

# A new on-line learning model

Shahar Mendelson  
Department of Mathematics  
Technion, Haifa 32000, Israel  
E-mail:shahar@tx.technion.ac.il

## Abstract

In this paper we introduce a new supervised learning model which is a non-homogeneous Markov process and investigate its properties. We are interested in conditions which ensure that the process converges to a “correct state”, which means that the system agrees with the teacher on every “question”. We prove a sufficient condition for almost sure convergence to a correct state and give several applications to the convergence theorem. In particular, we prove several convergence results for well known learning rules in neural networks.

**Key Words:** Supervised Learning, Neural Networks, Markov Processes, On line Gibbs learning

# 1 Introduction

In this paper we examine a model of abstract supervised learning. In a supervised learning model, the student is exposed to inputs and responds to them. If the response is incorrect (i.e., it does not agree with that of the teacher), the student “learns” and moves closer to the teacher in some sense. One of the first supervised learning models suggested was the Perceptron (see Minsky and Papert (1972), Ritter, Martinetz and Schulten (1992), Haykin (1994)). In this model, both the student  $S$  and the teacher  $T$  are halfspaces in  $\mathbb{R}^n$ . Given an input  $v \in \mathbb{R}^n$ , the student’s answer is incorrect if  $v$  is in  $S$  but not in  $T$ , or visa-versa. In this case, the student changes its position, and moves “closer” to the teacher by some learning rule. This learning model is on-line, i.e., at every “learning step” the system’s response depends only on its current state and on the input, and not, for example, on other parameters of the state space, which is a space containing all the possible students.

In a general supervised on-line learning model, the learning process is determined by an on-line error function denoted by  $\mathcal{E}(x, v)$ , where  $x$  is a student and  $v$  is an input. In the general case the teacher is not known, but makes its mark on the learning process via the on-line error function: The student  $x$  is asked a question (represented by the input  $v$ ), and  $\mathcal{E}(x, v)$  measures the magnitude of the mistake  $x$  made with regard to the input (question)  $v$ . In other words,  $\mathcal{E}(x, v)$  is a nonnegative bounded function such that for every state  $x$  and every input  $v$ ,  $\mathcal{E}(x, v) = 0$  if and only if  $x$  gives a correct response to the input  $v$ . Thus,  $\mathcal{E}(x, v) = 0$  if and only if  $x$  agrees with the teacher on the input  $v$ . The larger  $\mathcal{E}(x, v)$  is, the bigger the disagreement between  $x$  and the teacher on the input  $v$ .

An example of such a model is the on-line Gibbs learning, first suggested in Kim and Sompolinsky, (1996) and (1998). It is a Markov process in which the transition density from state to state – given the input at the  $n$ -th stage is determined by an “on-line” energy function. The  $n$ -th stage transition operator is defined so that on average, the transition from the  $n$ -th stage to the  $n+1$  stage reduces the energy. The energy function depends on the on-line error function  $\mathcal{E}(x, v)$  so that a reduction in the energy is equivalent to a smaller global error, which for every state  $x$  is the average of the error function on all possible inputs. Formally, the global error at  $x$  is  $\mathbb{E}_g(x) = \int_V \mathcal{E}(x, v) d\nu(v)$ , where  $V$  is the input set and  $\nu$  is the probability measure on  $V$  by which the inputs are selected.

In most supervised learning models suggested so far, the transition density from  $x$  to  $x'$  depended on both  $\mathbb{E}_g(x)$  and  $\mathbb{E}_g(x')$  – and not on the responses of  $x$  and  $x'$  to each input separately. Hence, those models do not enter into the category of on-line learning. The strength of the non on-line learning models is that at each learning step the global error decreases, therefore, it is easy to guarantee that the student converges to a minimum of  $\mathbb{E}_g(x)$ . In Kim and Sompolinsky, (1998) the authors claimed that the OLGA had some of the features of non on-line models, namely, that the OLGA converges in some sense to a minimum of  $\mathbb{E}_g$ .

The on-line model we suggest is a variation of the on-line Gibbs learning.

We examine it in two cases. First, when the learning step (i.e., the maximal distance between the  $n$ -th state and the  $n+1$  state) is a constant, and in the second case, the learning step decreases to 0. The main focus of this paper is on the first case, but we give an example of a convergence theorem for the second case under some additional assumptions on the on-line error function.

This paper is divided to five parts. In the first one, we define our model and state most of the convergence results concerning the process with a constant learning step. We also present several examples in which the process may be used. Two of those examples are the well known Perceptron and the multi layer Perceptron models (Ritter, Martinetz and Schulten (1992), Haykin (1994)). In the second part, we give an example of a convergence theorem in the case where the learning steps decrease to 0. In this example we take into account the possibility that at every stage the teacher has a probability  $p < 1/2$  to make a mistake. In the third part we prove the results stated in the first section and the fourth contains a comparison of our model and the OLG. In addition, we present a counterexample to one of the claims in Kim and Sompolinsky (1998). We end the paper with some concluding remarks.

Let us turn to some definitions and notations: throughout the paper  $(X, d, \mu)$  is a compact metric probability space with a metric  $d$  and a probability measure  $\mu$ . We assume that the measure  $\mu$  is compatible with the metric in the sense that every open set in  $X$  has a positive  $\mu$ -measure. If  $A \subset X$ , set  $d(x, A) = \inf_{a \in A} d(x, a)$ . We say that a sequence  $(x_n)$  converges to a set  $A$  if  $\lim_{n \rightarrow \infty} d(x_n, A) = 0$ . Let  $B_\lambda(x)$  be the closed ball of radius  $\lambda$  centered at  $x$ .  $X^n$  is the product space of  $n$  copies of  $X$  with the induced topology and measure.  $(V, \nu)$  is the compact metric space of all possible inputs, where  $\nu$  is a probability measure on  $V$ . Denote by  $\tau$  the product measure  $(\mu \times \nu)$  on  $X \times V$ .

For a random variable  $Y$ ,  $\mathbb{E}Y$  is the expectation of  $Y$  and  $\|Y\|_1 = \mathbb{E}|Y|$ . Given a set  $A$ ,  $\bar{A}$  denotes its closure and  $\chi_A$  is its characteristic function, i.e., the function which is 1 on  $A$  and vanishes elsewhere. Finally, we say that  $(a_n) \in \ell_p$ , if  $\sum |a_n|^p < \infty$ .

## 2 The model and some examples

In this section we define our model and list some results concerning it. Then, we give examples for ways in which this learning process may be used. We separate our discussion to two cases: one is when the error function is a 0-1 function and the other is when  $\mathcal{E}(x, v)$  is a nonnegative smooth function. For every  $x \in X$ , put  $C_x = \{v | \mathcal{E}(x, v) = 0\}$ . Thus,  $C_x$  is the set of inputs on which the student  $x$  agrees with the teacher.

Our process is a Markov process  $(X_n, V_n)$ , where  $(V_n)$  are i.i.d. which are distributed according to  $\nu$  and are independent of  $(X_n)$ , while  $X_n$  represents the state of the student at the  $n$ -th stage. We assume that  $X_1$  is distributed according to  $\mu$ , and that  $X_n$  is adapted as follows: let  $\lambda$  be the size of the maximal learning step and set  $T_n$  to be a sequence decreasing to 0. If  $d(x, x') > \lambda$  then  $x$  can not move to  $x'$ . Moreover, if  $x$  gives the correct response to  $v$ , then  $x$

remains stationary. If  $x$  does not give the correct response to  $v$  and if  $d(x, x') \leq \lambda$  then the transition density at the  $n$ -th stage from  $x$  to  $x'$  given  $v$  is

$$P_n(X_{n+1} = x' | X_n = x, V_n = v) = \frac{1}{c_n(x)} \mathcal{E}(x, v) e^{-\frac{\mathcal{E}(x', v)}{T_n}}, \quad (2.1)$$

where  $c_n(x)$  is the normalizing constant, which is determined by the fact that the probabilities that  $x$  remains stationary and that  $x$  moves to some  $x'$  should add up to 1. Thus,

$$c_n(x) = \nu(C_x) + \int_{B_\lambda(x)} \int_V \mathcal{E}(x, v) e^{-\mathcal{E}(x', v)/T_n} d\nu(v) d\mu(x').$$

The reason for this selection of a transition operator is simple. First, if  $x$  gives the correct response, there is no reason to change its location. In cases where  $x$  is wrong, the linear term  $\mathcal{E}(x, v)$  guarantees that if  $\mathcal{E}(x, v)$  is small, i.e., if the response of  $x$  is almost correct,  $x$  does not move by much. The exponent  $e^{-\mathcal{E}(x', v)/T_n}$  ensures that for small values of  $T_n$ ,  $x$  moves only to  $x'$  for which  $\mathcal{E}(x', v)$  is small too.

From here on we denote the conditional transition density  $P_n(X_{n+1} = x' | X_n = x, V_n = v)$  by  $P_n(x' | x, v)$  and set  $\mathcal{P}$  to be the measure induced on the orbits  $(X_n)$  of the process 2.1.

