CAN DOWNHILL WALKERS FROM THE TOP OF A HILL CLIMB IT AGAIN?

Akira Imada

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

Abstract

Thus far, using evolutionary computations, we have found many different solutions of the fully connected neural network model of associative memory. We now visualize them as peaks in fitness landscape defined on weight space. In this paper, we analyze our previous observation that the peaks located by an EP search are very different from the Hebbian peak.

1 Introduction

Although the fully connected neural network model of associative memory is somehow an old topics, 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 explore the fitness landscape defined on weight space of a fully connected neural network model of associative memory.

We employ 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 about shape 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 are trapped by or skip some of the small flat regions when walking down a side wall, and hence it was not possible to obtain a precise shape of a peak. Then, we proposed a new method of learning more precise profile of peaks utilizing the evolutionary programming (EP) capability of escaping the domain of attraction (Imada, 2001). In this paper, we study downhill walk paths obtained in that work more in detail. These studies are steps to learn a whole geometry of the fitness landscape.

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 behaviors of states of neurons 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.[1] 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:

$$s_i(t+1) = sgn\left(\sum_{j \neq i}^{N} w_{ij} s_j(t)\right), \tag{3}$$

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 (Sherrington 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

[1] The rule was at first advocated by Hebb (1949) and later formalized by Cooper (1973).

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.

We visualize these solutions on *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 our context, we have two possible fitness landscapes: the one defined on bipolar pattern space, and the other defined on synaptic weight space (see above). In both cases, top reagion is flat with the former top corresponding to *basin of attraction of a pattern* and the latter top to *plasticity of synaptic weights.* In this paper, we simply view fitness landscape as a surface determined by fitness value on all the possible weight configuration.

2 Methods

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

2.1 EP to Search for Peaks

To obtain a representative solution found by an evolutionary computation to be compared to the Hebbian weights, 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. Besides its N^2 coordinates, each point has additional N^2 variables σ_i to be added to the i-th coordinate of the point. Since these σ_i are chosen from the Gaussian distribution with an appropriate standard deviation, this is called Gaussiaon mutation. After the mutation, these σ_i are also modified as:

$$\sigma_i^{\text{new}} = \sigma_i^{\text{old}} + 0.01 \cdot \sigma_i^{\text{old}} \cdot N_i(0,1). \tag{4}$$

All the σ_i are initialized to σ_0 at the beginning of a run. 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.

2.2 Downhill Walk by EP

We call each individual's EP exploration 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. 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 above, 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 f (see below) is replaced with $1 - f$ so that the top of the peak corresponds to the lowest fitness value of zero.

2.3 Fitness Evaluation

Assume that p random bipolar patterns represented by the 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,

$$m^\nu(t) = \frac{1}{N}\sum_{i=1}^{N} \xi_i^\nu s_i^\nu(t), \tag{5}$$

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 $\langle m^\mu \rangle$ over all p memorized patterns. Namely, the fitness f is

$$f = \frac{1}{p \cdot t_0}\sum_{\nu=1}^{p}\sum_{t=1}^{t_0} m^\nu(t). \tag{6}$$

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.

3 Experiment

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

Here we have two peaks. One is the Hebbian peak determined by the Equation (2), and the other is the one found by an EP search in which all individuals start the search with the origin of the weight space,[2] that is, all zero weight configuration, and we found that this is very efficient initialization compared to usual randomized one.[3]

Then the two peaks are normalized on the surface of the hypersphere of radius one whose center is at the origin, because two weight matrices that differ only by scaling factor are equivalent in our problem. In other words, infinite number of equivalent solutions construct an infinitely extended straight *ridge* emanating from the origin, and the cross points of these ridges and the surface of the hypersphere construct *peaks* which represent solutions in a usual sense.

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 a minimal σ_0 that enables walkers to start downhill within 20 generations.[4]

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. The width of each region is represented by *standard deviation*.

Thus, after learning the width of the top, the best walker reaches 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.

4 Results and Discussion

In Figure 1, paths from each of the two tops are shown. Two peaks are very different in shape. Here we focus on how the difference affects the difficulty in searching for the location of these peaks.

