Combinatorics, Probability and Computing (1997) 6, 205–229
Copyright © 1997 Cambridge University Press

Stochastic Analysis of Convergence via Dynamic Representation for a Class of Line-search Algorithms

L. PRONZATO[1], H. P. WYNN[2] and A. A. ZHIGLJAVSKY[3]

[1] Laboratoire I3S, CNRS-URA 1376, 250 rue A. Einstein, bât. 4, Sophia Antipolis, 06560 Valbonne, France

[2] Department of Statistics, University of Warwick, Coventry CV4 7AL, UK

[3] Department of Mathematics, St. Petersburg University, Bibliotechnaya sq.2, St. Petersburg, 198904 Russia

Received 7 April 1995; revised 16 October 1995

Certain convergent search algorithms can be turned into chaotic dynamic systems by renormalisation back to a standard region at each iteration. This allows the machinery of ergodic theory to be used for a new probabilistic analysis of their behaviour. Rates of convergence can be redefined in terms of various entropies and ergodic characteristics (Kolmogorov and Rényi entropies and Lyapunov exponent). A special class of line-search algorithms, which contains the Golden-Section algorithm, is studied in detail. Their associated dynamic systems exhibit a Markov partition property, from which invariant measures and ergodic characteristics can be computed. A case is made that the Rényi entropy is the most appropriate convergence criterion in this environment.

1. Introduction

Random or stochastic methods in optimisation or search are often used as a way of generating alternatives at each iteration. Here the approach is different. It is shown that, for certain classes of algorithms, randomness in the ergodic sense is already embedded in the algorithm. Thus, algorithms which are normally considered as simply convergent can be shown to reveal ergodic behaviour. In this paper, certain line-search algorithms are studied.

The method is based on a simple renormalisation idea. Consider an algorithm which produces a series of nested sets

$$S_0 \supset S_1 \supset S_2 \supset \ldots$$

with the target x^* lying in each set and convergence consisting of the norm, in some sense, of S_n tending to zero as n goes to infinity. At each iteration renormalise S_n back to S_0 using, hopefully, some simple transformation such as a linear shift and rescaling. Then this renormalisation induces a dynamic system on the renormalised location of x^*. If

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

$g_n(S_n) = S_0$ is the renormalisation and $x_n = g_n(x^*) \in S_0$, then this system is given by the mapping $h_n : x_n \to x_{n+1}$.

In a series of papers [6, 10], the authors have used the theory of dynamic systems to study the asymptotic properties of a number of different line-search algorithms for finding the root or the minimum of a function. It was found that, with the additional property of symmetry of the function, the dynamic system, produced by renormalising the search interval, is time homogeneous. In such cases, the invariant measures of the process can yield new and valuable properties of the original algorithm. The properties can be summarised via the Lyapunov exponent and Kolmogorov entropy, or other types of entropy. These can then be related to rates of convergence of the algorithms. This amounts to a new kind of stochastic analysis of algorithms, which allow us to discover new algorithms with improved rates, even when the symmetry condition is relaxed to local symmetry.

In the line-search algorithms to which this paper is confined, the dynamic systems exhibit a Markov partition property which makes the computation of invariant measures reasonably tractable. A key suggestion of this paper is that Rényi entropy is a natural alternative to Kolmogorov entropy for algorithms based on contracting sets. This is because it reflects more closely the size of the set S_n (in the line-search case an interval). The technology of the paper, then, is to exhibit a class of line-search algorithms, describe the Markov process revealed after renormalisation, compute invariant measures and discuss competing rates of convergence, from both an ergodic and average point of view.

The starting point for both the present paper, and indeed, the entire research programme, is the Golden-Section and Fibonacci algorithms of Kiefer [4]. In [11] it was shown that under certain conditions the asymptotic rate of convergence of the Golden-Section algorithm can be improved from $\lambda = (\sqrt{5}-1)/2$ down to $\frac{1}{2}$ (see also Example 7 below) for almost all x^* in the initial search interval.

A second order method [4] is defined as follows. Suppose $f(x)$ is a uniextremal function on $[0, 1]$ with a minimum at an unknown point x^* in $[0, 1]$. Then $f(x)$ is decreasing (non-increasing) for $x \leqslant x^*$ and increasing (non-decreasing) for $x > x^*$. If we evaluate $f(x)$ at two points a and b in $[0, 1], 0 < a < b < 1$, then if

$$(L): \quad f(a) \geqslant f(b)$$

we delete $[0, a)$, and if

$$(R): \quad f(a) < f(b)$$

delete $(b, 1]$. Here (L) and (R) stand for left and right deletion. In each case one 'observation' point is carried forward for the next iteration, b for (L) and a for (R), a new observation point added and the process repeated.

The Golden Section (GS) is the case when $b = 1-a = \lambda = (\sqrt{5}-1)/2 \approx 0.6180$ and the lengths of subintervals stay in these ratios after successive iterations. Kiefer showed [4] that the finite sample size algorithm based on the use of Fibonacci numbers was worst-case optimal, in that it has the fastest rate of convergence for the worst function for fixed sample size among all algorithms using only function evaluations. The GS algorithm has this property asymptotically as the sample size goes to infinity.

Section 2 presents the basic idea of renormalisation, and introduces the dynamic system

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

associated with a line-search algorithm. A special class of algorithms, containing the GS algorithm, is described in Section 3 with some illustrative examples. The associated dynamic process is presented in Section 4, and the ergodic characteristics are considered in Section 5. Section 6 is devoted to some examples. The case of locally symmetric functions is briefly considered in Section 7. Section 8 gives a short conclusion.

2. Renormalisation

The key to the theory is to renormalise the interval, after (L) or (R) deletion, back to $[0, 1]$. The following transformation achieves this:

$$h(f, a, b, x) = \begin{cases} \dfrac{x-a}{1-a} & \text{if } (L) \\ \dfrac{x}{b} & \text{if } (R). \end{cases}$$

Under (L), $h(f, a, b, \cdot)$ maps $[a, 1]$ to $[0, 1]$ and under (R) it maps $[0, b]$ to $[0, 1]$. We can then always define the algorithm *after renormalisation* by function evaluation at a_n, b_n in $[0, 1]$ with either a_n or b_n being carried forward. If e_n is the carried forward point (from the previous iteration), we are at liberty to take a new observation at any point in $[0, 1]$ say e'_n, so that

$$a_n = \min(e_n, e'_n), \quad b_n = \max(e_n, e'_n).$$

The new iteration is then with $a = a_n$, $b = b_n$ and the renormalisation h above.

Now under renormalisation x^*, the minimiser, is continuously in the move, and we can track its path.

If at iteration n x^* is at x_n, then at iteration $n+1$ after renormalisation it is at

$$x_{n+1} = \begin{cases} \dfrac{x_n-a_n}{1-a_n} & \text{if } (L) \\ \dfrac{x_n}{b_n} & \text{if } (R). \end{cases}$$

At this point whether (L) or (R) is satisfied depends upon the function $f(\cdot)$ (which of course should be renormalised). This makes the study of the dynamic process in general extremely difficult. However, if $f(\cdot)$ is assumed to be symmetric about x^* (about x_n in the renormalised version), whether (L) or (R) is satisfied depends only upon the location of x^*, since $f(a_n) \geqslant f(b_n)$ is equivalent to $x_n \geqslant c_n$, with $c_n = (a_n + b_n)/2$. We thus obtain

$$x_{n+1} = h(a_n, b_n, x_n) = \begin{cases} \dfrac{x_n-a_n}{1-a_n} & \text{if } x_n \geqslant c_n, \\ \dfrac{x_n}{b_n} & \text{if } x_n < c_n. \end{cases} \tag{1}$$

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

Under (L) (that is, $x_n \geqslant c_n$) and (R) (that is, $x_n < c_n$), the point carried forward is b_n or a_n, respectively, which is transformed to

$$e_{n+1} = \begin{cases} \dfrac{b_n - a_n}{1 - a_n} & \text{if } x_n \geqslant c_n, \\ \dfrac{a_n}{b_n} & \text{if } x_n < c_n. \end{cases} \tag{2}$$

A new point e'_{n+1} is then chosen and the iteration is repeated with

$$(a_{n+1}, b_{n+1}) = (\min(e_{n+1}, e'_{n+1}), \max(e_{n+1}, e'_{n+1})).$$

We finally obtain a time-homogeneous dynamic system $(e_n, x_n) \to (e_{n+1}, x_{n+1})$. Given a_1, b_1 and the selection rule for $e'_n, n > 1$, the behaviour of the system is determined only by $x_1 = x^*$.

For the GS algorithm, the iteration is given by:

$$(e_{n+1}, x_{n+1}) = \begin{cases} (1-\lambda, (1+\lambda)x_n - \lambda) & \text{if } x_n \geqslant \frac{1}{2}, \\ (\lambda, (1+\lambda)x_n) & \text{if } x_n < \frac{1}{2}, \end{cases} \tag{3}$$

so that $e_n \in \{\lambda, 1-\lambda\}, \forall n$. As dynamic processes, the Golden Section and the new processes we are about to describe, have ergodic properties, and in particular invariant measures. We postpone this discussion to Section 5, where we also derive improved rates of convergence.

