

Bounds on Absorption Times of Directionally Biased Random Sequences

Bill Baritompa and Mike Steel*

Department of Mathematics and Statistics, University of Canterbury,
Christchurch, New Zealand

ABSTRACT

A sequence of random variables X_0, X_1, \dots with values in $\{0, 1, \dots, n\}$ representing a general finite-state stochastic process with absorbing state 0 is said to be *directionally biased towards 0*, if, for all $j > 0$, $\epsilon_j := \inf_{k > 0} \{j - \mathbb{E}[X_k | X_{k-1} = j]\} > 0$. For such sequences, let t be the expected value of the time to absorption at 0. For a fixed set of biases, the least upper bound for this time can be computed with an algorithm requiring $O(n^2)$ steps. Simple upper bounds are described. In particular, $t \leq \mathbb{E}[b_{X_0}]$, where $b_i = \sum_{j \neq i} 1/\bar{\epsilon}_j$ and $\bar{\epsilon}_j = \min_{l \neq j} \{\epsilon_l\}$. If all $\epsilon_j \leq \epsilon_{j+1}$ (so $\bar{\epsilon}_j = \epsilon_j$) and $\epsilon_n < 1$, this bound for t is the best possible. For certain finite stochastic processes which we term conditionally independent of $X_0 = i$, $b(i)$ bounds the expected time given $X_0 = i$. Similar results are given for lower bounds. The results of this paper were designed to be a useful tool for determining rates of convergence of stochastic optimization algorithms. © 1996 John Wiley & Sons, Inc.

Key Words: Convergence rates, Markov chains, simulated annealing, stochastic optimization, finite-state stochastic process.

1. INTRODUCTION AND MOTIVATION

Here we look at general finite-state stochastic processes absorbing at 0. Their expected times to absorption at 0 are related to the expected “progress” at each state. To illustrate this, consider the following simple Markov chain on $0, \dots, n$ with 0 being an absorbing state, and transitions at states $1, \dots, n$ being simply to move one step to the left with probability ϵ or stay at the present state. If the

* Supported in part by New Zealand Lotteries Board Grant.

process is at state j , the expected distance from 0 after one step is $j - \epsilon$, or in the terminology of this paper, the *directional bias* towards 0 is ϵ . We are interested in relating this measurement of progress to the expected completion times. For this illustration, starting at state i , the expected time to absorption (at 0) is i/ϵ .

It can be shown directly that a relationship like this holds for general finite-state stochastic processes if their expected progress is constant. Surprisingly, where constant progress is not made at each step, optimal bounds on "expected completion times" can be computed in terms of the expected progress (the directional biases) at each state. Note that no conditioning over the complete past is assumed. The results of this paper apply to completely general finite-state stochastic processes absorbing at 0. Special cases include (absorbing at 0) Markov processes and supermartingales.

The paper is organized as follows: Section 2. Notation and Terminology; Section 3. Main Result; Section 4. Relation to Stochastic Optimization; Section 5. Preliminary Results; Section 6. Proof of Main Result; Section 7. Remarks, Open Questions, and Future Work; and an Appendix giving details of an algorithm that computes the best bound.

2. NOTATION AND TERMINOLOGY

Given a sequence of random variables X_0, X_1, \dots , with values in $\{0, 1, \dots, n\}$, representing a general finite-state stochastic process that has 0 as an absorbing state, the *expected completion time* to get to 0 is the expected value of the first k for which $X_k = 0$. The *vector of expected conditional completion times* \mathbf{t} has components t_i , defined as the expected value of the first k for which $X_k = 0$, given that $X_0 = i$ (note $t_0 = 0$). In case $X_0 = i$ is impossible, we set $t_i = 1$.

Matrices of size $(n+1) \times (n+1)$ indexed by the states $\{0, 1, \dots, n\}$ play an important part in our techniques. An important example is the transition matrix for a Markov chain on $\{0, 1, \dots, n\}$. However, we wish to consider a type of "transition matrix" for more general processes, and this motivates the following definitions.

Given X_0, X_1, \dots , the *effective transition matrix* P_k at the k th step has (i, j) th entry equal to $\mathbb{P}[X_k = j | X_{k-1} = i]$, if the conditioning event $X_{k-1} = i$ is possible. Some care is required in assigning the entry in case the conditioning event is impossible. Here we adopt the following convention, which suffices for our calculations: We set the entry equal to $\mathbb{P}[X_l = j | X_{l-1} = i]$ for any l for which the conditioning event $X_{l-1} = i$ has nonzero probability. If no such l exists, the entry is set equal to 1 when $j = 0$ and 0 when $j > 0$. Note that, since the process absorbs at 0, $\mathbb{P}[X_k = 0 | X_{k-1} = 0] = 1$.

Similarly for $k > 1$, let the *effective transition matrices conditional on* $X_0 = l$, $P_{k|l}$, have (i, j) th entry of $\mathbb{P}[X_k = j | X_{k-1} = i \text{ and } X_0 = l]$, if the conditioning event is possible; otherwise it is taken as the (i, j) th entry of P_k . If, for all $k > 1$, $P_k = P_{k|l}$, we say the general process has *effective transitions independent of* $X_0 = l$. If this holds for all l , we say the general process has *effective transitions independent of* X_0 .

The vector of *effective times* is defined by the infinite series $\mathbf{t} = (\mathbf{I} + P_1 +$

$P_1P_2 + P_1P_2P_3 + \dots$) \mathbf{c} , where \mathbf{c} is the column vector of size $n + 1$ consisting of a zero in the first position and then n ones. Note, some components of \mathbf{t} may be ∞ .

For $j \geq 0, k > 0$, let μ_{jk} denote the j th component of the column vector $P_k(0, \dots, n)^T$. Thus, $\mu_{jk} = \mathbb{E}[X_k | X_{k-1} = j]$ whenever the conditioning event is possible (otherwise our convention above defines μ_{jk}). Define, for $k > 0, \epsilon_{jk} = j - \mu_{jk}$ to be the *directional biases* at the k th step. Note that $j - n \leq \epsilon_{jk} \leq j$ and $\epsilon_{0k} = 0$, since $\mathbb{E}[X_k | X_{k-1} = 0] = 0$. We say the finite-state stochastic process is *directionally biased towards 0* if all biases, except for ϵ_{0k} , are positive.

