Fitness Threshold Accepting over Extremal Optimization ranks

Karl Heinz Hoffmann* and Frank Heilmann
Institut für Physik, Technische Universität Chemnitz, D-09107 Chemnitz, Germany

Peter Salamon
Department of Mathematics and Statistics, San Diego State University, San Diego, California 92182, U.S.A
(Dated: July 23, 2004)

We treat the problem of selecting the next degree of freedom for update in an Extremal Optimization algorithm designed to find the ground state of a system with a complex energy landscape. We show that there exists a best distribution for selecting the next degree of freedom in order to optimize any linear function of the state probabilities, e. g. the expected number of visits to the ground state. We dub the class of algorithms using this best distribution in conjunction with Extremal Optimization *Fitness Threshold Accepting*. In addition, we construct an extended random walk and use it to show that Fitness Threshold Accepting is optimal also for several other measures of algorithm performance such as maximizing the expected probability of seeing the ground state and minimizing the expected value of the lowest energy seen.

I. INTRODUCTION

The need for finding ground states of complex systems exists in many areas of modern science and engineering [1]. In the realm of physics, such problems include spin glasses [2], neural networks [3], and protein folding [4]. Traditional stochastic optimization methods to tackle those are e. g. Simulated Annealing (SA) and Threshold Accepting (TA). The former uses Metropolis sampling combined with decreasing temperature [5, 6]. The latter avoids the expenses of evaluating a large number of exponential functions by introducing a time dependent "threshold" [7, 8]: if the energy difference between current and proposed state is smaller than the threshold, the proposed move is accepted, otherwise not.

Many of these problems have an additional structure that is not used by traditional methods such as SA or TA. The additional structure implies that each state is defined by the value of many degrees of freedom such as the value of many spins [9] and that a certain fitness can be associated with each degree of freedom. When this additional structure is available, we can take advantage of it by using an algorithm known as *Extremal Optimization* (EO) [10].

We prove a theorem concerning the optimal selection of the next degree of freedom to change in EO. To prove our theorem, we first show how to view EO as an algorithm of the random walk type. This enables us to apply the techniques developed in Franz et al. [11] to show that any measure of the algorithm performance which depends linearly on the state probabilities will use a Fitness Threshold Accepting (FTA) rule [12]. We here show that a construction analogous to Hoffmann et al. [13] can be used to show that FTA is optimal for many other mea-sures of algorithm performance including the expected value of the best energy seen during the walk.

II. EXTREMAL OPTIMIZATION

EO is a recently introduced heuristics to find ground states of complex physical systems. Moreover it provides a general scheme to find the global minimum or other low lying states in multi-minima optimization problems. EO is a stochastic optimization algorithm similar to SA and TA in that EO simulates a random walker in the state space. However, EO needs a special structure of the problem under consideration: every state is specified by several degrees of freedom (DoF) each of which can be assigned a fitness. While such a structure is not needed for SA or TA, it is present in a significant fraction of the problems treated by these methods. A typical example is spin glass problems, where the state is described by a spin configuration, and each of the spin variables represent one DoF. The local field at one spin multiplied by the value of the spin can be taken as the fitness of that spin. In that case the objective function for the problem – namely the energy – is additive over the fitnesses of the different DoF's.

EO takes advantage of this additional structure to achieve better typical performance on such problems by randomly selecting one DoF to change at each step. In EO the next DoF to change is selected by first ranking the DoF's according to their fitness values and then selecting a rank. The DoF with the selected rank is to be changed during the next step. This procedure is iterated, yielding repeated improvements of the DoF's fitness.

Problems which cannot be addressed by EO include those which offer only a number of abstract states without internal structure. An exploration of examples of such state spaces using SA is given in [14].

---

*Electronic address: hoffmann@physik.tu-chemnitz.de

III. DEFINITIONS

Technically the EO algorithm can be viewed as a random walk in the space $\Omega = \{\alpha\}$ of states of the problem [10, 15]. In addition there is an objective function $E = E(\alpha)$ (in physical systems often the energy), assigning every state a real number.