3. New algorithms and cycles

The new algorithms will share with the GS algorithm the property that we stay on a fixed collection of points as the algorithm proceeds, and in this sense they generalise the GS algorithm. The new point e'_{n+1} can be considered as a *control*, applied when the carried forward point is e_{n+1}, which forces the sequence to remain within the fixed set of points. The reader is invited to consider the algorithm described in Example 2 below, which has two pairs (e_n, e'_n) given by $(1/2, 3/4)$ and $(2/3, 1/3)$.

Definition 1. *The **state** of the process at iteration n is defined by the pair* $S_n = (e_n, e'_n)$.

For example, in the case of the GS algorithm there are only two states: $(1-\lambda, \lambda)$ and $(\lambda, 1-\lambda)$. For any state $S = (e, e')$, we define the transformation

$$L(S) = \frac{b-a}{1-a}, \quad R(S) = \frac{a}{b}, \tag{4}$$

with $a = \min(e, e'), b = \max(e, e')$. At iteration n, the point carried forward (2) for the process is thus with this new notation $e_{n+1} = L(S_n)$ or $R(S_n)$.

Definition 2. *A set of numbers* $\mathcal{Q} = \{e^{(j)}\}_{j=1}^N, 0 < e^{(1)} \leqslant e^{(2)} \leqslant \ldots \leqslant e^{(N)} < 1$ *will be called **section invariant** if there exists a set of controls* $\{e'^{(j)}\}_{j=1}^N$ *such that, for all states*

$$S^{(j)} = (e^{(j)}, e'^{(j)}),\ L(S^{(j)}) \in \mathcal{Q} \text{ and } R(S^{(j)}) \in \mathcal{Q}.$$

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

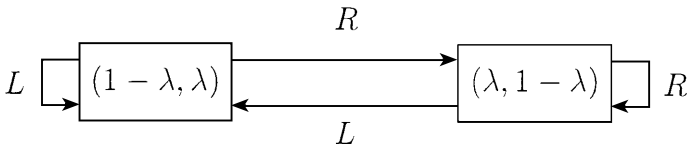


Figure 1 The directed graph of the GS agorithm.

It is important to note that the same value of $e^{(i)}$ may correspond to several states $S^{(j)}$, $S^{(k)}$, since different controls $e'^{(j)}, e'^{(k)}$ may be applied (see, e.g., Example 7). Also, the controls $e'^{(1)}, \ldots, e'^{(N)}$ are not required to be in $\mathcal{Q}$. This gives more freedom and generates some new processes even in the case $N = 2$ (see, e.g. Example 2 below).

Definition 3. *Let $\mathcal{Q} = \{e^{(j)}\}_{j=1}^{N}$ be a section-invariant set. A set of states*

$$\mathcal{S} = \{S^{(j)}\}_{j=1}^{N} = \{e^{(j)}, e'^{(j)}\}_{j=1}^{N}$$

such that $L(S^{(j)}) \in \mathcal{Q}$ and $R(S^{(j)}) \in \mathcal{Q}$ for all j will be called a ***section invariant process.***

Consider a section-invariant process $\mathcal{S} = \{S^{(j)}\}_{j=1}^{N}$ defined on a section invariant set $\mathcal{Q} = \{e^{(j)}\}_{j=1}^{N}$. For any $j \in \{1, \ldots, N\}$ we have $L(S^{(j)}) = e^{(\mathcal{L}(j))} \in \mathcal{Q}$ and $R(S^{(j)}) = e^{(\mathcal{R}(j))} \in \mathcal{Q}$, with $\mathcal{L}(j)$ and $\mathcal{R}(j) \in \{1, \ldots, N\}$ denoting the index of the destination state. The process thus defines a directed graph with N vertices. Each vertex j is the join of two edges directed from j to $\mathcal{L}(j)$ and $\mathcal{R}(j)$, possibly with a loop if either $\mathcal{L}(j)$ or $\mathcal{R}(j)$ equals j. Figure 1 presents the graph associated with the GS algorithm.

Definition 4. *A section invariant process $\mathcal{S} = \{S^{(j)}\}_{j=1}^{N}$ will be called an* ***N-state cycle*** *if no subset of $\mathcal{S}$ defines a section-invariant process.*

Note that each state of any N-state cycle is reachable from some other state, that is

$$e^{(i)} \in \left(\bigcup_{\substack{j=1 \\ j \neq i}}^{N} L(S^{(j)}) \right) \cup \left(\bigcup_{\substack{j=1 \\ j \neq i}}^{N} R(S^{(j)}) \right), \quad \forall i \in \{1, \ldots, N\},$$

and the graph associated with the cycle is connected, (see Figure 1 and Figure 2 in Example 4). Let us give some examples of cycles.

For the case $N = 2$ the general tableau is

e	e'	L	R
$e^{(1)}$	$e'^{(1)}$	$L(S^{(1)})$	$R(S^{(1)})$
$e^{(2)}$	$e'^{(2)}$	$L(S^{(2)})$	$R(S^{(2)})$

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

with $L(S^{(1)})$, $R(S^{(1)})$, $L(S^{(2)})$ and $R(S^{(2)}) \in \{e^{(1)}, e^{(2)}\}$. Examples 1–3 provide all two-state cycles (case $N = 2$), up to reflection around 1/2.

Example 1. (Golden Section).

e	e'	L	R
$1-\lambda$	λ	$1-\lambda$	λ
λ	$1-\lambda$	$1-\lambda$	λ

Example 2.

e	e'	L	R
1/2	3/4	1/2	2/3
2/3	1/3	1/2	1/2

Example 3.

e	e'	L	R
$e^{(1)}$	$e'^{(1)}$	$e^{(2)}$	$e^{(2)}$
$e^{(2)}$	$e'^{(2)}$	$e^{(1)}$	$e^{(1)}$

with $e^{(2)} = \psi$, where

$$\psi = \frac{1}{3}\left(\left(\frac{3\sqrt{69}+11}{2}\right)^{\frac{1}{3}} - 5\left(\frac{2}{3\sqrt{69}+11}\right)^{\frac{1}{3}} + 1\right) \approx 0.5698 \tag{5}$$

is a solution of the equation $(1-t)^2 = t^3$,

$$e^{(1)} = 1 - e^{(2)} \approx 0.4302, \quad e'^{(1)} = 1 - \psi + \psi^2 \approx 0.7549, \quad e'^{(2)} = 1 - e'^{(1)} \approx 0.2451.$$

The situation for $N > 2$ is more complex because of the range of different configurations available under (L) or (R). The authors have found all cycles for $N = 3$ by simple computer

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

enumeration. The general tableau is

e	e'	L	R
$e^{(1)}$	$e'^{(1)}$	$L(S^{(1)})$	$R(S^{(1)})$
$e^{(2)}$	$e'^{(2)}$	$L(S^{(2)})$	$R(S^{(2)})$
$e^{(3)}$	$e'^{(3)}$	$L(S^{(3)})$	$R(S^{(3)})$

Any given structure, that is when $\mathscr{L}(j)$ and $\mathscr{R}(j)$ are set to fixed elements in $\{1,2,3\}$, gives six polynomial equations of first or second degree with six unknowns (the elements of $\mathscr{Q} = \{e^{(1)}, e^{(2)}, e^{(3)}\}$ and the controls $e'^{(1)}, e'^{(2)}, e'^{(3)}$). Elimination of unknowns leads to one polynomial equation in one variable. This equation is typically of degree one, two or three. Its maximum degree, seven, is obtained when $\mathscr{L}(1) = \mathscr{R}(1) = 2$, $\mathscr{L}(2) = \mathscr{R}(2) = 3$ and $\mathscr{L}(3) = \mathscr{R}(3) = 1$. Since $e^{(1)} \leqslant e^{(2)} \leqslant e^{(3)}$, we have $e'^{(1)} > e^{(1)}$ and $e'^{(3)} < e^{(3)}$. Because of symmetry with respect to $1/2$, we restrict ourselves to the case $e'^{(2)} > e^{(2)}$. These conditions imply $\mathscr{R}(1) \in \{2,3\}, \mathscr{R}(2) = 3, \mathscr{L}(3) \in \{1,2\}$, and $\mathscr{L}(1), \mathscr{L}(2), \mathscr{R}(3) \in \{1,2,3\}$, which gives 108 possible cases. Reachability of $e^{(1)}$ (resp. of $e^{(2)}$) implies that 1 (resp. 2) should appear at least once in $\{\mathscr{L}(2), \mathscr{L}(3), \mathscr{R}(3)\}$ (resp. in $\{\mathscr{L}(1), \mathscr{R}(1), \mathscr{L}(3), \mathscr{R}(3)\}$). These conditions remove, respectively, 24 and 12 cases. Among the remaining 72 structures, 4 coincide with Example 1, 12 with Example 2 and 2 with Example 3 (which means that for 18 structures $e^{(1)} = e^{(2)}$ or $e^{(2)} = e^{(3)}$). Additionally, six structures yield $e^{(2)} < e^{(1)}$, and correspond to other admissible structures by permutation of indices. Finally, one structure has no solution, it corresponds to $\mathscr{L}(1) = 1, \mathscr{R}(1) = 2, \mathscr{L}(2) = 1, \mathscr{R}(2) = 3, \mathscr{L}(3) = 2, \mathscr{R}(3) = 1$. We thus end up with the 47 three-state cycles presented in the table in the appendix. The numerical values of $e^{(1)}, e^{(2)}, e^{(3)}, e'^{(1)}, e'^{(2)}, e'^{(3)}$ are obtained as roots of polynomials, and we work through one example to illustrate the construction of this table.