Given a vector ϵ indexed by $j = 1, \dots, n$, with j th component in $[j - n, j]$ for all j —which we will henceforth refer to as a *vector of biases*—let $\mathcal{F}(\epsilon), \mathcal{F}(\geq \epsilon)$ and $\mathcal{F}(\leq \epsilon)$ be the collection of all finite-state stochastic processes absorbing at 0 with, for $j > 0$, biases $\epsilon_{jk} = \epsilon_j, \epsilon_{jk} \geq \epsilon_j$, and $0 < \epsilon_{jk} \leq \epsilon_j$, respectively. Given a family of processes absorbing at 0 \mathcal{F} , $\text{lub}(\mathcal{F})$ denotes the least upper bound of the effective times for all processes in the family (note: vector comparisons are done component-wise). Note that $\mathcal{F}(\epsilon)$, and $\mathcal{F}(\geq \epsilon)$ are always nonempty, and $\mathcal{F}(\leq \epsilon)$ is nonempty provided $\epsilon > 0$, since we can find suitable Markov chains in each of them. To get the smallest ϵ so that a given finite-state stochastic process absorbing at 0 belongs to $\mathcal{F}(\geq \epsilon)$ set $\epsilon_j = \inf_k \{\epsilon_{jk}\}$. Let $\bar{\epsilon}_j = \min_{l \geq j} \{\epsilon_l\}$, and $\mu_j = j - \epsilon_j$.

For a Markov chain, all the effective transition matrices are equal, and only differ from the given transition matrix of the Markov chain at those entries where one is conditioning upon states that are never visited, either initially or at any later stage. In that case, however, we can replace the given transition matrix by any of the effective transition matrices, without altering the process.

Having made this substitution, ϵ_{jk} is independent of k so the process lies in $\mathcal{F}(\epsilon)$ where $\epsilon_j = \epsilon_{j1}$. Denote by $\epsilon(P)$ and $\mathbf{t}(P)$ the vector of biases and expected conditional completion times (resp.) associated with any Markov chain having effective transition matrix P . Note also that $\mu_j = \mathbb{E}[X_k | X_{k-1} = j]$ for all k . The absorption assumption implies that the first row of any transition matrices consists of a 1 followed by zeros.

By $P \in \mathcal{F}$, where \mathcal{F} is a family of finite-state stochastic processes absorbing at 0, we mean that any Markov chain with effective transition matrix P belongs to \mathcal{F} . A particularly simple Markov chain has at most two nonzero entries in each row of the transition matrix. We call such Markov chains *two-outcome* chains.

3. MAIN RESULT

We found an exact formula for the expected (conditional) completion times in terms of the directional biases only in the very special cases of constant biases with effective transitions independent of X_0 . In general the exact times are not simply a function of the biases. The main concern of this paper is to produce the best possible **bounds** on these times in terms of the biases. For general finite-state stochastic processes that are absorbing at 0 and directionally biased towards 0, we have the following.

Theorem 3.1. *Let $\epsilon > 0$ be a vector of biases*

$$(1) \text{lub}(\mathcal{F}(\epsilon)) = \text{lub}(\mathcal{F}(\geq \epsilon)).$$

- (2) $\text{lub}(\mathcal{F}(\epsilon))$ is monotone decreasing in ϵ .
- (3) $\text{lub}(\mathcal{F}(\epsilon))$ can be computed by a $O(n^2)$ algorithm.
- (4) $\text{lub}(\mathcal{F}(\epsilon))_i \leq 1 + 1/\bar{\epsilon}_1 + \cdots + 1/\bar{\epsilon}_{\lfloor \mu_i \rfloor} + (\mu_i - \lfloor \mu_i \rfloor)/\bar{\epsilon}_{\lfloor \mu_i \rfloor + 1}$
 $\leq 1/\bar{\epsilon}_1 + \cdots + 1/\bar{\epsilon}_i$, where $\mu_i = 1 - \epsilon_i$
- (5) If \mathcal{F} is $\mathcal{F}(\epsilon)$ or $\mathcal{F}(\geq \epsilon)$, then there exists a slowest two-outcome Markov chain P^* in \mathcal{F} , absorbing at 0, that has the largest times (for all i) and $\text{lub}(\mathcal{F}) = \mathbf{t}^* = \mathbf{t}(P^*)$. This chain is characterized by $P\mathbf{t}^* \leq P^*\mathbf{t}^*$ for all P in \mathcal{F} .
- (6) If $\epsilon_j = \epsilon$ for $j > 0$, then all processes in $\mathcal{F}(\epsilon)$ have effective times $\mathbf{t} = (\mathbf{i}/\epsilon)$.

Effective times relate to expected completion times as follows:

Corollary 3.1. Let \mathbf{b} be a vector (least) upper bound on the effective times given in the theorem. $\mathbb{E}[b_{X_0}]$ is the (least) upper bound on the expected completion time of the process (X_k) . For a process, with absorbing state 0, that has effective transitions independent of $X_0 = l$, the l th component of the expected conditional completion times and effective times are the same. If the process has effective transitions independent of X_0 , then both vectors are the same.

Although these results are for upper bounds, all the analogous results hold concerning lower bounds. For instance, the greatest lower bound analogue to (4) is

$$\begin{aligned} \text{glb}(\mathcal{F}(\epsilon))_i &\geq 1 + 1/\underline{\epsilon}_1 + \cdots + 1/\underline{\epsilon}_{\lfloor \mu_i \rfloor} + (\mu_i - \lfloor \mu_i \rfloor)/\underline{\epsilon}_{\lfloor \mu_i \rfloor + 1} \\ &\geq 1/\underline{\epsilon}_1 + \cdots + 1/\underline{\epsilon}_i, \quad \text{where } \underline{\epsilon}_j = \max_{l \geq j} \{\epsilon_l\}. \end{aligned}$$

Such results may be useful for bounding the biases if expected completion times are known.

The proof of these main results appears in Section 6; however, a general overview is as follows. Finding $\text{lub}(\mathcal{F}(\epsilon))$ is a nonlinear optimization problem over the set of all possible finite-state stochastic processes in $\mathcal{F}(\epsilon)$. Proposition 5.4 reduces this to a finite-dimensional problem by showing that it is necessary only to look at Markov chains. A further reduction (Proposition 5.7) implies solutions satisfy a finite dimensional linear program and are two-outcome Markov chains. A purely geometric result (Proposition 5.6) shows the existence of the optimal Markov chain (Proposition 5.8). In the Appendix an $O(n^2)$ algorithm is described which finds the required two-outcome chain (Proposition A.1).

