Available online at www.sciencedirect.com

ScienceDirect

Operations Research Letters 35 (2007) 573 – 580

Operations Research Letters

www.elsevier.com/locate/orl

Convergence properties of the cross-entropy method for discrete optimization

Andre Costa[a,*], Owen Dafydd Jones[b], Dirk Kroese[c]

[a]Centre of Excellence for Mathematics and Statistics of Complex Systems, University of Melbourne, 3010, Australia
[b]Department of Mathematics and Statistics, University of Melbourne, 3010, Australia
[c]Department of Mathematics, University of Queensland, 4072, Australia

Received 12 May 2006; accepted 20 November 2006
Available online 28 December 2006

Abstract

We present new theoretical convergence results on the cross-entropy (CE) method for discrete optimization. We show that a popular implementation of the method converges, and finds an optimal solution with probability arbitrarily close to 1. We also give conditions under which an optimal solution is generated eventually with probability 1.

Keywords: Cross-entropy method; Discrete optimization; Stochastic search

1. Introduction

The cross-entropy (CE) method was originally developed as an adaptive importance sampling scheme for estimating rare event probabilities via simulation. However, it was soon realized that the CE method could also be applied to a variety of optimization problems. The reader is referred to Rubinstein and Kroese [9] for a comprehensive overview and history of the CE method. In this paper, we focus on its application to *discrete optimization* problems, in which some objective function is maximized. We consider the deterministic setting, where exact objective function values are available, and where stochastic effects are introduced exclusively in the generation of candidate solutions, as follows. The CE method involves an iterative procedure consisting of two steps: *Step* 1: a random sample of candidate solutions is generated according to a parameterized probability distribution; and *Step* 2: the candidate solutions generated in Step 1 are evaluated using the objective function, and the parameters of the sampling distribution are updated in a manner which *increases* the probability that the best solutions found at the current iteration will occur in the next iteration.

* Corresponding author. Tel.: +61 03 8344 1619; fax: +61 3 8344 4599.

E-mail addresses: acosta@ms.unimelb.edu.au (A. Costa), O.D.Jones@ms.unimelb.edu.au (O.D. Jones), kroese@maths.uq.edu.au (D. Kroese).

0167-6377/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.orl.2006.11.005

Existing results on the convergence of the CE method for discrete optimization appear in [9,6], for a special case known as the “elite sample” version, whereby the sampling distribution is forced to favour the best solution obtained over *all* previous iterations, up to and including the current iteration. This differs from the more commonly used version of the CE method [9], which we study in this paper, whereby only the best solutions found at the *current* iteration are reinforced. As a result of this important difference, our convergence analysis requires a different technique to that employed in [9,6]. Given the common usage of the CE algorithm analysed in this paper, the convergence results presented here are of significant interest to practitioners and theoreticians of the CE method.

Our main contribution concerns the typical scenario where a constant “smoothing” parameter is used to update the sampling distribution; we show that in this case the CE method converges to an optimal solution, in the sense that the sampling distribution converges with probability 1 to a unit mass (via the convergence of its parameters), and that the probability that an optimal solution is found can be made arbitrarily close to 1 (at the expense of the rate of convergence of the sampling distribution). We note that the convergence properties of the CE method with a constant smoothing parameter have not been considered in any previous study. We also extend the methods of [9,6] to derive more general and easily checkable necessary conditions and sufficient conditions under which an optimal solution is generated eventually with probability 1, a property that can only be achieved by using a sequence of decreasing (as opposed to constant) smoothing parameters. We note that our methods of proof are independent of the objective function; as such, our results are quite general, but on the other hand, they do not yield explicit information regarding the sequence of objective function values that are generated by the algorithm.

The CE method can be placed within a broad group of stochastic search methods that includes the well-known simulated annealing [1], genetic algorithms [3], the method of Andradóttir [2] and many others (see [8,10] for recent surveys). In particular, a key feature of the CE method is that it is *model-based*, due to the fact that the algorithm revolves around the updating of a parameterized sampling distribution, which carries information about the best candidate solutions from one iteration to the next. In this respect, the CE method is most similar to estimation of distribution algorithms [5], which are also model-based. In contrast, *population-based* methods such as simulated annealing and genetic algorithms operate directly on a population of candidate solutions. It is not our aim to perform a comparative study of the CE method with alternative stochastic optimization techniques, nor is it our claim that the CE method is necessarily superior. Instead, our aim is to establish new theoretical results concerning its limiting properties. The reader is directed to [9] for extensive numerical experiments using the CE method.

The paper is structured as follows. In Section 2, we set up a discrete optimization framework, and present a generic CE algorithm. Our main results are presented in Section 3. Discussion and conclusions follow in Section 4.