Example 4. The tableau in this case is

e	e'	L	R
$e^{(1)}$	$e'^{(1)}$	$e^{(2)}$	$e^{(2)}$
$e^{(2)}$	$e'^{(2)}$	$e^{(1)}$	$e^{(3)}$
$e^{(3)}$	$e'^{(3)}$	$e^{(2)}$	$e^{(3)}$

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

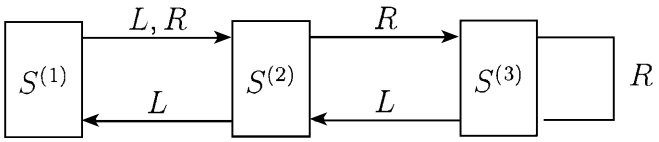


Figure 2 The directed graph in Example 4.

The corresponding graph is presented in Figure 2.

The analytical expressions for the components of the tableau are determined as follows. From the definitions of $L(S^{(j)})$ and $R(S^{(j)})$ (see (4)), we have

$$e^{(2)} = \frac{e'^{(1)} - e^{(1)}}{1 - e^{(1)}}, \quad e^{(2)} = \frac{e^{(1)}}{e'^{(1)}}.$$

Eliminating the control $e'^{(1)}$ we obtain

$$e^{(1)} = \frac{e^{(2)^2}}{e^{(2)^2} - e^{(2)} + 1}.$$

Taking $e'^{(2)} > e^{(2)}$ (the convention for the appendix), we obtain for the second line of the table

$$e^{(1)} = \frac{e'^{(2)} - e^{(2)}}{1 - e^{(2)}}, \quad e^{(3)} = \frac{e^{(2)}}{e'^{(2)}},$$

and for the third line

$$e^{(2)} = \frac{e^{(3)} - e'^{(3)}}{1 - e'^{(3)}}, \quad e^{(3)} = \frac{e'^{(3)}}{e^{(3)}}.$$

Solving for roots in $[0, 1]$, we get $e^{(2)} = 1 - \psi \approx 0.4302$, where ψ is defined by (5), and

$$e^{(1)} = \psi(1 - \psi) \approx 0.2451, \quad e^{(3)} = 1 - \psi + \psi^2 \approx 0.7549, \quad e'^{(1)} = e'^{(2)} = e'^{(3)} = \psi \approx 0.5698.$$

Notice that $e^{(2)}$ is solution of the equation $t^2 = (1-t)^3$, $e^{(1)}$ satisfies the equation $t^3 - 4t^2 + 5t - 1 = 0$ and the equation for $e^{(3)}$ is $t^3 + t^2 - 1 = 0$.

Families of cycles are far richer when $N > 3$. Many structures have no solution, but some give a continuum of solutions, as illustrated by the following example.

Example 5. The structure is

e	e'	L	R
$e^{(1)}$	$e'^{(1)}$	$e^{(1)}$	$e^{(3)}$
$e^{(2)}$	$e'^{(2)}$	$e^{(1)}$	$e^{(4)}$
$e^{(3)}$	$e'^{(3)}$	$e^{(1)}$	$e^{(4)}$
$e^{(4)}$	$e'^{(4)}$	$e^{(2)}$	$e^{(4)}$

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

where

$$e^{(1)} = 2 - \frac{1}{e^{(3)}}, \quad e'^{(1)} = \frac{2e^{(3)} - 1}{e^{(3)^2}}, \quad e^{(2)} = 1 - e^{(3)}, \quad e'^{(2)} = e^{(3)},$$

$$e^{(4)} = \frac{1}{e^{(3)}} - 1, \quad e'^{(4)} = \left(\frac{1}{e^{(3)}} - 1\right)^2, \quad e'^{(3)} = 1 - e^{(3)},$$

and $e^{(3)}$ can be considered as a free parameter, with $1/2 < e^{(3)} < \lambda$ to guarantee $0 < e^{(1)} < e^{(2)} < e^{(3)} < e^{(4)} < 1$.

The following tableau allows construction of examples for arbitrarily large N.

Example 6. Let $s_n = \sum_{r=0}^{n} x^r$ and consider the structure

e	e'	L	R
x	x^2	$x/(1+x)$	x
$\vdots$	$\vdots$	$\vdots$	$\vdots$
x^n	x^{n+1}	x^n/s_n	x
x^{n+1}	x^n	x^n/s_n	x
x^n/s_n	x^n	x^{n+1}	$1/s_n$
$1/s_n$	x^n/s_n	$1-x$	x^n
$\vdots$	$\vdots$	$\vdots$	$\vdots$

The way to 'close' the system is to set the term $1-x$ equal to one term in column 1. For example

(i) $1-x = x^n$, which gives $x = \frac{1}{2}$ when $n = 1, x = \lambda$ (Golden Section) when $n = 2$, etc.
(ii) $1-x = x^n/s_n$, that is $x^{n+1} + x^n - 1 = 0$, which gives $x = \lambda$ when $n = 1$, etc.

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

The following example was constructed by putting $1-x = x^n$ with $n = 1$, that is $x = \frac{1}{2}$, and allowing terms up to $n+1 = 3$.

e	e'	L	R
1/8	1/4	1/7	1/2
1/7	1/4	1/8	4/7
1/4	1/2	1/3	1/2
1/3	1/2	1/4	2/3
1/2	1/4	1/3	1/2
4/7	1/7	1/2	1/4
2/3	1/3	1/2	1/2

4. The dynamic process

Consider the evolution of x_n, the renormalised value of x^* in $[0, 1]$ for any given cycle, with $f(\cdot)$ symmetric with respect to x^*. Let h be defined by iteration (1) when

$$(a_n, b_n) = (\min(e^{(i)}, e'^{(i)}), \max(e^{(i)}, e'^{(i)})),$$

namely the iteration in the ith state $S^{(i)}$. The process is completely described by $(i, x), (i_n, x_n)$ at iteration n, which evolves in $\{1, \ldots, N\} \otimes [0, 1]$:

$$(i, x) \to \begin{cases} \left(\mathscr{L}(i), \dfrac{x - a^{(i)}}{1 - a^{(i)}}\right) & \text{if } x \geqslant c^{(i)} \\ \left(\mathscr{R}(i), \dfrac{x}{b^{(i)}}\right) & \text{if } x < c^{(i)}, \end{cases} \tag{6}$$

where $c^{(i)} = (a^{(i)} + b^{(i)})/2$, $a^{(i)} = \min(e^{(i)}, e'^{(i)})$, $b^{(i)} = \max(e^{(i)}, e'^{(i)})$ and $\mathscr{L}(i), \mathscr{R}(i)$ are the destination states.

To study the ergodic properties of the process, we shall define an equivalent dynamic process in $[0, 1]$. The interval $[0, 1]$ is divided into N subintervals

$$\Delta_i = [(i-1)/N, i/N], i = 1, \ldots, N,$$

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

with the state $S^{(i)}$ associated with Δ_i, that is, for each state $S^{(i)}$ we associate a new variable z living in Δ_i,

$$z = \frac{x}{N} + \frac{i-1}{N}.$$

By concatenating these domains Δ_i we can define a dynamic process $z_{n+1} = T(z_n)$ on the whole interval $[0, 1]$, with

$$T(z) = \sum_{i=1}^{N} \left(\frac{z - i/N}{1 - a^{(i)}} + \frac{\mathscr{L}(i)}{N} \right) I_{[c^{(i)}/N + (i-1)/N,\, i/N)}(z) + \sum_{i=1}^{N} \left(\frac{z - (i-1)/N}{b^{(i)}} + \frac{\mathscr{R}(i) - 1}{N} \right) I_{[(i-1)/N,\, c^{(i)}/N + (i-1)/N)}(z), \tag{7}$$

where $I_{\mathscr{A}}(\cdot)$ is the indicator function of the set $\mathscr{A}$. Both the original state $S^{(i_n)}$ and the position of x_n can be determined from the value of z_n. Note that in (7), the indicators $I_{[c^{(i)}/N + (i-1)/N,\, i/N)}(z)$ and $I_{[(i-1)/N,\, c^{(i)}/N + (i-1)/N)}(z)$ correspond, respectively, to the conditions $x \geqslant c^{(i)}, x < c^{(i)}$ in (6). Also, note that the convention used to define the mapping at the points $i/N, i = 1, \ldots, N$, is arbitrary, since these values of z can only be reached if the process is initialised at $x^* = 0$ or 1, and in this case $x_n = 0$ or 1 for all n. The process z_n associated with a given cycle is thus characterised by a piecewise linear mapping $T: [0, 1] \mapsto [0, 1]$.

Example 2. (continued). The mapping $T(\cdot)$ for the process z_n is presented in Figure 3.

In the next section we study the ergodic properties of the dynamic process z_n.

