Can Downhill Walkers from the top of a Hill Crimb it Again?

Akira Imada

Dokuz Eylül University
Computer Engineering Department
Bornova, Izmir, 35100 Turkey
akira@cs.deu.edu.tr

Abstract

Using various versions of evolutionary computation, we explore fitness landscape defined on continuous synaptic weight space of a fully connected neural network model of associative memory. Thus far, we have found many different solutions that give a function of associative memory to a network. However, the Hebbian peak, one of the most familiar solutions, has never been located in these explorations. To address this issue, we study shape of peaks on the fitness landscape by reverse hill-climbing walk taking advantage of the evolutionary programming. We guess why does the Hebbian peak resist to be found.

1 INTRODUCTION

Although the fully connected neural network model of associative memory, such as the Hopfield's model, is nowadays said to be impractical, many interesting issues from physics point of view are still unknown. The number and distribution of attractors in synaptic weight space, or the fate of specific trajectories when some of the weight values are slightly modified are these examples. To attack these issues, we use evolutionary computations that explore the fitness landscape defined on weight space of a fully connected neural network model of associative memory (see, e.g., (Imada et al., 1998)).

Techniques to learn statistically a geometry of fitness landscapes have been fairy well studied (e.g., (Manderick et al., 1976)). However, It was discrete spaces that were analyzed by those techniques, and analyses of those defined on a continuous domain remain few. We have not succeeded in applying these techniques to our problem, probably due to our real-valued weights and the lack of isotropic nature of the landscape.

We have also employed reverse hill-climbing to learn shape of peaks in our fitness landscape (Imada, 1999). By walking down a peak from the top, we can obtain information of its profile of the peak. However, because of a plasticity of synaptic weight of neural networks, peaks are made up of many flat regions of small area, and downhill walkers skip some of the small flat regions when walking down a side wall, and hence it was just a rough estimate of a peak shape. In this paper, we propose somewhat of a new method of learning more precise profile of peaks utilizing the evolutionary programming (EP) capability of escaping the domain of attraction.

This study is a step to learn a whole geometry of the fitness landscape, and also a step to challenge a real world problem.

1.1 ASSOCIATIVE MEMORY

Associative memory is a dynamical system which has a number of stable states with a domain of attraction around them (Komlós and Paturi 1988). If the system starts at any state in the domain, it will converge to the stable state. Hopfield (1982) proposed a fully connected neural network model of associative memory in which information is stored by being distributed among neurons, and is retrieved from dynamically relaxed neurons' states. In the model, some of the appropriate configurations of synaptic weights give the network a function of associative memory.

The Hopfield model consists of N neurons and N^2 synapses. Each neuron state is either active $(+1)$ or quiescent (-1). When an arbitrary N-bit *bipolar pattern*, a sequence of $+1$ and -1, is given to the network as an initial state, the dynamical behavior of neurons' states afterwards are characterized by the strengths of the N^2 synapses. The synaptic strengths are called *weights*, and the weight from neuron j to neuron i is denoted as w_{ij} in this paper. Provided the synaptic weights are determined appropriately, network can store some number of patterns as attractors. Hopfield employed the so-called Hebbian rule (Hebb, 1949) to prescribe the weights. That is, to store p bipolar patterns ξ^ν:

$$\xi^\nu = (\xi_1^\nu, \cdots, \xi_N^\nu), \qquad \nu = 1, \cdots, p, \tag{1}$$

the weight values are determined as:

$$w_{ij} = \frac{1}{N} \sum_{\nu=1}^{p} \xi_i^\nu \xi_j^\nu \quad (i \neq j), \quad w_{ii} = 0. \tag{2}$$

An instantaneous state of a neuron is updated asynchronously (one neuron at a time) as:

$$s_i\,(t+1) = sgn \left(\sum_{j \neq i}^{N} w_{ij} s_j(t) \right),$$

where $s_i(t)$ is a state of the i-th neuron at time t. If an initial state converges to one of the stored patterns ξ^ν as an equilibrium state, then the pattern is said to be *recalled.* Furthermore, if an initial state chosen from the stored patterns remains unchanged from the beginning, then the pattern is said to be stored as a *fixed point.*