2. A CE algorithm for discrete optimization

Suppose we wish to maximize some performance function $S(\mathbf{x})$ over all candidate solutions $\mathbf{x}$ belonging to a discrete finite set $\mathscr{X}$. In other words, we seek an optimal solution $\mathbf{x}^*$ satisfying $S(\mathbf{x}^*) \geqslant S(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{X}$. We do not require uniqueness of $\mathbf{x}^*$, and we let $\mathscr{X}^*$ denote the set of optimal solutions. We study the convergence properties of a general implementation of the CE method for discrete optimization, given below in Algorithm 1.

In order to implement the algorithm, we require a system for representing, or “encoding”, candidate solutions, and also a random mechanism for generating candidate solutions. The analysis presented in this paper is based on the following general approach. Candidate solutions are represented by a binary vector of length n, such that every $\mathbf{x} \in \mathscr{X}$ has a unique representation $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, where $x_i \in \{0, 1\}$. In principle, any discrete optimization problem can be encoded in this manner [7] (an example is given below). In the unconstrained case, we have $\mathscr{X} = \{0, 1\}^n$. For the constrained case, where $\mathscr{X} \subset \{0, 1\}^n$, we assume a penalty approach whereby $S(\mathbf{x}) = -\infty$ for all binary vectors $\mathbf{x} \in \{0, 1\}^n \notin \mathscr{X}$. In order to generate candidate solutions, the CEalgorithm maintains and updates a set of *reference parameters* $p_{t,i} \in [0, 1]$, $i = 1, \ldots, n$, where

$t \in \mathbb{N}$ is an iteration index. For each t, these are collected into a *reference vector*, $\mathbf{p}_t$. Candidate solutions are obtained by generating random vectors of the form $\mathbf{X}=(X_1, X_2, \dots, X_n)$, where the X_i, $i=1,\dots,n$, are independent Bernoulli random variables with parameters $p_{t,i}$, respectively. Thus, the vector $\mathbf{p}_t$ parameterizes a probability mass function $f\colon \{0,1\}^n \to [0,1]$, given by $f(\mathbf{x};\mathbf{p}_t) = \prod_{i=1}^n p_{t,i}^{x_i}(1-p_{t,i})^{(1-x_i)}$.

A canonical example of a discrete optimization problem is the "max-cut" problem, where the vertices of a graph with weighted edges must be partitioned into two sets $\mathscr{V}_1$ and $\mathscr{V}_2$ such that the resulting "cut" has maximum weight. Here, $x_i=1$ implies that vertex i belongs to $\mathscr{V}_1$, and $x_i=0$ implies that i belongs to $\mathscr{V}_2$. Furthermore, $\mathscr{X}=\{0,1\}^n \backslash \{(0,\dots,0),(1,\dots,1)\}$. We shall use this problem as the basis for illustrative numerical results which appear in Section 4. The reader is referred to [9] for a detailed description of this and other problems, and their associated binary encodings.

Algorithm 1 takes as its input the following parameters: an initial reference vector $\mathbf{p}_0$ satisfying $0 < p_{0,i} < 1$, $i=1,\dots,n$, a positive integer N, specifying the number of candidate solutions that are generated at each iteration of the algorithm, a real number $\rho \in (0,1)$, which determines the number N_b of candidate solutions at each iteration that are classed as the "best-performing", and a sequence of smoothing parameters $\{\alpha_t\}_{t=1}^\infty$, with $\alpha_t \in (0,1]$ for all t.

Algorithm 1 (*CE algorithm*).

(1) *Initialize* $\mathbf{p}_0$, N, ρ *and* $\{\alpha_t\}_{t=1}^\infty$, *and calculate* $N_\mathrm{b} = N+1-\lceil (1-\rho)N \rceil$. *Set* $t=1$ (*iteration counter*).

(2) *Generate a set of candidate solutions* $\mathbf{X}_t^{(k)}$, $k=1,\dots N$, *from the distribution* $f(\cdot;\mathbf{p}_{t-1})$, *and calculate the performances* $S(\mathbf{X}_t^{(k)})$ *for all* k, *ordering them from smallest to largest*: $S_{(1)} \leqslant S_{(2)} \leqslant \cdots \leqslant S_{(N)}$ (*ties are broken arbitrarily*). *Let* $\mathscr{B}_t$ *denote the set of indices k corresponding to the performances* $S_{(\lceil (1-\rho)N \rceil)}, \dots, S_{(N)}$.

(3) *For each* $i=1,\dots,n$, *calculate* $w_{t,i} = \sum_{k \in \mathscr{B}_t} X_{t,i}^{(k)}/N_\mathrm{b}$ *where* $X_{t,i}^{(k)}$ *represents the* ith *component of* $\mathbf{X}_t^{(k)}$. *Update the parameter vector according to*

$$p_{t,i} = (1-\alpha_t)p_{t-1,i} + \alpha_t w_{t,i}, \quad i=1,\dots,n. \tag{1}$$

(4) *If stopping criterion is reached then stop, otherwise set* $t=t+1$ *and reiterate from Step* 2.

A number of stopping criteria have been proposed (see [9]). The simplest is to stop when a fixed number T of iterations have been performed. The theoretical results presented in this paper address the limiting case $T \to \infty$.