5. Invariant measures, asymptotic rates and entropies

5.1. Existence of an ergodic invariant measure

The mapping $T(\cdot)$ defined by (7) is expanding, since $0 < a^{(i)} < 1$ and $0 < b^{(i)} < 1$ for all i. Thus, there exists a T-invariant absolutely continuous measure μ_z (see [1, p. 210]). Since for at least one i

$$(0, 1) \backslash T\left(\left(\frac{i-1}{N}, \frac{i}{N} \right) \right) \neq \varnothing,$$

its density $\phi_z(\cdot)$ is certainly discontinuous (see, e.g., [2, p. 195]). This density $\phi_z(\cdot)$ belongs to $L_1([0, 1], \mu)$, where μ is the Lebesgue measure, and it can therefore be normalised and considered as a probability density. In what follows we assume that $\int_0^1 \phi_z(z)\, dz = 1$. The invariant density $\phi_z(\cdot)$ is the eigenfunction of the Perron–Frobenius operator with the eigenvalue 1:

$$\phi_z(z) = \sum_{y_j \in \mathscr{J}(z)} \frac{1}{|T'(y_j)|} \phi_z(y_j), \tag{8}$$

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

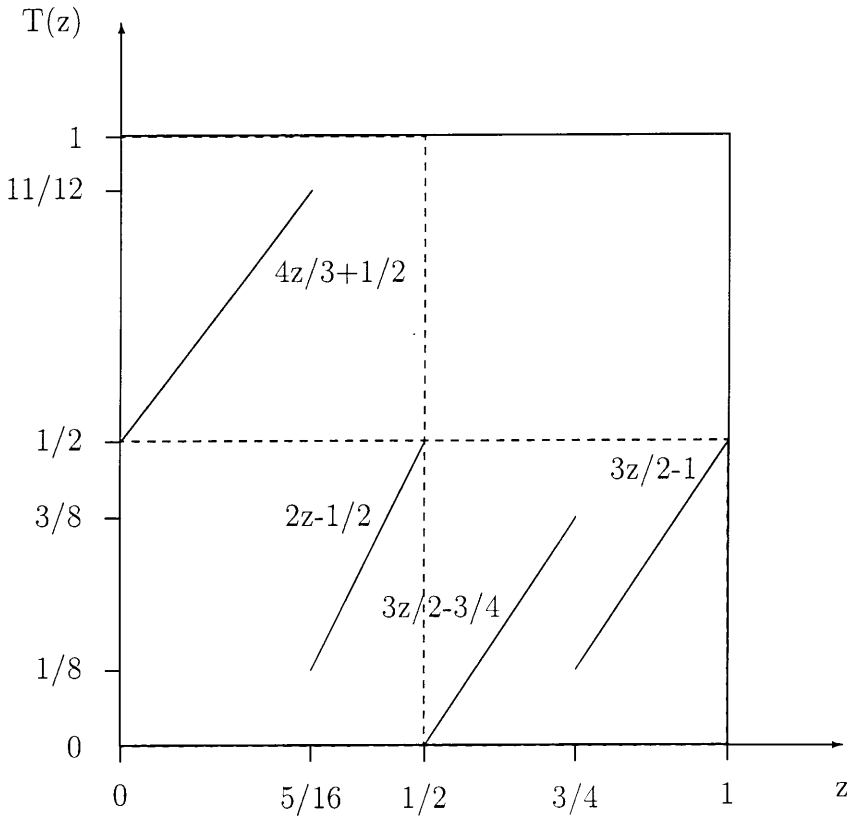


Figure 3 The graph of the mapping $z \to T(z)$ in Example 2. From $S^{(1)}$ ($z \in [0, 1/2]$) we go to $S^{(1)}$ or $S^{(2)}$ ($z \in [1/2, 1]$). From $S^{(2)}$ we go to $S^{(1)}$.

where $\mathscr{J}(z) = T^{-1}(z) = \{y_j | T(y_j) = z\}$, and $T'(\cdot)$ is the derivative of the mapping $T(\cdot)$. The invariant measure $\mu_z(\cdot)$ is ergodic, since for any set $\mathscr{A}$,

$$\mathscr{A} \subset [0, 1],\ \mathscr{A} = T^{-1}(\mathscr{A}) \Rightarrow \mu_z(\mathscr{A}) = 0 \text{ or } 1,$$

where $T^{-1}(\mathscr{A}) = \{T^{-1}(z),\ z \in \mathscr{A}\}$. This follows from the expanding property of $T(\cdot)$. Ergodicity implies that 1 is a simple eigenvalue of the Perron–Frobenius operator. Thus $\phi_z(\cdot)$ is the (unique) normalized *ergodic density* associated with the mapping $T(\cdot)$.

5.2. Partitions and Markov maps

Consider the set $\mathscr{A}$ defined by the endpoints of the intervals Δ_i, together with the points in the Δ_i's associated with the original $c^{(i)}$'s, that is

$$\mathscr{A} = \left(\bigcup_{i=1}^{N} \left\{ \frac{c^{(i)} + (i-1)}{N} \right\} \right) \cup \left(\bigcup_{i=0}^{N} \left\{ \frac{i}{N} \right\} \right).$$

The set $\mathscr{A}$ defines a finite partition of $[0, 1]$ into closed intervals $\mathscr{I}_i^0$, $i = 1, \ldots, M_0$. Assume that $x_1 = x^*$ has a uniform initial distribution, and that the initial state is fixed or randomly

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

chosen. The variable z_1 then inherits an initial distribution, piecewise constant on the intervals $\mathscr{I}_i^0$.

Applying the transformation $T(\cdot)$ given by (7) preserves the piecewise constant property, but now on the intervals $\mathscr{I}_i^1$, $i = 1, \ldots, M_1$, defined by the points in $\mathscr{S}^1 = A \cup T(\mathscr{A})$. Continuing in this way, we define $\mathscr{S}^\infty$ as

$$\mathscr{S}^\infty = \bigcup_{n=0}^{\infty} T^n(\mathscr{A}),$$

with $T^n(\mathscr{A}) = \{T^n(z),\ z \in \mathscr{A}\}$, which in turn defines a (possibly infinite) partition of $[0, 1]$. The intervals $\mathscr{I}_i$, $i = 1, \ldots, M \leqslant \infty$, of this partition form the states for a representation of the dynamic system z_n defined by (7) as a Markov chain, with initial distribution

$$p_i = \Pr(z_1 \in \mathscr{I}_i), \quad i = 1, \ldots, M.$$

For any $n > 1$ and any $i \in \{1, \ldots, M\}$, the distribution of z_n conditional to $z_n \in \mathscr{I}_i$ is uniform on $\mathscr{I}_i$. Also, by definition of $\mathscr{S}^\infty$,

$$T(\mathscr{I}_i) = \bigcup_{j \in \Omega_i} \mathscr{I}_j,$$

where $\Omega_i \subseteq \{1, \ldots, M\}$.

Example 2. (continued). The sets $\mathscr{A}$ and $\mathscr{S}^\infty$ are

$$\mathscr{A} = \{0, 5/16, 1/2, 3/4, 1\}, \quad \mathscr{S}^\infty = \{0, 1/8, 1/4, 5/16, 3/8, 1/2, 2/3, 3/4, 5/6, 11/12, 1\}.$$

The intervals $\mathscr{I}_i$ that define the partition of $[0, 1]$ are

$$\mathscr{I}_1 = [0, 1/8], \quad \mathscr{I}_2 = [1/8, 1/4], \quad \mathscr{I}_3 = [1/4, 5/16], \quad \mathscr{I}_4 = [5/16, 3/8], \quad \mathscr{I}_5 = [3/8, 1/2],$$
$$\mathscr{I}_6 = [1/2, 2/3], \quad \mathscr{I}_7 = [2/3, 3/4], \quad \mathscr{I}_8 = [3/4, 5/6], \quad \mathscr{I}_9 = [5/6, 11/12], \quad \mathscr{I}_{10} = [11/12, 1].$$

The transitions of the Markov chain can be read from Figure 3, and are defined as follows:

$$\mathscr{I}_1 \mapsto \mathscr{I}_6, \quad \mathscr{I}_2 \mapsto \mathscr{I}_7 \cup \mathscr{I}_8, \quad \mathscr{I}_3 \mapsto \mathscr{I}_9, \quad \mathscr{I}_4 \mapsto \mathscr{I}_2, \quad \mathscr{I}_5 \mapsto \mathscr{I}_3 \cup \mathscr{I}_4 \cup \mathscr{I}_5,$$
$$\mathscr{I}_6 \mapsto \mathscr{I}_1 \cup \mathscr{I}_2, \quad \mathscr{I}_7 \mapsto \mathscr{I}_3 \cup \mathscr{I}_4, \quad \mathscr{I}_8 \mapsto \mathscr{I}_2, \quad \mathscr{I}_9 \mapsto \mathscr{I}_3 \cup \mathscr{I}_4, \quad \mathscr{I}_{10} \mapsto \mathscr{I}_5. \qquad \square$$

We can now define the transition probability from $\mathscr{I}_i$ to $\mathscr{I}_j$ as:

$$p_{ij} = \begin{cases} \dfrac{\bar{\mathscr{I}}_j}{\sum_{k \in \Omega_i} \bar{\mathscr{I}}_k} & \text{if } j \in \Omega_i \\ 0 & \text{otherwise,} \end{cases} \tag{9}$$

where $\bar{\mathscr{I}}$ is the length of the interval $\mathscr{I}$. We shall denote by P the matrix with elements p_{ij}. This representation will be used in Section 5.4 to compute the entropies to be defined in Section 5.3.

