EUROPHYSICS LETTERS

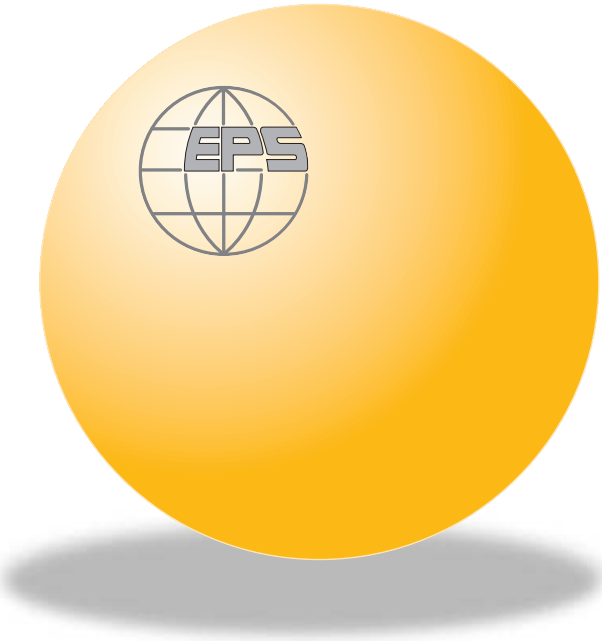
OFFPRINT

Vol. 66 • Number 3 • pp. 305–310

Best possible probability distribution over extremal optimization ranks

* * *

F. Heilmann, K. H. Hoffmann and P. Salamon



Published under the scientific responsibility of the
EUROPEAN PHYSICAL SOCIETY
Incorporating
JOURNAL DE PHYSIQUE LETTRES • LETTERE AL NUOVO CIMENTO

EDP SCIENCES

EUROPHYSICS LETTERS

Europhysics Letter was launched more than fifteen years ago by the European Physical Society, the Société Française de Physique, the Società Italiana di Fisica and the Institute of Physics (UK) and owned now by 17 National Physical Societies/Institutes.

Europhysics Letters aims to publish short papers containing non-trivial new results, ideas, concepts, experimental methods, theoretical treatments, etc. which are of broad interest and importance to one or several sections of the physics community.

Europhysics letters provides a platform for scientists from all over the world to offer their results to an international readership.

EUROPHYSICS LETTERS 1 May 2004

Europhys. Lett., **66** (3), pp. 305–310 (2004)
DOI: 10.1209/epl/i2004-10011-3

Best possible probability distribution over extremal optimization ranks

F. Heilmann[1], K. H. Hoffmann[1](*) and P. Salamon[2]
[1] *Institut für Physik, Technische Universität Chemnitz - D-09107 Chemnitz, Germany*
[2] *Department of Mathematics and Statistics, San Diego State University San Diego, CA 92182, USA*

(received 9 January 2004; accepted in final form 2 March 2004)

PACS. 02.50.Ga – Markov processes.
PACS. 02.60.Pn – Numerical optimization.
PACS. 05.10.-a – Computational methods in statistical physics and nonlinear dynamics.

Abstract. – We consider the problem of selecting the next degree of freedom (DoF) for update in an extremal optimization algorithm designed to find the ground state of a system with a complex energy landscape. We show that if we wish to minimize any linear function of the state probabilities, *e.g.* the final energy, then the best distribution for selecting the next DoF is a rectangular distribution with a cutoff for the fitness. We dub the family of algorithms using rectangular distributions in combination with extremal optimization *Fitness Threshold Accepting*.

Introduction. – Finding the ground state of a complex physical, chemical or combinatorial problem is a big challenge. The challenge is due in part to the typically huge cardinality of the state space and in part to the complex topology of the energy landscape defined on that space. The characteristic property of this energy landscape is that it is fraught with exponentially many local minima.

Generally, this causes deterministic algorithms to run very slowly. Alternatively, Monte Carlo (MC) methods like simulated annealing (SA) [1], Tsallis statistics [2, 3] and threshold accepting (TA) [4,5] are used. Despite the fact that finding the ground state in a finite number of simulation steps cannot be guaranteed, such methods are widely used because good results can be found in most cases.

