, Y_k\} \le y)$ as y varies. Second, we want to know the distribution of the number of iterations $N(y)$ immediately prior to reaching a preset level y, so we compute the probability masses $P(N(y) = k)$ as k varies. Note, that the first random variable is continuous and the second discrete.

A useful tool for finding these values and others proves to be a generating function. Given functions $a_k(y)$, the formal infinite series

$$G(y,z) = a_0(y) + a_1(y)z + a_2(y)z^2 + \cdots$$

is a generating function for them. They can be recovered by $a_k(y) = \frac{1}{k!}(\partial^k G/\partial z^k)(y,0)$. Our interest will be in the following generating functions:

(1) $G^{\mathrm{pdf}}(y,z)$ where $a_k(y) = f_k(y)$, the PDF for Y_k.
(2) $G^{\mathrm{cdf}}(y,z)$ where $a_k(y) = F_k(y)$, the CDF for Y_k.
(3) $G^{\mathrm{succ}}(y,z)$ where $a_k(y) = P(\min\{Y_0, Y_1, \ldots, Y_k\} \le y)$, the first above mentioned measure, the probability of the algorithm's success at reaching level y or below by the kth iteration.
(4) $G^{\mathrm{fm}}(y,z)$ where $a_k(y) = P(N(y) = k)$, the second above mentioned measure (this relates to convergence in the sense of arriving at level y, however, we use

[1] A similar recursion for the cdfs is possible directly. Note, that at a given state x the conditional CDF $F_x(y)$ is the piecewise linear function that starts at $(0,0)$ with slope $b(x)/x$ and at $y = x$ jumps to $1-w(x)$ and continues with slope $w(x)/(1-x)$. The new state's CDF is found as a mixture of these cdfs with respect to the old states, so $F_{n+1}(y) = \int_0^1 F_x(y)\mathrm{d}F_n(x)$. This gives the recursion for generating the cdfs, namely $F_0(y) = y$ and

$$F_{n+1}(y) = F_n(y) - (1-y)\int_0^y \frac{w(x)}{1-x} F_n'(x)\mathrm{d}x + y\int_y^1 \frac{b(x)}{x} F_n'(x)\mathrm{d}x.$$

"fm" as in due course, we show it to be the familiar factorial moment generating function of $N(y)$).

We state the main result concerning the explicit formulae for the key generating functions. These formulae are parts of Theorem 3.4 and Corollary 3.2.

This result and others in this section depend heavily on the following function.

$$E(y,z,c,d)=\frac{1}{1-zh(y)}\exp\left(z\left(\int_c^y \frac{w(t)}{(1-t)(1-zh(t))}\mathrm{d}t+\int_y^d \frac{b(t)}{t(1-zh(t))}\mathrm{d}t\right)\right) \quad (2)$$

For brevity, we denote $E(y,z)=E(y,z,0,1)$. We will see later that c and d may be chosen arbitrarily.

Theorem 3.1 *Let* $G^{\mathrm{base}}(u,z,x)=\int_x^u E(t,z,x,x)\mathrm{d}t$.

$$G^{\mathrm{cdf}}(y,z)=\frac{G^{\mathrm{base}}(y,z,0)}{(1-z)G^{\mathrm{base}}(1,z,0)},$$
$$G^{\mathrm{succ}}(y,z)=\frac{y/(1-z)}{y+(1-z)G^{\mathrm{base}}(1,z,y)},$$
$$G^{\mathrm{fm}}(y,z)=\frac{y}{y+(1-z)G^{\mathrm{base}}(1,z,y)}.$$

The reader may wish to skip the rest of this section and see the consequences of the main result.

3.1 Results and proofs

Our path to obtaining these generating functions begins by extending the horizon and finding the generating functions for the family of processes Y_k^x, which we associate with the original BAS process Y_k. These processes are identical to the original one in every way, except that their hesitation functions are set on $[0,x]$ to be one.[2] Y_k^x are "frozen" past level x, thus sample paths for these new processes remain at the first level hit, which is less than or equal to x.

We denote the corresponding generating functions of this family by $G^{x:\mathrm{pdf}}(y,z)$, $G^{x:\mathrm{cdf}}(y,z)$, $G^{x:\mathrm{succ}}(y,z)$ and $G^{x:\mathrm{fm}}(y,z)$. Note $G^{\mathrm{pdf}}(y,z)=G^{0:\mathrm{pdf}}(y,z)$, etc.

In this section, after establishing various lemmas, we provide the main results, which describe the family of generating functions and their interrelationships.

Lemma 3.1 (Integral equation). *Consider the generating function* $G^{\mathrm{pdf}}(y,z)$ *for the PDF for* Y_k.

$$\int_0^1 G^{\mathrm{pdf}}(y,z)\mathrm{d}y=\frac{1}{1-z}$$

and it satisfies

$$G(y,z)=1+zh(y)G(y,z)+z\int_0^y \frac{w(t)}{1-t}G(t,z)\mathrm{d}t+z\int_y^1 \frac{b(t)}{t}G(t,z)\mathrm{d}t. \quad (3)$$

[2] One could also use processes with w forced to be zero on $[0,x]$

Proof We have $G^{\mathrm{pdf}}(y,z) = f_0(y) + f_1(y)z + f_2(y)z^2 + \cdots$. So its integral is $1 + z + z^2 + \cdots$ as required. To produce the required integral equation, we combine the recursion for the probability density functions (1), multiplied by $1, z, z^2$, etc., as follows.