The 0-temperature process is the process whose transition operators are given by

$$P(x' | x, v) = \lim_{n \rightarrow \infty} P_n(x' | x, v),$$

where the limit is taken pointwise. Note that from the computational point of view, it is easier to run the process while decreasing  $T$  to 0 than to run the 0-temperature process. On the other hand, the analysis of the model is much easier at  $T = 0$ . We show that the “nice” properties of the process 2.1 at  $T = 0$  are obtained in the case where  $T_n$  is decreased to 0, and the schedule by which  $T_n$  is taken to 0 plays no essential role. A key part in the analysis of process 2.1 is the fact that for every  $A \subset X$ , the convergence of  $P_n(A | x) = \int_A \int_V P_n(x' | x, v) d\nu(v) d\mu(x')$  to  $P(A | x)$  is uniform. Therefore, we can approximate the behavior of this process by the 0-temperature process.

**Assumption 2.1** *Assume that  $\nu(C_x \Delta C_{x'})$  is continuous with respect to each variable separately, where  $A \Delta B = (A \cap B^c) \cup (B \cap A^c)$ . Assume further that there is a  $\delta > 0$  such that  $\nu(C_x) \geq \delta$  for every  $x$ .*

Though this last assumption seems innocent, it is rather strong and limiting. It implies that in some sense all the states in  $X$  are not too far away from the teacher. In all the examples we present, the only way to ensure that this assumption holds, is if we assume that the state space is a small enough neighborhood of the target concept. On the other hand, it is easy to construct many on-line error functions for which this assumption holds, in which the state space is not restricted to a neighborhood of the target concept.

**Definition 2.1** Let  $\mathcal{O} \subset X$  denote the set of local minima of the error function  $\mathbb{E}_g(x)$ . We say that  $A \subset X$  is a neighborhood of  $\mathcal{O}$ , if it is an open subset of  $X$  which contains  $\mathcal{O}$ . We denote by  $Q$  the set of states which agree with the teacher almost surely, i.e.,  $Q = \{x | \mathbb{E}_g(x) = 0\}$ .

Let us state our main results:

**Theorem 2.2** Assume that the error function is a 0-1 function. Then:

- a. If  $A$  is any neighborhood of  $\mathcal{O}$ , then for every  $\lambda > 0$  and every positive sequence  $(T_n) \rightarrow 0$ , the orbits  $(X_n)$  defined by the process 2.1 enter  $A$  infinitely often  $\mathcal{P}$ -almost surely.
- b. Assume that  $\mu(Q) > 0$ . Then, for  $\lambda = \text{diam}(X)$  and for every sequence  $(T_n) \rightarrow 0$ ,  $(X_n)$  converges  $\mathcal{P}$ -almost surely to  $Q$ .

Note that  $Q$  is an absorbing set, which means that the probability of leaving  $Q$  is 0.

In the case where the error function is not a 0-1 function we have to make an additional assumption:

**Assumption 2.2** Assume that  $\mathbb{E}_g$  is monotone in the sense that  $\mathbb{E}_g(y) \leq \mathbb{E}_g(x)$  when  $C_y \supset C_x$ . Assume further that for every  $x \notin \mathcal{O}$  and for every  $\varepsilon > 0$  there is some  $y \neq x$ ,  $y \in B_\varepsilon(x)$  such that  $C_y \supset C_x$ .

**Theorem 2.3** If the on-line error function satisfies Assumption 2.2 then the assertion of Theorem 2.2.a hold. The assertion of Theorem 2.2.b is true in the general case even without Assumption 2.2.

Note that the 0-1 case is much more easy to handle – for two reasons. First, there is an easy connection between  $\mathbb{E}_g(x)$  and  $\nu(C_x)$ , since  $\mathbb{E}_g(x) = 1 - \nu(C_x)$ . Therefore, as  $\nu(C_x)$  is increased, one moves closer to a minimal point of  $\mathbb{E}_g(x)$ . Second, it is easy to see that if  $d(x, y) \leq \lambda$  then there is a positive transition density from  $x$  to  $y$  if and only if  $\nu(C_y \setminus C_x) > 0$ . Because the error function is 0-1 and assuming that  $d(x, y) \leq \lambda$ , it suffices that  $\mathbb{E}_g(x) < \mathbb{E}_g(y)$  to ensure positive transition density from  $x$  to  $y$ .

On the other hand, for a general error function the situation is rather complicated. One does not know if there is any connection between  $C_y \setminus C_x$  and  $\mathbb{E}_g(y) - \mathbb{E}_g(x)$ . This fact is the reason for Assumption 2.2. In the final section we present an example (see Example 5.2), in which the error function is a smooth nonnegative bounded function, but the orbits can not leave the global maximum of  $\mathbb{E}_g(x)$ . In this example, even though for every  $x \in X$ ,  $\mathbb{E}_g(x) \leq \mathbb{E}_g(x_0)$ , it so happens that  $C_x \setminus C_{x_0}$  is empty.

Next, we present three examples in which model 2.1 may be used. In all three cases, the error function is a 0-1 function.

## 2.1 Example

The simplest case in which one can apply process 2.1 is the Perceptron learning rule (see Haykin (1994)). In this process we adapt a halfspace in  $\mathbb{R}^d$  using a teacher which is also a halfspace. Denote the teacher by  $T = \{y \in \mathbb{R}^d | x_0^*(y) > 0\}$  where  $x_0^*$  is a linear functional on  $\mathbb{R}^d$ , and set the student  $S = \{y \in \mathbb{R}^d | x_1^*(y) > 0\}$ . Assume that  $\|x_0^*\| = \|x_1^*\| = 1$ . In this case, the error function is 0 for inputs which are in  $T \cap S$  or in  $T^c \cap S^c$  and 1 otherwise. Assume further that the input set  $V$  is a finite union of balls, disjoint from the boundary of  $T$ , and that the probability measure  $\nu$  is given by a continuous density function supported on  $V$ . Recall that the Haar measure on the  $d$ -dimensional unit sphere  $S^{d-1}$  is simply the uniform distribution on that set. The space  $X$  will be a closed neighborhood of  $x_0^*$  in  $S^{d-1}$  such that for every  $x^* \in X$ ,  $\nu(C_{x^*}) > 0$ , endowed with the normalized Haar measure. Since the function  $x \rightarrow \nu(C_x)$  is continuous, it attains a positive minimum on  $X$ . Thus, there is some  $\delta > 0$  such that for every  $x^* \in X$ ,  $\nu(C_{x^*}) \geq \delta$ .

Clearly,  $\nu(C_x \Delta C_y)$  is continuous and the conditions of Theorem 2.2.b hold. Therefore, the orbits  $(X_n)$  defined by process 2.1 converge almost surely to the set  $\{x | \mathbb{E}_g(x) = 0\}$ .

Again, this example reveals the limitations of process 2.1: it may be applied only in the “local” setup, while states for which  $\nu(C_x) = 0$  are excluded from the state space.

## 2.2 Example

Here, we present two examples which deal with the uniform approximation of a continuous function  $f : V \rightarrow \mathbb{R}$  by a function selected from a family of continuous functions  $\{g_x | V \rightarrow \mathbb{R} | x \in X\}$ . Let  $X$  and  $V$  be compact subsets of  $\mathbb{R}^m$  and  $\mathbb{R}^n$ , while  $\mu$  and  $\nu$  are the normalized Lebesgue measures on  $X$  and  $V$ . Fix  $\eta > 0$ , and let

$$\mathcal{E}(x, v) = \begin{cases} 0 & \text{if } |f(v) - g_x(v)|, \\ 1 & \text{otherwise.} \end{cases}$$

Note that  $\mathbb{E}_g(x) = 0$  if and only if  $\|f - g_x\|_\infty < \eta$ . In the examples to follow there is a tradeoff between the accuracy parameter  $\eta$  and the selection of the state space  $X$ . In Example 2.2.a,  $X$  is a set of parameters which represent Multi-layer Perceptrons with a single hidden unit, while in Example 2.2.b each  $x \in X$  represents a polynomial. Both these classes are dense in the space of continuous functions  $C(V)$  with respect to the supremum norm. The fact that the set of polynomials is dense in  $C(V)$  is due to Weierstrass (Cheney, (1966)). As for MLPs with a single hidden layer, Leshno, Lin, Pinkus and Schocken (1993), proved the following:

**Theorem 2.4** *Let  $\sigma$  be a continuous function which is not an algebraic polynomial. Then,  $\text{span}\{\sigma(\langle v, w \rangle + \theta) | w \in \mathbb{R}^d, \theta \in \mathbb{R}\}$  is dense in  $C(V)$  for every compact set  $V \subset \mathbb{R}^d$ .*

Thus, given a function  $f \in C(V)$  which one wishes to approximate, the accuracy parameter determines the set of parameters in which one should seek the approximating function. As  $\eta$  decreases, the set of parameters needed to ensure that there is some state  $x$  for which  $\mathbb{E}_g(x) = 0$  increases in size. Conversely, if the parameter space  $X$  is given, there is some  $\eta$  which is the best possible accuracy in which one can approximate  $f$  using the functions  $\{g_x|x \in X\}$ . Hence, for smaller values of  $\eta$  there will be no  $x$  for which  $\mathbb{E}_g(x) = 0$ .

In the two examples presented here we assume that there is some  $g_x$  which approximates  $f$  up to an accuracy of  $\eta$ . Thus, there is some  $x \in X$  for which  $\mathbb{E}_g(x) = 0$ .