In analyzing the Hopfield model, there have been basically two different approaches: one is to explore *pattern space* searching for attractors under a specific weight configuration, and the other is to explore *weight space* searching for an appropriate weight configuration that stores a given set of patterns. To be more specific, the former is an analysis of the Hamiltonian energy as a function of all the possible configurations of bipolar pattern to be given to the network, where synaptic weights are pre-specified using a learning algorithm, usually the Hebb's rule, so that the network stores a set of p given patterns. In this context, the model for $p = 1$ corresponds to the Mattis model of spin-glass (Mattis, 1976), in which the Hamiltonian energy has two minima, while the model for infinitely large p corresponds to the Sherrington-Kirkpatrick model (Serrington and Kirkpatrick, 1975), in which the synaptic weights become Gaussian random variables. Analyses of the former type have been made in between these two extreme cases (Amit, 1989). The latter analysis, on the other hand, was addressed by Gardner (1988). She discussed the optimal weight configurations for a *fixed* number of given patterns in terms of the volume of the solutions in weight space, suggesting that the volume shrinks to vanish when p approaches to $2N$. In short, the former approach searches for the optimal pattern configurations which minimize the Hamiltonian energy *in pattern space* with the weights being fixed, while the latter searches for the weight configurations *in weight space* that optimally store a set of given fixed patterns.

In this paper, our concern is on weight space where some points give a network a capability to store a pre-specified one pattern ($p = 1$), and we call these points solutions.

1.2 FITNESS LANDSCAPE

The concept of the fitness landscape was first introduced by Wright (1932) to study biological evolutionary processes. Since then, this concept has been used not only in evolutionary biology but also in chemistry, physics, computer science and so on.

In chemistry, for example, a molecule can be represented as a string of N letters with each letter being chosen from an alphabet of size k (Macken et al., 1991). Twenty amino acids ($k = 20$) for proteins or four nucleotides ($k = 4$) for nucleic acids can be considered as examples of the alphabet. The k^N possible combinations of the letters construct a configuration space of the string. Then, for example, the free energy of RNA folding into secondary structures (Fontana et al., 1987), or the ability of peptides to bind to a particular substrate to catalyze a specific reaction (Maynard Smith, 1970) is assigned as a fitness value to each configuration. In physics, the Hamiltonian energy of Ising spins defines a fitness landscape on the configuration space of N spins, where each spin takes the value either 1 or -1 ($k = 2$). Bray and Moore (1980) argued about the number and distribution of meta-stable states (local optima) of the Hamiltonian energies.

To explore these fitness landscapes, we need a rule by which a point in the space moves to one of its neighbors. Then, consecutive movements of a point to the neighbors form a *walk* on the landscape. Macken et al. (1991) used *random point mutation* that changes a single letter in the string to specify neighbors of the string. Then, by sampling points along an *"evolutionary walk"* in which point moves to the *firstly found fitter* neighbor, they studied the statistical properties of the landscape defined by the chemical affinity of antibody for antigen in immune response. Weinberger (1990) used two different walks: *"gradient walk"* in which the walker steps to the *best* of its neighbors and *"random adaptive walk"* in which the next step is chosen *at random* from the set of better neighbors, to investigate the Kauffman's NK landscape which is a model formulated in more general form (Kauffman et al., 1987).

In this paper, we extend the concept of the discrete fitness landscape to a continuous one. Namely, a capability of a fully-connected neural network to store a set of bipolar patterns as associative memory assigns fitness on the configuration space of real-valued synaptic weights ($k = \infty$) of a network. A walker moves to its neighboring point determined by EP which uses Gaussian random mutation.

2 EXPERIMENT

Experiments were carried out on networks with 49 neurons, and we focus on $p = 1$ case.