$$\begin{aligned}
f_0(y) &= 1,\\
zf_1(y) &= z\left(f_0(y)h(y) + \int_0^y \frac{w(t)}{1-t}f_0(t)\mathrm{d}t + \int_y^1 \frac{b(t)}{t}f_0(t)\mathrm{d}t\right)\\
z^2f_2(y) &= z^2\left(f_1(y)h(y) + \int_0^y \frac{w(t)}{1-t}f_1(t)\mathrm{d}t + \int_y^1 \frac{b(t)}{t}f_1(t)\mathrm{d}t\right)\\
z^3f_3(y) &= z^3\left(f_2(y)h(y) + \int_0^y \frac{w(t)}{1-t}f_2(t)\mathrm{d}t + \int_y^1 \frac{b(t)}{t}f_2(t)\mathrm{d}t\right)\\
&\vdots
\end{aligned}$$

As the generating function is a formal series, summing provides the required integral equation. □

Note, that when using generating functions later, we need $z \in [0,1]$. For values at $z = 1$ or $z = 0$, we will use the appropriate limit.

Lemma 3.2 *For $z \in (0,1)$, the value of $(1-z)\int_0^1 E(x,z)\mathrm{d}x$ is finite.*

Proof As both worsening and bettering are bounded above by 1, it is easy to check that $(1-z)E(x,z) \le (x(1-x))^{-z}$ and that the integral of (i) is

$$\frac{2^{-1+2z}\sqrt{\pi}\,\Gamma(-z+1)}{\Gamma(-z+3/2)},$$

which is finite for z in the open interval. □

Lemma 3.3 *Both $G(y,z) = (1-zh(y))E(y,z)$ and $G(y,z) = (1-zh(y))G^{\mathrm{pdf}}(y,z)$ satisfy the differential equation:*

$$\frac{\partial}{\partial y}G(y,z) = \frac{z}{1-zh(y)}\left(\frac{w(y)}{1-y} - \frac{b(y)}{y}\right)G(y,z). \tag{4}$$

Proof Recalling that $b(y) + h(y) + w(y) = 1$, the first function is straightforwardly checked.[3] The second is an immediate consequence of Lemma 3.1 obtained by rearranging Eq. (3) and differentiating. □

Lemma 3.4 (Formula for G^{pdf}) *The generating function for the probability density function $f_k(y)$ of Y_k is given by*

$$G^{\mathrm{pdf}}(y,z) = \frac{E(y,z)}{(1-z)\int_0^1 E(t,z)\mathrm{d}t} = \frac{E(y,z,c,d)}{(1-z)\int_0^1 E(t,z,c,d)\mathrm{d}t}$$

and thus is independent of c and d.

[3] Differentiability of h is not required, as $1 - zh(y)$ is cancelled in the product defining $G(y,z)$

Proof We first show independence of c and d. Note

$$E(y,z,c,d) = E(y,z)\exp\left(-z\left(\int_0^c \frac{w(x)}{(1-x)(1-zh(x))}\mathrm{d}x + \int_d^1 \frac{b(x)}{x(1-zh(x))}\mathrm{d}x\right)\right)$$

so the common factor cancels.[4]

We have $(1-zh(y))G^{\mathrm{pdf}}(y,z)$ must be a multiple (by a function depending only on z) of $(1-zh(y))E(y,z)$ as the both satisfy the separable DE from Lemma 3.3. Thus $G^{\mathrm{pdf}}(y,z) = K(z)E(y,z)$. Now, integrating both sides and using the first part of Lemma 3.1, we have that $K(z)$ must be $\dfrac{1}{(1-z)\int_0^1 E(t,z)\mathrm{d}t}$. □

Lemma 3.5 $G^{\mathrm{pdf}}(y,z) = \dfrac{\partial}{\partial y}G^{\mathrm{cdf}}(y,z)$.

Proof $\frac{\partial}{\partial y}G^{\mathrm{cdf}}(y,z) = \frac{\partial}{\partial y}\sum_{k=0}^{\infty} z^k F_k(y) = \sum_{k=0}^{\infty} z^k f_k(y) = G^{\mathrm{pdf}}(y,z)$. □

Lemma 3.6 $G^{\mathrm{fm}}(y,z) = (1-z)G^{\mathrm{succ}}(y,z)$.

Proof As $N(y) = \min_k Y_k \le y$, the event $Y_i > y$ for $i = 0,\ldots,k$ is equivalent to $N(y) > k$. $P(\min\{Y_0, Y_1, \ldots, Y_k\} \le y) = P(N(y) \le k)$ and

$$\begin{aligned} G^{\mathrm{succ}}(y,z) &= P(N(y)\le 0) + zP(N(y)\le 1) + z^2P(N(y)\le 2) + \cdots \\ &= (1+z+z^2+\cdots)P(N(y)=0) + (z+z^2+z^3+\cdots)P(N(y)=1) + \cdots \\ &= \frac{1}{1-z}\left(P(N(y)=0) + zP(N(y)=1) + z^2P(N(y)=2) + \cdots\right) \\ &= \frac{1}{1-z}G^{fm}(y,z). \end{aligned}$$

□

The three main theorems that follow describe the generating function for the complete family $G^{x:\mathrm{pdf}}(y,z)$, $G^{x:\mathrm{cdf}}(y,z)$, $G^{x:\mathrm{succ}}(y,z)$ and $G^{x:\mathrm{fm}}(y,z)$.