The candidate solutions $\mathbf{X}_t^{(k)}$, $k=1,\dots,N$, $t \geqslant 1$, are random variables, and as such, the CE algorithm is a stochastic process. We emphasize the fact that $\mathbf{p}_t$, $t \geqslant 1$, are also random variables. In particular, they comprise a time-inhomogeneous Markov chain, since the probabilities governing the transition from $\mathbf{p}_t$ to $\mathbf{p}_{t+1}$ depend only on $\mathbf{p}_t$.

3. Convergence results

Algorithm 1 can be viewed as a stochastic process defined on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, where Ω is the set of all possible sample paths of the algorithm, $\mathscr{F} = \{\mathscr{F}_t, t \in \mathbb{N}\}$, where $\mathscr{F}_t$ is the σ-algebra generated by $\{\mathbf{X}_m^{(k)}, k=1,\dots,N, m=1,\dots,t\}$, and $\mathbb{P}$ is a probability measure on $(\Omega, \mathscr{F})$. We present convergence results for the CE algorithm in two parts. First, in Section 3.1, we give conditions under which $\mathbf{X}_t^{(k)} \in \mathscr{X}^*$ for at least one pair (k,t). We then present our main result in Section 3.2, which establishes limiting properties of the algorithm when a constant smoothing parameter is used, as is most common in practice [9].

3.1. Generating an optimal solution

Theorem 1 (*Necessary condition*). *An optimal solution is generated eventually by the CE algorithm with probability* 1 *only if the smoothing sequence* $\{\alpha_t\}_{t=1}^\infty$ *satisfies the condition* $\sum_{t=1}^\infty \prod_{m=1}^t (1-\alpha_m) = \infty$.

Proof. Without loss of generality, we assume that $x_1 = 1$ for all optimal $\mathbf{x} \in \mathscr{X}^*$ (this can always be achieved by adding a new variable to the start of the encoding vector, then mapping each previous candidate solution $(x_1,\dots,x_n)$ to $(1,x_1,\dots,x_n)$, with the same performance as before, and adding a new set of candidate solutions $\{(0,x_1,\dots,x_n)\}$, all with performance $-\infty$). For a given $\mathbf{p}_0$, N, ρ, $\{\alpha_t\}_{t=1}^\infty$, and a given i and t, the range of $p_{t,i}$ is a finite set. Let $p_{t,i}^{\min}$ be the minimum value in this set. From (1), observe that $p_{t,i}=p_{t,i}^{\min}$ when the event $\{w_{m,i}=0, m=1,\ldots,t\}$ occurs. Thus

$$p_{t,i}^{\min}=p_{0,i}\prod_{m=1}^{t}(1-\alpha_m) \tag{2}$$

for all $t\geqslant 0$, where henceforth we employ the convention $\prod_{m=1}^{0}(1-\alpha_m)=1$. Let $B_t=\{X_{m,1}^{(k)}\neq 1, k=1,\ldots,N, m=1,\ldots,t\}$, that is, the event that at every iteration up to and including iteration t, none of the candidate solutions contain the correct first component, $x_1=1$. Since B_{t-1} implies $p_{t,1}=p_{t,1}^{\min}$, we have that for each k, $\mathbb{P}(X_{t,1}^{(k)}=1\mid B_{t-1})=p_{0,1}\prod_{m=1}^{t}(1-\alpha_m)$. Since, by construction, the candidate solutions generated by the algorithm at a given iteration are conditionally independent given $\mathscr{F}_{t-1}$ and identically distributed, it follows that $\mathbb{P}(B_t\mid B_{t-1})=(1-p_{0,1}\prod_{m=1}^{t-1}(1-\alpha_m))^N$. Expanding $\mathbb{P}(B_T)$ as a product of conditional probabilities, we have

$$\mathbb{P}(B_T)=\mathbb{P}(B_1)\prod_{t=2}^{T}\mathbb{P}(B_t\mid B_{t-1})$$
$$=\left(\prod_{t=1}^{T}\left(1-p_{0,1}\prod_{m=1}^{t-1}(1-\alpha_m)\right)\right)^N,$$

where we have used the fact that $\mathbb{P}(B_1)=(1-p_{0,1})^N$. Define $E_t=\{\mathbf{X}_m^{(k)}\notin\mathscr{X}^*, k=1,\ldots,N, m=1,\ldots,t\}$. Since $B_T\subset E_T$, and thus $\mathbb{P}(E_T)\geqslant\mathbb{P}(B_T)$, it follows that $\lim_{T\to\infty}\mathbb{P}(E_T)=0$ only if

$$\lim_{T\to\infty}\left(\prod_{t=1}^{T}\left(1-p_{0,1}\prod_{m=1}^{t-1}(1-\alpha_m)\right)\right)^N=0. \tag{3}$$

Using standard results for infinite products [4], the product on the left-hand side of (3) diverges to zero only if the condition of Theorem 1 is satisfied (assuming, as stated in Section 2, that $p_{0,1}>0$). □

Theorem 2 (*Sufficient condition*). *An optimal solution is generated eventually by the CE algorithm with probability* 1 *if the smoothing sequence* $\{\alpha_t\}_{t=1}^{\infty}$ *satisfies the condition* $\sum_{t=1}^{\infty}\prod_{m=1}^{t}(1-\alpha_m)^n=\infty$.