The basic feature of MC methods is that a random walker in state space iterates moves from the current state to a randomly selected neighbouring state. The selection of the new state is usually performed in two steps. The first step selects a candidate move and the second step decides whether to accept this move or to stay at the current location. The decision whether to take a specific step is called the acceptance rule. A walk consists of a large but finite number of such steps. In each step, the acceptance rule may be altered due to a fixed schedule. The probability of taking the step is the product of the probability for choosing the neighbour and the probability for accepting the proposed move.

(*) E-mail: hoffmann@physik.tu-chemnitz.de

The acceptance rule and the schedule are to be chosen such that the random walker is brought as far down in the energy landscape as possible. The performance of such walks can be measured by analyzing the final probability distribution reached. For example, the mean final energy or the final probability for having the walker in the ground state are common measures of how well an algorithm performs. These should be, respectively, as low or high as possible. Under such circumstances, TA is provably the best possible choice for the mentioned acceptance rule [6].

In this paper, we examine the recently suggested heuristics called extremal optimization (EO) [7]. EO is a stochastic optimization algorithm similar to simulated annealing (SA) and threshold accepting (TA). EO also works by simulating random walkers, but 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. This typically arises by a special structure of the objective function for the problem which, *e.g.*, can be partially additive over the DoF. Examples of problems that have this structure include spin glasses and traveling-salesman problems. 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 according to their fitness values and then selecting one of the ranks according to a probability distribution defined on them. The DoF with the selected rank is to be changed during the next step. Based on the approach presented in [6] we show that there exists a provably best possible distribution over the ranks.

Definitions. – We begin by formalizing needed definitions and assumptions. A system is described by a discrete and finite set of states $\Omega = \{\alpha\}$ and an energy function $E = E(\alpha)$, assigning every state an energy value. In addition, the system possesses a neighbourhood relationship, with $N(\alpha) \subseteq \Omega$ denoting the set of states which are one step away from α.

Each state $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ is specified by the values of n DoF indexed by i, $1 \leq i \leq n$. The i-th degree of freedom has value $\alpha_i \in \alpha^{(i)} = \{\alpha_1^{(i)}, \ldots, \alpha_m^{(i)}\}$, where $\alpha^{(i)}$ is the set of possible values for the i-th DoF and n and m are finite. 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

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

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. Originally, a time-independent distribution $\propto k^{-\tau}$ was used, introducing the single parameter $\tau > 0$ [8,9].

Given the structure above, an EO random walk in Ω proceeds from an initial state $\beta = (\beta_1, \beta_2, \ldots, \beta_n)$ as follows [7,8]. First, 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. This basic step is iterated many times.

For a more in-depth discussion of EO in general, including motivation and issues related to defining fitnesses, we refer to the literature [8–10]. Here we only stress 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.

We adopt the following assumptions:

A1) Each step is independent of the former steps.

A2) At any epoch t, $1 \geq d^t(1) \geq d^t(2) \geq \cdots \geq d^t(n) \geq 0$, *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.

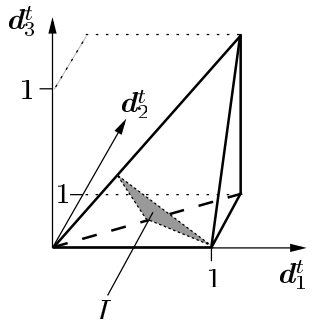


Fig. 1 – The simplex I in three dimensions.

These assumptions guarantee that we are dealing with a Markov process [11]. 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, $1 \le t \le S$. We are interested in controlling the walk so as to bring the walker energetically as low as possible. At the last step we measure the performance by some function of the final state probabilities p_α^S. Most commonly one of the following objectives is used [6]:

O1) The final mean energy should be as small as possible.

O2) The final probability of being in the ground state should be as large as possible.