Shapes of two peaks are compared. One is the Hebbian peak determined by the equation (2), and the other is the one found by an EP search where all individuals start the search with the origin of the weight space, that is, all zero weight configuration. Note that this is somewhat different than normal EPs where individuals in the first generation are randomized.

Then the two peaks are normalized on the surface of the hypersphere of radius one whose center is at the origin. The reason for this is that two weight matrices that differ only by scaling factor are equivalent in our problem. In other words, solutions are infinitely extended straight *ridges* emanating from the origin, instead of *peaks*.

In the following three subsections, ideas employed in the experiment are briefly described.

2.1 FITNESS EVALUATION

Assume that p random bipolar patterns represented by equation (1) are stored as an associative memory in the network. When one of these memorized patterns ξ^ν is given to the network as an initial state, states of neurons vary from time to time afterwards (unless ξ^ν is stored as a fixed point). In order for the network to function as an associative memory, the instantaneous network state must be similar to the input pattern ξ^ν. The similarity is defined by,

$$m^\nu(t) = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\nu s_i^\nu(t),$$

where $s_i^\nu(t)$ is the state of neuron i at time t. This $m^\nu(t)$ is referred to as *overlap*. The quality of retrieval of memorized patterns are represented by $\langle m^\mu \rangle$, a temporal average of $m^\nu(t)$ over certain time interval t_0. We evaluate the fitness value of each network by further averaging these $\langle m^\mu \rangle$'s over all p memorized patterns. Namely, the fitness f is

$$f = \frac{1}{p \cdot t_0}\sum_{\nu=1}^{p}\sum_{t=1}^{t_0} m^\nu(t). \qquad (3)$$

Then the fitness 1 implies all the p patterns are stored as fixed points, while all other cases have a fitness less than 1. In this paper, t_0 is set to $2N$, twice the number of neurons.

2.2 EP to SEARCH for PEAKS

To obtain a representative solution found by an evolutionary computation to be compared to the Hebbian weight, we employed EP which is a technique for optimization, proposed by L. Fogel, Owens and Walsh (1966). Although the population in EP is usually randomized at the beginning of a run, we start it with all identical points.

A population is constructed of μ individuals. Each of these μ points is mutated once, which produces μ mutants. Each point has additional variables σ_i for each coordinate, which are used as a standard deviation of a Gaussian random variable to be added to the i-th coordinate of the point. After the mutation, these σ_i are also modified as:

$$\sigma_i^{\text{new}} = \sigma_i^{\text{old}} + 0.01 \cdot \sigma_i^{\text{old}} \cdot N_i(0,1).$$

All the σ_i are initialized to σ_0 at the beginning of a walk. Note here that the value σ_0 plays an important role in this paper. Now, we have the original μ points and their μ mutants. The fitness value of each of the 2μ points are compared to those of q points which are chosen randomly at every time of the comparison from the whole 2μ points. Then the 2μ points are ranked according to the number of wins, and the best μ points survive (q-tournament selection). The cycle of reconstructing the new population with better fitness points and restarting the search is repeated until one of the global optima is found or a set maximum number of generation has been reached. In this paper, both λ and μ are set to 200 and q is chosen to be 10.

2.3 DOWNHILL WALK by EP

We call the EP search employed in this paper a *walk*, because (i) starting points are all initialized at an identical point, and (ii) EP does not use the so-called cross-over operation which is more like a *jump* than a *walk*.

In his Ph.D dissertation, Jones (1995) argued that *"a reverse hill-climbing algorithm allows the determination of details of the basin of attraction of points on a landscape."* The landscapes that Jones explored were defined on a discrete domain, that is, on vertices of a hyper cube. Here, we extended the idea to a landscape that is defined on a continuous domain.

In order for a walker to move around in *discrete* domain, a point mutation is employed. For example, a single letter constructed the point is changed to one of other letters if the domain is discrete. Since our landscape is defined on real-valued variables, on the other hand, the Gaussian mutation is employed. Though many versions of such a walk can be thought as candidates, we use EP here expecting its path to be almost the steepest one, and call the walk "*reverse EP hill-climbing*".