Clearly, if one wishes to use Theorem 2.2, we have to find some kind of continuity condition on the family  $\{g_x\}$ , to ensure that Assumption 2.2 holds. This is the object of the following Lemma:

**Lemma 2.5** *Assume that for every fixed  $s \in X$ ,  $\lim_{r \rightarrow s} \|g_r - g_s\|_1 = 0$ . Then  $\nu(C_x \Delta C_y)$  is continuous with respect to each variable separately.*

**Proof:** First, we show that for every  $s \in X$ ,  $\nu(C_s \setminus C_r)$  is continuous with respect to  $r$ . Since  $\nu$  is regular, there exist a compact set  $K \subset C_s$  such that  $\nu(C_s \setminus K) < \varepsilon/2$ . Therefore, there is some  $\delta > 0$  such that  $\sup_{v \in K} |g_s(v) - f(v)| \leq \eta - \delta$ . By Chebyshev's inequality,

$$\nu\{|g_s(v) - g_r(v)| \geq \frac{\delta}{2}\} \leq \frac{2 \|g_s - g_r\|_1}{\delta}.$$

Hence, if  $s$  and  $r$  are close enough then  $\nu(K \cap C_r^c) \leq \varepsilon/2$ . Therefore,

$$\nu(C_s \setminus C_r) = \nu(K \cap C_r^c) + \nu((C_s \setminus K) \cap C_r^c) \leq \varepsilon.$$

A similar argument shows that  $\lim_{r \rightarrow s} \nu(C_r \setminus C_s) = 0$ . ■

We are ready to present the two examples:

a) Multi Layer Perceptron (MLP)

The MLP is composed of several layers of Perceptrons and an output function  $\sigma$ . It is formed by using the output of each layer of Perceptrons as an input to the next one. Usually, the output function  $\sigma$  of each Perceptron in the MLP is assumed to be smooth, but we assume that  $\sigma$  is only a Lipschitz function and that the network has a single hidden layer. Let  $X', Y', V$  be compact sets in  $\mathbb{R}^{k \times d}$ ,  $\mathbb{R}^k$ ,  $\mathbb{R}^d$  respectively, and set  $Z' \subset \mathbb{R}^k$ . Assume that both  $X = X' \times Y' \times Z'$  and  $V$  are equipped with the normalized Lebesgue measures  $\mu$  and  $\nu$ . The MLP may be viewed as a function  $M : X' \times Y' \times Z' \times V \rightarrow \mathbb{R}$ : for every  $v \in \mathbb{R}^d$  and  $(x, y, z) \in X$  the response of the MLP  $(x, y, z)$  to the input  $v$

is  $M_{x,y,z}(v) = \sum_{j=1}^k y_j \sigma(\sum_{i=1}^d x_{ij} v_i + z_j)$ . If we view the MLP as a neural network,

$d$  is the number of cells in the input layer,  $k$  is the number of units in the second (hidden) layer,  $(x_{ij})_{i=1}^k$  represents the synaptic weights between the first layer

and the  $j$ -th cell in the second layer and  $z_j$  is the threshold of that cell. Hence,  $\sigma(\sum_{i=1}^d x_{ij}v_i + z_j)$  is the response of the  $j$ -th cell in the second layer.  $y_j$  is the synaptic weight between the  $j$ -th cell in the second layer and the output cell. Therefore, the response of the output cell is  $M_{x,y,z}(v)$ .

We assume that for the given accuracy parameter  $\eta$  there is some  $x_0, y_0, z_0$  in the interior of  $X$  such that  $\|M_{x_0,y_0,z_0} - f\|_\infty < \eta$ . Thus, the set of correct states  $Q$  has  $\mu$ -positive measure. Indeed, note that the set of states  $\{x, y, z\}$  for which  $M_{x,y,z}$   $\eta$ -approximates  $f$  is open: if  $M_{x,y,z} \in Q$ , then for some  $\varepsilon > 0$  the function  $M_{x,y,z}$   $\eta - \varepsilon$  approximates  $f$ . Since  $\sigma$  is Lipschitz then it is easy to see that

$$\sup_{v \in V} |M_{x_1,y_1,z_1}(v) - M_{x,y,z}(v)| \leq C(\|x - x_1\| + \|y - y_1\| + \|z - z_1\|),$$

where  $C$  is some constant which depends only on the Lipschitz constant of  $\sigma$ . Thus, a small perturbation in  $\{x, y, z\}$  also gives an  $\eta$ -approximation of  $f$ . By the same argument, if  $\{x_n, y_n, z_n\}$  converges to  $\{x, y, z\}$  then  $M_{x_n,y_n,z_n}$  converges to  $M_{x,y,z}$  uniformly, which implies convergence in the  $L_1$  norm. Hence, the conditions of Lemma 2.5 hold and  $\nu(C_x \Delta C_y)$  is continuous.

In this example too, the process has a local feature—since we require that  $X$  does not contain elements for which  $\nu(C_x) = 0$ . Here, one way to ensure that this is the case is to impose that the set  $\{g_x\}$  is contained in small  $L_1$  neighborhood of the target concept  $M_{x_0,y_0,z_0}$ . This does not make the approximation procedure trivial since a small neighborhood with respect to the  $L_1$  norm may be a large set with respect to the supremum norm. Thus, for most of the elements in such a neighborhood  $\mathbb{E}_g$  is large.

To summarize, the assumptions of Theorem 2.2.b hold, and the orbits  $(X_n)$  of the process 2.1 converge  $\mathcal{P}$ -almost surely to a function which  $\eta$ -approximates  $f$ .

b) Polynomial approximation of a Lipschitz function in  $[-1,1]$ .

This example is similar to the one presented above, so most of the details are omitted. There is a one-to-one correspondence between the set of polynomials of degree  $\leq d$  and  $\mathbb{R}^{d+1}$ . Note that it is possible to  $\eta$ -approximate every continuous function by a polynomial. However, the process requires to pre-determine the degree of polynomials we use, as well as the compact set from which the coefficients are selected. Assume that we have some additional information on the function  $f$  we wish to approximate – for example, its Lipschitz constant  $\lambda$ . In this case, the tradeoff between the accuracy parameter  $\eta$  and the set of parameters  $X$  may be estimated. For a bound on the degree of the approximating polynomial, we estimate  $E_d(f) = \inf_{a_0, \dots, a_{d-1}} \sup_{v \in [-1,1]} \left| \sum_{i=0}^{d-1} a_i v^i - f(v) \right|$ . By

Jackson's theorem (Cheney (1966)), if  $E_d(f) = \eta$  then  $d \leq C \frac{\lambda}{\eta}$ , where  $C$  is some absolute constant. Next, to estimate the size of the set from which the coefficient  $(a_i)$  are selected, we use Bernstein's inequality, which states that  $\|p'_d\|_\infty \leq d \|p_d\|_\infty$  where  $p_d$  is a polynomial of degree  $d$ . Note that  $|a_0| \leq \|p_d\|_\infty$ ,

$|a_1| = \left| p'_d(0) \right| \leq \left\| p'_d \right\|_\infty$ , and so on. Therefore, for every  $\eta$ , there are an integer  $d$  and a vector  $(a_1, \dots, a_d)$  such that  $\sup_{v \in [-1, 1]} \left| \sum_1^d a_i v^i - f(v) \right| < \eta$ , where both  $d$  and  $\|(a_1, \dots, a_d)\|_\infty$  depend only on  $\lambda$  and  $\eta$ . Adding the “locality” assumption, a similar argument to the one used in example (a) shows that the conditions of Theorem 2.2.b hold. Thus, the orbits of the process 2.1 converge to a correct state almost surely, which is a polynomial that  $\eta$ -approximates the target concept.

### 3 An example of a 0 temperature process

We investigate one example of a 0-temperature process in which the maximal learning step  $\lambda_n$  decreases to 0. We estimate the schedule by which  $(\lambda_n)$  should be taken to 0 to ensure that the orbits of this process converge to a correct state even if at every step there is a probability  $p < 1/2$  for the teacher to make a mistake.

Much like the results in Kim and Sompolinsky (1998), we deal with a situation in which the state space is a neighborhood of a local minimum of  $\mathbb{E}_g$ . However, we are able to prove convergence results which are stronger than those demonstrated for the OLGA. Indeed, in Kim and Sompolinsky (1998) it was shown that the average error  $\mathbb{E}(\mathbb{E}_g(X_n))$  tends to  $\mathbb{E}_g(x_{\min})$ , while in this section we show that  $X_n$  converges in probability to  $x_{\min}$ .