Note that these objectives are linear functions of the final state probabilities p_α^S. The arguments given below apply to any linear function of the state probabilities. We consider the state probabilities at time t and the linear objective function F as vectors $\boldsymbol{p}^t$ and $\boldsymbol{F}$. The measure of the performance of the random walk can then be written as

$$F\big(\boldsymbol{p}^S\big) = \boldsymbol{F}^{\mathrm{tr}} \cdot \boldsymbol{p}^S = \sum_{\alpha\in\Omega} F_\alpha p_\alpha^S\,, \tag{3}$$

with $(\cdot)^{\mathrm{tr}}$ denoting transpose.

The formal problem. – The transition probabilities of (2) are specified by the rules of EO to be

$$\Gamma_{\alpha\beta}^t = \begin{cases} \dfrac{1}{m-1} d^t(k_i) & \text{if } \alpha \text{ differs from } \beta \text{ only in the } i\text{-th DoF,} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Note that these transition probabilities are linear functions of $d^t(k_i)$. We will be concerned with selecting this distribution $d^t(k_i)$ so as to minimize

$$F\big(\boldsymbol{p}^S\big) = \boldsymbol{F}^{\mathrm{tr}} \cdot \boldsymbol{p}^S = \sum_{\alpha\in\Omega} F_\alpha p_\alpha^S \longrightarrow \min\,. \tag{5}$$

Following [6], we consider the distributions $d^t(k_i)$ as an n-dimensional vector $\boldsymbol{d}^t$ with entries $\boldsymbol{d}_i^t$ in $[0,1]$. As a consequence of our assumptions A2) and A3), the region I of admissible vectors $\boldsymbol{d}^t$ is defined by the $n+1$ linear inequalities in A2) along with one linear equality in A3). The

first inequality, $1 \geq \boldsymbol{d}_1^t$, follows from the others. Once we know that the $\boldsymbol{d}_i^t$ are all non-negative and sum to one, they must all be less than or equal to one. Of the remaining n inequalities, exactly $n-1$ of them must be set to equalities to find the extreme points (vertices) of the region I. Letting V denote the set of extreme points of I, the elements of V are exactly those vectors $\boldsymbol{d}^t$ that have an initial sequence of i entries equal to $1/i$, followed by the remaining $n-i$ entries equal to zeros. Explicitly, $V = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n\}$, where $\boldsymbol{v}_1 = (1,0,0,\ldots,0)^{\mathrm{tr}}$, $\boldsymbol{v}_2 = (1/2,1/2,0,0,\ldots,0)^{\mathrm{tr}}, \ldots, \boldsymbol{v}_i = (1/i,1/i,\ldots,1/i,0,0,\ldots 0)^{\mathrm{tr}}$, and $\boldsymbol{v}_n = (1/n,1/n,\ldots 1/n)^{\mathrm{tr}}$. Note that the elements of V are linearly independent. Figure 1 illustrates the situation for $n=3$, where the vertices are $(1,0,0)^{\mathrm{tr}}$, $(1/2,1/2,0)^{\mathrm{tr}}$, and $(1/3,1/3,1/3)^{\mathrm{tr}}$.

In fact I is exactly the convex hull of V,

$$C(V) = \left\{ \sum_{i=1}^{n} a_i \boldsymbol{v}_i = a_1 \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} + a_2 \begin{bmatrix} 1/2 \\ 1/2 \\ \vdots \\ 0 \end{bmatrix} + \ldots a_n \begin{bmatrix} 1/n \\ 1/n \\ \vdots \\ 1/n \end{bmatrix} ; a_i \in [0,1], \sum_{i=1}^{n} a_i = 1 \right\}, \tag{6}$$

which is a simplex. To see this, consider the l-th row of an element d^t of $C(V)$:

$$d_l^t = \sum_{i=l}^{n} a_i \frac{1}{i} = \sum_{i=l+1}^{n} a_i \frac{1}{i} + a_l \frac{1}{l} = d_{l+1}^t + a_l \frac{1}{l} \geq d_{l+1}^t \,, \tag{7}$$