Each of the states α is characterized by n degrees of freedom (DoF), which are each assigned real numbers called their *fitness*. For any of the $L = |\Omega|$ states $\alpha \in \Omega$, there is a neighborhood relationship that specifies $N(\alpha) \subseteq \Omega$, the set of states which can be reached from α by a change in one of the DoF's. More precisely a state $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ is specified by the values of n DoF's indexed by i, $1 \leq i \leq n$. The i-th degree of freedom has value $\alpha_i \in \alpha^{(i)} = \{\alpha_1^{(i)}, \ldots, \alpha_{m_i}^{(i)}\}$, where $\alpha^{(i)}$ is the set of possible values for the i-th DoF. The number of values m_i may differ for different DoF's, so $m_i = m(\alpha_i)$. We assume that $\infty > n > 1$, $\infty > m > 1$.

Each DoF is assigned a fitness $\lambda_i(\alpha_i)$, determining the ranking $k_i \in \mathbb{N}_n^* = \{1, 2, \ldots, n\}$, such that

$$k_i \leq k_j \text{ iff } \lambda_i \leq \lambda_j \quad \forall \text{ pairs } (i, j). \tag{1}$$

To complete the specification of the structure needed to perform an EO algorithm, we also need a time dependent probability distribution $d^t(k)$ over the ranks.

The EO algorithm proceeds as follows: Let us assume that the current state is $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$. First the DoF's of the current state β are ranked according to their fitness: the DoF with the smallest fitness has rank 1, the one with the highest fitness has rank n. Then a rank $k, 1 \leq k \leq n$ is selected with probability $d^t(k)$. This rank corresponds to a DoF, i, which is then changed by choosing with equal probability one of the possible values in $\alpha^{(i)} \setminus \{\beta_i\}$ so that the value of the k-th ranked DoF changes. The new state α is accepted unconditionally and β is set to α. This procedure is iterated until a stopping criterion is met.

Obviously the probability distribution used to select the rank of the DoF for the next state change plays a critical role for the performance of EO. Originally [15, 16], a distribution $\sim k^{-\tau}$ was used, introducing the single parameter $\tau > 0$. For a more in-depth discussion of EO in general, including motivation and issues related to defining fitnesses, we refer to the literature [15–17].

Here we focus on the dependence of the algorithm on a probability distribution over the ranks of the DoF and ask the question whether there exists a (provably) optimal choice for such a distribution. The class of distributions we will consider is characterized by the following rather weak conditions:

(A1) Each step is independent of the former steps.

(A2) At any epoch t, $1 \geq d^t(1) \geq d^t(2) \geq \ldots \geq d^t(n) \geq 0$, i. e. it is more probable to select a low rank (meaning a DoF with low fitness) than a high rank (meaning a DoF with high fitness).

(A3) $\sum_i d^t(k_i) = 1$: $d^t(k_i)$ is normalized.

IV. THE DYNAMICS

EO executes a random walk on the state space Ω. The probability of finding the random walker in state α at time step t is denoted by p_α^t. The conditions (A1) to (A3) guarantee that we are dealing with a Markov process [18]. Therefore the time development of p_α^t is described by the master equation

$$p_\alpha^t = \sum_{\beta \in \Omega} \Gamma_{\alpha\beta}^t p_\beta^{t-1} \tag{2}$$

with transition probabilities $\Gamma_{\alpha\beta}^t$. The random walk consists of a finite number of steps, $1 \leq t \leq S$. The transition probabilities of (2) are specified by the rules of EO to be

$$\Gamma_{\alpha\beta}^t = \begin{cases} \frac{1}{m-1} d^t(k_i) & \text{if } \alpha \text{ differs from } \beta \text{ only in} \\ & \qquad \text{the } i\text{-th DoF,} \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Note that these transition probabilities are linear functions of $d^t(k_i)$.

V. OPTIMAL ALGORITHMS

The goal of the random walk is to bring the walker as far down in the energy landscape as possible, controlling the random walk by choosing the probabilities $d^t(k)$ at each time step $t \in \{1, 2, \ldots, S\}$ in the algorithm with duration S steps. In order to achieve such control, a criterion is needed which quantifies this desire to come close to the global minimum of the energy function. Accordingly, we search for selection probabilities $d^t(k)$ which optimize some measure of how far down the random walker has gone.

The most common objective functions used to measure the quality of stochastic optimization procedures are:

(O1) the final mean energy $\langle E(\alpha(S)) \rangle$ should be as small as possible,

(O2) the final probability p_{GS}^S of ending up in the ground state should be as large as possible,

(O3) the expected number of visits to the ground state should be as large as possible,

(O4) the probability of visiting the ground state during the execution of an algorithm should be as large as possible,

(O5) the mean final BSF energy [19, 20] should be as small as possible. This so–called <u>B</u>est <u>S</u>o <u>F</u>ar energy of a given sequence or path $\alpha(t)$ up to step S is given as

$$E_{\text{BSF}}(S) = \text{Min}_{0 \leq t \leq S}\{E(\alpha(t))\} \tag{4}$$

and describes the lowest energy found along that path.

