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. email: acosta@ms.unimelb.edu.au*

[b] *Department of Mathematics and Statistics, University of Melbourne, 3010, Australia. email: O.D.Jones@ms.unimelb.edu.au*

[c] *Department of Mathematics, University of Queensland, 4072, Australia. email: kroese@maths.uq.edu.au*

Abstract

We present new theoretical convergence results on the Cross-Entropy method for discrete optimization. Our primary contribution is to show that a popular implementation of the Cross-Entropy method converges, and finds the optimal solution with probability arbitrarily close to 1. We also give necessary conditions and sufficient conditions under which the optimal solution is generated eventually with probability 1.

Key words: 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 realised that the CE method could also be applied to a variety of optimization problems. The reader is referred to Rubinstein and Kroese [1] 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. In particular, we assume that the optimal solution is unique. We consider the deterministic setting, where exact objective function values are available, and where stochastic effects are

* Corresponding author.

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.

Existing results on the convergence of the CE method for discrete optimization appear in [1,2], 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 general and commonly-used version of the CE method [1], 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 [1,2]. Furthermore, although the elite sample version possesses desirable limiting convergence properties [1,2], it has been found to exhibit poor performance in practice (that is, within a typical realistic number of iterations), as compared with the standard CE method [1]. Therefore, the convergence results presented in this paper 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 the optimal solution, in the sense that the sampling distribution converges with probability 1 to a unit mass, and that the probability that the 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 [1,2] to derive more general and easily-checkable necessary conditions and sufficient conditions under which the 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 (Aarts and Lenstra [3]), genetic algorithms (Holland [4]), the method of Andradóttir [5] and many others (see Pham and Karaboga [6] and Spall [7] 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 [8] and Ant Colony Optimisation [9] algorithms, 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 [1] 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 $\mathcal{X}$. In other words, we seek an optimal solution $\mathbf{x}^*$ satisfying $S(\mathbf{x}^*) \geq S(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. Here, we shall assume that the optimal solution is unique. While the CE method is able to identify multiple global optima (see [1] for numerical examples), the dynamics of the CE method in this case are more difficult to characterise theoretically, and we do not address this here. 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 $\mathbf{x} \in \mathcal{X}$ has a unique representation $\mathbf{x} = (x_1, x_2, \ldots, x_n)$, where $x_i \in \{0, 1\}$. In particular, the optimal solution has the representation $\mathbf{x}^* = (x_1^*, x_2^*, \ldots, x_n^*)$. In principle, any discrete optimization problem can be encoded in this manner [10]. A canonical example is the "max-cut" problem [1], where the vertices of a graph with weighted edges must be partitioned into two sets $\mathcal{V}_1$ and $\mathcal{V}_2$ such that the resulting "cut" has maximum weight. Here, $x_i = 1$ implies that vertex i belongs to $\mathcal{V}_1$, and $x_i = 0$ implies that i belongs to $\mathcal{V}_2$. The reader is referred to [1] for a detailed description of this and other problems, and their associated binary encodings. In order to generate candidate solutions, the CE algorithm maintains and up-

dates 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*, $\mathbf{p}_t$. A natural way to generate candidate solutions is to generate random vectors of the form $\mathbf{X} = (X_1, X_2, \ldots, X_n)$, where the $X_i, i = 1, \ldots, n$, are independent Bernoulli random variables with parameters $p_{t,i}$, respectively. Thus, the vector $\mathbf{p}_t$ parameterizes a probability mass function $f(\mathbf{x}; \mathbf{p}_t) : \mathcal{X} \to [0,1]$. The simplest way to handle constraints on the components $x_i, i = 1, \ldots, n$ is to select components independently as described above, and then perform acceptance-rejection of the samples [1]. The analysis in this paper addresses this scenario. For example, for the basic "max-cut" problem on a fully-connected graph, we require at least one component to be different to the others to ensure that $\mathcal{V}_1$ and $\mathcal{V}_2$ are non-empty. Thus, the vectors $(1, \ldots, 1)$ and $(0, \ldots, 0)$ would be rejected, but all others would be accepted. We note that it is not necessary to select each component of a candidate solution independently of the others; indeed, for some applications, such as the travelling salesman problem, it is more efficient to perform "conditional sampling", as described in [1]. We do not consider this here.