4. RELATION TO STOCHASTIC OPTIMIZATION

The motivation for this paper came from studying stochastic algorithms such as simulated annealing [1], threshold accepting [3], and others [6] which try to find the value and location of a global optimum of a function which may have many nonglobal local optima. These algorithms may not always make progress at each step. In fact, retrograde steps often prevent the algorithm from getting stuck at a

local optimum. The question that inspired this work concerned what could be said about a stochastic algorithm that “expects to make progress” at each iteration.

The limiting behavior of stochastic optimization algorithms has been extensively investigated, particularly for methods based on the simulated annealing regime [5, 11, 15]. Some bounding results on the probability of success for simulated annealing are discussed in [7]. However, generally, the problem of determining rates of convergence appears to have received less attention. This paper presents results which may be useful in analysing such problems in a general setting. As in Berg [2], a stochastic algorithm would be regarded as a type of random walk on the objective function’s values rather than on its domain. Now, for a wide class of NP-hard discrete optimization problems, the number of different function values is small compared to the (exponentially growing) cardinality of the domain; furthermore, our bounds on convergence rates are determined by a single vector ϵ , which itself could be bounded by studying **jointly** the “landscape” of the objective function **and** the stochastic algorithm. Thus, estimating the rates of convergence for a problem could be, within bounds, reduced to the problem of determining how these two factors influence ϵ .

Pure adaptive search (PAS) introduced by Smith and Zabinsky [14] provides a goal for optimization algorithms as it has complexity that is “linear in dimension”; however, it is impractical to implement. Stochastic algorithms that “approximate” PAS might be expected to have nice complexity. In [16], we exactly analyzed the behavior of PAS for global optimization on functions with a finite range of size n and showed the expected number of iterations is $\sum_{j=1, \dots, n} 1/j$, which is $O(\log(n))$. When writing [16], we knew that each iteration of PAS on average halved the region known to contain the global optimum. Our intuition was that this expected halving should imply $O(\log(n))$ behavior on average. The main result of this paper provides the technique. The context of optimizing on $\{0, \dots, n\}$, PAS has biases $\epsilon_j = (j+1)/2$. The simple bound in (4) of the main result shows the time is bounded by $\sum_{j=1, \dots, n} 2/(j+1)$, which is $O(\log(n))$.

J_n
A

5. PRELIMINARY RESULTS

This section gives results about times of finite-state stochastic processes. The following straightforward lemma often used for Markov chains that absorb at state 0 [10, Theorem 3.3.5], but applicable in the more general setting is worth noting.

Lemma 5.1. *For processes absorbing at 0, the expected completion time is $\sum_{k \geq 0} \mathbb{P}[X_k \neq 0]$ and the expected conditional completion time is $t_i = \sum_{k \geq 0} \mathbb{P}[X_k \neq 0 | X_0 = i]$.*

Proof. Let $Z_k = 1$ if $X_k \neq 0$ and $Z_k = 0$ otherwise. Now, the expected completion time equals $\mathbb{E}[\sum_{k \geq 0} Z_k] = \sum_{k \geq 0} \mathbb{E}[Z_k] = \sum_{k \geq 0} \mathbb{P}[X_k \neq 0]$. The expected conditional completion times are computed similarly. \square

Effective times for a general process with effective transitions P_k are precisely the conditional times for the nonstationary Markov process that has P_k as the

transition matrix at step k . The following proposition establishes their use in computing expected completion times.

Proposition 5.1. *Consider a general finite-state stochastic process absorbing at 0 with effective times \mathbf{t} . The expected completion time is the expected value of the effective times with respect to the distribution of X_0 , i.e., $(\pi_j) \cdot \mathbf{t}$, where $\pi_j = \mathbb{P}[X_0 = j]$. If the process has transitions independent of $X_0 = i$, then t_i is the expected conditional completion time given $X_0 = i$.*

Proof. Let P_k be the effective transition matrix for the given general process at the k th step. Recall that the effective time is $\mathbf{t} = (I + P_1 + P_1P_2 + P_1P_2P_3 + \cdots)\mathbf{c}$.

For any sequence of random variables taking values in the set $\{0, \dots, n\}$, we have

$$\mathbb{P}[X_k = j] = \sum_{m=0, \dots, n} \mathbb{P}[X_k = j | X_{k-1} = m] \mathbb{P}[X_{k-1} = m],$$

which can be rewritten as the (row) vector equality

$$(\mathbb{P}[X_k = j]) = (\mathbb{P}[X_{k-1} = m])P_k,$$

and by induction this gives

$$(\mathbb{P}[X_k = j]) = (\mathbb{P}[X_0 = m])P_1 \cdots P_k.$$

Since $(\mathbb{P}[X_0 = m])P_1 \cdots P_k \mathbf{c}$ gives the probability the k th step of the process is different from 0, by Lemma 5.1 the infinite sum gives the expected completion time which equals $(\mathbb{P}[X_0 = j]) \cdot \mathbf{t}$.

Similarly, the expected conditional completion time when $X_0 = i$ is

$$\mathbf{e}_i(I + P_1 + P_1P_{2|i} + P_1P_{2|i}P_{3|i} + \cdots)\mathbf{c},$$

where \mathbf{e}_i is the row with 1 in the i th position. If the effective transitions are independent of $X_0 = i$, $P_{k|i} = P_k$ for all $k \geq 2$ so t_i is the expected conditional completion time.

5.1. Facts about Markov Chains

If there are negative or zero biases, it may be possible for a chain to have infinite expected times. The following characterizes Markov chains absorbing at 0 with infinite times.

Proposition 5.2. *Given a vector of biases $\boldsymbol{\epsilon}$, there exists a Markov chain in $\mathcal{F}(\boldsymbol{\epsilon})$ with unbounded time if and only if there exist $0 < i \leq j$ such that both μ_i and μ_j belong to $[i, j]$ (i.e. $\epsilon_i \in [i - j, 0]$ and $\epsilon_j \in [0, j - i]$).*

Proof. If such i and j exist, it is possible to describe a Markov chain with the given biases that has $\{i, j\}$ as an ergodic set. Let the distribution for X_0 be uniform, and let the transitions out of states i and j be only i and j themselves (the probabilities assigned to produce μ_i and μ_j). Clearly t_i and t_j are infinite and the vector of times is unbounded. Conversely, if a Markov chain has infinite

expected time to absorption by 0, there must be some ergodic set of states S not containing 0. Let i and j be the smallest and biggest states in S . \square

For the rest of this paper we assume $\epsilon > 0$, thus the classes $\mathcal{F}(\epsilon)$ and $\mathcal{F}(\geq \epsilon)$ will consist only of finite-state stochastic processes absorbing at 0 and directionally biased towards 0.