Proof. Without loss of generality, assume that $\mathbf{x}^*=(x_1^*,x_2^*,\ldots,x_n^*)=(1,1,\ldots,1)\in\mathscr{X}^*$. This type of relabelling is always possible, since the update (1) takes exactly the same form if we replace $p_{t,i}$ and $w_{t,i}$ with $1-p_{t,i}$ and $1-w_{t,i}$, respectively (for any combination of the components i). In other words, Algorithm 1 is insensitive to arbitrary one-to-one mappings of the encoding set $\{0,1\}^n$ onto itself. For $t\geqslant 1$, we write $\phi_t=\mathbb{P}(\mathbf{X}_t^{(k)}\in\mathscr{X}^*\mid\mathscr{F}_{t-1})$ for the conditional probability that *an arbitrary candidate k* generated at iteration t belongs to the set of optimal solutions $\mathscr{X}^*$, and we note that ϕ_t is a $\mathscr{F}_{t-1}$-measurable random variable. We also have $\mathbb{P}(X_{t,i}^{(k)}=x_i^*\mid\mathscr{F}_{t-1})=p_{t-1,i}$ for an arbitrary candidate k, since the components of each candidate are generated independently (conditional on $\mathscr{F}_{t-1}$). Thus $\phi_t\geqslant\mathbb{P}(\mathbf{X}_t^{(k)}=\mathbf{x}^*\mid\mathscr{F}_{t-1})=\prod_{i=1}^{n}p_{t-1,i}$ with probability 1. It follows that

$$\phi_t\geqslant\prod_{i=1}^{n}p_{t-1,i}^{\min}=\left(\prod_{i=1}^{n}p_{0,i}\right)\prod_{m=1}^{t-1}(1-\alpha_m)^n \tag{4}$$

with probability 1. For $t\geqslant 2$, let $\mathbb{P}(\mathbf{X}_t^{(k)}\in\mathscr{X}^*\mid E_{t-1})$ denote the probability that an arbitrary candidate k generated at iteration t belongs to the set of optimal solutions $\mathscr{X}^*$, conditional on E_{t-1}. Then using (4), and recalling that $\mathbf{x}^*=(1,1,\ldots,1)\in\mathscr{X}^*$, we have $\mathbb{P}(\mathbf{X}_t^{(k)}\in\mathscr{X}^*\mid E_{t-1})\geqslant(\prod_{i=1}^{n}p_{0,i})\prod_{m=1}^{t-1}(1-\alpha_m)^n$, and thus $\mathbb{P}(\mathbf{X}_t^{(k)}\notin\mathscr{X}^*|E_{t-1})\leqslant 1-(\prod_{i=1}^{n}p_{0,i})\prod_{m=1}^{t-1}(1-\alpha_m)^n$. Since the candidate solutions generated by the algorithm at a given iteration are (conditionally) independent and identically distributed, it follows that

$$\mathbb{P}(E_t\mid E_{t-1})=[\mathbb{P}(\mathbf{X}_t^{(k)}\notin\mathscr{X}^*\mid E_{t-1})]^N$$
$$\leqslant\left[1-\left(\prod_{i=1}^{n}p_{0,i}\right)\prod_{m=1}^{t-1}(1-\alpha_m)^n\right]^N.$$

Now, expanding $\mathbb{P}(E_T)$ as a product of conditional probabilities, we obtain $\mathbb{P}(E_T)=\mathbb{P}(E_1)\prod_{t=2}^{T}\mathbb{P}(E_t\mid E_{t-1})$. Combining these results, we obtain

$$\mathbb{P}(E_T)\leqslant\mathbb{P}(E_1)\prod_{t=2}^{T}\left[1-\left(\prod_{i=1}^{n}p_{0,i}\right)\prod_{m=1}^{t-1}(1-\alpha_m)^n\right]^N. \tag{5}$$

Then $\lim_{T\to\infty} \mathbb{P}(E_T) = 0$ if the infinite product $\prod_{t=2}^{\infty}[1 - (\prod_{i=1}^{n} p_{0,i})\prod_{m=1}^{t-1}(1-\alpha_m)^n]$ diverges to zero, which in turn occurs if the condition of Theorem 2 is satisfied. □

Remark 1. The sufficient condition of Theorem 2 holds if $\sum_{t=1}^{\infty} \alpha_t < \infty$.

Remark 2. For a given set of parameters N, $\{\alpha_t\}_{t=1}^{\infty}$, $\mathbf{p}_0$, expression (5) provides a lower bound on the probability that an optimal solution is generated at least once in T iterations.