Theorem 3.2 (Formula for $G^{x:\mathrm{pdf}}$) *The generating function for the PDF $f_k^x(y)$ of Y_k^x is given by*

$$G^{x:\mathrm{pdf}}(y,z) = \begin{cases} \dfrac{1/(1-z)}{x + (1-z)G^{\mathrm{base}}(1,z,x)} & \text{for } y \le x, \\ \dfrac{E(y,z,x,x)}{x + (1-z)G^{\mathrm{base}}(1,z,x)} & \text{for } y > x, \end{cases}$$

where

$$G^{\mathrm{base}}(u,z,x) = \int_x^u E(t,z,x,x)\mathrm{d}t.$$

[4] The integrals may diverge if $c = 1$ or $d = 0$, giving a ratio of "0/0", so to avoid taking limits these values are not used in practice.

Proof $G^{x:\mathrm{pdf}}(y,z)$ is just $G^{\mathrm{pdf}}(y,z)$ for the different bettering, hesitation and worsening functions appropriate to Y_k^x. Using superscripts to denote these, formally, h^x is the discontinuous function that is 1 on $[0,x]$ and h on $(x,1]$ (similarly define b^x and w^x to be zero on the initial closed interval). Denote by $E^x(y,z,c,d)$ the analogue of (2) using h^x, b^x and w^x. The result follows by using Lemma 3.4 with $c=d=x$ and noting that

$$E^x(y,z,x,x)=\begin{cases}1/(1-z) & \text{for } y\le x,\\ E(y,z,x,x) & \text{for } y>x.\end{cases}$$ □

Note that for $y \le x$, since, the distribution of the first record less than or equal to y is uniform on $[0,y]$, the generating function in the theorem statement does not depend on y.

Our, next theorem shows how to move from any one of the three generating functions to any other.

Theorem 3.3

(1) $G^{x:\mathrm{succ}}(y,z)=\begin{cases}G^{x:\mathrm{cdf}}(y,z) & \textit{for } y\le x,\\ G^{y:\mathrm{cdf}}(y,z) & \textit{for } y\ge x,\end{cases}$

(2) $G^{x:\mathrm{cdf}}(y,z)=\int_0^y G^{x:\mathrm{pdf}}(t,z)\mathrm{d}t,$

(3) $G^{x:\mathrm{fm}}(y,z)=(1-z)G^{x:\mathrm{succ}}(y,z).$

Proof For result (1) when $y \le x$, the failure event $Y_k^x > y$ for all $k=0,\ldots,n$ is equivalent to $Y_n^x > y$. This follows since, the process first frozen at x and then frozen at y is the same as the original process frozen at x. When $y \ge x$, the failure event $Y_k^x > y$ for all $k=0,\ldots,n$ is equivalent to $Y_n^y > y$. This follows since, the process first frozen at x and then frozen at y is the same as the original process frozen at y.

The last two results follow immediately from Lemmas 3.5 and 3.6. □

An immediate consequence provides the generating function for the original BAS process.

Corollary 3.1 $G^{\mathrm{succ}}(y,z)=G^{y:\mathrm{cdf}}(y,z).$

Proof $G^{\mathrm{succ}}(y,z)=G^{0:\mathrm{succ}}(y,z)=G^{y:\mathrm{cdf}}(y,z).$ □

Our final results in this section encompasses the main theorem and further explicit formulae in terms of the quantity

$$G^{\mathrm{base}}(u,z,x)=\int_x^u E(t,z,x,x)\mathrm{d}t$$

used in Theorem 3.2.

Theorem 3.4

(1) $G^{\mathrm{cdf}}(y,z)=\dfrac{G^{\mathrm{base}}(y,z,0)}{(1-z)G^{\mathrm{base}}(1,z,0)}$

(2) $G^{x:\mathrm{cdf}}(y,z)=\begin{cases}\dfrac{y/(1-z)}{x+(1-z)G^{\mathrm{base}}(1,z,x)} & \textit{for}\quad y\le x,\\ \dfrac{x/(1-z)+G^{\mathrm{base}}(y,z,x)}{x+(1-z)G^{\mathrm{base}}(1,z,x)} & \textit{for}\quad y\ge x.\end{cases}$

Proof Formula (1) for G^{cdf} follows from Lemmas 3.4 and 3.5. To get formula (2) for $G^{x:\text{cdf}}$, working with h^x, b^x and w^x define $G^{x:\text{base}}(y,z,a) = \int_a^y E^x(t,z,a,a)\mathrm{d}t$, so by the first formula

$$G^{x:\text{cdf}}(y,z) = \frac{G^{x:\text{base}}(y,z,0)}{(1-z)G^{x:\text{base}}(1,z,0)}.$$

Note by checking cases $t \le x$ and $t \ge x$ and using that w^x and b^x are zero on $[0,x]$, we have $E^x(t,z,0,0) = E^x(t,z,x,x)$. Now, $\int_0^1 E^x(t,z,0,0)\mathrm{d}t = \int_0^x E^x(t,z,x,x)\mathrm{d}t + \int_x^1 E^x(t,z,x,x)\mathrm{d}t$ which, using the formula for E^x in the proof of Theorem 3.2, gives

$$G^{x:\text{base}}(1,z,0) = x/(1-z) + G^{\text{base}}(1,z,x).$$

Similarly

$$G^{x:\text{base}}(y,z,0) = \begin{cases} y/(1-z) & \text{for } y \le x, \\ x/(1-z) + G^{\text{base}}(y,z,x) & \text{for } y \ge x \end{cases}$$