The reverse EP hill-climbing is essentially the same as the EP search described in the previous subsection, with the difference being the following two points: (i) the starting point is not the origin, but the top of the peak, or a point in a flat region where we want to learn its width as will be described later, and (ii) fitness function is replaced with $1 - f$ where f is defined by equation (3), so that the top of the peak corresponds to the lowest fitness value of zero.

3 RESULTS and DISCUSSION

A weight configuration that stores a pattern as associative memory is searched for by an EP. The fitness versus generation during the evolution is shown in Figure 1 (left). The EP search is started with all individuals in the first population being initialized at the origin, and we found that this is very efficient initialization compared to usual randomized one.[1]

Then this weight configuration, together with the Hebbian one, is normalized on the surface of hypersphere of radius one whose center is the origin, by which we can call these two configurations *peaks* instead of *ridges* (see previous section).

To see a downhill path from the top of the peak, EP reverse hill-climbing starts with all walkers starting at the top of the peak. The top is surrounded by a flat region of the same height which corresponds to the *synaptic plasticity* of neural networks. Hence, in order to start a downhill walk, the initial standard deviation of the EP search should be larger than a critical value. We conjecture that this critical value reflects the width of the flat region where walkers start. Starting with a small enough value of σ_0 with which walkers cannot start downhill, we increase the value little by little to find σ_0 that enables walkers to start downhill within 20 generations.[2] To obtain a rough image of a hill shape, we may trace fitness values along this reverse hill climbing with an appropriately larger value of σ_0 than its critical value. An example are shown in Figure 1 (right). The figure was obtained by plotting fitness values along the downhill walk from the top of the peak whose evolution was shown in Figure 1 (left). Horizontal parts of the path indicate that walkers needed multiple trials to reach the next step, which suggests that the side wall of the peak is also made up

[1] In his book, Kauffman (1993) pointed out that "*In short, the landscape is very smooth for small synaptic weights, and hence the optimum is easily sought.*" We doubt, however, this is because the origin is the cross point of all ridges of solutions, though the origin itself is a singular.

[2] The probability of escaping a region is never be zero due to the property of the Gaussian mutation. So, we wait for 20 steps here.

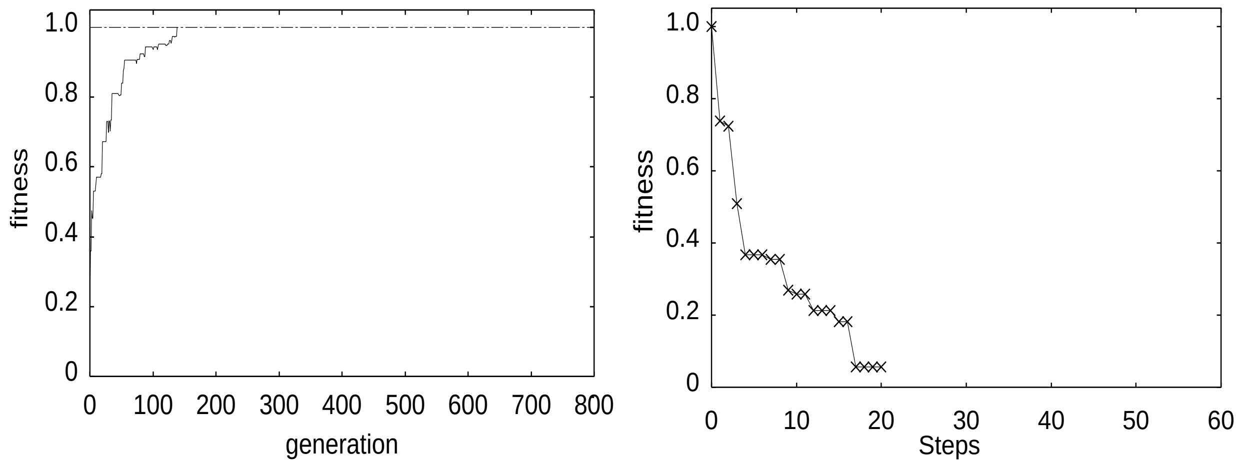


Figure 1: A peak found by EP (left), and a downhill path from the top of the peak (right).