Algorithm 1 takes as its input the following parameters: an initial reference vector $\mathbf{p}_0$, which is chosen so that $f(\mathbf{x}; \mathbf{p}_0)$ is the uniform distribution (this is a natural choice, in the absence of prior information regarding the identity of the optimal solution), a positive integer N, specifying the number of candidate solutions that are generated at each iteration of the algorithm, a positive integer T, specifying the total number of iterations to be performed, 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 (Cross-entropy algorithm)

(1) Initialize $\mathbf{p}_0, N, T, \rho$ *and* $\{\alpha_t\}_{t=1}^{\infty}$. *Set* $t = 1$ *(iteration counter).*

(2) Generate a set of candidate solutions $\mathbf{X}_t^{(k)}, k = 1, \ldots N$, *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:* $S_{(1)} \leq S_{(2)} \leq \ldots \leq S_{(N)}$ *(ties are broken arbitrarily). Compute the sample* $(1-\rho)$*-quantile of the performances, given by* $\hat{\gamma}_t = S_{(\lceil (1-\rho)N \rceil)}$, *and let* $\mathcal{B}_t$ *denote the set of indices* k *for which* $S(\mathbf{X}_t^{(k)}) \geq \hat{\gamma}_t$. *Let* $N_b = |\mathcal{B}_t|$ *(note that this is independent of* t, *and depends only on* ρ*).*

(3) For each $i = 1, \ldots, n$, *calculate* $w_{t,i} = \frac{\sum_{k \in \mathcal{B}_t} X_{t,i}^{(k)}}{N_b}$ *where* $X_{t,i}^{(k)}$ *represents the* i^{th} *component of* $\mathbf{X}_t^{(k)}$. *Update the parameter vector according to*

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

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

The candidate solutions $\mathbf{X}_t^{(k)}, k = 1, \ldots N, t \geq 1$, are random variables, and as such, the CE algorithm is a *stochastic process*. We emphasize the fact that $\mathbf{p}_t, t \geq 1$, 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 $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is the set of all possible sample paths of the algorithm, $\mathcal{F} = \{\mathcal{F}_t, t \in \mathbb{N}\}$, where $\mathcal{F}_t$ is the σ-algebra generated by $\{\mathbf{X}_m^{(k)}, k = 1, \ldots N, m = 1, \ldots, t\}$, and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$. We present convergence results for the CE algorithm in two parts. First we give conditions under which $\mathbf{X}_t^{(k)} = \mathbf{x}^*$ for at least one pair (k, t). We then present our main result, which establishes limiting properties of the algorithm when a constant smoothing parameter is used, as is most common in practice [1].

3.1 Generating the optimal solution

By construction, the candidate solutions $\mathbf{X}_t^{(k)}$ generated at iteration t are conditionally independent given $\mathcal{F}_{t-1}$, and are identically distributed with distribution $f(\cdot; \mathbf{p}_{t-1})$. We therefore write $\phi_t = \mathbb{P}(\mathbf{X} = \mathbf{x}^* \mid \mathcal{F}_{t-1})$ for the conditional probability that *an arbitrary candidate* generated at iteration t is equal to the optimal solution, and we note that ϕ_t is a $\mathcal{F}_{t-1}$-measurable random variable. We now establish a relationship between ϕ_t and the reference vector $\mathbf{p}_{t-1}$.