and the result follows. □

Corollary 3.2

$$(1)\quad G^{x:\text{succ}}(y,z) = \begin{cases} \dfrac{y/(1-z)}{x + (1-z)G^{\text{base}}(1,z,x)} & \textit{for} \quad y \le x, \\ \dfrac{y/(1-z)}{y + (1-z)G^{\text{base}}(1,z,y)} & \textit{for} \quad y \ge x, \end{cases}$$

$$(2)\quad G^{\text{succ}}(y,z) = \frac{y/(1-z)}{y + (1-z)G^{\text{base}}(1,z,y)},$$

$$(3)\quad G^{x:\text{fm}}(y,z) = \begin{cases} \dfrac{y}{x + (1-z)G^{\text{base}}(1,z,x)} & \textit{for} \quad y \le x, \\ \dfrac{y}{y + (1-z)G^{\text{base}}(1,z,y)} & \textit{for} \quad y \ge x, \end{cases}$$

$$(4)\quad G^{\text{fm}}(y,z) = \frac{y}{y + (1-z)G^{\text{base}}(1,z,y)}.$$

Note, when $b = 1 - h$, G^{fm} reduces to the formula for the factorial moment generating function for HAS given in [5].

4 Applications

This section, contains some illustrations of the application of the theory of the previous section, then a general result concerning the performance of BAS.

4.1 Factorial moment generating function

It is well known that the factorial moment generating function of $N(y), E(z^{N(y)})$, is $p_0(y) + p_1(y)z + p_2(y)z^2 + \cdots$ where $p_k(y) = P(N(y) = k)$ and hence is $G^{\text{fm}}(y,z)$. So the nth factorial moment is obtained as $\mu_{(n)}(y) = E(n!\binom{N(y)}{n}) = (\partial^n G^{\text{fm}}/\partial z^n)(y,1)$. As an immediate consequence of this we have:

Proposition 4.1 *For $N(y)$, the expected number of iterations immediately prior to reaching y,*

(1) $E(N(y)) = \frac{1}{y}\int_y^1 X_1(y,t)\mathrm{d}t$

(2) $Var(N(y)) = \frac{1}{y^2}\left(\int_y^1 X_1(y,t)\mathrm{d}t\right)^2 - \frac{2}{y}\int_y^1 X_1(y,t)B_2(y,t)\mathrm{d}t + \frac{1}{y}\int_y^1 \frac{1+h(t)}{1-h(t)}X_1(y,t)\mathrm{d}t,$

where

$$B_1(y,t) = \int_y^t \left\{\frac{w(x)}{1-x} - \frac{b(x)}{x}\right\}\frac{\mathrm{d}x}{b(x)+w(x)},$$
$$B_2(y,t) = \int_y^t \left\{\frac{w(x)}{1-x} - \frac{b(x)}{x}\right\}\frac{\mathrm{d}x}{(b(x)+w(x))^2},$$
$$X_1(y,t) = E(t,1,y,y) = \frac{\exp(B_1(y,t))}{b(t)+w(t)}.$$

Proof $E(N(y)) = \mu_{(1)}(y)$ and $\mathrm{Var}(N(y)) = \mu_{(2)}(y) + \mu_{(1)}(y) - (\mu_{(1)}(y))^2$. Using the formula of the factorial moment generating function in Corollary 3.4, Maple gives the required result. □

Note that a non-obvious rearrangement (verified by Maple) for $E(N(y))$ is

$$\frac{1-y}{y}\int_y^1 \frac{1}{(1-t)(1-h(t))}\exp\left(\int_y^t \frac{-b(x)}{x(1-x)(1-h(x))}\mathrm{d}x\right)\mathrm{d}t.$$

By differentiating it with respect to y, it reduces to the differential equation,

$$y(1-y)(1-h(y))(E(N(y)))' + w(y)E(N(y)) + 1 - y = 0$$

with initial condition $E(N(1)) = 0$, shown independently in [3].

Also if the BAS has absorbing states other than the global optimum, the moments can be infinite.

4.2 Examples

We give some explicit examples. The use of Maple is acknowledged. The first relates to PAS.

Example 4.1 For $b(x) = \gamma$ and $w(x) = 0$,

$$G^{\mathrm{succ}}(y,z) = \frac{1}{1-z}y^{\frac{1-z}{1-z+z\gamma}},$$
$$G^{\mathrm{fm}}(y,z) = y^{\frac{1-z}{1-z+z\gamma}},$$
$$E(N(y)) = -\frac{\ln y}{\gamma}.$$

Another simple version of HAS.

Example 4.2 For $b(x) = \beta x$ and $w(x) = 0$,

$$G^{\text{cdf}}(y,z) = \frac{y(z\beta + 1 - z)}{(1 - z + z\beta y)(1 - z)},$$
$$G^{\text{pdf}}(y,z) = \frac{z\beta + 1 - z}{(1 - z + z\beta y)^2},$$
$$G^{\text{succ}}(y,z) = \frac{y(z\beta + 1 - z)}{(1 - z + z\beta y)(1 - z)},$$
$$G^{\text{fm}}(y,z) = \frac{y(z\beta + 1 - z)}{1 - z + z\beta y},$$
$$E(N(y)) = \frac{1 - y}{\beta y}.$$

The next specific example of backtracking has very complicated generating functions, but the expected number of iterations is simply expressed.

Example 4.3 For $b(x) = 1/2$ and $w(x) = 1/2$,

$$E(N(y)) = \frac{\sqrt{-y+1}\left(\pi - 2\arcsin\left(\sqrt{y}\right)\right)}{\sqrt{y}}.$$

Although the explicit forms of the various generating functions have been given, we now give two differential equations that they solve. They can be verified by substituting the given functions into the stated differential equations.