of flat regions of some width. So, in this figure the flat regions are represented by *number of steps required to escape the region.*

In this method, however, walkers miss some of the flat regions because σ_0 is determined according to the width of the top area, not to the width of each flat region. Hence, in this paper, we propose a slightly different way of learning a downhill path as follows. At each step of the downhill walk, all the walkers start at the point where the best-fit-walker reached in the previous step. And σ_0 is also renewed according to the width of the current region. In this case, the width of each reagion is represented by *standard deviation* instead of the above mentioned number of steps.

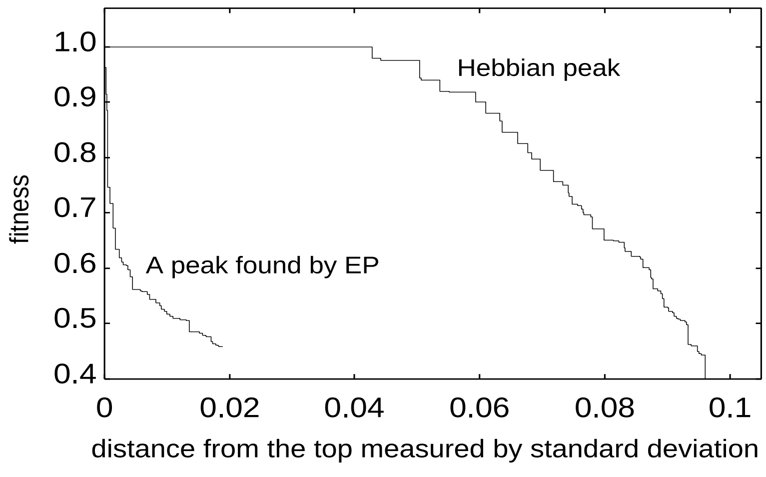


Figure 2: A more precise shape of peaks: the Hebbian peak and a peak found by EP.

Thus, after learning the width of the top, the best walker reach a point in the first flat region just beneath the top. The procedure then is repeated to learn the width of this flat region, that is, the new critical value for σ_0 to escape this region is sought. By repeating this procedure step by step, we can obtain a shape of the whole path from the top. In Figure 2, paths from the top of two peaks, the Hebbian peak and the one found by EP, are shown.

We found that some of the flat regions close to the Hebbian top are much wider than those of the peak found by EP. An example of these values of the first 5 steps along a downhill path are shown in Table 1, together with the value of the top area.

Table 1: The minimum σ_0 that enables walkers to escape each region.

	Peak found by EP	Hebbian Peak
Top	0.000049	0.042855
1st step	0.000186	0.001300
2nd step	0.000085	0.006180
3rd step	0.000065	0.000297
4th step	0.000045	0.002980
5th step	0.000023	0.001492

We also show the distribution of the width of 30 flat regions around the top, which are taken from five runs of the downhill walk. The result is shown in Figure 3 (b) and (c). For comparison purpose, the distribution of the width of 30 flat regions where randomly chosen 30 points are fallen is also shown in Figure 3 (a).

We can observe that the slope of the peak found by EP is constructed by flat regions of small width almost comparable as those in background level, while the Hebbian peak has wide variety of width including very large ones (especially near the top). From its comparatively large values beneath the top of the Hebbian peak, the width of the flat regions fluctuates, gradually becomes smaller, and eventually approaches a distribution similar to the back ground level (not shown in the figure).