We begin by formulating the general assumptions regarding our model. The basic assumption is that if  $d(x, y) > \lambda_n$  then the probability that  $x$  moves to  $y$  during the  $n$ -th step is 0. Since we assume that the teacher can make a mistake with probability  $p$ ,  $x$  remains stationary in two cases: if  $\mathcal{E}(x, v) = 0$  and the teacher is correct, or, if  $\mathcal{E}(x, v) = 1$  and the teacher is wrong. Thus, up to a normalizing constant, the probability that  $x$  does not move during the  $n$ -th step is  $(1 - p)\nu(C_x) + p(1 - \nu(C_x))$ . We assume that if the teacher makes an error, then it provides the wrong answer both for  $\mathcal{E}(x, v)$  and for  $\mathcal{E}(y, v)$ . Therefore, for every  $y \neq x$ ,  $y \in B_{\lambda_n}(x)$  the conditional transition density from  $x$  to  $y$  is (up to a normalizing constant)

$$p(1 - \mathcal{E}(x, v))e^{-\frac{1 - \mathcal{E}(y, v)}{T_m}} \chi_{\{\mathcal{E}(x, v)=0\}} + (1 - p)\mathcal{E}(x, v)e^{-\frac{\mathcal{E}(y, v)}{T_m}} \chi_{\{\mathcal{E}(x, v)=1\}} = (*).$$

Note that the first term arises from the case in which the teacher makes simultaneous mistakes, while the second term is the case in which the teacher is right. The normalizing constant is given by

$$d_{m, n} = (1 - p)\nu(C_x) + p(1 - \nu(C_y)) + \int_{B_{\lambda_n}(x)} \int_V (*) d\nu(v) d\mu(y).$$

The 0-temperature process is defined by taking  $T_m \rightarrow 0$ . Therefore, The probability that  $x$  remains stationary is (up to the normalizing constant)  $(1 - p)\nu(C_x) +$

$p(1 - \nu(C_x))$ , while the conditional transition density during the  $n$ -th step from  $x$  to  $y \neq x$ ,  $y \in B_{\lambda_n}(x)$  is (up to the normalizing constant)

$$p\chi_{\{\mathcal{E}(x,v)=0\} \cap \{\mathcal{E}(y,v)=1\}} + (1-p)\chi_{\{\mathcal{E}(x,v)=1\} \cap \{\mathcal{E}(y,v)=0\}}.$$

Let  $X \subset \mathbb{R}^d$  be a compact set and put  $V = [0, 1]$ , both equipped with the normalized Lebesgue measure  $\mu$  and  $\nu$ . Assume that for every  $x \in X$ ,  $C_x = [0, f(x)]$ , where  $f \in C^2(X)$  and  $0 < f(x) \leq 1$ . Thus, the error function  $\mathcal{E}(x, v)$  is:

$$\mathcal{E}(x, v) = \begin{cases} 1 & \text{if } v > f(x) \\ 0 & \text{if } v \leq f(x) \end{cases}. \quad (3.1)$$

This assumption on the structure of the error function will allow us to estimate expressions of the form  $\nu(C_x \setminus C_y)$  easily.

Here, the conditional transition density from  $x$  to  $y \neq x$ ,  $y \in B_{\lambda_n}(x)$  is

$$p\chi_{\{f(y) \leq v \leq f(x)\}} + (1-p)\chi_{\{f(x) \leq v \leq f(y)\}}.$$

Therefore, by integrating over  $V$ , the transition operators of the 0 temperature process are:

$$\begin{aligned} G_n(y|x) &= \frac{1}{c_n(x)} Q_n(y|x) \\ &= \frac{1}{c_n(x)} \begin{cases} (1-p)(f(y) - f(x)) & \{f(y) > f(x)\} \cap B_{\lambda_n}(x) \\ p(f(x) - f(y)) & \{f(y) < f(x)\} \cap B_{\lambda_n}(x) \\ 0 & \text{otherwise} \end{cases}. \end{aligned} \quad (3.2)$$

The normalizing constant  $c_n(x)$  is determined by the fact that the probabilities that  $x$  remains stationary and that  $x$  moves to  $y$  should add up to 1. Hence,

$$c_n(x) = (1-p)f(x) + p(1-f(x)) + \int_{B_{\lambda_n}(x)} Q_n(y|x) d(y).$$

Denote by  $(X_n)$  the orbits of the process 3.2 at the  $n$ -th stage.

**Theorem 3.1** *Let  $\lambda_n$  be the sequence of the learning steps and assume that  $(\lambda_n) \in \ell_{d+3} \setminus \ell_{d+2}$ . If  $f$  has a unique critical point in  $X$  and if that point is a global maximum, then the orbits  $(X_n)$  converge in probability to that point. If  $p = 0$  the convergence is  $\mathcal{P}$ -almost surely.*

Put  $\rho_n = \lambda_n^{d+2}$ ,  $Y_n(x) = (X_{n+1} - X_n)/\rho_n$  given that  $X_n = x$ , denote by  $H(x)$  the Hessian of  $f$  at  $x$  and set  $U_n(x) = \mathbb{E}(\langle \nabla f(x), Y_n(x) \rangle | X_n = x)$ . Let  $o$  be the unique point for which  $\nabla f(o) = 0$  and assume that  $o$  is a global maximum of  $f$ .

The following Lemma describes the properties of the random variables defined above.

**Lemma 3.2** 1. Assume that  $K$  is a compact set such that  $d(K, o) > 0$ . Then, there are an integer  $N$  and some  $L > 0$  such that for every  $n > N$ ,

$$\inf_{x \in K} U_n(x) \geq L.$$

2. There is an integer  $N$  such that for every  $n > N$ ,  $\mathbb{E}U_n \geq 0$ .

3. If  $a_n = \sup_x \rho_n^2 \mathbb{E} \|Y_n(x)\|^2$  then  $(a_n) \in \ell_1$ .

**Proof:** Let  $S$  be the unit sphere in  $\mathbb{R}^d$  and denote by  $|\cdot|$  the Haar measure on  $S$ . Put  $S_x = \{s \in S | \langle \nabla f(x), s \rangle > 0\}$  and  $S_{x,r}^+ = \{s \in S | f(x+rs) > f(x)\}$ .  $S_{x,r}^-$  is defined in a similar way with the reversed inequality. Since  $c_n(x)$  tends to  $(1-p)f(x) + p(1-f(x))$  uniformly and since  $U_n(x) = \frac{1}{\rho_n} \langle \nabla f(x), \mathbb{E}(X_{n+1} - X_n | X_n = x) \rangle$  it is enough to prove the first claim for the function

$$\begin{aligned} U_n(x)c_n(x) &= \frac{1}{\rho_n} \left\langle (1-p) \int_{A_n(x)} (y-x)(f(y)-f(x))dy + \right. \\ &\quad \left. + p \int_{B_n(x)} (y-x)(f(x)-f(y))dy, \nabla f(x) \right\rangle = (*), \end{aligned}$$

where  $A_n(x) = \{f(y) > f(x)\} \cap B_{\lambda_n}(x)$  and  $B_n(x) = \{f(x) > f(y)\} \cap B_{\lambda_n}(x)$ . By Taylor's formula,  $f(y) - f(x) = \langle \nabla f(x), y-x \rangle + O(\|y-x\|^2)$ , hence

$$\begin{aligned} (*) &= \frac{1}{\rho_n} \left( (1-p) \int_{A_n(x)} \langle \nabla f(x), y-x \rangle^2 dy - p \int_{B_n(x)} \langle \nabla f(x), y-x \rangle^2 dy + \right. \\ &\quad \left. + \int_{B_{\lambda_n}(x)} O(\|y-x\|^3) dy \right). \end{aligned}$$

A simple calculation shows that the third term converges uniformly on  $K$  to 0. Therefore, it is enough to estimate the first and second terms. Note that

$$\begin{aligned} &\int_{A_n(x)} \langle \nabla f(x), y-x \rangle^2 dy - p \int_{B_n(x)} \langle \nabla f(x), y-x \rangle^2 dy = \\ &\int_0^{\lambda_n} r^{d+1} \left( (1-p) \int_{S_{x,r}^+} \langle \nabla f(x), s \rangle^2 h(s) ds - p \int_{S_{x,r}^-} \langle \nabla f(x), s \rangle^2 h(s) ds \right) dr, \end{aligned}$$

where  $r^{d-1}h(s)$  is the Jacobian of the transformation to spherical coordinates. Set  $g(x, r) = (1-p) \int_{S_{x,r}^+} \langle \nabla f(x), s \rangle^2 h(s) ds - p \int_{S_{x,r}^-} \langle \nabla f(x), s \rangle^2 h(s) ds$ . Then,

$$\begin{aligned} g(x, r) &= \\ &\int_{S_{x,r}^+} \langle \nabla f(x), s \rangle^2 h(s) ds - p \int_S \langle \nabla f(x), s \rangle^2 h(s) ds = \\ &g_1(x, r) - pg_2(x). \end{aligned}$$

To finish the proof of the first claim, it is enough to find  $R$  and  $l$  such that for every  $r < R$  and every  $x \in K$ ,  $g(x, r) \geq l$ . Indeed, once we establish the

existence of such  $R$  and  $l$ , and if  $\lambda_n < R$ , then

$$\frac{1}{\rho_n} \int_0^{\lambda_n} r^{d+1} g(x, r) dr \geq \frac{1}{\rho_n} \int_0^{\lambda_n} r^{d+1} l \geq \frac{l}{d+2}.$$

Let  $D_{x,t} = \{s \in S \mid \langle \nabla f(x), s \rangle > t\}$ . It is easy to see that for every  $x$ , the sets  $D_{x,t}$  increase to  $S_x$  as  $t$  tends to 0. Since  $|D_{x,t}|$  and  $|S_x|$  are both continuous functions of  $x$ , then by Dini's Theorem,  $|D_{x,t}|$  converges uniformly to  $|S_x|$  on  $K$ . For every  $\varepsilon > 0$  there is some  $t > 0$  such that for every  $x$ ,  $|S_x \setminus D_{x,t}| = |S_x| - |D_{x,t}| < \frac{\varepsilon}{2M}$ , where  $M = \sup_x \|\nabla f(x)\|$ . Since  $\nabla f$  is continuous, there is some  $\delta_x > 0$  such that for every  $y \in B_{\delta_x}(x)$ ,  $D_{x,t} \subset D_{y,t/2}$  and  $\nu(C_x \Delta C_y) < \frac{\varepsilon}{2M}$ . Note that if  $\|y - x\| < \delta_x$ ,  $s \in D_{x,t}$  and  $z = y + rs$ , there is some absolute constant  $C$  such that

$$f(z) - f(y) = \langle \nabla f(y), s \rangle r + o(r^2) \geq \frac{tr}{2} - Cr^2.$$

Hence, there is some  $R > 0$  such that if  $r < R$  and  $y \in B_{\delta_x}(x)$ , then  $D_{x,t} \subset S_{y,r}^+ \cap D_{y,t/2} \subset S_y$ .