The expected conditional completion times for a Markov chain are easily characterized in terms of the transition matrix [10].

Proposition 5.3. *For a Markov chain P , absorbing at 0, $\mathbf{t}(P)$ equals $(I + P + P^2 + \dots)\mathbf{c}$ and satisfies the recursion $\mathbf{t}(P) = \mathbf{c} + P\mathbf{t}(P)$. Conversely, if $\mathbf{t} < \infty$, $t_0 = 0$ and $\mathbf{t} = \mathbf{c} + P\mathbf{t}$, then $\mathbf{t} = \mathbf{t}(P)$.*

Proof. From [10], the i th entry of $P^k\mathbf{c}$ gives the probability the k th step of the chain is different from 0 having started with $X_0 = i$, so, by Lemma 5.1, the series gives the expected conditional completion times. Clearly $\mathbf{t}(P)$ satisfies the recursion. Conversely, for $\mathbf{t} < \infty$ and $t_0 = 0$, $\mathbf{t} = \mathbf{c} + P\mathbf{t}$ gives

$$\mathbf{t} = \mathbf{c} + P\mathbf{t} = \mathbf{c} + P\mathbf{c} + P^2\mathbf{t} = \mathbf{c} + P\mathbf{c} + P^2\mathbf{c} + P^3\mathbf{t} = \dots$$

So the partial sums in the series for $\mathbf{t}(P)$ are all bounded above by \mathbf{t} . Thus $\mathbf{t}(P) < \infty$ and $P^k(\mathbf{c}) \rightarrow \mathbf{0}$ as $k \rightarrow \infty$. Now, for some $\alpha > 0$, $0 \leq \mathbf{t} < \alpha\mathbf{c}$, so $P^k\mathbf{t} \rightarrow \mathbf{0}$ as $k \rightarrow \infty$ and $\mathbf{t} = \mathbf{t}(P)$. \square

5.2. Reduction to Markov Chains

It is possible to test a candidate for the least upper bound on the effective times of a collection of finite-state stochastic processes. The following provides a sufficient condition for a Markov chain to be the least upper bound.

Proposition 5.4. *Let \mathcal{F} be a family of finite-state stochastic processes absorbing at 0. If there exists a Markov chain P^* in \mathcal{F} with $\mathbf{t}^* = \mathbf{t}(P^*)$ and, for all Markov chains P in \mathcal{F} , $P\mathbf{t}^* \leq P^*\mathbf{t}^*$, then \mathbf{t}^* is the least upper bound of the effective times for all general finite-state stochastic processes having each of their effective transitions P_k in \mathcal{F} .*

Proof. Given such a P^* , take a general process having effective transitions in \mathcal{F} with effective conditional completion times \mathbf{t} . Define $\mathbf{t}_0 = \mathbf{t}^*$ and for $k > 0$,

$$\begin{aligned} \mathbf{t}_k &= (I + P_1 + P_1P_2 + \dots + P_1 \dots P_k + P_1 \dots P_kP^* + P_1 \dots P_k(P^*)^2 + \dots)\mathbf{c} \\ &= (I + P_1 + P_1P_2 + \dots + P_1 \dots P_{k-1})\mathbf{c} + P_1P_2 \dots P_k\mathbf{t}^* . \end{aligned}$$

We now show $\mathbf{t}^* = \mathbf{t}_0 \geq \mathbf{t}_1 \geq \mathbf{t}_2 \leq \dots$. To show $\mathbf{t}_{k+1} \leq \mathbf{t}_k$ we use $P_{k+1}\mathbf{t}^* \leq P^*\mathbf{t}^*$, which implies $P_1 \dots P_kP_{k+1}\mathbf{t}^* \geq P_1 \dots P_kP^*\mathbf{t}^*$. Thus

$$\begin{aligned} \mathbf{t}_{k+1} &= (I + P_1 + P_1P_2 + \dots + P_1 \dots P_k)\mathbf{c} + P_1P_2 \dots P_{k+1}\mathbf{t}^* \\ &\leq (I + P_1 + P_1P_2 + \dots + P_1 \dots P_k)\mathbf{c} + P_1P_2 \dots P_kP^*\mathbf{t}^* \\ &= (I + P_1 + P_1P_2 + \dots + P_1 \dots P_{k-1})\mathbf{c} + P_1P_2 \dots P_k\mathbf{t}^* = \mathbf{t}_k \end{aligned}$$

since $\mathbf{c} + P^*\mathbf{t}^* = \mathbf{t}$ by Proposition 5.3.

Now the partial sums in the series for \mathbf{t} satisfy $(I + P_1 + P_1P_2 + \dots + P_1P_2 \dots P_{k-1})\mathbf{c} \leq \mathbf{t}_k \leq \mathbf{t}^*$, so $\mathbf{t} \leq \mathbf{t}^*$. So we get that \mathbf{t}^* is an upper bound. Since P^* , together with the uniform distribution for X_0 in such a process, \mathbf{t}^* is the least upper bound. \square

If all biases are equal to a constant, a modification of the above proof gives the following exact result, which is reminiscent of a version of Wald's equation (c.f. [8] Vol. 9, p. 522).

Proposition 5.5. *Given a general finite-state stochastic process absorbing at 0, such that for all $k > 0, j > 0, \epsilon_{jk} = \epsilon > 0$, then the vector of effective times is (\mathbf{i}/ϵ) , the same as the conditional completion times for the Markov chain absorbing at 0 with constant biases ϵ . In particular the expected completion time of the general process is $\mathbb{E}[X_0]/\epsilon$.*