3.2. Constant smoothing parameter

We now present our main result, which establishes a limiting property of the CE method for the case of a constant smoothing parameter. Indeed, this is how the CE method is most commonly implemented in practice [9].

Theorem 3. *If the smoothing sequence is a constant, with $\alpha_t = \alpha$, $\alpha \in (0, 1]$, and $p_{0,i} \in (0, 1)$ for all i, then the sequence of probability mass functions $f(\mathbf{x}; \mathbf{p}_t)$, $t \geqslant 1$, converges with probability* 1 *to a unit mass located at some* (*random*) *candidate $\mathbf{x} \in \{0, 1\}^n$. Furthermore*, *the probability that an optimal solution is generated can be made arbitrarily close to* 1 *by selecting a sufficiently small value of α.*

Proof. Define $Z_{t,i} = p_{t,i} - p_{t-1,i}$, for $t = 1, 2, \ldots$, and let $\tau_{k,i}$ be the (random) iteration number at which $Z_{t,i}$ changes sign for the kth time. Note that those iterations t for which $Z_{t,i} = 0$ are *not* included in this collection. We establish that each $p_{t,i}$ converges by showing that $Z_{t,i}$ changes sign a finite number of times with probability 1. We then show that $\{0, 1\}$ are the only feasible limits for the $p_{t,i}$, which implies that $f(\mathbf{x}; \mathbf{p}_t)$ converges to a unit mass located at some (random) vector $\mathbf{x} \in \{0, 1\}^n$. To simplify the exposition of the proof, we fix the component i, and suppress it by writing p_t, w_t, Z_t and τ_k. The following analysis applies independently for each i, and therefore applies to the entire vector $\mathbf{p}_t$. Let $p_t^{\max}$ be the largest possible value of p_t. In particular, $p_t = p_t^{\max}$ when $\{w_m = 1, m = 1, \ldots, t\}$. Using (1), given an initial value $p_0 \in (0, 1)$, it can be shown by induction that $p_t^{\max} = \prod_{m=1}^{t}(1-\alpha_m)p_0 + \sum_{j=1}^{t} \alpha_j \prod_{m=j+1}^{t}(1-\alpha_m)$. Writing $\alpha_j = 1 - (1 - \alpha_j)$, we obtain

$$p_t^{\max} = \prod_{m=1}^{t}(1-\alpha_m)p_0 + \sum_{j=1}^{t}\left(\prod_{m=j+1}^{t}(1-\alpha_m) - \prod_{m=j}^{t}(1-\alpha_m)\right)$$

$$= 1 - (1-p_0)\prod_{m=1}^{t}(1-\alpha_m). \tag{6}$$

The change times have the following important properties: for all k,

(i) $\tau_k = \infty \Longrightarrow \tau_{k+1} = \infty$,
(ii) $Z_{\tau_k} < 0 \Longrightarrow p_{\tau_k} < 1 - \alpha/N_{\mathrm{b}} < 1$,
(iii) $Z_{\tau_k} > 0 \Longrightarrow p_{\tau_k} > \alpha/N_{\mathrm{b}} > 0$.

For fixed $N \geqslant 1$, define the function $g_\alpha(u) = \prod_{t=0}^{\infty}(1 - (1-u)(1-\alpha)^t)^N$. Note that $g_\alpha(0) = 0$, $g_\alpha(1) = 1$, and that $g_\alpha(u)$ is non-decreasing and strictly positive on (0, 1], since $\sum_{t=0}^{\infty}(1-\alpha)^t < \infty$. Observe that $\mathbb{P}(w_t = 1 \mid \mathscr{F}_{t-1}) \geqslant p_{t-1}^N$. Using (6), it follows that for each iteration l, and each $t > l$,

$$\mathbb{P}(w_t = 1 \mid w_m = 1, l \leqslant m \leqslant t-1, \mathscr{F}_{l-1}) \geqslant (1 - (1-p_{l-1})(1-\alpha)^{t-l})^N$$

so that

$$\mathbb{P}(w_t = 1, t \geqslant l \mid \mathscr{F}_{l-1}) \geqslant \prod_{t=l}^{\infty}(1 - (1-p_{l-1})(1-\alpha)^{t-l})^N = g_\alpha(p_{l-1}). \tag{7}$$

Similarly, $\mathbb{P}(w_t = 0 \mid \mathscr{F}_{t-1}) \geqslant (1 - p_{t-1})^N$, and using (2) we obtain

$$\mathbb{P}(w_t = 0, t \geqslant l \mid \mathscr{F}_{l-1}) \geqslant \prod_{t=l}^{\infty} \times (1 - p_{l-1}(1-\alpha)^{t-l})^N = g_\alpha(1 - p_{l-1}). \tag{8}$$