Let $\mathbf{1}\{\cdot\}$ denote the indicator function. Define the random variables $\phi_{t,i} = p_{t-1,i}\mathbf{1}\{x_i^* = 1\} + (1 - p_{t-1,i})\mathbf{1}\{x_i^* = 0\}$, for all $i = 1, \ldots, n$, and $t \geq 1$. Then with probability 1, $\phi_t = \prod_{i=1}^n \phi_{t,i}$. For a given $\mathbf{p}_0$, $N, \rho, \{\alpha_t\}_{t=1}^\infty$, and a given i and t, the range of $p_{t,i}$ is a finite set. Let $p_{t,i}^{\min}$ and $p_{t,i}^{\max}$ be the minimum and maximum values in this set, respectively. These values (derived below) will form the basis for bounds used in the theorem proofs. The ranges of $\phi_{t,i}$ and ϕ_t are therefore also finite sets, which we denote $\mathcal{R}_{t,i}$ and $\mathcal{R}_t$, respectively. Let $\phi_{t,i}^{\min}$ be the minimum value in $\mathcal{R}_{t,i}$, and let $\phi_t^{\min}$ be the minimum value in $\mathcal{R}_t$. Observe that $\phi_{t,i}^{\min} = p_{t-1,i}^{\min}\mathbf{1}\{x_i^* = 1\} + (1 - p_{t-1,i}^{\max})\mathbf{1}\{x_i^* = 0\}$. From (1), we obtain

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

for all $t \geq 0$, where henceforth we employ the convention $\prod_{m=1}^0 (1 - \alpha_m) = 1$. From Algorithm 1, we observe that $p_{t,i} = p_{t,i}^{\min}$ when the event $\{w_{m,i} = 0, m = 1, \ldots, t\}$ occurs. Using (1), given an initial value $p_{0,i} \in (0, 1)$, it can be shown

by induction that $p_{t,i}^{\max} = \prod_{m=1}^{t}(1-\alpha_m)p_{0,i} + \sum_{j=1}^{t}\alpha_j \prod_{m=j+1}^{t}(1-\alpha_m)$. Writing $\alpha_j = 1-(1-\alpha_j)$, we obtain

$$\begin{aligned} p_{t,i}^{\max} &= \prod_{m=1}^{t}(1-\alpha_m)p_{0,i} + \sum_{j=1}^{t}\left(\prod_{m=j+1}^{t}(1-\alpha_m) - \prod_{m=j}^{t}(1-\alpha_m)\right) \\ &= 1-(1-p_{0,i})\prod_{m=1}^{t}(1-\alpha_m). \end{aligned} \tag{3}$$

In particular, $p_{t,i} = p_{t,i}^{\max}$ when $\{w_{m,i} = 1, m = 1,\ldots,t\}$. It follows that $\phi_{t,i}^{\min} = \phi_{1,i}\prod_{m=1}^{t-1}(1-\alpha_m)$ for $t \geq 1$, and thus $\phi_t \geq \phi_t^{\min} = \prod_{i=1}^{n}\phi_{t,i}^{\min} = \phi_1\prod_{m=1}^{t-1}(1-\alpha_m)^n$ with probability 1.

Theorem 1 *(Necessary condition) The 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: Let $B_t = \{X_{m,1}^{(k)} \neq x_1^*, k = 1,\ldots N, m = 1,\ldots,t\}$, 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^*. Let $\mathbb{P}(X_{t,1} = x_1^* \,|\, B_{t-1})$ denote the probability that *an arbitrary candidate* generated at iteration t contains the correct first component, x_1^*, as it appears in the optimal solution $\mathbf{x}^*$, conditional on the event B_{t-1}. Since B_{t-1} implies $\{\phi_{t,1} = \phi_{t,1}^{\min}\}$, we have $\mathbb{P}(X_{t,1} = x_1^* \,|\, B_{t-1}) = \mathbb{P}(X_{t,1} = x_1^* \,|\, \phi_{t,1} = \phi_{t,1}^{\min}) = \phi_{t,1}^{\min} = \phi_{1,1}\prod_{m=1}^{t-1}(1-\alpha_m)$, where we have used the fact that $\mathbb{P}(X_{t,1} = x_1^* \,|\, \phi_{t,1} = r) = r$. Since the candidate solutions generated by the algorithm at a given iteration are independent and identically distributed, it follows that $\mathbb{P}(B_t \,|\, B_{t-1}) = \left(1-\phi_{1,1}\prod_{m=1}^{t-1}(1-\alpha_m)\right)^N$. Expanding $\mathbb{P}(B_T)$ as a product of conditional probabilities, we have