Proof. Let $\mathbf{t}^* = (\mathbf{i}/\epsilon)$. For any Markov chain P with absorbing state 0 and $\epsilon(P) = (\epsilon, \epsilon, \dots, \epsilon)$ we have $\epsilon P\mathbf{t}^* = P(\mathbf{0} \mathbf{1} \dots \mathbf{n})^T = (\mu_i)^T = (\mathbf{0}, \mathbf{1} - \epsilon, \dots, \mathbf{n} - \epsilon)^T$ so $P\mathbf{t}^* = \mathbf{t}^* - \mathbf{c}$. Let P^* be a given Markov chain with these biases. By Proposition 5.3, $\mathbf{t}^* = \mathbf{t}(P^*)$. Let t and t_k be as in the previous proof. So all the effective transition matrices have $\epsilon(P_k) = (\epsilon, \epsilon, \dots, \epsilon)$. Since, for all $k, P_k\mathbf{t}^* = P^*\mathbf{t}^* = \mathbf{t}^* - \mathbf{c}$, the proof yields $\mathbf{t}^* = \mathbf{t}_0 = \mathbf{t}_1 = \dots \leq \mathbf{t}$. Since $\mathbf{t}^* < \infty$, then $\mathbf{t} < \infty, P_1 \dots P_k \mathbf{c} \rightarrow \mathbf{0}$ as $k \rightarrow \infty$, so $P_1 \dots P_k \mathbf{t}^* \rightarrow \mathbf{0}$ as $k \rightarrow \infty$ and $\lim_{k \rightarrow \infty} \mathbf{t}_k = \mathbf{t}$. Thus, $\mathbf{t}^* = \mathbf{t}$. Proposition 5.1 gives the completion time. \square

5.3. Reduction to Two-Outcome Markov Chains

Convexity plays a key role for the results in this section. Convexity results like the following are often used in statistical estimation [9] and to find optimal stopping rules for a Markov decision process [4, 12, 13]. Two-outcome Markov chains are used in describing birth and death processes in [4, p. 155].

We require a purely geometric result. A vector $\mathbf{t} = (t_i)$ can be viewed as a function on $\{0, \dots, n\}$. The *least concave upper envelope* h of \mathbf{t} is the concave piecewise linear function $h: [0, n] \rightarrow \mathbb{R}$, with $h(i) \geq t_i$, is less than or equal to all other such envelopes. Its *nodes* are those numbers in $\{0, \dots, n\}$ below the endpoints of the line segments making up the function's graph. Given a vector of biases ϵ (and hence $\mu_i = i - \epsilon_i$), the vector \mathbf{t} that satisfies $t_i = 1 + h(\mu_i)$ is the key to constructing the slowest Markov chain.

Figure 1 illustrates a geometric fact that the required vector \mathbf{t}^* and its upper envelope h satisfy. The equation $t_i = 1 + h(\mu_i)$ means that the right triangle with base length ϵ_i and height 1 touching the upper envelope at vertex $(\mu_i, h(\mu_i))$ lies under or on the upper envelope, and if i is a node of h , touches, at its top vertex as well.

Proposition 5.6. *Given a vector of biases $\epsilon > 0$, there exists a vector $\mathbf{t}^* = (t_i)$ with least concave upper envelope h such that $t_0 = 0$ and, for $i > 0, t_i = 1 + h(\mu_i)$. Additionally $t_i \leq 1 + 1/\bar{\epsilon}_1 + \dots + 1/\bar{\epsilon}_{\lfloor \mu_i \rfloor} + (\mu_i - \lfloor \mu_i \rfloor)/\bar{\epsilon}_{\lfloor \mu_i \rfloor + 1} \leq 1/\bar{\epsilon}_1 + \dots + 1/\bar{\epsilon}_i$. Furthermore, h is increasing for $x \leq \max(\mu_i)$.*

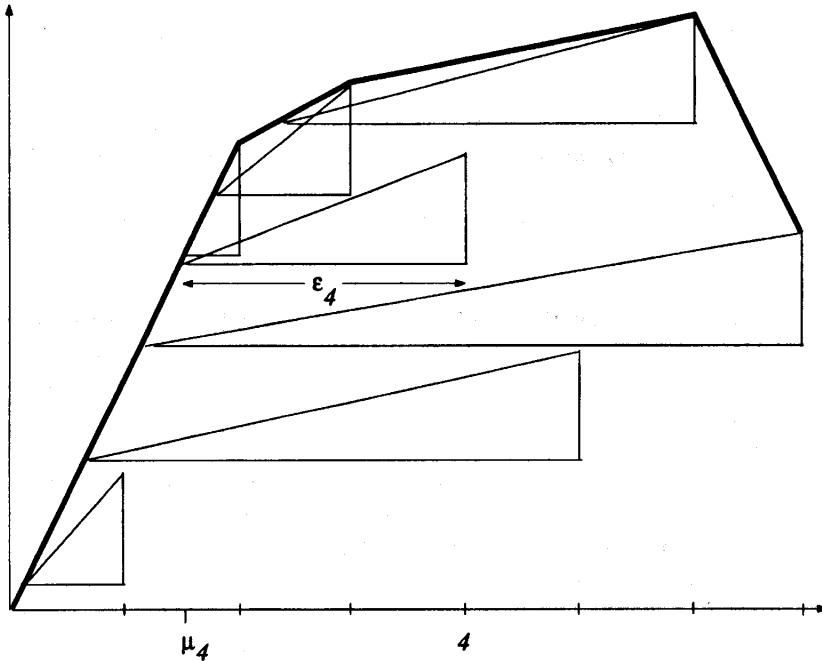


Fig. 1. Geometric meaning of $t_i = 1 + h(\mu_i)$.

Proof. The basic idea of the proof is to approximate the vector and envelope from above by a succession of approximations found by a fixed point iteration process. We start with h_0 and find \mathbf{t}_1 using the key equation, its least concave upper envelope gives h_1 , which in turn gives \mathbf{t}_2 and so on.

Let h_0 be the least concave upper envelope of $(\sum_{j=1, \dots, i} 1/\bar{\epsilon}_j)$. Note that this a piecewise linear function with $h_0(0) = 0$ with slopes of $1/\bar{\epsilon}_j$ on $[j - 1, j]$. Clearly $h_0 \geq 0$. Define \mathbf{t}_1 by $(\mathbf{t}_1)_i = 1 + h_0(\mu_i)$ for $i > 0$ and $(\mathbf{t}_1)_0 = 0$. As $h_0 \geq 0$ so $\mathbf{t}_1 \geq \mathbf{c}$.

From the definition of $\bar{\epsilon}_i$, observe that $\bar{\epsilon}_i \leq \epsilon_i$ and $\bar{\epsilon}_1 \leq \dots \leq \bar{\epsilon}_n$. This implies that the slope $1/\epsilon_i$ of the hypotenuse of the triangle with vertices $(\mu_i, h_0(\mu_i))$, $(i, h_0(\mu_i))$ and $(i, (\mathbf{t}_1)_i)$ is less than or equal to the slopes $\{1/\bar{\epsilon}_1, 1/\bar{\epsilon}_2, \dots, 1/\bar{\epsilon}_i\}$ of all line segments of the graph of h_0 to the left of i . Geometrically this means all such triangles are under the graph, so $\mathbf{t}_1 \leq (h_0(i))$.