Lemma 4.1 *Let G' denote the partial derivative of G with respect to y, $p = (y(1-y)(1-zh(y)))$, $q = zw(y) + (1-z)(1-y)$ and $r = -zw(y)$. Then*

(1) *$G = G^{\text{fm}}(y,z)$ satisfies the differential equation $pG' = qG + rG^2$ with initial condition $G(1,z) = 1$, and*
(2) *$G = G^{\text{succ}}(y,z)$ satisfies the differential equation $pG' = qG + r(1-z)G^2$ with initial condition $G(1,z) = 1/(1-z)$.*

For HAS where $h = 1 - b$, the differential equation for G^{fm} reduces to the one in [5].

An unexpected discovery came from looking at the success probabilities of a number of examples. Any algorithm can easily be modified to accept only improving points (hence forcing $h(x) = 1 - b(x)$.) We found that sometimes this improves the algorithm, sometimes it has no effect and sometimes it even degrades performance. We explore this further in Sect. 5. The following consequence of our results shows that BAS with $b(x)/x$ constant has performance independent of hesitation and worsening.

Proposition 4.2 *For $b(x) = \beta x$ and any h and w, $G^{\text{succ}}(y,z)$ and $G^{\text{fm}}(y,z)$ are as in Example 4.2.*

Proof Substitute G^{succ} in the differential equation of Lemma 4.1. □

Note, however, that G^{pdf} and G^{cdf} do depend on h and w.

A special case of the above example relates to Pure Random Search (PRS) where each iteration is chosen independently, and thus $b(x) = x$. In this case

$$G^{\text{succ}}(y,z) = \frac{y}{(1 - z + zy)(1 - z)}.$$

So the probability of success $P(\min\{Y_0, Y_1, \ldots, Y_k\} \leq y) = 1 - (1-y)^k$ from kth term of the generating function. As noted the cdf depends on h and w. PRS as usually implemented, simply accepts each iteration, thus $w(x) = 1 - x$. In this case $G^{\text{cdf}}(y, z) = y/1 - z$ and gives cdfs of $P(Y_k \leq y) = y$ at each iteration. However, by never accepting worsening iterations, PRS can be implemented as HAS with $h(x) = 1 - x$. For any HAS, $G^{\text{cdf}}(y, z) = G^{\text{succ}}(y, z)$, so in this case gives cdf of $1 - (1-y)^k$ for iteration k. This illustration reinforces that the probability of success is the useful measure of performance.

The following examples show that in certain cases it is better to avoid hesitating. It compares the two extremes of BAS. This first is an HAS.

Example 4.4 For $b(x) = x^2$ and $w(x) = 0$,

$$G^{\text{succ}}(y, z) = \frac{y/(1-z)}{\sqrt{1 - z + zy^2}} \quad \text{and} \quad E(N(y)) = \frac{1 - y^2}{2y^2}.$$

The first few values of success probabilities $P(\min\{Y_0, Y_1, \ldots, Y_k\} \leq y)$ are: y, $-1/2\left(y^2 - 3\right)y, 1/8\left(3y^4 - 10y^2 + 15\right)y$, and $-1/16\left(5y^6 - 21y^4 + 35y^2 - 35\right)y$.

The following has the same bettering function, but never hesitates.

Example 4.5 For $b(x) = x^2$ and $w(x) = 1 - x^2$,

$$G^{\text{succ}}(y, z) = \frac{zy/(1-z)}{zy - ze^{z(1-y)} + z + e^{z(1-y)} - 1} \quad \text{and} \quad E(N(y)) = \frac{1}{y}(e^{1-y} - 1).$$

The first few values of success probabilities $P(\min\{Y_0, Y_1, \ldots, Y_k\} \leq y)$ are: y, $-1/2\left(y^2 - 3\right)y, 1/12\left(3y^4 + 2y^3 - 12y^2 - 6y + 25\right)y$,
and $-1/24\left(3y^6 + 4y^5 - 14y^4 - 20y^3 + 35y^2 + 32y - 64\right)y$.

Figure 1 supports that the second example dominates the first. In the next section, we prove this is always the case for a bettering, which has $b(x)/x$ increasing.

There is some interest in the case, where there is no hesitation (i.e. $h(y) = 0$) since if such a process is understood, it can be "stopped" to explore the more general BAS with non-zero hesitation. The various theorems and propositions become slightly simpler when $h(x) = 0$ and $w(x) = 1 - b(x)$.

Applied to global optimisation, these results provide a standardized measure of performance. The graph of success versus number of iterations, however, is still flawed, in that iterations are not directly comparable. One algorithm's iteration may be more costly than that of another. A method can be slowed down to compensate for a cost, and the expected number of iterations changes proportionally (although the other moments do not). We have,

Proposition 4.3 *Let $c > 1$. Let $Y_k^{(c)}$ be a new process using w/c and b/c for the worsening and bettering functions. Then $E(N^{(c)}(y)) = cE(N(y))$.*

Proof Replace $w(x)$ and $b(x)$ by $w(x)/c$ and $b(x)/c$ in the formula for $E(N(y))$ in Proposition 4.1. □

Empirically, we observe that often, for each level of success, $N^{(c)}(y)$ is approximately c times $N(y)$. It is easy, however, to create examples that show this is only an approximation.

Empirical tests led to the discovery of Proposition 4.2. Those tests support a further conjecture: that two hesitant adaptive searches have similar performance if the slope of the bettering is the same at zero. This is shown in Figs. 2 and 3.