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

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

$$\lim_{T\to\infty}\left(\prod_{t=1}^{T}\left(1-\phi_{1,1}\prod_{m=1}^{t-1}(1-\alpha_m)\right)\right)^N = 0. \tag{4}$$

Using standard results for infinite products [11], the product on the left-hand side of (4) diverges to zero only if the condition of Theorem 1 is satisfied (assuming, as is standard practice, that the algorithm is initialized so that $\phi_{1,1} > 0$). $\square$

Theorem 2 *(Sufficient condition) The 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: We retain the notation of Theorem 1. For $t \geq 2$, let $\mathbb{P}(\mathbf{X}_t = \mathbf{x}^* \,|\, E_{t-1})$ denote the probability that an arbitrary candidate generated at iteration t is equal to the optimal solution, conditional on E_{t-1}. Then

$$\begin{aligned}\mathbb{P}(\mathbf{X}_t = \mathbf{x}^* \,|\, E_{t-1}) &= \sum_{r\in\mathcal{R}_t} \mathbb{P}(\mathbf{X}_t = \mathbf{x}^* \,|\, \phi_t = r, E_{t-1})\mathbb{P}(\phi_t = r \,|\, E_{t-1}) \\ &\geq \min_{r\in\mathcal{R}_t} r, \end{aligned} \tag{5}$$

using the fact that, by construction of the CE algorithm, $\mathbb{P}(\mathbf{X}_t = \mathbf{x}^* \,|\, \phi_t = r, E_{t-1}) = r$. In particular, the minimum in (5) is attained when $\phi_t = \phi_t^{\min}$, so that $\mathbb{P}(\mathbf{X}_t = \mathbf{x}^* \,|\, E_{t-1}) \geq \phi_1 \prod_{m=1}^{t-1}(1-\alpha_m)^n$, and $\mathbb{P}(\mathbf{X}_t \neq \mathbf{x}^* | E_{t-1}) \leq 1 - \phi_1 \prod_{m=1}^{t-1}(1-\alpha_m)^n$. Now, expanding $\mathbb{P}(E_T)$ as a product of conditional probabilities, we obtain $\mathbb{P}(E_T) = \mathbb{P}(E_1)\prod_{t=2}^{T}\mathbb{P}(E_t \,|\, E_{t-1})$. Since the candidate solutions generated by the algorithm at a given iteration are independent and identically distributed, it follows that

$$\mathbb{P}(E_t \,|\, E_{t-1}) = [\mathbb{P}(\mathbf{X}_t \neq \mathbf{x}^* \,|\, E_{t-1})]^N \leq \left[1 - \phi_1 \prod_{m=1}^{t-1}(1-\alpha_m)^n\right]^N. \tag{6}$$

Combining these results, we obtain

$$\mathbb{P}(E_T) \leq \mathbb{P}(E_1)\prod_{t=2}^{T}\left[1 - \phi_1 \prod_{m=1}^{t-1}(1-\alpha_m)^n\right]^N. \tag{7}$$

Then $\lim_{T\to\infty}\mathbb{P}(E_T) = 0$ if the infinite product $\prod_{t=2}^{\infty}\left[1 - \phi_1\prod_{m=1}^{t-1}(1-\alpha_m)^n\right]$ 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, $\{\alpha_t\}_{t=1}^{\infty}$, T *and* $\mathbf{p}_0$, *expression (7) provides a lower bound on the probability that the optimal solution is generated at least once in* T *iterations. Alternatively, this expression can be used to determine a combination of parameter values which yield a desired minimum probability of generating the optimal solution.*

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 [1].