Now let h_1 be the least concave upper envelope of \mathbf{t}_1 . Clearly $h_1 \geq 0$. Define \mathbf{t}_2 by $(\mathbf{t}_2)_i = 1 + h_1(\mu_i)$ for $i > 0$ and $(\mathbf{t}_2)_0 = 0$. Observe $\mathbf{t}_1 \leq (h_1(i)) \leq (h_0(i))$ as h_0 is a concave upper envelope, so $\mathbf{t}_2 \leq \mathbf{t}_1 \leq (h_1(i))$.

We get a (possibly infinite) sequence of decreasing \mathbf{t}_i and h_i bounded below by \mathbf{c} , so the limits \mathbf{t}^* and h exist and satisfy the required equation $t_i = 1 + h(\mu_i)$. Note that h is the piecewise linear extension of $\{h(i) | i = 0 \dots n\}$. As the h_i are concave, it follows easily that h is concave. It is the least concave upper envelope of \mathbf{t} , since assuming a smaller concave upper envelope gives a contradiction. The required inequality is just $\mathbf{t}^* \leq \mathbf{t}_1 \leq (h_0(i))$.

As h is concave, to show that h is increasing over the required region, it suffices to show the slope at $\max(\mu_i)$ is positive. Let $\mu_j = \max(\mu_i)$ say. Since $t_j = 1 + h(\mu_j)$ and $t_j \leq h(j)$, the triangle with vertices $(\mu_j, h_0(\mu_j))$, $(j, h_0(\mu_j))$, and

$(j, (t_1)_j)$ is under the graph and the slope at μ_j exceeds the slope $1/\epsilon_j$ of its hypotenuse. □

Proposition 5.7. *Given a function $t = (t_i)$ on $\{0, \dots, n\}$, let $h: [0, n] \rightarrow \mathcal{R}$ be the least concave upper envelope of t . Of all distributions p on $0, \dots, n$ having mean $\mu \in [0, n]$, p_{\max} with all its weight at the single nodes of h to either side of μ maximizes $\mathbb{E}[t | p]$. For this distribution, $\mathbb{E}[t | p_{\max}] = h(\mu)$. (Note that if μ is a node of h , all the mass will be at one point).*

Proof. We first show that the maximizing distribution has at most two nonzero probabilities. The problem $\max_p \mathbb{E}[t | p]$ is the linear program $\max_{(p_i)} \sum_{i=0, \dots, n} t_i p_i$ subject to $p_i \geq 0$, $\sum p_i = 1$, and $\sum i p_i = \mu$. The two independent equality constraints eliminate two variables; thus the problem is $(n - 1)$ -dimensional. So the maximum will occur at a vertex which must satisfy at least $n - 1$ of the inequality constraints with equality. Thus at most two of the inequalities are strictly satisfied and only distributions with at most two nonzero masses need be considered. For such a distribution having nonzero probabilities at j and k , $k \geq \mu \geq j$, the expected value of t is the intersection of the line from (j, t_j) and (k, t_k) with the vertical line $x = \mu$. Thus $\mathbb{E}[t | p_{\max}] = h(\mu)$ and any other at most two point distribution will have a no larger expected value. □

The slowest process absorbing at 0 with a given ϵ can be found now.

Proposition 5.8. *Given a vector of biases $\epsilon > 0$. Let h and t^* be as in Proposition 5.6. Let P^* with $\epsilon(P^*) = \epsilon$ be the transition matrix of the two-outcome Markov chain with rows having nonzero probabilities determined by the nodes of h as in Proposition 5.7. P^* gives the slowest finite-state stochastic process in $\mathcal{F}(\geq \epsilon)$ and $\mathcal{F}(\epsilon)$, $t(P^*) = t^* = (t_i^*)$ and t^* is the least upper bound for those families.*

Proof. Proposition 5.6 gives properties of h and t^* . We have, for any $P \in \mathcal{F}(\epsilon)$, $Pt^* \leq P^*t^*$, by using Proposition 5.7 on the rows. Take $\epsilon' \geq \epsilon$ and denote $\mu'_i = i - \epsilon'_i$. Now, if $P \in \mathcal{F}(\epsilon')$, then by Proposition 5.7 again $Pt^* \leq (h(\mu'_1)) \leq (h(\mu_1)) = P^*t^*$, since h is increasing for $x \leq \max(\mu_i)$. Since $P^*t^* = (h(\mu_1)) = t^* - c$ and $t(P^*) < \infty$ from Proposition 5.2, Proposition 5.3 gives $t(P^*) = t^*$. By Proposition 5.4 with $\mathcal{F} = \mathcal{F}(\epsilon)$ or $\mathcal{F} = \mathcal{F}(\geq \epsilon)$, the result follows. □

Note, the above proposition implies that h and t^* of Proposition 5.6 are unique. Also for $\mathcal{F} = \mathcal{F}(\epsilon)$ or $\mathcal{F} = \mathcal{F}(\geq \epsilon)$ the converse of Proposition 5.4 holds.

6. PROOF OF THE MAIN RESULTS—THEOREM 3.1 AND COROLLARY 3.1

Only those propositions relating to upper bounds have been presented in the preceding section. All the analogous results hold concerning lower bounds.

Proof. In the theorem (1) and (4) follow from Propositions 5.6 and 5.8, and (2) follows from (1). (5) follows from Propositions 5.4 and 5.8. (6) is Proposition 5.5

(3) follows from Proposition A.1 in the Appendix. The corollary follows from Proposition 5.1. □

7. REMARKS, OPEN QUESTIONS AND FUTURE WORK

In special cases simpler description of the bounds are possible.

- If $\epsilon_j \leq \epsilon_{j+1}$, then $\bar{\epsilon}_j = \epsilon_j$.
- If $0 < \epsilon_j \leq 1$, then the nodes n_i of h are easily described: n_1 is the biggest index giving the smallest bias; n_2 is the biggest index giving the next smallest bias past n_1 and so on. So $\bar{\epsilon}_1 = \dots = \bar{\epsilon}_{n_1} = \epsilon_{n_1} > \bar{\epsilon}_{n_1+1} = \dots = \bar{\epsilon}_{n_2} = \epsilon_{n_2} \dots$. Thus, for $n_k \leq i \leq n_{k+1}$, $t_i = n_1/\epsilon_{n_1} + (n_2 - n_1)/\epsilon_{n_2} + \dots + (i - n_k)/\epsilon_{n_{k+1}} + (\epsilon_{n_{k+1}} - \epsilon_i)/\epsilon_{n_{k+1}}$.
- In particular if $0 < \epsilon_j \leq 1$ and $\epsilon_j < \epsilon_{j+1}$, the nodes of h are $\{0, \dots, n\}$ and $t_i = h(i) = 1/\epsilon_1 + \dots + 1/\epsilon_i$.