Ergodicity of $T(\cdot)$ is equivalent to the uniqueness of the absorbing class for the Markov chain. The eigenvector associated with the simple eigenvalue 1 for the matrix P^T gives the

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

invariant distribution for the Markov chain, denoted by $\bar{p}_i$, $i = 1, \ldots, M$, with $\sum_{i=1}^{M} \bar{p}_i = 1$. The invariant measure for z is then

$$\phi_z(z) = \sum_{i=1}^{M} \phi_i I_{\mathscr{I}_i}(z),$$

with $\phi_i = \bar{p}_i / \bar{\mathscr{I}}_i$. This measure can also be obtained through the solution of the Perron–Frobenius equation as follows. Since $T(\cdot)$ is piecewise linear, $\phi_z(\cdot)$ is piecewise constant on the intervals $\mathscr{I}_i$, and is defined by $\phi_z(z) = \phi_i$, $z \in \mathscr{I}_i$. Equation (8) then reduces to a linear equation of the form $\Phi_z = Q\Phi_z$, with $\Phi_z = (\phi_1, \ldots, \phi_M)^T$, and the ϕ_i's are therefore obtained as the solutions of an eigenvalue problem.

From the density for z_n we deduce the density for x_n. Define $\phi_x^{(i)}(\cdot)$ as follows:

$$\phi_x^{(i)}(\cdot) = \Pr\{S = S^{(i)}\} \times \phi_x(\cdot \mid S = S^{(i)}),$$

that is, the conditional density of x in the state $S^{(i)}$ multiplied by the invariant probability of being in this state. We have

$$\phi_x^{(i)}(x) = \frac{1}{N} \phi_z\left(\frac{x + (i-1)}{N}\right),$$

and

$$\phi_x(x) = \sum_{i=1}^{N} \phi_x^{(i)}(x) = \frac{1}{N} \sum_{i=1}^{N} \phi_z\left(\frac{x + (i-1)}{N}\right).$$

From this we can construct the matrix Π of transition probabilities for the states $S^{(i)}$, $i = 1, \ldots, N$. Let $\mathscr{L}(i) = j$ and $\mathscr{R}(i) = k$, the transition probabilities π_{ij} from $S^{(i)}$ to $S^{(j)}$ and π_{ik} from $S^{(i)}$ to $S^{(k)}$ are simply given by

$$\pi_{ij} = \frac{\displaystyle\int_{(c^{(i)}+(i-1))/N}^{i/N} \phi_z(z)\, dz}{\displaystyle\int_{(i-1)/N}^{i/N} \phi_z(z)\, dz} \quad \text{and} \quad \pi_{ik} = \frac{\displaystyle\int_{i-1/N}^{(c^{(i)}+(i-1))/N} \phi_z(z)\, dz}{\displaystyle\int_{(i-1)/N}^{i/N} \phi_z(z)\, dz}.$$

Example 2. (continued). The matrix Q is defined from the slopes of the transformation $T(\cdot)$ and can be read from Figure 3,

$$Q = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 2/3 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 1/2 & 0 & 2/3 & 0 & 2/3 & 0 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 0 & 2/3 & 0 & 2/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 0 & 2/3 & 0 & 2/3 & 0 \\
0 & 0 & 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 2/3 \\
3/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 3/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 3/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 3/4 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix},$$

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

from which we determine

$$\Phi_z = (0, 16/7, 16/7, 16/7, 0, 0, 12/7, 12/7, 12/7, 0)^T.$$

The density $\phi_x(\cdot)$ is thus

$$\phi_x(x) = \tfrac{8}{7} I_{[\frac{1}{4}, \frac{3}{4}]}(x) + \tfrac{6}{7} I_{[\frac{1}{3}, \frac{5}{6}]}(x),$$

and the matrix of transition probabilities for this 2-state cycle is given by

$$\Pi = \begin{pmatrix} 1/4 & 3/4 \\ 1 & 0 \end{pmatrix}.$$

The dynamic processes associated with the cycles considered here generally have the *strong mixing property* [1, p. 224], that is, their transition matrix Π satisfies $(\Pi^n)_{ij} > 0 \, \forall (i,j)$, for some $n > 0$. However, some cycles generate ergodic dynamic systems which do not have the strong mixing property. This is the case, for instance, when $N = 2$ for Example 3, when $N = 3$ with

$$\mathscr{L}(1) = \mathscr{R}(1) = 2, \quad \mathscr{L}(2) = \mathscr{R}(2) = 3, \quad \mathscr{L}(3) = \mathscr{R}(3) = 1$$

(see Section 3). It is also the case when the cycle corresponds to the symmetric algorithms considered in [10].

In many cases, M is finite, and $T(\cdot)$ becomes a Markov map (see [2, p. 196]). The computation of invariant distributions is then straightforward. However, the fact that partitions $\mathscr{S}^\infty$ might be infinite can be proved by an example.

Example 5. (continued). Take c as any non-algebraic number, $1/2 < c < \lambda$, then the finiteness of $\mathscr{S}^\infty$ would imply that c satisfies some polynomial equation.

5.3. Asymptotic rates and entropies

Define the rate at iteration n, from (i_n, x_n) to (i_{n+1}, x_{n+1}), as

$$r_n = \begin{cases} 1 - a_n & \text{if } x_n \geqslant c_n, \\ b_n & \text{if } x_n < c_n, \end{cases}$$

with $c_n = (a_n + b_n)/2$, $(a_n, b_n) = (\min(e^{(i_n)}, e'^{(i_n)}), \max(e^{(i_n)}, e'^{(i_n)}))$. The true length of the unnormalised interval after the nth iteration is thus

$$L_n = \prod_{k=1}^{n} r_k. \tag{10}$$

We define the (logarithmic) asymptotic convergence rate of the search algorithm as

$$\rho = -\lim_{n \to \infty} \frac{1}{n} \log L_n = -\lim_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} \log r_k,$$

and the limit exists for cycles, since the function $\log r(x)$ is piecewise constant on $[0, 1]$. For

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

instance, in state $S^{(i)}$ one has

$$r(x) = \frac{1}{\left| T'\left(\frac{x+(i-1)}{N}\right)\right|}.$$

This implies that ρ is also the Lyapunov exponent Λ of the dynamic process z_n, where for a general mapping $T(\cdot)$ the Lyapunov exponent can be defined as

$$\Lambda = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \log |T'(z_k)|,$$

if this limit exists and is the same for almost all z_1. A less important alternative is

$$r = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} r_k.$$

Note that larger values of ρ and smaller values of r are better.

Taking now the Bayesian viewpoint, we can assume that x^* has a uniform prior density on the initial interval $[0, 1]$. For any given x^*, each iteration reduces, by left or right deletion, the length of the uncertainty interval $T_n(x^*)$ in which x^* lies at iteration n. The interval $T_n(x^*)$ has length $L_n = L_n(x^*)$ defined by (10), and a natural performance criterion is given by its expected length

$$EL_n = E_{x^*}\{L_n(x^*)\}.$$

A less natural possible alternative is

$$E \log L_n = E_{x^*}\{\log L_n(x^*)\}.$$

The expected length EL_n and the expected log-length $E \log L_n$ tend, respectively, to 0 and $-\infty$ as n tends to infinity, so that we shall consider instead the characteristics

$$E_\rho = -\lim_{n\to\infty} \frac{1}{n} \log EL_n, \quad H = -\lim_{n\to\infty} \frac{1}{n} E \log L_n,$$

under the condition that these limits exist. Note that larger values of E_ρ and H are better. Moreover, Jensen's inequality implies

$$-E_{x^*}\{\log L_n(x^*)\} \geqslant -\log E_{x^*}\{L_n(x^*)\},$$

so that $H \geqslant E_\rho$.

Assume that at iteration $n \geqslant 1$ the uncertainty interval is $T_n(x^*) \subseteq [0, 1)$ with x^* having a uniform density in a cell $C_n = [s_n, t_n) \subseteq T_n$. It is straightforward to check that, at iteration $n+1$, $C_{n+1} = [s_{n+1}, t_{n+1}) \subseteq [s_n, t_n)$, and $C_{n+1} \subset T_{n+1} \subset T_n$. The updating rules for s_n and t_n are obtained from left and right deletion rules under the symmetry assumption for $f(\cdot)$, as in (1).

All x^* in a cell C_n lead to the same sequence of left and right deletions and intervals T_n and $C_n = [s_n, t_n)$. Moreover, the intervals $C_n(x^*)$ and $C_n(x^{*\prime})$ corresponding to different

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

sequences of left and right deletions do not intersect. The interval $[0,1)$ can thus be partitioned into a union of cells $C_n^{(i)}$ in which the behaviour of the algorithm is the same up to iteration n. Since the initial density is uniform, the probability that x^* is in a particular cell $C_n^{(i)} = [s_n^{(i)}, t_n^{(i)})$ is $p_n^{(i)} = t_n^{(i)} - s_n^{(i)}$, with

$$p_n^{(i)} > 0, \quad \sum_{i \in \mathscr{I}_n} p_n^{(i)} = 1.$$

The index set $\mathscr{I}_n$ at iteration n satisfies $\#\mathscr{I}_n \leqslant 2^n$, since each cell is divided into two parts at most at each iteration. We thus have for the expected log-length

$$E \log L_n = \sum_{i \in \mathscr{I}_n} p_n^{(i)} \log L_n^{(i)},$$

and for the expected length

$$EL_n = \sum_{i \in \mathscr{I}_n} p_n^{(i)} L_n^{(i)},$$

with $L_n^{(i)} = L_n(x^*)$ for all x^* in $C_n^{(i)}$.