Now, $\{w_t = 0, t \geqslant 1\} \cup \{w_t = 1, t \geqslant 1\} \subseteq \{\tau_1 = \infty\}$, so $\mathbb{P}(\tau_1 = \infty | p_0) \geqslant g_\alpha(p_0) + g_\alpha(1 - p_0) = a_\alpha$, where a_α is a constant that depends on p_0 and α, and which is strictly positive under the assumptions of the theorem. Thus $\mathbb{P}(\tau_1 < \infty | p_0) \leqslant 1 - a_\alpha$. For $k \geqslant 2$, observe that $\{w_t = 0, t \geqslant \tau_{k-1} + 1\} \subseteq \{\tau_k = \infty\}$. Therefore, setting $l = \tau_{k-1} + 1$ in (8) and using property (ii) yields

$$\mathbb{P}(\tau_k = \infty | \tau_{k-1} < \infty, Z_{\tau_{k-1}} < 0) \geqslant g_\alpha \left(\frac{\alpha}{N_\mathrm{b}} \right) = d_\alpha, \tag{9}$$

where d_α is a strictly positive constant that depends on α (and N_b). Similarly, $\{w_t = 1, t \geqslant \tau_{k-1} + 1\} \subseteq \{\tau_k = \infty\}$. Setting $l = \tau_{k-1} + 1$ in (7) and using property (iii) yields

$$\mathbb{P}(\tau_k = \infty | \tau_{k-1} < \infty, Z_{\tau_{k-1}} > 0) \geqslant g_\alpha \left(\frac{\alpha}{N_\mathrm{b}} \right) = d_\alpha. \tag{10}$$

Now observe that

$$\begin{aligned}
&\mathbb{P}(\tau_k = \infty | \tau_{k-1} < \infty) \\
&\quad = \mathbb{P}(\tau_k = \infty | \tau_{k-1} < \infty, Z_{\tau_{k-1}} < 0) \\
&\qquad \times \mathbb{P}(Z_{\tau_{k-1}} < 0 | \tau_{k-1} < \infty) \\
&\qquad + \mathbb{P}(\tau_k = \infty | \tau_{k-1} < \infty, Z_{\tau_{k-1}} > 0) \\
&\qquad \mathbb{P}(Z_{\tau_{k-1}} > 0 | \tau_{k-1} < \infty) \\
&\quad \geqslant d_\alpha(\mathbb{P}(Z_{\tau_{k-1}} < 0 | \tau_{k-1} < \infty) \\
&\qquad + \mathbb{P}(Z_{\tau_{k-1}} > 0 | \tau_{k-1} < \infty)) \\
&\quad = d_\alpha.
\end{aligned}$$

It follows that $\mathbb{P}(\tau_k < \infty | \tau_{k-1} < \infty) \leqslant 1 - d_\alpha$ for all k, and thus $\mathbb{P}(\cap_{k=1}^{\infty} \{\tau_k < \infty\}) \leqslant (1 - a_\alpha) \prod_{k=2}^{\infty} (1 - d_\alpha) = 0$, where we have used the fact that p_t (and hence Z_t) is a Markov chain, and the fact that $d_\alpha > 0$. It follows that $\mathbb{P}(\cup_{k=1}^{\infty} \{\tau_k = \infty\}) = 1$, that is, Z_t changes sign a finite number of times with probability 1. This implies that p_t is eventually monotonic and thus converges to a limit p^*. From (1), and the fact that $w_t \in \{0, 1/N_\mathrm{b}, 2/N_\mathrm{b}, \ldots, 1\}$, we must have $p^* = j/N_\mathrm{b}$ for some $j \in \{0, 1, 2, \ldots, N_b\}$, and $w_t = p^*$ for all $t \geqslant t_0$, for some t_0. However, for $p^* \neq 0, 1$, we have $\mathbb{P}(\{w_t = 0, t \geqslant t_0\} \cup \{w_t = 1, t \geqslant t_0\}) \geqslant g_\alpha(p^*) + g_\alpha(1 - p^*) > 0$, so we must have $p^* = 0$ or 1. Thus, we have established that $f(\mathbf{x}; \mathbf{p}_t)$ converges with probability 1 to a unit mass located at some (random) candidate $\mathbf{x} \in \{0, 1\}^n$.

To conclude the proof, we set $\alpha_m = \alpha$ for all m in (5), so that $\mathbb{P}(E_T) \leqslant \mathbb{P}(E_1) \prod_{t=2}^{T} (1 - \prod_{i=1}^{n} p_{0,i} (1 - \alpha)^{(t-1)n})^N$. Using the fact that $(1 - u)^N \leqslant \mathrm{e}^{-Nu}$, for $0 \leqslant u \leqslant 1$ and $N \geqslant 0$, we obtain $\mathbb{P}(E_T) \leqslant \mathbb{P}(E_1) \prod_{t=2}^{T} \exp(-N(\prod_{i=1}^{n} p_{0,i})(1 - \alpha)^{(t-1)n}) = \mathbb{P}(E_1) \exp(-N(\prod_{i=1}^{n} p_{0,i}) \sum_{t=1}^{T-1} (1 - \alpha)^{tn})$. Thus

$$\lim_{T \to \infty} \mathbb{P}(E_T) \leqslant \mathbb{P}(E_1) \exp(-N \left(\prod_{i=1}^{n} p_{0,i} \right) h(\alpha)),$$

where $h(\alpha) = 1/(1 - (1 - \alpha)^n) - 1$. Since $h(\alpha) \to \infty$ as $\alpha \to 0$, $\lim_{T \to \infty} \mathbb{P}(E_T)$ can be made arbitrarily close to zero by selecting a sufficiently small value of α. □

We conclude this section with a simple but informative necessary condition for the convergence of the sequence $f(\mathbf{x}; \mathbf{p}_t)$ to a unit mass.