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* $f(\mathbf{x}; \mathbf{p}_t), t \geq 1$, *converges with probability 1 to a unit mass located at some (random) candidate* $\mathbf{x} \in \mathcal{X}$. *Furthermore, the probability that the 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 k^{th} 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) candidate $\mathbf{x} \in \mathcal{X}$. 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 $\mathbf{p}_t$. The change times have the following important properties: for all k,

(i) $\tau_k = \infty \implies \tau_{k+1} = \infty$,
(ii) $Z_{\tau_k} < 0 \implies p_{\tau_k} < 1 - \frac{\alpha}{N_b} < 1$,
(iii) $Z_{\tau_k} > 0 \implies p_{\tau_k} > \frac{\alpha}{N_b} > 0$.

For fixed $N \geq 1$, define the function $g_\alpha(u) = \prod_{t=0}^{\infty} \left(1 - (1-u)(1-\alpha)^t\right)^N$. 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 $\mathbb{P}(w_t = 1 \,|\, \mathcal{F}_{t-1}) \geq p_{t-1}^N$. Using (3), it follows that for each iteration l, and each $t > l$,

$$\mathbb{P}(w_t = 1 \,|\, w_m = 1, l \leq m \leq t-1, \mathcal{F}_{l-1}) \geq \left(1 - (1 - p_{l-1})(1-\alpha)^{t-l}\right)^N$$

so that

$$\mathbb{P}(w_t = 1, t \geq l \,|\, \mathcal{F}_{l-1}) \geq \prod_{t=l}^{\infty} \left(1 - (1 - p_{l-1})(1-\alpha)^{t-l}\right)^N = g_\alpha(p_{l-1}). \qquad (8)$$

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

$$\mathbb{P}(w_t = 0, t \geq l \,|\, \mathcal{F}_{l-1}) \geq \prod_{t=l}^{\infty} \left(1 - p_{l-1}(1-\alpha)^{t-l}\right)^N = g_\alpha(1 - p_{l-1}). \qquad (9)$$

Now, $\{w_t = 0, t \geq 1\} \cup \{w_t = 1, t \geq 1\} \implies \{\tau_1 = \infty\}$, so $\mathbb{P}(\tau_1 = \infty | p_0) \geq g_\alpha(p_0) + g_\alpha(1 - p_0) = a_\alpha$, 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 $\mathbb{P}(\tau_1 < \infty | p_0) \leq 1 - a_\alpha$. For $k \geq 2$, observe that $\{w_t = 0, t \geq \tau_{k-1} + 1\} \implies \{\tau_k = \infty\}$. Therefore, setting $l = \tau_{k-1} + 1$ in (9) and using property (ii) yields

$$\mathbb{P}(\tau_k = \infty | \tau_{k-1} < \infty, Z_{\tau_{k-1}} < 0) \geq g_\alpha\left(\frac{\alpha}{N_b}\right) = d_\alpha, \tag{10}$$

where d_α is a strictly positive constant that depends on α (and N_b). Similarly, $\{w_t = 1, t \geq \tau_{k-1} + 1\} \implies \{\tau_k = \infty\}$. Setting $l = \tau_{k-1} + 1$ in (8) and using property (iii) yields

$$\mathbb{P}(\tau_k = \infty | \tau_{k-1} < \infty, Z_{\tau_{k-1}} > 0) \geq g_\alpha\left(\frac{\alpha}{N_b}\right) = d_\alpha. \tag{11}$$