Our original approach was an attempt to find the maximum of the conditional times subject to the constraint (X_k) being in $\mathcal{F}(\epsilon)$. Other bounds could be found if weaker constraints were used. In particular, considering first moments, we have

$$\mathbb{E}[X_k] = \mathbb{E}[X_{k-1}] - \sum_{j=1}^n \epsilon_{jk} \mathbb{P}[X_{k-1} = j]$$

and $\lim_{p \rightarrow \infty} \sum_{k=1 \dots p} (\mathbb{E}[X_k] - \mathbb{E}[X_{k-1}]) = -\mathbb{E}[X_0]$ so we get the identity

$$\sum_{k>0} \sum_{j=1}^n \epsilon_{jk} \mathbb{P}[X_{k-1} = j] = \mathbb{E}[X_0].$$

(Note that this identity and Lemma 5.1 gives an alternative proof of the result in Proposition 5.5 that $t = \mathbb{E}[X_0]/\epsilon$ in case $\epsilon_{jk} = \epsilon$ for $k > 0$ and $j > 0$.) By considering second-order quantities (such as $\mathbb{E}[X_k^2]$) a further (inequality) constraint arises. By looking at simple examples, these first- and second-order constraints produce weaker bounds than those provided by Corollary 3.1.

The techniques we have used are tantalizingly similar to methods used in [4, 13] to find optimal stopping rules for Markov decision processes. In this paper we find an extreme process for the fixed rule of stopping when first absorbed. Perhaps there is some duality between the two problems.

Can these results be generalized by defining biases with respect to other weighting of states? Can these results be extended to general processes with more than one absorbing/ergodic states? What is the analog for continuous state stochastic processes?

APPENDIX—ALGORITHMIC CONSTRUCTION OF A SLOWEST TWO-OUTCOME MARKOV CHAIN

Proposition 5.6 proves the existence (as the limit of a possibly infinite fixed point iteration procedure) of t^* needed to construct the slowest two-outcome Markov

chain. It is possible to find this vector via a direct algorithm, which we now describe. Given a vector of biases $\epsilon > 0$, consider the following construction:

Algorithm A.1.

Initialize:

$$t_0 = 0, \quad \mu_0 = 0, \quad n_0 = 0, \quad \text{and} \quad k = 1$$

$$t_i = 1, \quad \mu_i = i - \epsilon_i \quad \text{for} \quad i > 0.$$

Loop:

Define the next node, next slope, the new part of h , and new t_i where μ_i is in the region to the next node.

$$\text{Let } S_k = \{1/\epsilon_j \mid \mu_j > n_{k-1}\}.$$

$$\text{Let } C_k = \{(t_j - t_{n_{k-1}})/(j - n_{k-1}) \mid j > n_{k-1} \text{ and } \mu_j \leq n_{k-1}\}.$$

$$\text{Let } s_k = \max(S_k \cup C_k).$$

Let n_k be the biggest of the indices j where the maximum is attained.

Let h^k be the piecewise linear function with $h^k(0) = 0$ having nodes at n_0, \dots, n_{k-1}, n and slopes s_1, \dots, s_k .

Let $t_i = 1 + h^k(\mu_i)$ for those i with $n_{k-1} < \mu_i \leq n_k$.

Pictorially, the algorithm "floats" right triangles (as shown in Fig. 2 of heights 1 and base ϵ_j extending left from $(j, 0)$, upwards in a controlled way while forming the least concave upper envelope by bending down approximations to it at successively found nodes.

Initially all the triangles are on the axis and are floated upwards until the lines through their hypotenuses contain $(0, 0)$. The top most of these lines becomes the graph of h^1 , and the node n_1 is the index of the right most triangle touching this line.

More generally at the $(k + 1)$ th step the triangles are floated further upwards until either the triangle's left vertex hits the graph of h^k to the left of n_k , or $(n_k, h^k(n_k))$ is on the line containing its hypotenuse. The topmost of the rays from $(n_k, h^k(n_k))$ to the tops of the triangles becomes the new part of h^{k+1} and n_{k+1} is the index of the right most triangle with top vertex on this ray.

Proposition A.1. *The above algorithm terminates after a finite number of steps $m \leq n$ with $n_m = n$ and requires $O(n^2)$ operations. The function $h(x) = h^m(x)$ is the least concave upper envelope of $\mathbf{t}^* = (\mathbf{0}, \mathbf{t}_1, \dots, \mathbf{t}_n)$ and $t_i = 1 + h(\mu_i)$.*

Proof. Clearly at most n inductive steps are needed and the last node $n_m = n$. At each step to find the next slope at most n values need to be inspected and possibly redefined, so the total number of operations is $O(n^2)$. Note that h^k and $h = h^m$ agree on $[0, n_k]$. So clearly $t_i = 1 + h(\mu_i)$.

First examine the situation at the end of the k th step when a new node n_k has been found, s_k determined, and h^k extended from h^{k-1} .

The key geometric fact to note is that $S_k \cup C_k$ consists of the slopes of the chords from $(n_{k-1}, h^{k-1}(n_{k-1}))$ to the tops of the triangles. S_k relates to those triangles with left vertices bigger than n_{k-1} and C_k relates to the rest.

The node n_k is the biggest index of the maximum value in $S_k \cup C_k$ by construction. So s_k is bigger than or equal to all values in $S_k \cup C_k$ (i.e., the tops of