For the objectives (O1) and (O2) which are linear functions of the final state probabilities p_α^S, we showed [12] that Fitness Threshold Accepting is the best strategy to use. In the following we will prove that this applies for any objective which is a linear function of the state probabilities p_α^t during the whole process and not only at $t = S$. This extends the theorem to cover objective (O3).

Furthermore, we show that this is true not only for the given random walk but also for a class of Markov chains which can be constructed from the random walk. This construction will enable us to include objectives (O4) and (O5) in the cases to which the theorem applies. These last two goals are the crucial quantities of interest if EO is employed as we are interested in finding the ground state, or at least very low lying states, not at the *end* of the run but at any step *during* the run.

VI. THE PROBABILITY DISTRIBUTION FOR THE BSF ENERGY

In order to determine the mean final BSF energy we need the probability $B^S(E)$ to have seen an energy E or better up to time S. This probability can be obtained by considering a modified random walk which turns states at or below energy E into absorbing states [21]. This is achieved by introducing a modified transition probability matrix $\Gamma_{\alpha\beta;E}^t$

$$\Gamma_{\alpha\beta;E}^t = \begin{cases} \delta(\alpha,\beta) & \text{if } E(\beta) \leq E, \\ \Gamma_{\alpha\beta}^t & \text{if } E(\beta) > E, \end{cases} \tag{5}$$

where $\delta(\alpha,\beta)$ is Kronecker's delta. Note that these modifications still keep the dependence on the selection probabilities $d^t(k_i)$ linear in all of the $\Gamma_{\alpha\beta;E}^t$.

In the modified stochastic process a random walker reaching a state with energy less than or equal to E is trapped at that state. Evolving the associated probability distribution $p_{\alpha;E}^t$

$$p_{\alpha;E}^t = \sum_{\beta \in \Omega} \Gamma_{\alpha\beta;E}^t p_{\beta;E}^{t-1}, \tag{6}$$

gives the probability of being in state α of the modified chain after t steps. For $E(\alpha) > E$ this is the same as the probability of being in state α in the unmodified random walk and not having visited any states with an energy less than or equal to E before time t. The probability to have visited a state with energy less than or equal to E up to time S in the unmodified random walk is thus

$$B^S(E) = \sum_{\alpha:E(\alpha)\leq E} p_{\alpha;E}^S. \tag{7}$$

For the full distribution of the BSF energy we need to repeat this modified random walk for all possible energy values in the system. Due to the finiteness of the

state space we can sort the (finite) number of different energy values in ascending order and label them E_r, $r \in \{1, 2, \ldots, R\}$ to get $E_1 < E_2 < \ldots < E_R$. Then for every r the corresponding $B^S(E_r)$ is determined and the probability that the lowest energy visited is E_r is given by

$$b^S(E_r) = B^S(E_r) - B^S(E_{r-1}), \tag{8}$$

where we introduce for convenience an additional energy E_0, which is an arbitrary energy value lower than the ground state energy E_1. Since E_0 is less than any energy reachable by the chain, we must have $p_{\alpha;E_0}^t = p_\alpha^t$, $\Gamma_{\alpha\beta;E_0}^t = \Gamma_{\alpha\beta}^t$, and $B^S(E_0) = 0$.

Then from eq. (8) the mean BSF energy $\langle E_{\text{BSF}}(S) \rangle$ is obtained as

$$\langle E_{\text{BSF}}(S) \rangle = \sum_{r=1}^R b^S(E_r) E_r. \tag{9}$$

Summarizing the above, for each r the master equation with corresponding modified transition probabilities $\Gamma_{\alpha\beta;E_r}^t$ needs to be iterated. This can be presented in a compact way by introducing a vector/matrix notation for the original master equation (2):

$$\boldsymbol{p}^t = \boldsymbol{\Gamma}^t \cdot \boldsymbol{p}^{t-1} \tag{10}$$

where $\boldsymbol{p}^t$ is the vector of probabilities p_α^t representing the state of the random walk at time t and $\boldsymbol{\Gamma}^t$ is the transition matrix with entries $\Gamma_{\alpha\beta}^t$. Similarly, (6) is expressed as

$$\boldsymbol{p}_{E_r}^t = \boldsymbol{\Gamma}_{E_r}^t \cdot \boldsymbol{p}_{E_r}^{t-1}. \tag{11}$$