Now observe that

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

It follows that $\mathbb{P}(\tau_k < \infty | \tau_{k-1} < \infty) \leq 1 - d_\alpha$ for all k. Thus, the probability that Z_t changes sign infinitely often is given by

$$\begin{aligned}
\mathbb{P}(\cap_{k=1}^{\infty} \tau_k < \infty) &= \mathbb{P}(\tau_1 < \infty | p_0) \prod_{k=2}^{\infty} \mathbb{P}(\tau_k < \infty | \tau_j < \infty, j = 1, \ldots, k-1) \\
&= \mathbb{P}(\tau_1 < \infty | p_0) \prod_{k=2}^{\infty} \mathbb{P}(\tau_k < \infty | \tau_{k-1} < \infty) \\
&\leq (1 - a_\alpha) \prod_{k=2}^{\infty} (1 - d_\alpha) = 0,
\end{aligned} \tag{12}$$

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 from (12) 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 $w_t \in \{0, \frac{1}{N_b}, \frac{2}{N_b}, \ldots, 1\}$, we must have $p^* = \frac{j}{N_b}$ for some $j \in \{0, 1, 2, \ldots, N_b\}$, and $w_t = p^*$ for all $t \geq t_0$, for some t_0. However, for $p^* \neq 0, 1$, we have $\mathbb{P}(\{w_t = 0, t \geq t_0\} \cup \{w_t = 1, t \geq t_0\}) \geq g_\alpha(p^*) + g_\alpha(1 - p^*) > 0$, 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 $\mathbf{x} \in \mathcal{X}$.

To conclude the proof, we set $\alpha_m = \alpha$ for all m in (7), so that $\mathbb{P}(E_T) \leq \mathbb{P}(E_1) \prod_{t=2}^{T} \left(1 - \phi_1 (1-\alpha)^{(t-1)n}\right)^N$. Using the fact that $(1-u)^N \leq e^{-Nu}$, for $0 \leq u \leq 1$ and $N \geq 0$, we obtain $\mathbb{P}(E_T) \leq \mathbb{P}(E_1) \prod_{t=2}^{T} \exp(-N\phi_1 (1-\alpha)^{(t-1)n}) = \mathbb{P}(E_1) \exp(-N\phi_1 \sum_{t=1}^{T-1} (1-\alpha)^{tn})$. Thus

$$\lim_{T \to \infty} \mathbb{P}(E_T) \leq \mathbb{P}(E_1) \exp(-N\phi_1 h(\alpha)),$$

where $h(\alpha) = \frac{1}{1-(1-\alpha)^n} - 1$. 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. $\square$

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 $f(\mathbf{x}; \mathbf{p}_t), t \geq 1$, converges with probability 1 to a unit mass located at some candidate $\mathbf{x} \in \mathcal{X}$ only if $\sum_{t=1}^{\infty} \alpha_t = \infty$.*

Proof: $f(\mathbf{x}; \mathbf{p}_t)$ converges to a unit mass located at some $\mathbf{x} \in \mathcal{X}$ only if eventually $p_{t,i} \to 0$ or 1 for each i, which, given (2) and (3), occurs only if $\prod_{m=1}^{\infty}(1-\alpha_m) = 0$, which implies the result. $\square$

4 Discussion and conclusion

The CE algorithm is most-commonly implemented using a constant smoothing parameter [1], that is, $\alpha_t = \alpha$ for all t, where $\alpha \in (0, 1]$. In general, this yields 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. 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 $\mathbf{x} \in \mathcal{X}$, and that the limiting probability of generating the 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 rate of convergence of $f(\mathbf{x}; \mathbf{p}_t)$ from the initial uniform distribution to a unit mass. Therefore, when using a constant smoothing parameter, there exists a tension between achieving the optimal solution with high probability, and achieving a fast rate of convergence of the sampling distribution. To illustrate the former, Figure 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, using an illustrative instance of the "max-cut" problem [1] with $n = 8$ vertices.