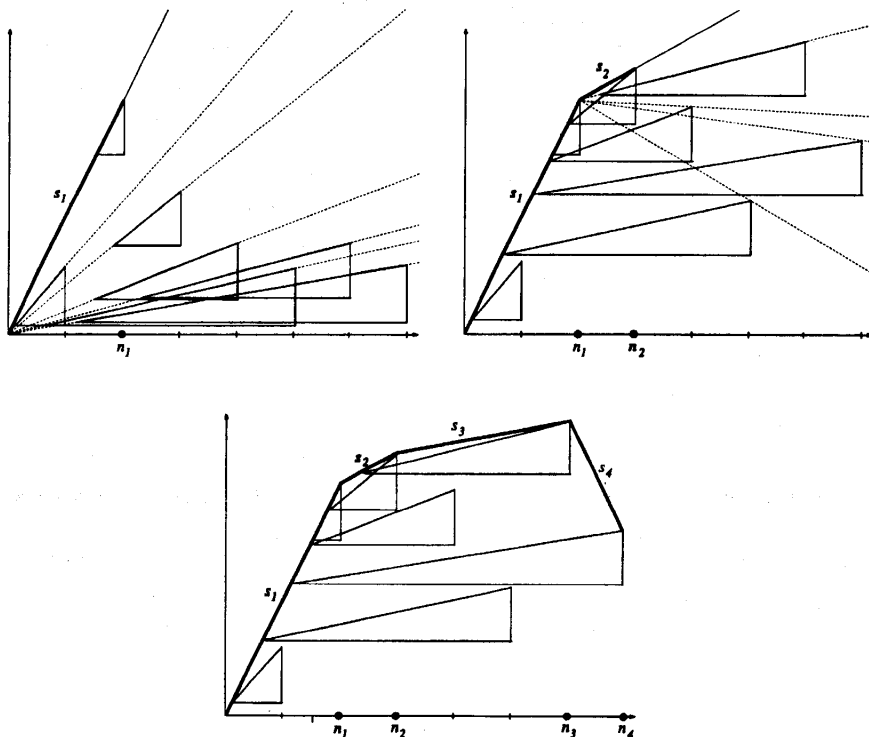


Fig. 2. Floating the triangles skyward.

all the triangles are on or below the graph of h^k) and *strictly* bigger than those values indexed by $j > n_k$ (i.e., these tops are strictly below the graph).

In the algorithm new t_i are defined if $\mu_i \in (n_{k-1}, n_k]$. Geometrically this corresponds to further floating the triangles upward [i.e., (i, t_i) is the top vertex of the triangles that now have left vertex touching the graph of h^k to the left of n_k].

At the new node t_{n_k} is defined and $t_{n_k} = h^k(n_k) = h(n_k)$, as the triangle with vertex over $(n_k, 0)$ determined it.

We now show that, for $j \leq n_k$, $t_j \leq h^k(j)$ (i.e., the tops of the corresponding triangles lie on or below the graph). If for some $j \leq n_k$ this is not the case, then $\mu_j > n_{k-1}$ (as otherwise t_j would be the height of the triangle before it was further raised, which was on or below the graph). So the line containing the hypotenuse of the corresponding triangle before it was just raised contains $(n_{k-1}, h^{k-1}(n_{k-1}))$ and so its slope is $\leq s_k$. Thus when the triangle is raised its top t_j stays on or below the graph of h^k . This means that for $j \leq n_k$ that $t_j \leq h(j)$. This means h is an upper envelope.

A similar argument shows that the tops of the triangles for $j > n_k$ lie strictly below the graph of h^k . So all slopes described in $S_{k+1} \cup C_{k+1}$ are strictly less than s_k . So $s_{k+1} < s_k$ which means that h is concave.

Clearly h agrees with t^* at the nodes, so h is the least concave upper envelope. □

A simple modification of the algorithm for finding the best bounds usually works much better than $O(n^2)$. This version is $O(n\beta \log n)$ if there exists a β such that for each $n > 0$, all entries of ϵ are positive and less than $\beta \log n$. We suspect a version of the algorithm exists that is this order for all cases. A matlab implementation of the algorithm is available from the first author.

For some ϵ that have negative entries, all finite-state stochastic processes with these biases converge in finite time. The existence proof of Proposition 5.6 works in these cases if an appropriate starting upper envelope for h_0 is found. Future work relates to finding a computational algorithm for general vector of biases ϵ .

ACKNOWLEDGMENTS

This work was motivated by discussions with Graham Wood and Zelda Zabinsky and follows on from our joint work. Special thanks to Graham for his comment "it sounds like convexity to me" and reference to G. Kelly's thesis [9]. Thanks to Bill Taylor for the picture that inspired Proposition 5.7. Thanks to Murray Smith for his helpful comments, and John Hannah for his help with the figures. We thank the referee who corrected the proof of Proposition 5.3.

REFERENCES

- [1] S. Antily and A. Federgruen, Simulated annealing methods with general acceptance probabilities, *J. Appl. Probab.* **24**, 657–667 (1987).
- [2] B.A. Berg, Locating global minima in optimization problems by a random-cost approach, *Nature*, **361**, 708–710 (1993).
- [3] G. Dueck and T. Scheuer, Threshold accepting: a general purpose optimization algorithm appearing superior to simulated annealing, *J. Comput. Phys.* **90**, 161–175 (1990).
- [4] E.B. Dynkin and A. A. Yushkevich, *Markov Processes—Theorems and Problems*, Plenum, New York, 1969.
- [5] U. Faigle and W. Kern, Note on the convergence of simulated annealing algorithms, *SIAM J. Control Optim.* **29**, 153–159 (1991).
- [6] U. Faigle and W. Kern, Some convergence results for probabilistic tabu search, *ORSA J. Comput.*, **4**(1), 33–37 (1992).
- [7] A. Ferreira and J. Zerovnik, Bounding the probability of success of stochastic methods for global optimization, *Comput. Math. Appl.* **25**, 1–8 (1993).
- [8] N. L. Johnson, and S. Kotz, *Encyclopedia of Statistical Sciences*, Wiley, New York, 1988.
- [9] G. R. Kelly, Three optimization problems in mathematical statistics, Master Thesis, University of Canterbury, 1980.
- [10] J. G. Kemeny and J. L. Snell, *Finite Markov Chains*, Springer-Verlag, New York, 1983.
- [11] M. Lundy and A. Mees, Convergence of the annealing algorithm, *Math. Prog.*, **34**, 111–124 (1986).

- [12] A.S. Manne, Linear programming and sequential decisions, *Manage. Sci.*, **6**(3), 259–267 (1960).
- [13] S.M. Ross, *Applied Probability Models with Optimization Applications*, Holden-Day, 1970.
- [14] R. Smith and Z. Zabinsky, Pure adaptive search in global optimization, *Math. Prog.*, **53**, 323–338 (1992).
- [15] F. J. Solis and R. J. B. Wets, Minimization by random search techniques, *Math. Op. Res.* **6**(1), 19–30 (1981).
- [16] Z. Zabinsky, G. R. Wood, M. Steel, and W. Baritomba, Pure adaptive search for finite global optimization, *Math Prog.*, **69**(3), 443–448 (1995).

Received January 4, 1995

Revised December 21, 1995

Accepted March 5, 1996