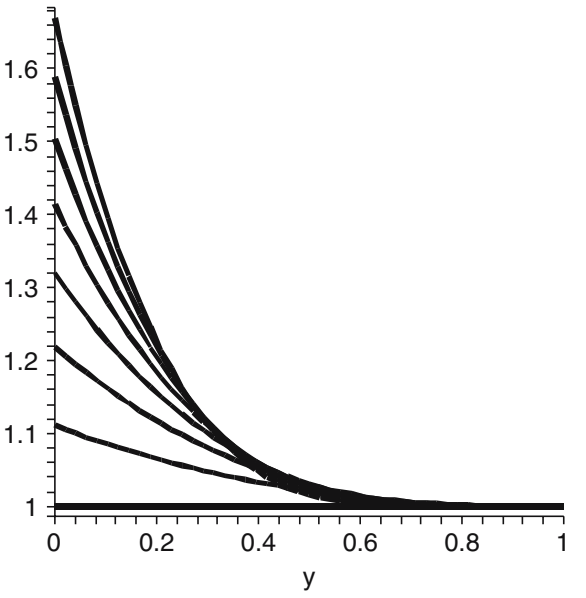

Fig. 1 Ratios of success probabilities of example 4.5 compared to example 4.4 (for $k = 1, 2, \ldots, 9$ from bottom to top)

5 A general performance result

In this section, we imagine the bettering function b to be fixed, and investigate how the allocation of the remaining probability $1-b$ between the hesitation and worsening functions h and w affects performance. We saw in Proposition 4.2 that a BAS with bettering such that $b(x)/x$ is constant (i.e., $b(x) = \beta x$) has performance independent of w and h. Ordinarily one imagines that it is advantageous to hesitate rather than to worsen. In fact, this partly depends on the function $b(x)/x$: it turns out that if $b(x)/x$ is decreasing, then hesitation is preferable, but if $b(x)/x$ is increasing, then "worsening" becomes preferable to hesitation.

The following relates the success probabilities of BAS methods with the same bettering.

Theorem 5.1 *Let* $\mathrm{BAS}(b, h, w)$ *denote the Markov chain in the range* $[0, 1]$ *resulting from the standardised BAS algorithm with probability functions* b, h *and* w *defined on* $[0, 1]$. *Let* $N_h(y)$, $N(y)$ *and* $N_w(y)$ *be the numbers of iterations required respectively by* $\mathrm{BAS}(b, h + w, 0)$, $\mathrm{BAS}(b, h, w)$ *and* $\mathrm{BAS}(b, 0, h + w)$ *to have sampled a value less than or equal to* y. *For each iteration* k *as follows:*

(1) *if* $b(x)/x$ *is decreasing, then hesitation is preferable to worsening, in that*

$$P(N_h(y) \leq k) \geq P(N(y) \leq k) \geq P(N_w(y) \leq k);$$

(2) *if* $b(x)/x$ *is constant, then*

$$P(N_h(y) \leq k) = P(N(y) \leq k) = P(N_w(y) \leq k);$$

(3) *if* $b(x)/x$ *is increasing, then worsening is preferable to hesitation, in that*

$$P(N_h(y) \leq k) \leq P(N(y) \leq k) \leq P(N_w(y) \leq k).$$

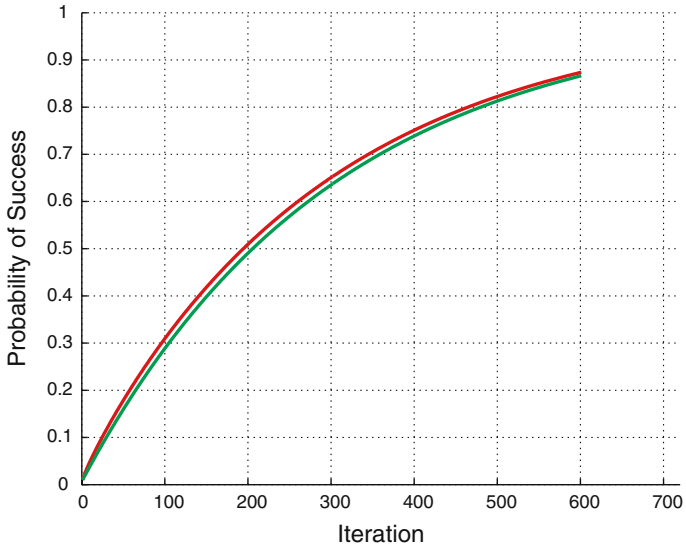

Fig. 2 Similar performance of two HASs (the upper with $b(x) = xe^x/3$ and the lower with $b(x) = x/3$)

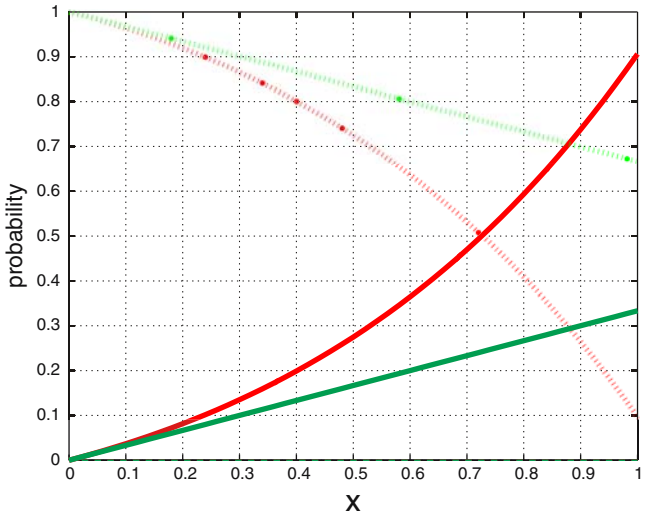

Fig. 3 The associated betterings functions are tangent at $x = 0$ (solid lines give bettering functions, dotted lines give worsening functions)

Note that $P(N(y) \leq k)$ is just the coefficient of z^k in G^{succ} for the corresponding BAS process.