When the renormalisation is used, $T_n^{(i)}$ is renormalised in $[0,1]$, $L_n^{(i)}$ to 1 and for a cycle the length $l^{(i)}$ of the renormalised cell $C_n^{(i)}$ can only take a finite number of values. The ratio $L_n^{(i)}/p_n^{(i)}$ thus satisfies

$$1 < \frac{L_n^{(i)}}{p_n^{(i)}} \leqslant c = \max_i \frac{1}{l^{(i)}},$$

so that

$$E_\rho = -\lim_{n \to \infty} \frac{1}{n} \log \sum_{i \in \mathscr{I}_n} (p_n^{(i)})^2, \tag{11}$$

when the limit exists. Following the same argument, we get

$$H = -\lim_{n \to \infty} \frac{1}{n} \sum_{i \in \mathscr{I}_n} p_n^{(i)} \log p_n^{(i)}. \tag{12}$$

The criterion H thus coincides with the Kolmogorov entropy (see [1, p. 214]) of the dynamic process of Section 4. The Kolmogorov entropy is based on the Shannon entropy of the partition $\bigcup_{i \in \mathscr{I}_n} C_n^{(i)}$. Similarly, the criterion E_ρ (11) coincides with the entropy of the dynamic process based on the (second-order) Rényi entropy [9] of the same partition (see also [3]). The next subsection is devoted to the computation of the criteria ρ, r and E_ρ.

5.4. Computation of rates and entropies

From Birkhoff's ergodic theorem (see [2, p. 44]), the ergodic rates ρ and r for the process can be computed directly from

$$\rho = -\sum_{i=1}^{M} \bar{p}_i \log r_{(i)}, \quad r = \sum_{i=1}^{M} \bar{p}_i r_{(i)}, \tag{13}$$

where $r_{(i)}$ is the rate for z in the interval $\mathscr{I}_i$, $r_{(i)} = 1/|T'(z)|$ for $z \in \mathscr{I}_i$.

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

We shall use the Markov chain representation of Section 5.2 to show the equivalence between different ergodic characteristics of the process.

Theorem 1. *Assume that the Markov chain defined in Section* 5.2 *has only one absorbing class. Then*

$$H = \rho = \Lambda = -\sum_{i=1}^{M}\sum_{j=1}^{M} \bar{p}_i p_{ij} \log p_{ij},$$

where the p_{ij}'s *are defined by* (9) *and* $\{\bar{p}_i\}$ *is the invariant distribution for the Markov chain.*

Proof. The equality $\rho = \Lambda$ has already been shown in Section 5.3. Since the Markov chain has only one absorbing class, the invariant distribution $\{\bar{p}_i\}$ is uniquely defined. The property

$$H = -\sum_{i=1}^{M}\sum_{j=1}^{M} \bar{p}_i p_{ij} \log p_{ij}$$

is then a well-known result for the Kolmogorov entropy for Markov shifts (see Theorem 4.21 in [1]). Note that we use here the partition of $[0, 1]$ into intervals $\mathscr{I}_i$, as defined in Section 5.2, as the initial partition when computing H. Using

$$p_{ij} = \begin{cases} \dfrac{\bar{\mathscr{I}}_j}{\sum_{k\in\Omega_i} \bar{\mathscr{I}}_k} = \dfrac{\bar{\mathscr{I}}_j}{\dfrac{1}{r_{(i)}}\bar{\mathscr{I}}_i} & \text{if } j \in \Omega_i \\ 0 & \text{otherwise,} \end{cases}$$

we then get $\sum_{j=1}^{M} p_{ij} \log p_{ij} = \log r_{(i)}$ by simple manipulation. Using the expression (13) for ρ, one then gets the result. □

Remark 1. The equality $H = \rho$ also follows from the Shannon–McMillan–Breiman theorem (see [1, p. 214]), which implies that for almost all x^* in $[0, 1]$, $-\dfrac{1}{n}\log L_n(x^*)$ tends to H as n tends to infinity.

We can also obtain a closed-form expression for E_ρ.

Theorem 2. *Assume that* M *is finite, and that the Markov chain defined by* P *has a unique absorbing class. Let* P' *be the transition matrix restricted to this class. Assume that* P' *has the strong mixing property* (*see Section* 5.2 *for a definition*), *and define* P'_2 *as the matrix with elements* $(P'_2)_{ij} = (P'_{ij})^2$. *Assume, moreover, that the initial probability of being in the absorbing class is equal to* 1. *Then*

$$E_\rho = -\log \lambda_{\max}(P'_2). \tag{14}$$

Proof. We consider the partition of $[0, 1]$ in intervals $\mathscr{I}_i$ as defined in Section 5.2. Let $\mathscr{I}'_i$, $i = 1, \ldots, M'$ be the intervals corresponding to the absorbing class for P. The probability

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

$\Pr(z_1 \in \mathscr{I}'_{i_1}, z_2 \in \mathscr{I}'_{i_2}, \ldots, z_n \in \mathscr{I}'_{i_n})$ is equal to $p'_{i_1} p'_{i_1 i_2} \ldots p'_{i_{n-1} i_n}$, where $(i_1, \ldots, i_n) \in \{1, \ldots, M'\}^n$ and $p'_i, i = 1, \ldots, M'$ defines the initial distribution. Then, E_ρ defined by (11) can be written as

$$E_\rho = -\lim_{n\to\infty} \frac{1}{n} \log \left(\sum_{i_1=1}^{M'} \ldots \sum_{i_n=1}^{M'} (p'_{i_1} p'_{i_1 i_2} \ldots p'_{i_{n-1} i_n})^2 \right)$$

$$= -\lim_{n\to\infty} \frac{1}{n} \log ((\boldsymbol{p}'_2)^T (P'_2)^n \mathbf{1}),$$

where $(\boldsymbol{p}'_2)^T = ((p'_1)^2, \ldots, (p'_{M'})^2)$ and $\mathbf{1}^T = (1, \ldots, 1)$. Since P'_2 is strong mixing, it follows from Perron–Frobenius theorem (see [2, Theorem 7.25, p. 205]) that its maximal eigenvalue $\lambda_{\max}(P'_2)$ is simple and

$$E_\rho = -\lim_{n\to\infty} \frac{1}{n} ((\boldsymbol{p}'_2)^T \boldsymbol{u})(\boldsymbol{v}^T \mathbf{1}) \lambda_{\max}(P'_2),$$

where $\boldsymbol{u}$ is the eigenvector associated with $\lambda_{\max}(P'_2)$ and $\boldsymbol{v}$ is the eigenvector of $(P'_2)^T$ associated with the same eigenvalue. Both vectors have strictly positive elements and satisfy $\boldsymbol{u}^T \boldsymbol{v} = 1$. We thus get (14). □

The criterion E_ρ corresponds here to the limiting behaviour of the expected length of the uncertainty interval. It is interesting to note that it does not seem to have been considered in the study of dynamic systems, although it is more natural from a practical point of view than the criterion ρ, which corresponds here to the limiting behaviour of the expectation of the logarithm of the length of the uncertainty interval.

Note that the condition of Theorem 2 of being initially in the absorbing class can always be achieved by a proper choice of the initial state and uncertainty interval. Indeed, if the initial state of the algorithm is $S^{(i)}$, the probability of being in the states of the Markov chain (i.e. the intervals $\mathscr{I}_j$) associated with the interval Δ_i is one. An adjustment of the prior distribution of z_1 over the $\mathscr{I}_j$'s $\subset \Delta_i$, that is restricting the support to some of the $\mathscr{I}_j$'s, is then equivalent to a suitable expansion of the uncertainty interval for x^*, keeping its density uniform over the original interval (see [8] for elaboration of this point). Such a choice of the initial distribution p_i, $i = 1, \ldots, M$ over the states of the Markov chain may be critical for the asymptotic worst-case behaviour of the algorithm. Indeed, it appears that the worst rates may be associated with the transient states of the chain (see [8]). The fact that E_ρ is affected by the choice of the p_i's, $i = 1, \ldots, M$, whereas ρ is not another indication that E_ρ is a more refined characteristic of the asymptotic behaviour of the algorithm.

We shall now consider a collection of examples.

6. Examples

Example 1. (Golden Section, continued). There are two states only, with

$$\{e^{(1)}, e^{(2)}\} = \{1-\lambda, \lambda\},\ e'^{(1)} = e^{(2)},\ e'^{(2)} = e^{(1)},$$

and the dynamic process is defined by (3). The invariant measure has the step-function density shown in Figure 4.