Combining all the probability vectors $\boldsymbol{p}_{E_r}^t$ ($r \in \{0, 1, \ldots, R\}$) into one vector $\boldsymbol{q}^t$, we can write

$$\begin{aligned} \boldsymbol{q}^{t+1} &= \begin{pmatrix} \boldsymbol{p}_{E_0}^{t+1} \\ \boldsymbol{p}_{E_1}^{t+1} \\ \vdots \\ \boldsymbol{p}_{E_R}^{t+1} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Gamma}_{E_0}^t & 0 & \cdots & 0 \\ 0 & \boldsymbol{\Gamma}_{E_1}^t & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{\Gamma}_{E_R}^t \end{pmatrix} \begin{pmatrix} \boldsymbol{p}_{E_0}^t \\ \boldsymbol{p}_{E_1}^t \\ \vdots \\ \boldsymbol{p}_{E_R}^t \end{pmatrix} \\ &= \tilde{\boldsymbol{\Gamma}}^t \cdot \boldsymbol{q}^t. \end{aligned} \tag{12}$$

Thus for $\alpha \in \{1, \ldots, L\}$ and $r \in \{0, \ldots, R\}$ we have $q_{Lr+\alpha}^t = p_{\alpha;E_r}^t$. Hence the time development of the unmodified chain is contained in $q_\gamma^t, \gamma = 1, \ldots, L$. The mean BSF energy can be expressed as

$$\begin{aligned} \langle E_{\text{BSF}}(S) \rangle &= \sum_{r=1}^R E_r \left(B^S(E_r) - B^S(E_{r-1}) \right) \\ &= \sum_{r=1}^R E_r \left(\sum_{\alpha:E(\alpha)\leq E_r} p_{\alpha;E_r}^S - \sum_{\alpha:E(\alpha)\leq E_{r-1}} p_{\alpha;E_{r-1}}^S \right) \\ &= \sum_{r=1}^R E_r \left(\sum_{\alpha:E(\alpha)\leq E_r} q_{Lr+\alpha}^S - \sum_{\alpha:E(\alpha)\leq E_{r-1}} q_{L(r-1)+\alpha}^S \right). \end{aligned}$$

Note that all our objective functions (O1) - (O5) are linear functions of the probabilities q_γ^t, $\gamma \in \{1, \ldots, L(R+1)\}$, $t \in \{1, \ldots, S\}$, a fact which is central to the arguments below.

VII. THE THEOREM

In an earlier publication [12] we proved that for selection probabilities with properties (A1)-(A3), the optimal strategy is to use Fitness Threshold Accepting. Here we extend that work by investigating also objectives (O3) - (O5), and more generally, any objective function which is a linear function of q_γ^t.

THEOREM: For any optimization algorithm in an EO framework satisfying properties (A1) to (A3), and any objective function which is linear in the probabilities q_γ^t, $\gamma \in \{1, \ldots, L(R+1)\}$, $t \in \{1, \ldots, S\}$ of the extended random walk constructed above, the optimal selection probability is Fitness Threshold Accepting.

VIII. THE PROOF

Fundamental to our proof is the fact that all of the entries in the transition matrix $\tilde{\mathbf{\Gamma}}^t$ are linear functions of the selection probabilities $d^t(k_i)$. In the following it is convenient to introduce a vector notation. Just as we changed from denoting the vector of probabilities p_α^t as an L-dimensional vector $\boldsymbol{p}^t$, and the $L(R+1)$ dimensional state vectors by $\boldsymbol{q}^t$, we associate a sequence of vectors $\boldsymbol{F}^t$ of $L(R+1)$ elements to any linear function of the state vectors $\boldsymbol{q}^t, t = 1, \ldots, S$. Our theorem concerns any objective function linear in the $\boldsymbol{q}^t$, i. e. minimizing any function of the form

$$F(\boldsymbol{q}^1, \boldsymbol{q}^2, \ldots, \boldsymbol{q}^S) = \sum_{t=1}^{S} (\boldsymbol{F}^t)^{\mathrm{tr}} \cdot \boldsymbol{q}^t = \sum_{t=1}^{S} \sum_{i=1}^{L(R+1)} F_i^t q_i^t$$
$$\rightarrow \min, \qquad (13)$$

where $(\cdot)^{\mathrm{tr}}$ denotes transpose and the minimum is taken over all possible sequences of selection probabilities $d^t(k_i)$, $t = 1, \ldots, S$.