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

$$\sum_{l=1}^{n} d_l^t = \sum_{l=1}^{n} \sum_{i=l}^{n} a_i \frac{1}{i} = \sum_{l=1}^{n} l a_l \frac{1}{l} = \sum_{l=1}^{n} a_l = 1, \tag{8}$$

showing that A3) is also fulfilled. Thus, $C(V) \subset I$. Conversely, consider an arbitrary point $\boldsymbol{p} \in I$. Since the vertices $\boldsymbol{v}_i$ are linearly independent, we can use them as a basis and write $\boldsymbol{p}$ as a linear combination:

$$\boldsymbol{p} = \sum_{i=1}^{n} b_i \boldsymbol{v}_i. \tag{9}$$

For the l-th component p_l this gives

$$p_l = \sum_{i=l}^{n} b_i \frac{1}{i} = p_{l+1} + b_l \frac{1}{l} \,, \tag{10}$$

which by A2) implies

$$p_l \geq p_{l+1} \Longrightarrow p_l - p_{l+1} = b_l \frac{1}{l} \geq 0 \Longrightarrow b_l \geq 0. \tag{11}$$

Summing up all p_l and using A3) gives

$$\sum_{l=1}^{n} p_l = \sum_{l=1}^{n} l b_l \frac{1}{l} = \sum_{l=1}^{n} b_l = 1 \Longrightarrow b_l \leq 1. \tag{12}$$

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.* $I \subset C(V)$.

The solution. – We are now prepared to tackle the optimization problem (5). Following the approach in [6], we apply the Bellman principle of dynamic programming [12], and work our way backwards starting with the last step. The output of the last step $\boldsymbol{p}^S$ is used to determine the optimality criterion (5).

In the last step S, we have to solve the optimization problem (5) for a given input $\boldsymbol{p}^{S-1}$. Using (2) we get

$$F(\boldsymbol{p}^S) = \sum_{\alpha,\beta\in\Omega} F_\alpha \Gamma^S_{\alpha\beta} p^{S-1}_\beta \longrightarrow \min \tag{13}$$

with the matrix elements $\Gamma^S_{\alpha\beta}$ given by (4). Hence we have to find the minimum of a linear function on a simplex. The minimum is found by taking the distribution $\boldsymbol{d}^S$ which selects the DoF that should be changed on the last step equal to one of the vertices $v_i \in V$. We denote the corresponding optimal transition probabilities by Γ^S.

Now, consider the second to last step $S-1$. For any given input $\boldsymbol{p}^{S-2}$ we have to solve

$$\boldsymbol{F}^{\mathrm{tr}} \cdot \boldsymbol{p}^S = \left(\boldsymbol{F}^{\mathrm{tr}} \cdot \Gamma^S\right) \cdot \left(\sum_{\alpha,\beta\in\Omega} \Gamma^{S-1}_{\alpha\beta} p^{S-2}_\beta \right) \longrightarrow \min . \tag{14}$$

Defining $F^{(2)} = \boldsymbol{F}^{\mathrm{tr}} \cdot \Gamma^S$ as the new objective function, we can apply the same arguments for determining $\boldsymbol{d}^{S-1}$ and denote the resulting transition matrix by Γ^{S-1}. Hence also the optimal transition probabilities for Γ^{S-1} are found by taking $\boldsymbol{d}^{S-1}$ to be an element of V. We process all remaining steps in a similar manner, finding that optimality can be achieved at every step by choosing d^t to be one of the vertices in V.

The proof shows that a rectangular distribution over some of the "least fit" DoF gives the best implementation of EO. We name the resulting class of algorithms using rectangular distributions in connection with EO *fitness threshold accepting* (FTA), because, in analogy to TA, all moves triggered by selecting ranks which lie under a certain fitness threshold are selected with equal probability.

Uniqueness. – The proof above is based on the fundamental theorem of linear programming, which states that a linear function defined on a simplex assumes its minimum at a vertex. Our proof does not state that *all* optimal strategies are of the given form. Other strategies may do equally well, but not better.