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

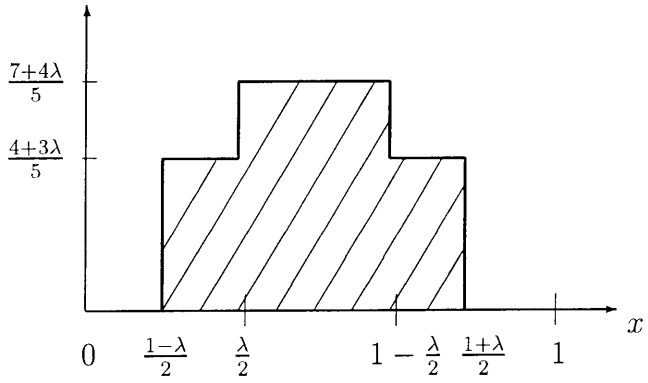


Figure 4 The graph of the invariant density of the mapping (3).

The asymptotic rates are $r = \lambda \approx 0.6180$ and $\rho = E_\rho = -\log\lambda \approx 0.4812$. The GS cycle is the best among two-state cycles.

As we shall see soon for any $N > 2$ there are cycles that are better than Golden Section (see also [5] where several 5- and 6-state cycles are detailed).

Example 2. (continued). From (13), where the values of $T'(z)$ are found in Figure 2, we get the ergodic rates $\rho = 4/7\log 2$ and $r = 19/28$. The value of E_ρ is $-\log\lambda_{\max}$, with $\lambda_{\max}$ the largest real root of $16t^4 - 8t^2 - t + 1$, that is $\lambda_{\max} \approx 0.6748$ and $E_\rho \approx 0.3933$.

Example 4. (continued). The invariant density for x is given by:

$$\phi_x = \phi_x^{(1)}(x) + \phi_x^{(2)}(x) + \phi_x^{(3)}(x),$$

with

$$\begin{aligned}
\phi_x^{(1)}(x) &= KI_{[\frac{a}{2},\frac{1}{2}]}(x),\\
\phi_x^{(2)}(x) &= K[2(\psi^2+1)\,I_{[\frac{b}{2},\frac{a+b}{2}]}(x) + (\psi^2+\psi+1)\,I_{[\frac{a+b}{2},\frac{1}{2}]}(x)\\
&\quad + (\psi^2-\psi+2)\,I_{[\frac{1}{2},\frac{1+b}{2}]}(x)],\\
\phi_x^{(3)}(x) &= K[2(1-\psi+\psi^2)\,I_{[\frac{c}{2},\frac{1}{2}]}(x) + 2(\psi^2+1)\,I_{[\frac{1}{2},\frac{b+c}{2}]}(x)\\
&\quad + (2\psi^2+\psi+1)\,I_{[\frac{b+c}{2},\frac{\psi+c}{2}]}(x) + (3-\psi+2\psi^2)\,I_{[\frac{\psi+c}{2},\frac{\psi+1}{2}]}(x)\\
&\quad + (\psi^2-\psi+3)\,I_{[\frac{\psi+1}{2},\frac{c+1}{2}]}],
\end{aligned}$$

and the asymptotic rates are given by

$$\begin{aligned}
\rho &= -\frac{K}{2}[(4\psi^2-3\psi+3)\log\psi + (3\psi^2-3\psi+\log(1-\psi)\\
&\quad + (5\psi^2-11\psi+6)\log(1-\psi+\psi^2)] \approx 0.5575,\\
r &= \frac{K}{2}(5\psi^2+6\psi-2) = \frac{1017\psi^2-1529\psi+1753}{2075} \approx 0.5841,
\end{aligned}$$

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

where ψ is defined through (5) and

$$K = \frac{2}{12\psi^2 - 17\psi + 11} = \frac{688 - 84\psi + 482\psi^2}{2075} \approx 0.3839$$

is the normalisation constant.

The value of E_ρ equals $-\log \lambda_{\max}$, where $\lambda_{\max}$ is the unique real root of the polynomial

$$t^9 + (3\psi^2 - 2\psi)\, t^7 + (2 - 3\psi - \psi^2)\, t^6 + (4\psi - 4\psi^2 - 1)\, t^5 + (7 - 14\psi + 3\psi^2)\, t^3 + (15 - 40\psi + 24\psi^2)\, t^2 + (24 - 33\psi - 16\psi^2)\, t + (-16 + 56\psi - 49\psi^2).$$

This gives $E_\rho \approx 0.549\,22$. The cycle of this example is the best among 2- and 3-state cycles according to ρ- and r-criteria, and in particular is significantly better than the Golden Section.

Example 7. In this example the asymptotic log-rate ρ tends to the optimal one

$$(\rho = \log 2 \approx 0.693\,147)$$

when the number of states is increased to infinity. We take $e'^{(k)} = e^{(k)}/2$ when $e^{(k)} = 1/(1+2^k)$ and $e'^{(k)} = (1+e^{(k)})/2$ when $e^{(k)} = 1 - 1/(1+2^k)$. We postpone the discussion on the initialisation of the process to the end of the example. The states and controls are completely defined by the index k, except when $k = 0$. We then simply have to keep trace of the previous state and action: if the previous value of e was such that $e_{n-1} < 1/2$ (i.e. $e_{n-1} = 1/(1+2^k)$, $k \geqslant 1$) with the decision (R), then $e_n = 1/2$ and $e'_n = 3/4$. The symmetrical situation is handled in the same way. We thus introduce a binary variable $s \in \{0, 1\}$, and the states and control are defined as in the following tableau:

(k, s)	e	e'	L	R
$(0, 0)$	$e^{(0,0)} = \frac{1}{2}$	$e'^{(0,0)} = \frac{1}{4}$	$(1, 0)$	$(0, 1)$
$(0, 1)$	$e^{(0,1)} = \frac{1}{2}$	$e'^{(0,1)} = \frac{3}{4}$	$(0, 0)$	$(1, 1)$
$(k, 0)$	$e^{(k,0)} = \frac{1}{1+2^k}$	$e'^{(k,0)} = \frac{u_{k,0}}{2}$	$(k+1, 0)$	$(0, 1)$
$(k, 1)$	$e^{(k,1)} = \frac{2^k}{1+2^k}$	$e'^{(k,1)} = \frac{1+u_{k,1}}{2}$	$(0, 0)$	$(k+1, 1)$

The collection of pairs $(e^{(k,s)}, e'^{(k,s)})$ can be made finite by using a modified control when e is close to 0 or 1, i.e. when k is large. Let K be the maximal admissible value of k, we define

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

the following controls when $k = K$:

	e	e'	L	R
$(K,0)$	$\frac{1}{1+2^K}$	$\frac{1+2^{K-1}}{1+2^K}$	$(0,0)$	$(K-1,0)$
$(K,1)$	$\frac{2^K}{1+2^K}$	$\frac{2^{K-1}}{1+2^K}$	$(K-1,1)$	$(0,1)$

An important feature of this cycle is that the invariant measures $\phi_x^{(k,s)}(\cdot)$ conditioned on any given state (k,s) are uniform between $e^{(k,s)}$ and $e'^{(k,s)}$. The ergodic probabilities of the events (L) and (R) are then equal to $1/2$ for $k < K$. When $k = K$, the ergodic probability of the transition $(K,s) \to (K-1,s)$ is one. The asymptotic rates can then be calculated analytically,

$$r(K) = \frac{1}{1+2^K}\left(2^{K-1} + \frac{1}{2+2^K} + \frac{1+2^{K-1}}{1+2^K}\right) + \frac{2^{K-3}}{1+2^K}\sum_{k=0}^{K-2}\frac{1}{2^k(1+2^k)},$$

and

$$\rho(K) = \frac{2^K}{1+2^K}\log 2,$$

which tends to $\log 2$ as K increases. The value of $E_\rho = -\log \lambda_{\max}$ is obtained from $\lambda_{\max} = t^*/4$ where t^* is the largest real root of the polynomial

$$t^{K+2} - 2t^{K+1} - 4t^K + 8t^{K-1} + t - 4.$$

This satisfies $\lambda_{\max} > 1/2$ and tends to $1/2$ as K increases.

The process can be initialised with $(e^{(\alpha)}, e'^{(\alpha)}) = (0,1)$. Consider the pairs of states and controls $(e^{(\beta)}, e'^{(\beta)}) = (1, 1/2), (e^{(\gamma)}, e'^{(\gamma)}) = (0, 1/2)$. We use the following tableau

e	e'	(L)	(R)
$e^{(\alpha)}$	$e'^{(\alpha)}$	$e^{(\beta)}$	$e^{(\gamma)}$
$e^{(\beta)}$	$e'^{(\beta)}$	$e^{(\beta)}$	$e^{(0,1)}$
$e^{(\gamma)}$	$e'^{(\gamma)}$	$e^{(0,0)}$	$e^{(\gamma)}$

We can easily check that this initialisation has no influence on the asymptotic characteristics of the algorithm.

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

7. Asymptotic behaviour for locally symmetric functions

Previous asymptotic results hold under the assumption that the function $f(\cdot)$ is uniextremal and symmetric. A natural question concerns the generalisation of these results to the case where $f(\cdot)$ is locally symmetric:

$$f(x) = f(x^*) + C_1|x - x^*|^\gamma + O(|x - x^*|^{\beta+\gamma}) \quad \text{for some } \beta > 0, \tag{15}$$

with $C_1 > 0$, $\gamma > 0$. Note that if $f(\cdot)$ is smooth at x^*, then (15) holds with $\gamma = 2$. In [7], the asymptotic behaviour of the Golden-Section algorithm is then shown to be the same as for symmetric functions. The proof is based on a *self-correcting property* of the associated dynamic process, namely if at iteration n we delete the wrong part of the interval because of non-symmetry of $f(\cdot)$, that is if we go from z_n to z'_{n+1}, then in a finite number of iterations the iterates of z'_{n+1} will coincide with those of z_{n+1} that would obtain if $f(\cdot)$ had been symmetric. This property is satisfied for most of the cycles considered in this paper. In particular, it holds for Example 7, which has the important consequence that there exist cycles with rates ρ and E_ρ arbitrarily close to the optimal value $\log 2$ for locally symmetric functions. An algorithm that achieves the optimal ergodic rate of $\rho = \log 2$ for functions satisfying (15) is detailed in [11].