For such  $r$  and  $y$  we see that

$$\begin{aligned} g_1(y, r) &= \int_{S_{y,r}^+} \langle \nabla f(y), s \rangle^2 h(s) ds \geq \int_{D_{x,t}} \langle \nabla f(y), s \rangle^2 h(s) ds = \\ &= \int_{S_y} \langle \nabla f(y), s \rangle^2 h(s) ds - \int_{S_y \setminus D_{x,t}} \langle \nabla f(y), s \rangle^2 h(s) ds > \\ &> \int_{S_y} \langle \nabla f(y), s \rangle^2 h(s) ds - M(|S_y \setminus S_x| + |S_x \setminus D_{x,t}|) > \\ &> \int_{S_y} \langle \nabla f(y), s \rangle^2 h(s) ds - \varepsilon. \end{aligned}$$

On the other hand, for every  $y$ ,

$$pg_2(y) = p \int_S \langle \nabla f(y), s \rangle^2 h(s) ds = 2p \int_{S_y} \langle \nabla f(y), s \rangle^2 h(s) ds.$$

Therefore,  $g(y, r) \geq (1 - 2p) \int_{S_y} \langle \nabla f(y), s \rangle^2 h(s) ds - \varepsilon$ . Since  $f \in C_2(X)$  and  $\nabla f(x) \neq 0$  on  $K$  then  $\int_{S_x} \langle \nabla f(x), s \rangle^2 h(s) ds \geq C$  on  $K$ , implying that  $g(y, r) > (1 - 2p)C - \varepsilon$ . From here, the claim follows using a standard compactness argument.

Let us turn to the second assertion. By the same argument used above,

$$\liminf_{r \rightarrow 0} g(x, r) \geq (1 - 2p) \int_{S_x} \langle \nabla f(x), s \rangle^2 h(s) ds = \eta(x).$$

Therefore, by Fatou's lemma

$$\begin{aligned} \liminf_{n \rightarrow \infty} U_n(x)c_n(x) &= \\ &= \liminf_{n \rightarrow \infty} \frac{1}{\rho_n} \int_0^{\lambda_n} r^{d+1} g(x, r) dr \geq \frac{1}{\rho_n} \int_0^{\lambda_n} r^{d+1} \eta(x) dr = \\ &= \frac{\eta(x)}{d+2}. \end{aligned}$$

Hence,  $\liminf_{n \rightarrow \infty} \mathbb{E}U_n \geq \mathbb{E} \frac{\eta(x)}{(d+2)c(x)} > 0$ .

Turning to the final claim, note that for every integer  $n$  and every  $x$ ,

$$\begin{aligned} \rho_n^2 \mathbb{E} \|Y_n(x)\|^2 &\leq \\ &= \frac{1}{(1-p)f(x) + p(1-f(x))} \int_{A_n(x) \cup B_n(x)} \|y-x\|^2 |f(y) - f(x)| dy \leq \\ &\leq C \int_{A_n(x) \cup B_n(x)} \|y-x\|^2 \left| \langle \nabla f(x), y-x \rangle + O(\|y-x\|^2) \right| dy \leq \\ &\leq C' \int_{B_{\lambda_n}(x)} \|y-x\|^3 dy \leq C'' \lambda_n^{d+3}, \end{aligned}$$

where  $C, C'$  and  $C''$  are absolute constants. Thus,

$$\sum_{n=1}^{\infty} a_n \leq C'' \sum_{n=1}^{\infty} \lambda_n^{d+3} < \infty.$$

■

**Corollary 3.3** *If  $\mathbb{E}U_n$  converges to 0 then  $\mathbb{E}|U_n|$  converges to 0 too.*

**Proof:** For every  $\varepsilon > 0$ , fix a compact set  $K \subset X$  such that  $\mu(X \setminus K) < \varepsilon$  and  $d(K, o) > 0$ . By the Lemma, for an integer  $n$  large enough, the functions  $U_n$  are nonnegative on  $K$ . Also,  $U_n$  are uniformly bounded and w.l.o.g we assume that they are bounded by 1. Therefore,

$$\begin{aligned} \mathbb{E}|U_n| &= \int_{X \setminus K} |U_n| d\mu + \int_K U_n d\mu \leq \\ &= \mathbb{E}U_n + 2 \int_{X \setminus K} |U_n| d\mu < \mathbb{E}U_n + 2\varepsilon. \end{aligned}$$

■

**Proof of Theorem 3.1:** The idea behind the proof is to use Taylor's formula to approximate the differences  $f(X_{n+1}) - f(X_n)$ . Indeed, given  $X_n$ , we see that

$$\begin{aligned} f(X_{n+1}) &= \\ f(X_n + \rho_n Y_n) &= f(X_n) + \rho_n \langle \nabla f(X_n), Y_n \rangle + \frac{1}{2} \rho_n^2 \langle Y_n, H(X_n + \theta \rho_n Y_n) Y_n \rangle. \end{aligned}$$

If  $V_n(x) = \mathbb{E}(\langle Y_n, H(X_n + \theta \rho_n Y_n) \rangle | X_n)$  then the conditional expectation of the expression for  $f(X_{n+1})$  is

$$\mathbb{E}(f(X_{n+1}) | X_n) = f(X_n) + \rho_n U_n(x) + \frac{1}{2} \rho_n^2 V_n(x).$$

Taking expectations on both sides and iterating it follows that

$$\mathbb{E}(f(X_{n+1})) = \mathbb{E}f(X_1) + \sum_1^n \rho_i \mathbb{E}U_i + \sum_1^n \frac{1}{2} \rho_i^2 \mathbb{E}V_i. \quad (3.3)$$

Since  $f$  is bounded by 1 then  $|\mathbb{E}(f(X_{n+1}))| \leq 1$ . By the Cauchy-Schwarz inequality  $|V_i| \leq C \sup_x \mathbb{E} \|Y_i(x)\|^2$ , which implies that

$$\rho_i^2 \mathbb{E} |V_i| \leq C \rho_i^2 \sup_x \mathbb{E} \|Y_i(x)\|^2 = C a_i.$$

Recall that by Lemma 3.2  $(a_i) \in \ell_1$ , hence  $\sum_1^n \rho_i^2 \mathbb{E}V_i$  converges. Again, by the Lemma,  $\sum_{i=1}^n \rho_i \mathbb{E}U_i$  is a nonnegative series sufficiently large  $i$ . By 3.3 and the above this series is bounded, thus it converges, which implies that  $\mathbb{E}(f(X_n))$  converges too.

Since  $(\rho_n) \notin \ell_1$ , there is a subsequence  $\mathbb{E}U_{n_j}$  which tends to 0, and by Corollary 3.3  $\mathbb{E} |U_{n_j}|$  tends to 0 too. Using Chebyshev's inequality,  $U_{n_j}$  converges in probability to 0. Therefore, there is a subsequence of  $U_{n_j}$ , also denoted by  $U_{n_j}$ , which converges almost surely to 0. According to Lemma 3.2,  $U_n$  are uniformly bounded away from 0 on every compact set not containing  $o$ . Therefore  $(X_{n_j})$  must converge almost surely to  $o$ . On the other hand,  $\mathbb{E}(f(X_n))$  converges and it is a continuous function of  $X_n$ . Hence, it must converge to  $\mathbb{E}f(o) = f(o)$ . Since  $o$  is a unique maximum of  $f$ , then for every  $\varepsilon > 0$  there is some  $\delta > 0$  such that  $\{\|X_n - o\| \geq \varepsilon\} \subset \{f(o) - f(X_n) \geq \delta\}$  and by Chebyshev's inequality the measure of the later set tends to 0.

To prove the second claim, note that if  $p = 0$  then  $f(X_n)$  is increasing almost surely. Therefore,  $(f(X_n))$  converges almost surely. Since  $\mathbb{E}(f(X_n))$  converges to  $f(o)$ ,  $f(X_n)$  must converge to  $f(o)$  almost surely, and since  $o$  is a unique maximum,  $X_n$  converges to  $o$  almost surely. ■

## 4 Proofs of the results from part 1

In this section our aim is to prove the results stated in the first section.