Corollary 1 (*Necessary condition*). *The sequence of probability mass functions* $f(\mathbf{x}; \mathbf{p}_t), t \geqslant 1$, *converges with probability* 1 *to a unit mass located at some candidate* $\mathbf{x} \in \{0, 1\}^n$ *only if* $\sum_{t=1}^{\infty} \alpha_t = \infty$.

Proof. $f(\mathbf{x}; \mathbf{p}_t)$ converges to a unit mass located at some $\mathbf{x} \in \{0, 1\}^n$ only if eventually $p_{t,i} \to 0$ or 1 for each i, which, given (2) and (6),occurs only if $\prod_{m=1}^{\infty} (1 - \alpha_m) = 0$, which implies the result. □

4. Discussion and conclusion

The CE algorithm is most-commonly implemented using a constant smoothing parameter [9], that is, $\alpha_t = \alpha$ for all t, where $\alpha \in (0, 1]$. This has been found empirically to yield a significantly faster rate of convergence of the sampling distribution $f(\mathbf{x}; \mathbf{p}_t)$ compared with decreasing smoothing schemes, which is the main reason for its popularity [9]. For this special but important case, our main result (Theorem 3) shows that the sampling distribution always converges to a unit mass located at a random candidate $\mathbf{x} \in \{0, 1\}^n$, and that the limiting probability of generating an optimal solution can be made arbitrarily close to 1 by selecting a sufficiently small value of α. We note that using a smaller value of α effectively reduces the empirical rate of convergence of $f(\mathbf{x}; \mathbf{p}_t)$ from the initial distribution to a

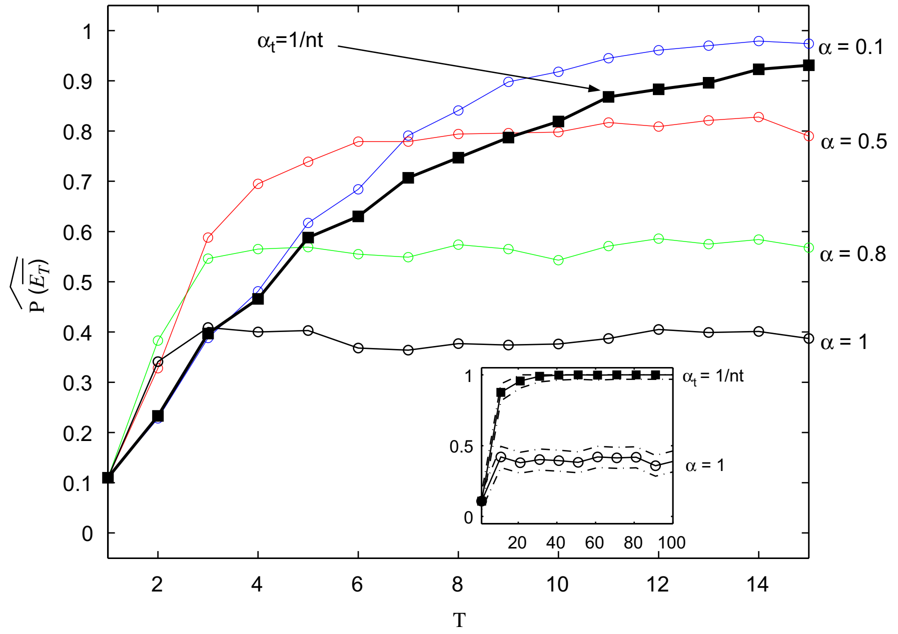

Fig. 1. Illustrative empirical results. The main figure shows transient behaviour of Algorithm 1. The insert shows limiting results for $\alpha = 1$ and $\alpha_t = 1/nt$ with associated 95% confidence intervals (dotted lines) which are omitted from the main figure for clarity.

unit mass. Therefore, when using a constant smoothing parameter, there exists a tension between achieving an optimal solution with high probability, and achieving a fast rate of convergence of the sampling distribution. To demonstrate the former, we take an illustrative instance of the "max-cut" problem (see Section 2) with $n = 8$ vertices and a unique optimal solution. Fig. 1 shows empirical estimates of $\mathbb{P}(\overline{E_T})$ for a range of values of α and T, where $\overline{E_T}$ is the event that $\mathbf{X}_t^{(k)} = \mathbf{x}^*$ for at least one pair (k, t), $k = 1, \ldots, N$, $t = 1, \ldots, T$. We see that the limiting probability of obtaining the optimal solution can be made arbitrarily close to 1. These results were generated by performing 100 independent replications of Algorithm 1 for each fixed α and T.