8. Conclusions

The ergodic theory of dynamic systems has been used for a new probabilistic analysis of the convergence of a class of line-search algorithms. The construction of the associated dynamic systems and Markov partitions, and the calculation of their ergodic characteristics has been detailed and illustrated by a series of examples. The fact that convergence criteria coincide with various kinds of entropy has been demonstrated. In particular Rényi entropy appears to be the most appropriate convergence criterion in this context.

Acknowledgements

This work was supported by a French–British Alliance grant (nx94002), by the UK Engineering and Physical Science Research Council grant for the second author, and by a Visiting Fellowship from CNRS for the third author.

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

Appendix: table of all three-state cycles for $\epsilon'^{(2)} > \epsilon^{(2)}$

$e^{(1)}$	$e'^{(1)}$	$\mathcal{L}(1)$	$\mathcal{R}(1)$	$e^{(2)}$	$e'^{(2)}$	$\mathcal{L}(2)$	$\mathcal{R}(2)$	$e^{(3)}$	$e'^{(3)}$	$\mathcal{L}(3)$	$\mathcal{R}(3)$	ρ	r
0.245	0.570	2	2	0.430	0.570	1	3	0.755	0.570	2	3	0.557	0.584
0.317	0.534	1	2	0.594	0.817	3	3	0.682	0.217	2	1	0.317	0.742
0.317	0.783	3	2	0.406	0.594	1	3	0.682	0.466	2	3	0.483	0.622
0.363	0.637	2	3	0.430	0.755	3	3	0.570	0.325	1	3	0.421	0.658
0.366	0.698	2	2	0.524	0.826	3	3	0.634	0.232	2	1	0.321	0.732
0.382	0.618	1	3	0.500	0.809	3	3	0.618	0.236	2	1	0.365	0.701
0.382	0.764	3	2	0.500	0.809	3	3	0.618	0.382	1	3	0.393	0.681
0.382	0.764	3	2	0.500	0.809	3	3	0.618	0.236	2	1	0.423	0.662
0.382	0.618	1	2	0.618	0.854	2	3	0.724	0.276	2	1	0.336	0.726
0.382	0.618	1	2	0.618	0.764	1	3	0.809	0.500	2	2	0.481	0.638
0.393	0.723	2	2	0.544	0.839	3	3	0.648	0.420	1	3	0.395	0.682
0.400	0.800	3	2	0.500	0.750	2	3	0.667	0.445	1	3	0.410	0.671
0.406	0.682	2	3	0.466	0.783	3	3	0.594	0.241	2	1	0.348	0.712
0.414	0.657	1	2	0.631	0.892	3	3	0.707	0.500	1	3	0.428	0.668
0.409	0.737	2	2	0.555	0.802	2	3	0.692	0.479	1	3	0.420	0.667
0.414	0.828	3	2	0.500	0.707	1	3	0.707	0.500	1	3	0.420	0.670
0.423	0.667	1	2	0.634	0.866	2	3	0.732	0.536	1	3	0.438	0.663
0.430	0.755	2	2	0.570	0.755	1	3	0.755	0.570	1	3	0.451	0.656
0.430	0.755	2	2	0.570	0.755	1	3	0.755	0.430	2	2	0.423	0.667
0.430	0.755	2	2	0.570	0.815	2	3	0.699	0.301	2	1	0.366	0.705
0.445	0.692	1	2	0.643	0.802	1	3	0.802	0.643	1	3	0.479	0.650
0.447	0.724	2	3	0.500	0.809	3	3	0.618	0.309	1	2	0.385	0.686
0.453	0.829	3	2	0.547	0.794	2	3	0.688	0.312	2	1	0.382	0.695
0.456	0.704	1	3	0.544	0.839	3	3	0.648	0.352	1	2	0.414	0.667
0.459	0.848	3	2	0.541	0.752	1	3	0.720	0.389	2	2	0.390	0.687
0.460	0.817	3	2	0.563	0.852	3	3	0.661	0.372	1	2	0.351	0.716
0.466	0.783	2	2	0.594	0.871	3	3	0.682	0.406	1	2	0.347	0.722
0.475	0.724	1	2	0.656	0.905	3	3	0.724	0.475	1	2	0.362	0.717
0.476	0.751	2	3	0.524	0.826	3	3	0.634	0.302	1	1	0.378	0.692
0.500	0.854	3	2	0.586	0.828	2	3	0.707	0.414	1	2	0.349	0.721
0.500	0.809	2	2	0.618	0.854	2	3	0.724	0.447	1	2	0.353	0.721
0.500	0.809	2	2	0.618	0.809	1	3	0.764	0.382	2	1	0.366	0.711
0.500	0.866	3	2	0.577	0.789	1	3	0.732	0.366	2	1	0.352	0.718
0.500	0.750	1	2	0.667	0.889	2	3	0.750	0.500	1	2	0.360	0.722
0.500	0.750	1	2	0.667	0.833	1	3	0.800	0.400	2	1	0.362	0.723
0.520	0.853	3	2	0.610	0.880	3	3	0.693	0.360	1	1	0.329	0.737
0.532	0.879	3	2	0.605	0.815	1	3	0.742	0.449	1	2	0.338	0.732
0.538	0.836	2	2	0.644	0.899	3	3	0.717	0.386	1	1	0.321	0.746
0.544	0.839	2	2	0.648	0.839	1	3	0.772	0.500	1	2	0.357	0.724
0.547	0.876	3	2	0.624	0.859	2	3	0.727	0.397	1	1	0.317	0.748
0.555	0.802	1	2	0.692	0.863	1	3	0.802	0.555	1	2	0.368	0.724
0.563	0.852	2	2	0.661	0.885	2	3	0.747	0.421	1	1	0.317	0.751
0.570	0.895	3	2	0.637	0.844	1	3	0.755	0.430	1	1	0.307	0.757
0.570	0.815	1	2	0.699	0.926	3	3	0.755	0.430	1	1	0.311	0.756
0.586	0.828	1	2	0.707	0.914	2	3	0.773	0.453	1	1	0.313	0.757
0.594	0.871	2	2	0.682	0.871	1	3	0.783	0.466	1	1	0.316	0.756
0.618	0.854	1	2	0.724	0.894	1	3	0.809	0.500	1	1	0.314	0.761

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press

References

[1] Barndorff-Nielsen, O. E., Jensen, J. L. and Kendall, W. S. (eds.) (1993) *Networks and Chaos – Statistical and Probabilistic Aspects*. Chapman & Hall.

[2] Bedford T., Keane, M. and Series, C. (eds.) (1991) *Ergodic Theory*, *Symbolic Dynamics and Hyperbolic Spaces*. Oxford University Press.

[3] Grassberger, P. and Procaccia, I. (1983) Estimating the Kolmogorov entropy from a chaotic signal. *Phys. Rev.* **A 28** 2591.

[4] Kiefer, J. (1953) Sequential minimax search for a maximum. *Proc. American Mathematical Society* **4** 502–506.

[5] Pronzato, L., Wynn, H. P. and Zhigljavsky, A. A. (1993) Section invariant numbers and ergodically fast line search. Internal Report 93-50, 13S, CNRS-URA 1376, bât. 4, 250 rue A. Einstein, Sophia-Antipolis, 06560 Valbonne, France.

[6] Pronzato, L., Wynn, H. P. and Zhigljavsky, A. A. (1995) Dynamic systems in search and optimisation. *CWI Quarterly* **8** 201–236.

[7] Pronzato, L., Wynn, H. P. and Zhigljavsky, A. A. (1995) Improving on Golden-Section optimisation for locally symmetric functions. In: *MODA* 4 – *Advances in Model-Oriented Data Analysis*, Ch. P. Kitsos and W. G. Müller (eds.) Physica-Verlag, pp. 285–297.

[8] Pronzato, L., Wynn, H. P. and Zhigljavsky, A. A. (1995) A generalised Golden Section algorithm for line-search. Internal Report 95-55, 13S, CNRS-URA 1376, bât. 4, 250 rue A. Einstein, Sophia-Antipolis, 06560 Valbonne, France. *IMA J. Mathematical Control and Information* (to appear).

[9] Rényi, A. (1965) On the foundations of information theory. *Rev. Inst. Internat. Stat.* **33** 1–14.

[10] Wynn, H. P. and Zhigljavsky, A. A. (1993) Chaotic behaviour of search algorithms. *Acta Applicandae Mathematicae* **32** 123–156.

[11] Wynn, H. P. and Zhigljavsky, A. A. (1995) Beating the Golden Section: a fast line-search algorithm for locally smooth functions. *Journal of Complexity* **11** 196–226.

https://doi.org/10.1017/S096354839600288X Published online by Cambridge University Press