Recall that for every set  $A$  of positive measure,  $P_n(A|x)$  is the transition probability from  $x$  to  $A$  in the  $n$ -th stage. Clearly, for every  $x$  and any such set  $A$ ,

$$P_n(A|x) = \frac{1}{c_n(x)} \int_{A \cap B_\lambda(x)} \int_V \mathcal{E}(x, v) e^{-\mathcal{E}(x', v)/T_n} d\mu(x') dv(v) = \frac{f_n^A(x)}{c_n(x)} \quad (4.1)$$

converges pointwise to

$$P(A|x) = \frac{\int_{A \cap B_\lambda(x)} \int_V \mathcal{E}(x, v) \chi_{C_x} d\nu(v) d\mu(x')}{\nu(C_x) + \int_{B_\lambda(x)} \int_V \mathcal{E}(x, v) \chi_{C_x} d\nu(v) d\mu(x')} = \frac{f^A(x)}{c(x)}. \quad (4.2)$$

Let  $\mathcal{P}^0$  be the probability measure induced by the orbits  $(X_n)$  of the 0-temperature process 4.2 and set  $\mathcal{P}$  to be the induced measure by the orbits of process 4.1.

We will show that the convergence of  $P_n$  to  $P$  is uniform both in  $x$  and in  $A$ . First, we prove that  $f_n^A(x)$  converges uniformly to  $f^A(x)$ . With a similar argument, it is possible to show that  $c_n(x)$  converges uniformly to  $c(x)$ . The desired convergence follows since  $c_n(x)$  and  $c(x)$  are bounded away from 0.

**Lemma 4.1** *For every  $A \subset X$  of positive measure, let  $f^A$  and  $f_n^A$  be as in 4.1 and 4.2. Then,  $(f_n^A)$  converges to  $f^A$  uniformly both in  $x$  and in  $A$ .*

**Proof:** We may assume that  $\mathcal{E}(x, v)$  is bounded by 1.

Set  $E = \{(x', s) | \mathcal{E}(x', s) > 0\}$  and note that if  $(x', v) \notin E$  then  $e^{-\mathcal{E}(x', v)/T_n} = \chi_{C_{x'}}(v)$ . Also, for every  $\varepsilon > 0$  there is a set  $K \subset E$  and some  $\beta > 0$  such that  $\tau(E \setminus K) < \varepsilon$  and  $\mathcal{E}(x', v) > \beta$  on  $K$ . Therefore,

$$\begin{aligned} \sup_{x \in X} |f_n^A - f^A| &\leq \sup_{x \in X} \int_{X \times V} \mathcal{E}(x, v) \left| e^{-\mathcal{E}(x', v)/T_n} - \chi_{C_{x'}}(v) \right| d\tau(x', v) = \\ &= \sup_{x \in X} \int_E \mathcal{E}(x, v) \left| e^{-\mathcal{E}(x', v)/T_n} - \chi_{C_{x'}}(v) \right| d\tau(x', v) = (*). \end{aligned}$$

Since the integrands are uniformly bounded by 2,  $\chi_{C_x}(v)$  vanishes on  $E$  and  $\tau(E \setminus K) < \varepsilon$  it follows that

$$(*) \leq 2\varepsilon + \int_K \mathcal{E}(x, v) e^{-\beta/T_n} d\tau \leq 2\varepsilon + e^{-\beta/T_n}.$$

Clearly the estimates above are uniform in the set  $A$ , which proves our claim.  $\blacksquare$

Another observation which follows using a similar computation, is the following:

**Lemma 4.2** *For every set  $A \subset X$ , the function  $P(A|x)$  is continuous. Moreover, the function  $P(X_n \in A_n, \dots, X_2 \in A_2 | X_1 = x)$  is continuous function of  $x$  for every measurable sets  $A_2, A_3, \dots, A_n$ .*

To prove this fact, one has to use the assumption that for every  $x$ ,  $\nu(C_x \Delta C_{x'})$  is a continuous function of  $x'$ . The proof of the Lemma is straightforward, and is omitted.

Our aim is to use information concerning the 0-temperature process 4.2 to derive similar results about process 4.1.

If  $(X_n)$  denotes an orbit then for every set  $O \subset X$ , put  $O_i = \{(X_n) | X_i \in O \text{ for } n = i\}$ ,  $L_n^0(x, O) = \mathcal{P}^0\{X_i \in O \text{ for some } i \geq n | X_n = x\}$ ,  $L_n(x, O) = \mathcal{P}\{X_i \in O \text{ for some } i \geq n | X_n = x\}$  and  $L^0(x, O)$  is the  $\mathcal{P}^0$ -probability to enter  $O$  infinitely often given that  $X_1 = x$ .

**Lemma 4.3** *Let  $O \subset X$  for which there are  $\alpha > 0$  and an integer  $N$  such that for every  $x \in X$ ,  $\mathcal{P}^0\{\cup_1^N O_i | X_1 = x\} > \alpha$ . Then, there is some integer  $M_0$  such that for every  $m > M_0$  and every  $x \in X$ ,  $L_m(x, O) \geq \frac{\alpha}{2}$ .*

**Proof:** Recall that the 0-temperature is a homogeneous Markov process. Also, since for every  $m$  and every  $x$ ,  $\mathcal{P}^0\{\cup_m^{m+N} O_i | X_m = x\} > \alpha$ , and since  $P_n$  converges uniformly to  $P$ , then for  $m$  large enough and every  $x$ ,

$$\mathcal{P}\{\cup_m^{m+N} O_i | X_m = x\} > \frac{\alpha}{2}.$$

Hence, for  $m$  large enough and for every  $x$ ,  $L_m(x, O) > \frac{\alpha}{2}$ . ■

**Lemma 4.4** *Assume that there are  $\alpha > 0$  and an integer  $N$  such that for every  $n > N$  and for every  $x$ ,  $L_n(x, O) > \alpha$ . Then, the orbits of the process 4.1 enter  $O$  infinitely often  $\mathcal{P}$ -almost surely.*

This result appears in Orey (1971) in a slightly weaker form. The proof uses the same idea as the one presented in Orey (1971), and is brought for the sake of completeness.

**Proof:** The first part of the proof is a version of a 0-1 law which is due to P. Lévy (see Loève (1963)). Let  $Y_1, Y_2, \dots$ , be a sequence of random variables and put  $Y$  to be a random variable defined on  $Y_1, Y_2, \dots$  such that  $\mathbb{E}|Y| < \infty$ . Clearly,  $Z_n = \mathbb{E}(Y | Y_1, \dots, Y_n)$  forms a martingale, hence, by the martingale convergence theorem (Loève (1963), pg. 393),  $Z_n$  converges almost surely to  $Y$ . By setting  $B_i = \{X_i \in O\}$ ,  $B = \{X_n \in O \text{ i.o.}\}$ ,  $Y_n = X_n$  and  $Y = \chi_B$ , it follows that  $\mathcal{P}(B | Y_1, \dots, Y_n) = \mathbb{E}(Y | Y_1, \dots, Y_n)$  converges almost surely to  $\chi_B$  and  $\mathcal{P}(\cup_k^\infty B_i | Y_1, \dots, Y_n)$  tends to  $\chi_{\cup_k^\infty B_i}$  for every fixed  $k$ .

On the other hand, for every  $k \leq n$ ,

$$\mathcal{P}(\cup_k^\infty B_i | Y_1, \dots, Y_n) \geq \mathcal{P}(\cup_n^\infty B_i | Y_1, \dots, Y_n) \geq \mathcal{P}(B | Y_1, \dots, Y_n),$$

thus, by taking  $n \rightarrow \infty$ ,

$$\chi_{\cup_k^\infty B_i} \geq \limsup_{n \rightarrow \infty} \mathcal{P}(\cup_n^\infty B_i | Y_1, \dots, Y_n) \geq \liminf_{n \rightarrow \infty} \mathcal{P}(\cup_n^\infty B_i | Y_1, \dots, Y_n) \geq \chi_B.$$

Taking  $k \rightarrow \infty$ , the left side converges almost surely to  $\chi_B$ . Hence,

$$\mathcal{P}(\cup_n^\infty B_i | Y_1, \dots, Y_n) \rightarrow \chi_B.$$

Denote by  $X_\infty$  the set of all the orbits of the process. Since

$$L_n(X_n, O) = \mathcal{P}(\cup_n^\infty B_i | X_1, \dots, X_n),$$

then by the 0-1 law presented above,  $L_n(X_n, O)$  tends to the characteristic function of the set  $\{X_n \in O \text{ i.o.}\}$ . By our assumption, for  $n$  large enough and

for every  $x$ ,  $L_n(x, O) > \alpha$ . For such  $n$ ,  $L_n(X_n, O) > \alpha$  almost surely. Therefore, excluding a set of 0-probability,

$$X_\infty \subset \{\limsup_{n \rightarrow \infty} L_n(X_n, O) > 0\} = \{\lim_{n \rightarrow \infty} L_n(X_n, O) = 1\} = \{X_n \in O \text{ i.o.}\}.$$

Hence,  $X_n \in O$  infinitely often  $\mathcal{P}$ -almost surely. ■

By a similar method, one shows the following:

**Lemma 4.5** *Assume that there are  $\alpha > 0$  and an integer  $N$ , such that for every  $n > N$  and every  $x \in A$ ,  $L_n(x, B) > \alpha$ . Then,  $\mathcal{P}$ -almost surely, orbits which visits  $A$  infinitely often also visit  $B$  infinitely often.*