The vectors $\boldsymbol{F}^t$ may be any arbitrary $L(R+1)$-tuples of numbers. For instance for the criteria (O1) to (O5) these are

(O1) for minimizing the mean final energy,

- $F_\gamma^t = 0$ for $t < S$,
- $\dfrac{F_\gamma^S}{E(\gamma)} =$ for $\gamma \leq L$,
- $F_\gamma^S = 0$ for $\gamma > L$;

(O2) for maximizing the final ground state probability,

$$\sum_{\gamma \leq L, E(\gamma) = E_1} p_\gamma^S,$$

$F_\gamma^t = 0$ unless $t = S$, $\gamma \leq L$ and $E(\gamma) = E_1$ in which case $F_\gamma^S = -1$;

(O3) for maximizing the expected number of visits to the ground state, $F_\gamma^t = 0$ unless $\gamma \leq L$ and $E(\gamma) = E_1$, in which case $F_\gamma^t = -1$;

(O4) for maximizing the probability of visiting the ground state during the execution of the algorithm, we need to maximize $b^S(E_1) = B^S(E_1)$. Thus $F_\gamma^t = 0$ unless $t = S$, $L < \gamma \leq 2L$, and $E(\gamma) = E_1$, in which case $F_\gamma^S = -1$.

(O5) for minimizing the mean BSF energy,

- $F_\gamma^t = 0$ for $t < S$,
- $F_{Lr+\alpha}^S = 0$ for $r \in \{0, \ldots, R-1\}$, $\alpha \in \{1, \ldots, L\}$ and $E(\alpha) > E_r$,
- $F_{Lr+\alpha}^S = E_r - E_{r+1}$ for $k \in \{0, \ldots, R-1\}$, $\alpha \in \{1, \ldots, L\}$ and $E(\alpha) \leq E_r$,
- $F_{LR+\alpha}^S = E_R$ for $\alpha \in \{1, \ldots, L\}$.

Let us consider the distributions $d^t(k_i)$ as an n-dimensional vector $\boldsymbol{d}^t$ with entries d_i^t in $[0, 1]$. We note some consequences of our condition (A2): The possible range for the vector $\boldsymbol{d}^t$ is a simplex X in an n-dimensional space. The vertices of the simplex are those vectors $\boldsymbol{d}^t$ which have an initial sequence of ones, followed by zeros. Condition (A3) defines a hyperplane H in the same n-dimensional space. The intersection $I = X \cap H$, having dimension $n-1$, is the final set of allowed selection distributions $d^t(k)$. Figure 1 shows the construction for $n = 3$. Note that I is itself a simplex; we denote the set of the vertices of I with $V = \{\boldsymbol{v}_i\}$.
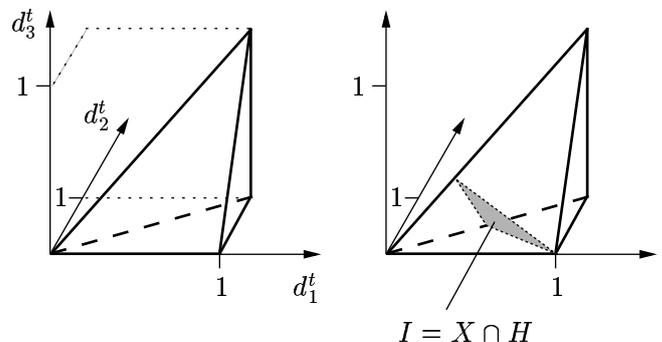

FIG. 1: Construction of the simplex X in three dimensions. The axes on the left side denote the entries of $\boldsymbol{d}^t$. Condition (A2) (see text) defines X. On the right side the effect of condition (A3) is shown: X is cut by H. The vertices of $X \cap H$ can be computed.

Based on this construction we are able to compute the elements of V explicitly. According to (A2), the point $(1, 0, \ldots, 0)^{tr}$ is an element of V. Let us start a search for the rest of the vertices of I. We can do this by decreasing $d^t(1)$ and increasing $d^t(2)$ while leaving all other $d^t(k_i)$

untouched until we cannot go any further. This results in the point $(1/2, 1/2, 0, \ldots, 0)^{tr}$. From there we decrease the first two coordinates while increasing the third, again leaving the rest untouched, until we cannot go further again. The result is the point $(1/3, 1/3, 1/3, 0, \ldots, 0)^{tr}$. We repeat this procedure until the last vertex is found: the point $(1/n, 1/n, \ldots, 1/n)^{tr}$. Note that the elements of V are pairwise distinct and linearly independent.