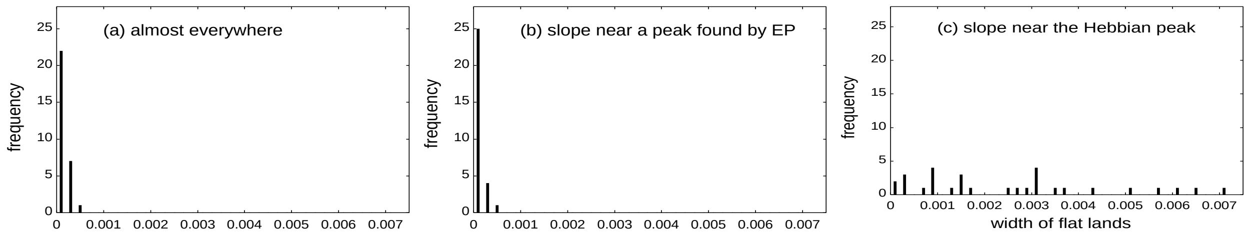


Figure 3: Width distribution of 30 points on flat-lands taken from: (a) background, (b) slope near hill-top found by EP, and (c) slope near the Hebbian hill-top.

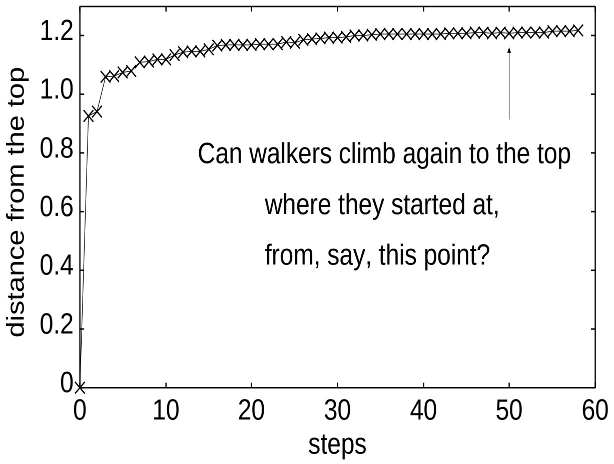
Two peaks are very different in shape. Figure 2 suggests that when walkers are to climb the Hebbian peak they must increase their step sizes as they approach the top area, while the other way around in climbing the peak found by EP. If we consider that adaptive mutation usually implies a decrease of step sizes as they approach the optimum, this shape of the Hebbian peak gives us one of the reasons why it is difficult to be climbed.

Walking down the hill, walkers seem to spread apart from the top where they start. In Figure 4 (left), the Euclidean distances from the Hebbian top during the downhill walk are shown. The data are taken from the downhill walk shown in Figure 2.

Then, an interesting question arises. Can walkers climb the slope again and reach the original top? To answer this question, a hill-climbing EP (not downhill) was tried with starting points being, say, at a point in the 50th flat region down from the original top. The result is shown in Figure 4 (right). The distance from the original top increases as walkers climb the hill again. This suggests that the walkers climb a different peak which is located another side of the original Hebbian peak. Although since the distances are Euclidean, not the distances on the hypersphere surface, and the original top is less provable that it was the center of the flat region of the top area, this result supports our guess more or less. That is, other more easily reachable global optima rather than the original one exists around this area of the 50th step.

4 CONCLUSIONS

We have described comparisons of two peaks in the fitness landscape defined on weight space of the fully connected neural network model of associative memory. And we have guessed why the Hebbian peak has not



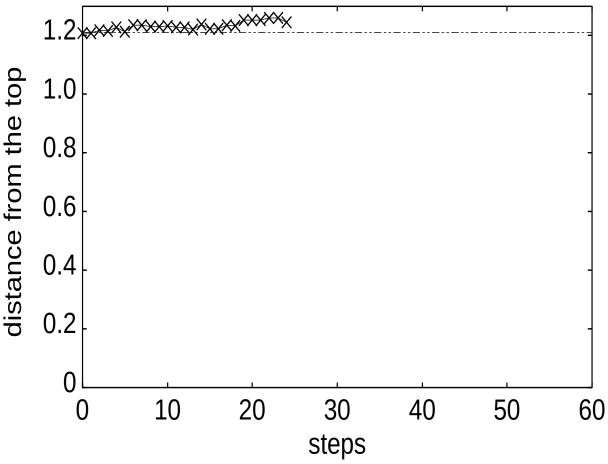


Figure 4: Distance between each of the downhill steps and the original top (left), and distance between each of the re-hillcliming steps and the original top (right).