Proof Part (2) is simply a restatement of Proposition 4.2. Part (1) can be established by induction in a straightforwards manner, although an analogue of the more complicated argument we present for (3) will also work.

Now consider the second inequality of (3). For each $n \in \{0, 1, 2, \ldots\}$, let $(Y_k^{(n)})$ be an inhomogeneous Markov chain with Y_0 distributed uniformly on $[0, 1]$, and such that

the n transitions from Y_0 to Y_n are according to the kernel BAS$(b, 0, h + w)$, and all subsequent transitions are according to the kernel BAS(b, h, w). Now $(Y_k^{(0)})$ is simply BAS(b, h, w), whereas in the limit as $n \to \infty$, $(Y_k^{(n)})$ becomes BAS$(b, 0, h + w)$. For each $y \in [0, 1]$ and $n \in \{0, 1, 2, \ldots, \}$, let $N^{(n)}(y)$ be the random variable

$$\min\{k : Y_k^{(n)} \leq y\}$$

so that $N^{(0)}(y) = N(y)$ and $N^{(k)}(y) \to N_w(y)$ as $k \to \infty$. We will demonstrate that

$$P\left[N^{(n+1)}(y) \leq k\right] \geq P\left[N^{(n)}(y) \leq k\right], \tag{5}$$

implying the second inequality of (3).

The chains $(Y_k^{(n)})$ and $(Y_k^{(n+1)})$ are identical up to iteration n, inclusively, so that (5) is trivial if $k \leq n$. Thus assume $k > n$. Note, that $P[N^{(n)}(y) \leq n]$ and $P[N^{(n+1)}(y) \leq n]$ are equal; denote this shared probability by π. Further, the distributions of $Y_n^{(n)}$ conditioned on $N^{(n)}(y) > n$ and of $Y_n^{(n+1)}$ conditioned on $N^{(n+1)}(y) > n$ are identical; denote this shared distribution by F.

Now $P[N^{(n+1)}(y) \leq k]$

$$\begin{aligned}
&= \pi + (1 - \pi)P\left(N^{(n+1)}(y) \leq k \,\middle|\, N^{(n+1)}(y) > n\right) \\
&= \pi + (1 - \pi)\int_{t\in(y,1]} P\left(N^{(n+1)}(y) \leq k \,\middle|\, N^{(n+1)}(y) > n \text{ and } Y_n^{(n+1)} = t\right) \mathrm{d}F(t) \\
&= \pi + (1 - \pi)\int_{t\in(y,1]} \\
&\quad \left((h(t) + w(t))P\left(N^{(n+1)}(y) \leq k \,\middle|\, N^{(n+1)}(y) > n \text{ and } Y_{n+1}^{(n+1)} > Y_n^{(n+1)} = t\right) \right. \\
&\quad \left. + \; b(t)P\left(N^{(n+1)}(y) \leq k \,\middle|\, N^{(n+1)}(y) > n \text{ and } Y_{n+1}^{(n+1)} < Y_n^{(n+1)} = t\right) \right) \mathrm{d}F(t).
\end{aligned}$$

Letting (Y_k) denote BAS(b, h, w), we can rewrite the above as

$$\begin{aligned}
&\pi + (1 - \pi)\int_{t\in(y,1]} \left((h(t) + w(t))P\left(N(y) \leq k \,\middle|\, N(y) > n \text{ and } Y_{n+1} > Y_n = t\right) \right. \\
&\quad \left. + \; b(t)P\left(N(y) \leq k \,\middle|\, N(y) > n \text{ and } Y_{n+1} < Y_n = t\right) \right) \mathrm{d}F(t)
\end{aligned}$$

because the chains (Y_k) and $(Y_k^{(n+1)})$ have identical kernels for all transitions from iteration $n + 1$ onwards. Similarly $P[N^{(n)}(y) \leq k]$ is

$$\begin{aligned}
&\pi + (1 - \pi)\int_{t\in(y,1]} \left(h(t)P\left(N(y) \leq k \,\middle|\, N(y) > n \text{ and } Y_{n+1} = Y_n = t\right) \right. \\
&\quad + w(t)P\left(N(y) \leq k \,\middle|\, N(y) > n \text{ and } Y_{n+1} > Y_n = t\right) \\
&\quad \left. + \; b(t)P\left(N(y) \leq k \,\middle|\, N(y) > n \text{ and } Y_{n+1} < Y_n = t\right) \right) \mathrm{d}F(t).
\end{aligned}$$

Thus $P[N^{(n+1)}(y) \le k] - P[N^{(n)}(y) \le k]$

$$= (1-\pi)\int_{t\in(y,1]} h(t)\Big(P\left(N(y)\le k \,\middle|\, N(y) > n \text{ and } Y_{n+1} > Y_n = t\right) - P\left(N(y)\le k \,\middle|\, N(y) > n \text{ and } Y_{n+1} = Y_n = t\right)\Big)\,\mathrm{d}F(t).$$