We show that the convex hull of the $\boldsymbol{v}_i$

$$C(V) = \sum_{i=1}^n a_i \boldsymbol{v}_i = a_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + a_2 \begin{bmatrix} 1/2 \\ 1/2 \\ \vdots \\ 0 \end{bmatrix} + \ldots + a_n \begin{bmatrix} 1/n \\ 1/n \\ \vdots \\ 1/n \end{bmatrix} \tag{14}$$

with the coefficients a_i obeying $a_i \in [0, 1]$, $\sum_i a_i = 1$ is the largest possible simplex in $(n-1)$-dimensional space fulfilling (A2, A3).

The lth row d_l^t of (14) is

$$d_l^t = \sum_{i=l}^n a_i \frac{1}{i} = \sum_{i=l+1}^n a_i \frac{1}{i} + a_l \frac{1}{l} = d_{l+1}^t + a_l \frac{1}{l} \geq d_{l+1}^t, \tag{15}$$

so (A2) is fulfilled. Summing up the rows of $C(V)$ gives

$$\sum_{l=1}^n d_l^t = \sum_{l=1}^n \sum_{i=l}^n a_i \frac{1}{i} = \sum_{l=1}^n l a_l \frac{1}{l} = \sum_{l=1}^n a_l = 1, \tag{16}$$

showing that (A3) is also fulfilled: $C(V) \subset I$.

Now we look at an arbitrary point $\boldsymbol{p} \in I$. Using the basis $\{\boldsymbol{v}_i\} \subset I$ for describing $\boldsymbol{p}$'s coordinates p_l we get

$$p_l = \sum_{i=l}^n b_i \frac{1}{i} = p_{l+1} + b_l \frac{1}{l} \tag{17}$$

and due to (A2)

$$p_l \geq p_{l+1} \Rightarrow p_l - p_{l+1} = b_l \frac{1}{l} \geq 0 \Rightarrow b_l \geq 0. \tag{18}$$

Summing up all p_l and using (A3) gives

$$\sum_{l=1}^n p_l = \sum_{l=1}^n l b_l \frac{1}{l} = \sum_{l=1}^n b_l = 1 \Rightarrow b_l \leq 1. \tag{19}$$

So we have $b_l \geq 0$ and $b_l \leq 1$, therefore $\boldsymbol{p} \in C(V) \, \forall \boldsymbol{p} \in I$, i. e. $I \subset C(V)$.

IX. THE SOLUTION

The optimization task (13) for the dynamic process described by (12) is a discrete control problem, where the controls are the selection probability vectors $\boldsymbol{d}^t$. We apply the Bellman principle of dynamic programming [22], and work our way backwards starting with the last step. The scheme of our dynamic programming problem is illustrated in figure 2.
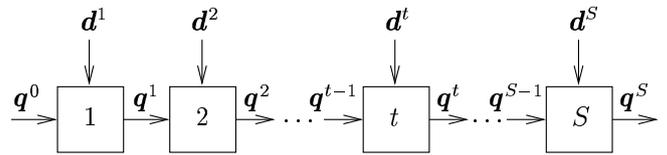

FIG. 2: The dynamic optimization process. An input $\boldsymbol{q}^{t-1}$ is transformed into an output $\boldsymbol{q}^t$ under the influence of some control $\boldsymbol{d}^t$ (see text).

In every step t, an input $\boldsymbol{q}^{t-1}$ is transformed into the output $\boldsymbol{q}^t$ under the influence of the control $\boldsymbol{d}^t$. Finally, the output for all the steps is used to determine the optimality criterion $F(\boldsymbol{q}^1, \boldsymbol{q}^2, \ldots, \boldsymbol{q}^S)$. Let us first consider the last step S. For given inputs $\boldsymbol{q}^t$, $1 \leq t \leq S-1$, we have to solve the optimization problem

$$\begin{aligned} \sum_{t=1}^S (\boldsymbol{F}^t)^{\text{tr}} \cdot \boldsymbol{q}^t &= \sum_{t=1}^{S-1} (\boldsymbol{F}^t)^{\text{tr}} \cdot \boldsymbol{q}^t + (\boldsymbol{F}^S)^{\text{tr}} \cdot \boldsymbol{q}^S \\ &= \text{const.} + (\boldsymbol{F}^S)^{\text{tr}} \tilde{\boldsymbol{\Gamma}}^S \boldsymbol{q}^{S-1} \\ &\rightarrow \min, \end{aligned} \tag{20}$$