been so far located by our evolutionary searches. Of course, the peak is just a needle among vast amount of other needles, and this is one of the main reason, but we guessed the other two important reasons. One is the shape of the Hebbian peak. The slope gradually becomes less steep as walkers approached the top area, while a peak which was easily found by EP has a slope the other way around. The other reason is existence of other more reachable peaks in the vicinity of the Hebbian peak. We found that when walkers try to climb again the hill from a point on the slope, they rather select other easier-to-climb-peak.

Many other issues await our explorations. It is possible that more tolerant unknown solutions exist in weight space. Will a run with specific σ_0 be able to locate such a peak whose top is much wider than the top of the already known peaks? Or, what is the distribution of these solutions? We think we will be able to shed new light on these issues using evolutionary computations.

Reference

Amit, D. J. (1989) *Modeling Brain Function: The World of Attractor Neural Networks.* Cambridge University Press.

Bray, A., and M. Moore (1980) *Metastable States in Spin Glasses.* Journal of Physics C: Solid State Physics 13, pp.L469–L476.

Fogel, L. J., A. J. Owens, and M. J. Walsh (1966) *Artificial Intelligence through Simulated Evolution.* Wiley.

Fontana, W., and P. Schuster (1987) *A Computer Model of Evolutionary Optimization.* Biophysical Chemistry 26, pp.123–147.

Gardner, E. (1988) *The Phase Space of Interactions in Neural Network Models.* Journal of Physics, 21A, pp.257-270.

Hopfield, J. J. (1982) *Neural Networks and Physical Systems with Emergent Collective Computational Abilities.* Proceedings of the National Academy of Sciences, USA 79, pp.2554-2558.

Hebb, D. O. (1949) *The Organization of Behavior.* Wiley.

Imada, A, and K. Araki (1997) *Evolution of Hopfield Model of Associative Memory by the Breeder Genetic Algorithm.* Proceedings of he 7th International Conference on Genetic Algorithms, Morgan Kaufman, pp.784–791.

Imada, A (1999). *Downhill Walk from the Top of a Hill by Evolutionary Programming.* Proceedings of the Congress on Evolutionary Computation, pp.1414–1418.

Jones, T. (1995) *Landscapes and Search.* Ph.D dissertation, The University of New Mexico.

Kauffman, S. A., and S. Levin (1987) *Towards a General Theory of Adaptive Walks on Rugged Landscapes.* Journal of Theoretical Biology 128, pp.11–45.

Kauffman, S. A. (1993) *The Origin of Order: Self-organization and Selection in Evolution.* Oxford University Press.

Komlós, J., and R. Paturi (1988) *Convergence Results in an Associative Memory Model.* Neural Networks 1, pp.239–250.

Macken, C., P. Hagan, and A. Perelson (1991) *Evolutionary Walks on Rugged Landscapes.* SIAM Journal of Applied Mathematics 51, pp.799–827.

Manderick, B., M. de Weger, and P. Spiessens (1993) *The GA and the Structure of the Fitness Landscape.* Proceedings of the 4th International Conference on Genetic Algorithm, pp.128–135.

Mattis, D.C. (1976) *Solvable Spin Systems with Random Interactions.* Physics Letters, 56A, pp.421–422.

Maynard Smith, J. (1970) *Natural Selection and the Concept of a Protein Space.* Nature 225. pp.563–564.

Sherrington, D., and S.Kirkpatrick (1975) *Solvable Model of a Spin Glass.* Physical Review Letters 35, pp.1792–1796.

Weinberger, E. D. (1990) *Correlated and Uncorrelated Fitness Landscapes and How to Tell the Difference.* Biological Cybernetics 63, pp.325–336.

Wright, S. (1932) *The Roles of Mutation, Inbreeding, Crossbreeding and Selection in Evolution.* Proceedings of the 6th International Congress of Genetics, pp.356–366.