Therefore (5) follows if, we can show that

$$P\left(N(y)\le k \,\middle|\, N(y) > n \text{ and } Y_{n+1} > Y_n = t\right) \ge P\left(N(y)\le k \,\middle|\, N(y) > n \text{ and } Y_{n+1} = Y_n = t\right)$$

whenever $k > n$ and $t > y$. Since, (Y_k) is homogeneous, this is equivalent to

$$P\left(N(y)\le \ell \,|\, Y_0 > t\right) \ge P\left(N(y)\le \ell \,|\, Y_0 = t\right)$$

for $\ell \ge 0$. Let $N(t-)$ denote $\min\{k\colon\ Y_k < t\}$. With $Y_o \ge t$, the value $Y_{N(t-)}$ is independent of Y_0, $N(t-)$ and $Y_{N(t-)-1}$; straightforwards calculation shows that

$$P(Y_k \le y_k | Y_0 = y_0 \text{ and } N(t-) = k \text{ and } Y_{k-1} = y_{k-1}) = y_k/t$$

whenever $y_k < t \le y_{k-1}$. Thus, by the Markov property, the sequence $(Y_{N(t-)}, Y_{N(t-)+1}, \ldots)$ is independent of Y_0 and $N(t-)$, and in particular $N(y) - N(t-)$ is independent of Y_0 and $N(t-)$. Thus, it suffices to show that

$$P\left(N(t-) > \ell \,|\, Y_0 > t\right) \le P\left(N(t-) > \ell \,|\, Y_0 = t\right) \tag{6}$$

(we have reversed the order of inequality for convenience).

For each k, let $\pi_k = P\left(Y_k \ge t \,\middle|\, Y_0 > t \text{ and } Y_1, \ldots, Y_{k-1} \ge t\right)$. By the definition of BAS and the fact that $b(t)/t$ is increasing, $\pi_k \le 1 - b(t)$ for each k. Now

$$P\left(N(t-) > \ell \,|\, Y_0 > t\right) = \pi_1\pi_2\cdots\pi_\ell.$$

On the other hand, $P\left(N(t-) > \ell \,|\, Y_0 = t\right)$

$$\begin{aligned}
&= P\left(Y_\ell = \cdots = Y_1 = t \,|\, Y_0 = t\right) \\
&\quad + \sum_{k=1}^{\ell} P\left(Y_{k-1} = \cdots = Y_1 = t \text{ and } Y_k > t \text{ and } Y_{k+1}, \ldots, Y_\ell \ge t \,\middle|\, Y_0 = t\right) \\
&= (1-b(t)-w(t))^\ell + \sum_{k=1}^{\ell} w(t)(1-b(t)-w(t))^{k-1}\pi_1\cdots\pi_{\ell-k} \\
&\ge \pi_1\cdots\pi_\ell\left(\left(\frac{1-b(t)-w(t)}{1-b(t)}\right)^\ell + \frac{w(t)}{1-b(t)}\sum_{j=0}^{\ell-1}\left(\frac{1-b(t)-w(t)}{1-b(t)}\right)^j\right) \\
&= \pi_1\cdots\pi_\ell,
\end{aligned}$$

implying (6) and thence (5) and the second inequality of (3).

The first inequality of (3) can be established by a similar argument, in which the family of inhomogeneous chains $(Y_k^{(n)})$ is defined to tend to BAS$(b, h+w, 0)$ rather than BAS$(b, 0, h+w)$. □

6 Summary

An alternative analysis of (standardized) backtracking adaptive search (Sect. 2) has been presented. Generating functions (Sect. 3) have been established for three quantities of interest: the distribution of the objective function value at the kth iteration, the distribution of the best value at the kth iteration and the number of iterations to convergence. Examples, to illustrate the power of the methods have been presented (Sect. 4) and a general result given (Sect. 5) relating algorithm improvement propensity to the need to backtrack.

Acknowledgements The authors would like to thank the Marsden Fund of the Royal Society of New Zealand for support of this research.

References

1. Bulger, D.W., Wood, G.R.: Hesitant adaptive search for global optimisation. Math. Program. **79**, 89–102 (1998)
2. Bulger, D.W., Baritompa, W.P., Wood, G.R.: Implementing Pure Adaptive Search with Grover's Quantum Algorithm. J. Optim. Theory Appl. **116**, 517–529 (2003)
3. Bulger, D.W., Alexander, D.L.J., Baritompa, W.P., Wood, G.R., Zabinsky, Z.B.: Expected Hitting Times for Backtracking Adaptive Search. Massey University Technical Report, IIST (2002)
4. Reaume, D.J., Romeijn, H.E., Smith, R.L.: Implementing pure adaptive search for global optimization using Markov chain sampling. J. Global Optim. **20**(1), 33–47 (2001)
5. Wood, G.R., Zabinsky, Z.B., Kristinsdottir, B.P.: Hesitant adaptive search: the distribution of the number of iterations to convergence. Math. Program. **89**, 479–486 (2001)
6. Wood, G.R., Alexander, D.L.J., Bulger, D.W.: Approximation of the distribution of convergence times for stochastic global optimisation. J. Global Optim. **22**(1), 271–284 (2002)
7. Wood, G.R., Bulger, D.W., Baritompa, W.P., Alexander, D.L.J.: Backtracking adaptive search: the distribution of the number of iterations to convergence. J. Optim. Theory Appl. **128** (2006), to appear.
8. Zabinsky, Z.B., Smith, R.L.: Pure adaptive search in global optimization. Math. Program. **53**, 323–338 (1992)