where, as noted above, the matrix elements $\tilde{\Gamma}_{ij}^S$ given in (12) depend linearly on the control vector $\boldsymbol{d}^S$. The possible range for $\boldsymbol{d}^S$ is the simplex described in the previous section. Hence we have to find the minimum of a linear function on a simplex. By the fundamental theorem of linear programming [23], this minimum is found at one of the vertices in V, i. e. at a Fitness Threshold Acceptance function [12]. Call this vertex $\boldsymbol{v}^S$. Of course this vertex $\boldsymbol{v}^S$ depends on the input $\boldsymbol{q}^{S-1}$, i. e. $\boldsymbol{v}^S = \boldsymbol{v}^S(\boldsymbol{q}^{S-1})$.

Now let us continue with the second to last step $S-1$. For given inputs $\boldsymbol{q}^t$, $1 \leq t \leq S-2$, we have to solve the optimization problem

$$\begin{aligned} \text{const.} \ + \ &(\boldsymbol{F}^{S-1})^{\text{tr}} \cdot \tilde{\boldsymbol{\Gamma}}^{S-1} \boldsymbol{q}^{S-2} \\ + \ &(\boldsymbol{F}^S)^{\text{tr}} \tilde{\boldsymbol{\Gamma}}^S (\boldsymbol{v}^S) \tilde{\boldsymbol{\Gamma}}^{S-1} \boldsymbol{q}^{S-2} \\ \rightarrow \ &\min, \end{aligned} \tag{21}$$

where we now already know that $\boldsymbol{\Gamma}^S(\boldsymbol{v}^S)$ is a transition matrix corresponding to Fitness Threshold Acceptance. For fixed $\boldsymbol{v}^S$ the optimization problem (21) is again a linear problem with the same structure as (20) over the same range. Thus the optimal control is found at one of the vertices in V, which we call $\boldsymbol{v}^{S-1}$. This vertex $\boldsymbol{v}^{S-1}$ depends on the input $\boldsymbol{q}^{S-2}$ and on the vertex $\boldsymbol{v}^S$, i. e. $\boldsymbol{v}^{S-1} = \boldsymbol{v}^{S-1}(\boldsymbol{q}^{S-2}, \boldsymbol{v}^S)$. Since the vertex set V is finite, there is a vertex $\boldsymbol{v}^S$ which gives the minimum over all $|V|$ possible minimum values in problem (21). In a similar way, we process the remaining steps of the dynamical optimization problem from the end to the beginning. At each step we find a linear optimization problem over the same simplex range which attains its minimum at one of the vertices thereby completing the proof.

The proof shows that a uniform distribution over some of the "least fit" DoF's gives the best implementation of

EO. The resulting class of algorithms – EO in combination with this special distribution $d^t(k)$ – was dubbed *Fitness Threshold Accepting* (FTA), because in analogy to TA all moves selecting ranks which lie under a certain fitness threshold are selected with equal probability.

We remark that our proof does not state that *all* optimal strategies are of the given form. In principle other strategies may do equally well (but certainly not better). Other equally good strategies can only occur if an edge or a face of the simplex does equally well as one of the vertices belonging to it. The optimality of such an edge corresponds to selecting the least r ranks with equal probability doing as well as selecting the least $r-1$ ranks. Finally we remark that the optimality of a strictly monotonic distribution such as $d^t(k) \propto \tau^{-k}$ would imply that all the vertices in V do equally well, a case which can only happen for rather trivial problems.

X. CONCLUSIONS

We considered certain measures of algorithm performance associated with the problem of finding the ground state of a complex system by using the heuristic known as Extremal Optimization. We used a master equation to describe the corresponding dynamics of random walkers on state space and formulated some straightforward conditions on the probability distribution for selecting the degree of freedom (DoF) to change at the next step.

Our goal was to find selection probabilities which optimally control the movements of the random walkers. We found that a special distribution of selection probabilities is provably optimal provided the performance of the random walk is measured by a linear function in the state probabilities. This includes minimizing the expected final energy or maximizing the probability of being in the ground state at the final time. By constructing an extended dynamics, we were able to show that linear measures include the expected probability of visiting the ground state and the expected value of the best energy seen. We named the resulting optimal control Fitness Threshold Accepting since it always selects with equal probability from those degrees of freedom with fitness values below a certain threshold.