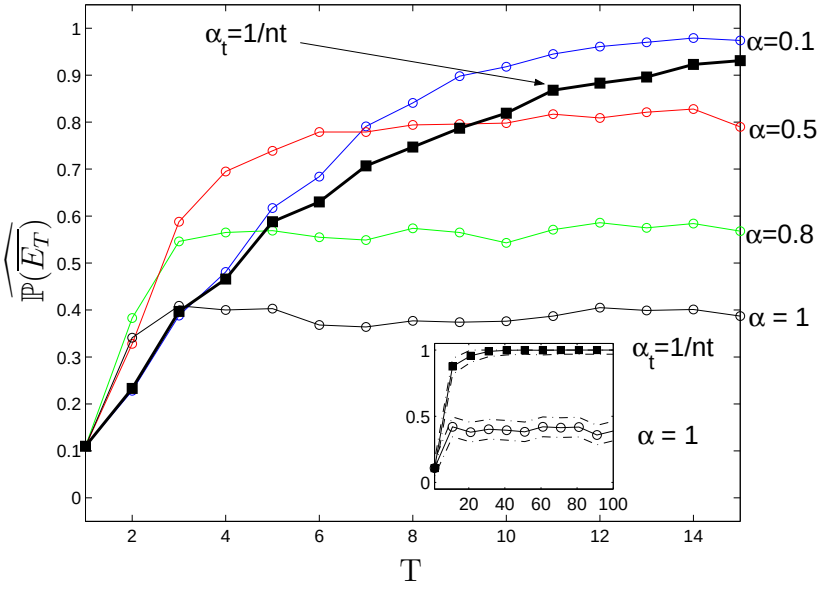

Fig. 1. Illustrative empirical results. The main figure shows transient behaviour of Algorithm 1. The inset shows limiting results for $\alpha = 1$ and $\alpha_t = \frac{1}{nt}$ with associated 95% confidence intervals (dotted lines) which are omitted from the main figure for clarity.

Examples of smoothing sequences which eventually generate the optimal solution with probability 1 (that is, which satisfy the sufficient condition of Theorem 2) include $\alpha_t = \frac{1}{(t+1)^\beta}$ and $\alpha_t = \frac{1}{(t+1)\log(t+1)^\beta}$, when $\beta > 1$, as well as $\alpha_t = \frac{1}{nt}$ where n is the "problem size" parameter introduced in Section 2. Figure 1 illustrates that the sequence $\alpha_t = \frac{1}{nt}$ 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 optimisation 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 $\mathbf{x} \in \mathcal{X}$).

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 the optimal solution with probability 1. For example, the smoothing sequence $\alpha_t = \frac{1}{nt}$ 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 the 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. If true, this would constitute a significant difference compared with the elite sample version of

the CE method analysed in [1,2], where a sufficient condition for both (a) and (b) is given.

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 the optimal solution is obtained in the short term, that is, in the first few iterations before the limiting regime is reached. Our results make a first step towards such a result (see Remark 2), but there is much scope for further research on transient behaviour of the CE method.

Acknowledgements

The authors are grateful to Alan Jones for his assistance, and to Felisa Vázquez-Abad and Reuven Rubinstein for helpful discussions. The authors also wish to thank the editors and anonymous referees for their efforts which helped us to improve this paper. 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] R. Y. Rubinstein, D. P. Kroese, The Cross-Entropy Method: A Unified Approach to Combinatorial Optimisation, Monte-Carlo Simulation, and Machine Learning, Springer, 2004.

[2] L. Margolin, On the convergence of the Cross-Entropy Method, Annals of Operations Research 134 (2004) 201–214.

[3] E. Aarts, J. Lenstra, Local Search in Combinatorial Optimisation, Wiley, Chichester, U.K., 1997.

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

[5] S. Andradóttir, A Global Search Method for Discrete Stochastic Optimization, SIAM Journal on Optimization 6 (1996) 513–530.

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

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

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

[9] M. Dorigo, T. Stutzle, Ant Colony Optimization, MIT Press, Cambridge, 2004.

[10] G. Nemhauser, L. Wolsley, Integer and Combinatorial Optimization, Wiley, 1988.

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