The Hebbian peak is the one very difficult to be located by EP. The figure suggests that when walkers are to climb the Hebbian peak they must increase their step sizes as they approach the top area, while opposit when walkers climb the peak found by EP. Considering 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.

In order to check if the Hebbian peak is more challenging than the one found by EP search, walkers climb again the both slopes starting, say, with a point in the 10th and 40th flat regions beneath the top area. Although we observe that the Hebbian peak

[2]Note that this is somewhat different than the standard EP where individuals in the first generation are randomized.

[3]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.

[4]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.

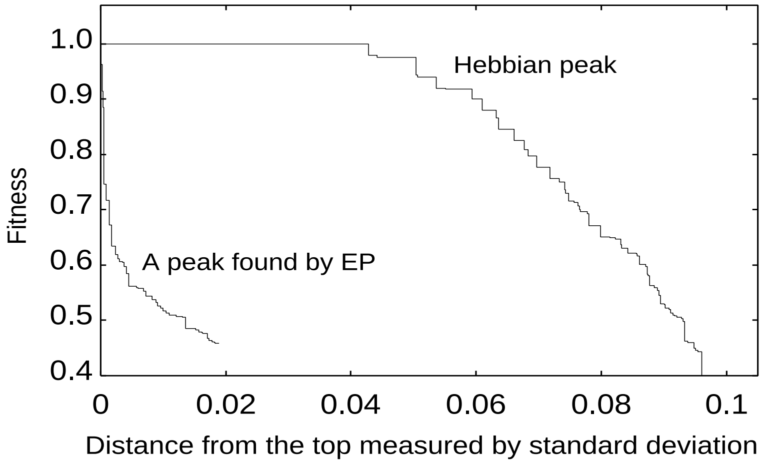

Figure 1: Downhill paths from the Hebbian peak and a peak found by EP.

gradually becomes more difficult to be attacked than the peak found by EP as starting point becomes farther from the top, the walkers can manage to climb it again (the data are not shown here). But is the point where the climber reaches really the top area of the same peak? To ascertain this, the Eucledian distance of the walker from the original top of the peak[5] as walkers approach the top are plotted in Figure 3.

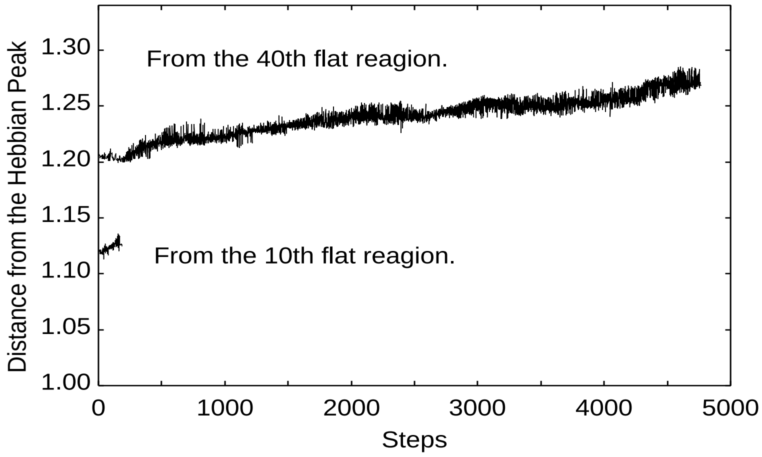

Figure 2: Eucledian distance between walker and the top as walkers approach it.

On the contrary to our intuition, the distance from the original top increases as walkers climb the hill. That is, walkers walk to a direction that is opposit to the original top. This holds even if they start just one step beneath the top (not shown here though). Hence we suspect the top region includes many halls, and the path that walkers walked down in Figure 1 was along a wall of one of these halls. Or, it is possible that the walkers climb

[5]By the original top we mean the point in weight space calculated by the Equation (2).

a different peak. Do other more easily reachable global optima exist around the starting point?