Examples of smoothing sequences which eventually generate an optimal solution with probability 1 (that is, which satisfy the sufficient condition of Theorem 2) include $\alpha_t = 1/(t+1)^\beta$ and $\alpha_t = 1/((t+1)\ \log(t+1))^\beta$, when $\beta > 1$, as well as

$$\alpha_t = \frac{1}{nt}, \tag{11}$$

where n is the "problem size" parameter introduced in Section 2. Indeed, the insert in Fig. 1 demonstrates that $\mathbb{P}(\overline{E_T})$ approaches 1 for large T when the sequence (11) is used. The main panel of Fig. 1 illustrates that (11) yields similar transient behaviour of $\mathbb{P}(\overline{E_T})$ to the case of constant $\alpha = 0.1$. We have found this behaviour to be typical for such decreasing sequences, and for a range of different optimization problems and problem sizes. The necessary condition of Corollary 1 is useful as it shows that the first two of the above decreasing sequences cannot also yield convergence of the sampling distribution to a unit mass, since for these cases α_t decreases too rapidly(in fact, the limiting distribution, if it exists, has a strictly positive mass on every candidate $\mathbf{x} \in \{0, 1\}^n$).

It remains an open theoretical problem to establish whether there exists a smoothing sequence which yields convergence to a unit mass that is located at an optimal solution with probability 1. For example, the smoothing sequence $\alpha_t = 1/nt$ satisfies both the sufficient condition of Theorem 2 and the necessary condition of Corollary 1, and might thus appear to be a likely candidate. However, our experience with the

CE method suggest that this is not the case for Algorithm 1, and that the two properties: (a) convergence to a unit mass with probability 1, and (b) eventually generating an optimal solution with probability 1, are in fact mutually exclusive. This conjecture is supported by the fact that the conditions in Remark 1 and Corollary 1 are mutually exclusive, and remains a topic for further investigation. We note that if Algorithm 1 is modified so that a record is kept of the best candidate found over all iterations, and if the sampling distribution is forced to favour this candidate (the "record") at each iteration, then one recovers the "elite sample" version of the CE method studied in [9,6], for which it can be shown that the sequence $\alpha_t = 1/nt$ achieves both (a) and (b). On a theoretical level, it is interesting to note that this "memory" property in fact facilitates the convergence analysis (see [9,6]), however, it is not clear whether the elite sample version offers practical advantages over Algorithm 1. In our experience with the CE method, we have not found evidence of improved (transient) performance, however, a comprehensive numerical comparison is currently lacking in the literature, and is a potential area for further research.

Finally, we note that the results presented in this paper pertain to the algorithm's limiting properties, whereas results concerning the transient properties of the CE algorithm would be extremely useful from a practical point of view. For instance, although the influence of the smoothing parameter dominates that of the other parameters in the limit $T \to \infty$, the practitioner may wish to know how to *jointly* set all of the parameters ρ, N and $\{\alpha_t\}_{t=1}^{\infty}$, so as to maximize the probability that an optimal solution is obtained in the short term, that is, in the first few iterations before the limiting regime is reached. There is much scope for further research on transient behaviour of the CE method.

Acknowledgements

The authors wish to thank the editors and anonymous referees for their valuable suggestions, which helped us to improve this paper. The authors are grateful to Alan Jones for his assistance, and to Felisa Vázquez-Abad and Reuven Rubinstein for helpful discussions. Andre Costa would like to acknowledge the support of the Australian Research Council Centre of Excellence for Mathematics and Statistics of Complex Systems. Dirk Kroese would like to acknowledge the support of the Australian Research Council, under Grant number DP0558957.

References

[1] E. Aarts, J. Lenstra, Local Search in Combinatorial Optimisation, Wiley, Chichester, UK, 1997.

[2] S. Andradóttir, A global search method for discrete stochastic optimization, SIAM J. Optim. 6 (1996) 513–530.

[3] J. Holland, Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence, MIT Press, Cambridge, MA, 1992.

[4] K. Knopp, Infinite Sequences and Series, Dover Publications, New York, 1956.

[5] P. Larranaga, J. Lozano, Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation, Kluwer Academic Publishers, Dordrecht, MA, 2001.

[6] L. Margolin, On the convergence of the cross-entropy method, Ann. Oper. Res. 134 (2004) 201–214.

[7] G. Nemhauser, L. Wolsley, Integer and Combinatorial Optimization, Wiley, New York, 1988.

[8] D.T. Pham, D. Karaboga, Intelligent Optimisation Techniques, Springer, Berlin, 2000.

[9] R.Y. Rubinstein, D.P. Kroese, The Cross-Entropy Method: A Unified Approach to Combinatorial Optimisation, Monte-Carlo Simulation, and Machine Learning, Springer, Berlin, 2004.

[10] J. Spall, Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control, Wiley, New York, 2003.