**Corollary 4.6** *Combining Lemma 4.3 with Lemma 4.4, it follows that in order to prove Theorem 2.2.a, it is enough to show that for every neighborhood  $A$  of  $O$ , there are  $\alpha > 0$  and an integer  $N$  such that for every  $x$ ,  $\mathcal{P}^0(\cup_1^N A_i | x) > \alpha$ .*

**Proof of Theorem 2.2:** We begin with the proof of (b). Let  $A$  be an open set containing  $Q = \{x | \mathbb{E}_g(x) = 0\}$ . Since  $\lambda = \text{diam} X$  and since  $Q$  has a  $\mu$ -positive measure, every  $x$  has a positive  $\mathcal{P}^0$ -probability to enter  $Q$ . Since  $A^c$  is compact, a simple continuity argument shows that there is some  $\alpha > 0$  such that for every  $x \notin A$ ,  $P(Q|x) \geq \alpha$ . But, since  $P_n \rightarrow P$  uniformly, then for  $n$  large enough  $\inf_{x \notin A} P_n(Q, x) \geq \frac{\alpha}{2}$ . Hence, by Lemma 4.5, orbits which visit  $A^c$  i.o. must enter  $Q$   $\mathcal{P}$ -almost surely. This is impossible since  $Q$  is an absorbing set. Thus,  $\mathcal{P}$ -almost every orbit enters  $A^c$  a finite number of times, implying that the orbits of the process 2.1 converge  $\mathcal{P}$ -a.s. to  $Q$ .

To prove (a) we use Corollary 4.6. Assume that we limit the size of the learning step to  $\lambda$  and let  $A$  be a neighborhood of  $O$ . Set  $A_i = \{(X_n) | X_i \in A \text{ for } n = i\}$  and recall that

$$\mathcal{P}^0(A^i \setminus \cup_1^{i-1} A_j | X_1 = x) = \mathcal{P}^0(X_i \in A, X_{i-1} \in A^c, \dots, X_2 \in A^c | X_1 = x)$$

is a continuous function of  $x$ . Clearly, for every integer  $n$ ,

$$\mathcal{P}^0(\cup_1^n A_i | X_1 = x) = \sum_1^n \mathcal{P}^0(A_i \setminus \cup_1^i A_j | X_1 = x).$$

Therefore, for every  $n$ ,  $h_n(x) = \mathcal{P}^0(\cup_1^n A_i | X_1 = x)$  is a continuous functions too.

Note that if for every  $x$  there is some integer  $n = n(x)$  such that  $h_n(x) = \varepsilon_x > 0$ , then by the continuity of  $h_n$  there is a neighborhood  $U_x$  of  $x$  on which  $h_n(x) > \frac{\varepsilon}{2}$ . Since  $X$  is compact and  $(h_n(x))$  is a monotone increasing sequence, there exists a finite sub-cover  $(U_{x_i})$  of  $X$ , such that for every  $x \in U_{x_i}$  and every  $n > n(x_i)$ ,  $h_n(x) \geq \frac{\varepsilon_{x_i}}{2}$ . Therefore, there are  $\alpha > 0$  and an integer  $N$  such that for every  $x \in X$ ,  $h_N(x) > \alpha$ . By Corollary 4.6, this suffices to prove our claim.

Moreover, note that it suffices to show that  $L^0(x, A) > 0$  for every  $x$ , since this implies that for every  $x$  there is some integer  $N(x)$  such that for every  $n > N(x)$ ,  $h_n(x) > 0$ .

Finally, to show that indeed, for every  $x \in X$   $L^0(x, A) > 0$ , recall that since the error function is 0-1, then if  $\mathbb{E}_g(x) < \mathbb{E}_g(y)$  and  $d(x, y) \leq \lambda$ , there is a positive transition density from  $x$  to  $y$ .

For every  $x$ , let  $y_x$  be a point in  $B_{\lambda/2}(x)$  in which the relative maximum of  $\nu(C_x)$  is attained in  $B_{\lambda/2}(x)$ . Define a sequence  $x_1 = x, x_2 = y_x, x_3 = y_{y_x}$  and so on. A simple compactness argument shows that  $x_i = x_0$  for  $i$  larger than some  $n$ . Thus,  $x_0$  must be a local maximum of  $\nu(C_x) = 1 - \mathbb{E}_g(x)$ . Note that by the above, there is a positive transition density from  $x$  to  $y_x$ . Moreover, since  $\nu(C_{y_x} \Delta C_z)$  is a continuous function of  $z$ , there is some  $0 < \varepsilon < \frac{\lambda}{2}$  such that for every  $z \in B_\varepsilon(y_x) \subset B_\lambda(x)$ ,  $x$  has a positive transition density to  $z$ . Since  $B_\varepsilon(y_x)$  has a nonempty interior, it has a  $\mu$ -positive measure, implying that the transition from  $x$  into that set is positive. Using a similar argument one shows that there is a positive probability that  $B_\varepsilon(y_x)$  is mapped into an arbitrarily small ball centered in  $y_{y_x}$ . It takes a finite number of steps to reach a neighborhood of  $x_0$  which is a subset of  $A$ , thus  $L^0(x, A) > 0$ . ■

**Theorem 2.3 – Sketch of proof:** The proof of Theorem 2.3 goes along the same lines as the proof of Theorem 2.2. The only difference is in the definition of  $y_x$ . The idea is to equip  $B_\lambda(x)$  with the partial order  $\leq$  defined by:  $x \leq y$  if and only if  $C_x \subset C_y$ . Then, use Zorn's lemma to find a maximal element in  $B_\lambda(x)$  and let  $y_x$  be that maximal element. Again, define the sequence  $(x_n)$  and use a compactness argument to show both that for  $i > n$ ,  $x_i = x_0$  and that  $x_0$  is the maximal element in  $B_\lambda(x_0)$ . Note that by Assumption 2.2, the only elements which are maximal in their neighborhood are elements of  $\mathcal{O}$ .

The proof of Theorem 2.3.b is identical to that of Theorem 2.2.b and does not require any additional assumptions. ■

## 5 Discussion

This investigation was motivated by the work of Kim and Sompolinsky, (1996) and (1998), who introduced the On-line Gibbs algorithm.

The on-line Gibbs learning is slightly different than the model we defined. In the OLGA the conditional transition density is given by

$$P_n(X_{n+1} = x' | X_n = x, V_n = v) = \frac{1}{c} e^{-\frac{E_n(x', x, v)}{2T_n}}, \quad (5.1)$$

where  $E_n(x', x, v) = \mathcal{E}(x', v) + \frac{1}{2}\lambda_n \|x - x'\|^2$ .

In the 0-temperature process,  $x$  moves during the  $n$ -th stage to the nearest point  $x'$  which gives a correct response to the input  $v$  – assuming that  $\|x - x'\| \leq \sqrt{2/\lambda_n}$ , and remains stationary otherwise. Note that the part that  $\lambda$  plays in process 2.1 is similar to the role of  $\lambda^{-1}$  in the OLGA.

There are several differences between process 2.1 and the OLGA. For  $T \neq 0$ , the processes differ on two main aspects. First, is the way the bounded change

enters the game: in process 2.1 the bounded change is simply a cut-off function which prevents the orbits to move “to far”. In the OLGA rule this is done in a more smooth fashion, which is the addition the term  $\frac{1}{2}\lambda_n \|x - x'\|^2$  to the energy function.

The second and more important difference is the additional linear term  $\mathcal{E}(x, v)$  appearing in the transition densities of process 2.1. This feature prevents large changes if the student gives an “almost correct” answer, and prevents change if the answer is correct. Note that in the OLGA rule for  $T \neq 0$ , changes may occur even if the student’s answer is correct.

In the 0-temperature process the OLGA changes the current state to the nearest state which yields a correct answer to the given input, provided that the change is not “too large”. For example, if the answer is correct, no change is made.

As in the OLGA, in our version of the 0-temperature process the change is made only if the student gave the wrong answer to the given input. However, unlike the OLGA, the transition is made to any state which gives the correct answer and is close enough to the current state. Hence, given the current state  $x_n$  and such an input  $v$ ,  $x_{n+1}$  will be equally distributed on the set  $\{x | d(x, x_n) \leq \lambda, \mathcal{E}(x, v) = 0\}$ .

This definition has two advantages compared with the 0-temperature OLGA rule. One is that this is the “natural” limit of the process where  $T \rightarrow 0$ , and the other is that one does not have to solve a difficult optimization problem needed to run the OLGA at  $T = 0$ .

Our selection of the 0-temperature process has its analytical benefits too. The fact that the transition operators as  $T \rightarrow 0$  converge uniformly to the 0-temperature operators allowed us to use the simpler 0-temperature process to estimate the orbits of the process as  $T \rightarrow 0$ .