In order to learn what the top region of the peak look like, a single EP walker explores on the region starting from the original top until the walker found the point whose fitness value is less than one. This exploration is repeated 100 times, and each time when the walker reaches the point where fitness is less than one, the point is recorded as an edge point. In Figure 3, a series of fitness values in between a pair of two points chosen out of these 100 points.[6] No hole exists!

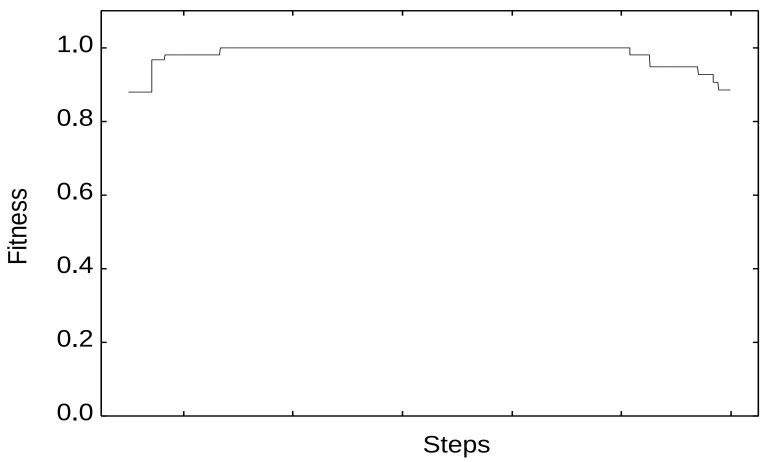

Figure 3: An exploration of a single EP walker on the Hebbian top region.

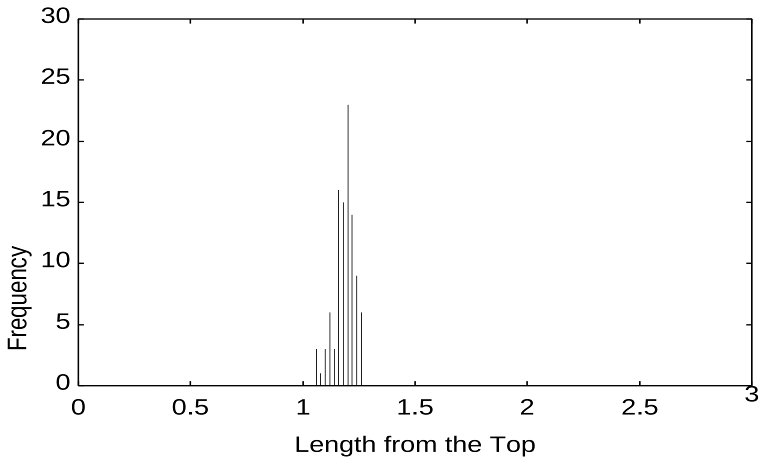

Figure 4: Distribution of traveling distances of the waker starting from the Hebbian peak to an edge of the top region.

[6]Although the two points whose distance is maximum are chosen here as an example, we observe a similar result with other set of two points including the minimum distance pair of points.

The distribution of the hundred distances is shown in Figure 4. The distances walker walks to the edges of the Hebbian top region range from 1.049653 to 1.259438. We can see that the points climbers eventually reach in Figure 3 are all within this range.

5 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. To learn why the Hebbian top is more difficult to be found than the other peaks found by EP, walkers climbed it again from a point on its slope. We observed somewhat peculiar behaviors of the walkers, but the reason is still under consideration with many other issues, which awaits our explorations.

Reference

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

Cooper, L. N. (1973) *A Possible Organization of Animal Memory and Learning.* Proc. of the 24th Nobel Symposium on Collective Properties of Physical Systems, pp.252–264.

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

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.* Proc. of the National Academy of Sciences, USA 79, pp.2554-2558.

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

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

Imada, A (2001). *EP Downhill Walkers Explore a Landscape in Weight Space of a Fully-connected Neural Network Model: Why does Hebbian Peak Resist to be Found?* Proc. of the International Conference on Neural Network and Artificial Intelligence, (to appear).

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

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.

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

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

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