If there exists an optimal strategy other than FTA, it follows that an edge or a face of the simplex does equally well. The optimality of such an edge corresponds to selecting the least r ranks with equal probability doing equally well as selecting the least $r-1$ ranks.

Much more unlikely is the possibility that a strictly monotonic distribution such as $d^t(k) \propto \tau^{-k}$ can possibly be optimal [6]. This would in fact imply that all the vertices in V do equally well. It should be clear that this can only happen for rather trivial problems.

Conclusions. – In this paper we considered the problem of finding the ground state of a complex system by using the heuristics known as extremal optimization. We used a master equation to describe the corresponding dynamics of random walkers on state space and formulated some straightforward assumptions on the probability distribution for selecting the DoF to change at the next step.

Our goal was to find transition probabilities which assure the optimum control of the random walkers' movements. We found that a special distribution of transition probabilities, which we named fitness threshold accepting, 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.

While we cannot show that fitness threshold accepting is the *only* optimal way to implement extremal optimization, our proof shows that a strictly monotonic distribution over ranks k, such as $d^t(k) \propto \tau^{-k}$ advocated by Boettcher *et al.* [7], can only match the optimal performance if *all* distributions perform equally well.

Knowledge that best performance can be achieved using fitness threshold accepting is only of limited use, since the thresholds to be used are not known *a priori*. Therefore, we performed preliminary numerical experiments indicating a better performance of FTA compared to the original implementation of EO. In particular, we have implemented FTA and have tried different cooling schedules which narrow the rectangular distribution over the least-fit DoF in every step. The indications are that FTA outperforms the DoF selection rule advocated by Boettcher and Percus. We are presently performing a careful and extensive numerical comparison which will be the topic of a future publication. Based on these results, our opinion is that the FTA algorithm should be used.

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. As an example, the *best-so-far* energy E_{bsf} as a measure is beyond the scope of the proof presented here. While it can also be shown that E_{bsf} is minimized (within an EO framework) by employing an FTA scheme, the requisite arguments for that proof are too lengthy for this letter and will soon be published elsewhere.

Further, our proof had to assume a finite state space. We postpone the exploration of continuous state spaces to a future effort, but point out that the realities of discrete arithmetic on digital computers make any state space effectively finite. Finally, we note that our discussion considered only algorithms based on extremal optimization. The possibility of better algorithms not based on extremal optimization remains. But within the given field, the arguments presented here establish the structure of a provably optimal strategy which furthers the study of heuristic approaches to global minimization.

REFERENCES

[1] Kirkpatrick S., Gelatt C. D. and Vecchi M. P., *Science*, **220** (1983) 671.
[2] Penna T. J. P., *Phys. Rev. E*, **51** (1995) R1.
[3] Tsallis C. and Stariolo D. A., *Physica A*, **233** (1996) 395.
[4] Dueck G. and Scheuer T., *J. Comput. Phys.*, **90** (1990) 161.
[5] Moscato P. and Fontanari J. F., *Phys. Lett. A*, **146** (1990) 204.
[6] Franz A., Hoffmann K. H. and Salamon P., *Phys. Rev. Lett.*, **86** (2001) 5219.
[7] Boettcher S. and Percus A. G., *Phys. Rev. Lett.*, **86** (2001) 5211.
[8] Boettcher S. and Percus A., *Artif. Intell.*, **119** (2000) 275.
[9] Boettcher S. and Percus A. G., *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '99)* (Morgan Kaufmann Publishers) 1999, p. 826.
[10] Boettcher S., Percus A. G. and Grigni M., *Proceedings of the Sixth International Conference on Parallel Problem Solving from Nature* (Springer-Verlag, Berlin, Heidelberg) 2000, p. 447.
[11] van Kampen N. G., *Stochastic Processes in Physics and Chemistry* (Elsevier, Amsterdam) 1997.
[12] Bellman R. E. and Dreyfus S. E., *Applied Dynamic Programming* (Princeton University Press, Princeton) 1962.