We assumed that the set of possible values of each degree of freedom and the number of degrees of freedom are the same for each state. This assumption can be dropped without affecting the arguments or the conclusions – the only effect would be the necessity of a much more cumbersome notation.

We did not show that Fitness Threshold Accepting is the *only* optimal way to implement Extremal Optimization. Our proof shows that using any linear performance measure, including the measures (O1) to (O5), would make a strictly monotonic distribution over ranks k, such as $d^t(k) \propto \tau^{-k}$ advocated by Boettcher et al. [10], optimal only if *all* selection distributions perform equally well.

Without knowing what the optimal thresholds are, the knowledge that best performance can be achieved using Fitness Threshold Accepting is of limited use. In particular, Fitness Threshold Accepting with the wrong thresholds might be outperformed by other well adapted selection distributions.

We considered only algorithms based on Extremal Optimization. The possibility of better algorithms not based on Extremal Optimization remains. Within the given field, however, the arguments presented here establish the structure of a provably optimal strategy, which can benefit the further study of heuristic approaches to global minimization.

Our proof had to assume a finite state space. We postpone the exploration of continuous state spaces to future efforts but note that discrete arithmetic on digital computers make state spaces effectively finite. Finally, our proof was based on the assumption that the objective measuring the performance of the Extremal Optimization is a linear function of the state probabilities. While this includes most desirable measures, it does not include them all.

[1] P. Salamon, P. Sibani, and R. Frost, *Facts, Conjectures, and Improvements for Simulated Annealing*, vol. 7 of *Monographs on Mathematical Modeling and Computation* (SIAM, 2002), 1st ed.

[2] K. H. Fischer and J. A. Hertz, *Spin Glasses* (Cambridge University Press, 1991).

[3] J. A. Hertz and R. Palmer, *Introduction to the Theory of Neural Computation* (Addison-Wesley, Redwood City, CA, 1991).

[4] K. D. Ball, R. S. Berry, R. E. Kunz, F. Y. Li, A. Proykova, and D. J. Wales, Science **271**, 963 (1996).

[5] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).

[6] S. Kirkpatrick, C. Gelatt, and M. Vecchi, Science **220**, 671 (1983).

[7] G. Dueck and T. Scheuer, J. Comput. Phys. **90**, 161 (1990).

[8] P. Moscato and J. Fontanari, Phys. Lett. A **146**, 204 (1990).

[9] A. K. Hartmann, Phys. Rev. B **64**, 224430/1 (2001).

[10] S. Boettcher and A. G. Percus, Phys. Rev. Lett. **86**, 5211 (2001).

[11] A. Franz, K. Hoffmann, and P. Salamon, Phys. Rev. Lett. **86**, 5219 (2001).

[12] F. Heilmann, K. H. Hoffmann, and P. Salamon, Europhys. Lett. **66**, 305 (2004).

[13] K. H. Hoffmann, A. Franz, and P. Salamon, Phys. Rev. E **66**, 046706 (2002).

[14] A. Fachat, K. Hoffmann, and A. Franz, Comp. Phys.

Comm. **132**, 232 (2000).

[15] S. Boettcher and A. Percus, Artificial Intelligence **119**, 275 (2000), research note.

[16] S. Boettcher and A. G. Percus, in *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '99)* (1999), pp. 826–832.

[17] S. Boettcher, A. G. Percus, and M. Grigni, in *Proceedings of the Sixth International Conference on Parallel Problem Solving from Nature* (2000), vol. 1917 of *Lecture Notes in Computer Science*, pp. 447–456.

[18] N. G. van Kampen, *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam, 1997).

[19] M. O. Jakobsen, K. Mosegaard, and J. M. Pedersen, *Model Optimization in Exploration Geophysics II,* edited by A. Vogel (Friedr. Vieweg & Son, Braunschweig, 1988).

[20] K. Hoffmann, P. Sibani, J. Pedersen, and P. Salamon, Appl. Math. Lett. **3**, 53 (1990).

[21] J. G. Kemeny and J. L. Snell, *Finite Markov Chains* (Van Nostrand, Princeton, 1960).

[22] R. E. Bellman and S. E. Dreyfus, *Applied Dynamic Programming* (Princeton University Press, Princeton, 1962).

[23] D. G. Luenberger, *Linear and Nonlinear Programming* (Addison-Wesley, Reading, MA, 1984).