Much like the OLGA, process 2.1 has a local nature. In process 2.1 the local nature is expressed in the assumption that each state has “many” inputs on which it agrees with the teacher. In the case of the OLGA, the local nature is more explicit. In most of the examples investigated in Kim and Sompolinsky (1998) for  $T = 0$  the results are obtained by expansions near a local minimum while assuming that the learning steps decrease to 0. The philosophy behind this idea is that the orbits are eventually trapped in a neighborhood of a local minimum and the event of transition from the neighborhood of one local minimum to the other is rare. The authors focused on estimating the fluctuations of the *average* error  $\mathbb{E}(\mathbb{E}_g(X_n))$  from  $\mathbb{E}_g(x_{\min})$ , where  $x_{\min}$  is the local minimum. For  $T \neq 0$  and  $\lambda \rightarrow \infty$  it was claimed that the OLGA converges to the Gibbs distribution where  $\mathbb{E}_g$  plays the role of the energy function. Thus, like other annealing processes, to ensure convergence in distribution to a global minimum of  $\mathbb{E}_g$  one has to decrease the temperature slowly while decreasing the learning steps.

We were able to prove convergence results for process 2.1. Namely, we showed that the orbits come arbitrarily close to a local minimum of  $\mathbb{E}_g$  almost surely without assuming that the state space is a small neighborhood of a local minimum. Moreover, if there are “many” correct states, then by Theorem 2.3.b

the orbits of process 2.1 converge almost surely to a correct state.

Note that both process 2.1 and the OLGA may not converge to a global minimum of  $\mathbb{E}_g$  when  $T = 0$ . To show that we present two examples with a common feature: both of them are 0-temperature processes in which the process does not converge to a global minimum of  $\mathbb{E}_g$ . In the first example the orbits are trapped in a local minimum and in the second example the situation is even worse – the global *maximum* of  $\mathbb{E}_g$  is absorbing.

**Example 5.1** Set  $X = [-1/2, 2]$ ,  $V = [-1, 1]$  and assume that the error function is 0-1. We define  $\mathcal{E}(x, v)$  which induces the 0-temperature process using figure 1.

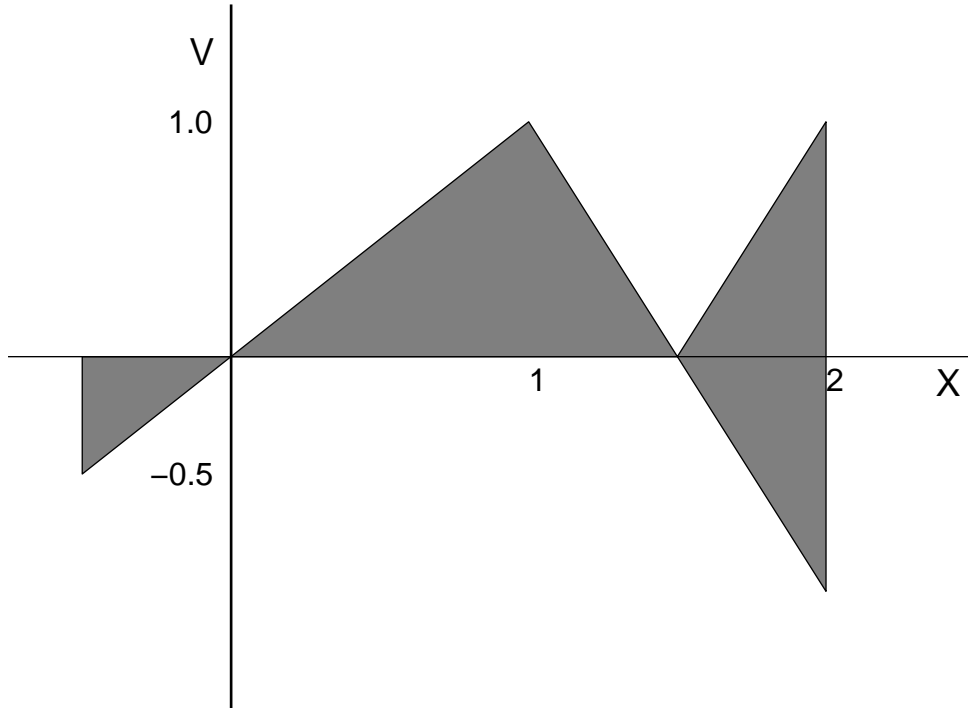


Figure 1, Mendelson, Manuscript number 1963

Here,  $\mathcal{E}(x, v) = 0$  on the shaded area and 1 otherwise. Note that the global minimum of  $\mathbb{E}_g$  in  $X$  is attained at  $x = 2$  and in this state, the student is in complete agreement with the teacher. However, if  $x < -\sqrt{2/\lambda_n}$ , it can not move towards 2. Indeed, if  $x < v < 0$ ,  $x$  gives a correct response to  $v$  so

it remains stationary. For  $v > 0$ ,  $x$  does not move because the nearest point which yields a correct response to  $v$  is too far away. Hence, if for example, the initial distribution  $\mu$  is supported in  $[-1/2, -1/4]$  and  $\lambda_n > 32$ , the orbits of the 0-temperature process converge almost surely to  $x = -1/2$ .

In a similar fashion, for every learning step sequence  $\sqrt{2/\lambda_n}$  tending to 0, there are initial distributions such that the orbits of the 0-temperature do not converge in distribution to the global minimum of  $\mathbb{E}_g(x)$ . Moreover, for some initial distributions and for every sequence  $\lambda_n$  such that each  $\lambda_i > M$ , the process 5.1 converges almost surely to a local minimum of  $\mathbb{E}_g$  which is not the global minimum.

The following example serves as a counterexample to one of the claims appearing in Kim and Sompolinsky (1998). It is a smooth on-line error function for which the OLGA at  $T = 0$  and  $\lambda \rightarrow \infty$  does not converge to a local minimum of  $\mathbb{E}_g$  (even in the weak sense of convergence of the average error), since the global maximum of  $\mathbb{E}_g$  is an absorbing state. This emphasizes the local nature needed for the OLGA: if one does not limit the state space to a neighborhood of the local minimum, the orbits may be trapped in another domain of attraction, which, in this case, is a global maximum of  $\mathbb{E}_g$ .

**Example 5.2** Here we construct a continuous function  $\mathcal{E}(x, v)$  on the set  $D = X \times V = [-1/2, 1/2] \times [0, 2]$  with respect to the normalized Lebesgue measure, such that the global maximum of  $\mathbb{E}_g$  is an absorbing state.

Define the error function on  $D$  by

$$\mathcal{E}(x, v) = \begin{cases} \frac{2(1-x^2)(v+x^2-1)}{x^2+1} & \text{if } v > 1 - x^2 \\ 0 & \text{if } v \leq 1 - x^2. \end{cases}$$

Clearly, for every  $x$ ,  $C_x = \{v | 0 \leq v \leq 1 - x^2\}$ . Thus,  $C_o \supset C_x$ ,  $\mathcal{E}(x, v) \geq 0$  on  $D$  and  $x = 0$  is an absorbing state. Indeed, for every input  $v > 1$  there is no state for which  $\mathcal{E}(x, v) = 0$  and for  $0 \leq v \leq 1$ ,  $\mathcal{E}(0, v) = 0$ . Moreover, it is clear that the 0-temperature OLGA converges to  $x = 0$ . Indeed, note that if  $|x| > |y|$  then the transition probability from  $y$  to  $x$  is 0, because  $C_y \supset C_x$ . Hence, if  $y$  given the wrong answer on the input  $v$ , so will  $x$  - and no transition will be made. Therefore, the dynamics of the processes "pushes" any initial state towards  $x = 0$ .

On the other hand,  $\mathbb{E}_g(x) = (1 - x^2) \int_{1-x^2}^2 \frac{2(v+x^2-1)}{x^2+1} dv$ . Changing the integration variable to  $z = \frac{2(v+x^2-1)}{x^2+1}$  we see that  $\mathbb{E}_g(x) = \frac{1-x^4}{2} \int_0^2 z dz = 1 - x^4$ . Thus  $\mathbb{E}_g(x)$  attains a global maximum in  $x = 0$ .

In a similar fashion, it is possible to construct such a function with any degree of smoothness.

This counterexample seems to indicate that deriving "global" results based on the behavior near local minima has its problems. Indeed, in the analysis of the OLGA (see Kim and Sompolinsky (1998), pg. 2339) the process was analyzed using the Fokker-Planck equation. Neglecting terms which are  $O(\frac{1}{\lambda^2})$ , the

Fokker-Planck equation for the transition operators is

$$\frac{\partial}{\partial \tau} P(x, \tau) = \nabla \cdot [\nabla \mathbb{E}_g(x) P(x, \tau)] + \frac{1}{\lambda} \sum_{i,j} \partial_i \partial_j [T_{ij} P(x, \tau)],$$

where  $\tau$  is the time variable and  $T_{ij}$  is the diffusion matrix. Next, the diffusion term (which is  $O(\frac{1}{\lambda})$ ) is neglected – under the assumption that  $\nabla \mathbb{E}_g$  is not small. However, this does not guarantee a descent towards a local minimum. This would be the case if one assumes that there is a unique critical point in the state space, which is a local minimum, i.e., a strong “locality” assumption. As the example above shows, if there are other critical points the orbits of the OLG may converge to those points and not to the local minimum.

Roughly speaking, the reason for the behaviour demonstrated in Example 5.2 is that the OLG does not take into account the “size” of the error made by  $x$  on the input  $v$ . The transition is made based on the fact that the new state is correct on that input. Hence, the orbits of the process are driven towards a state  $x_0$  which was the largest set of correct answers. Now, if the error function is 0-1, this implies that  $\mathbb{E}_g$  attains a minimum in  $x_0$  and that the OLG will eventually converge to  $x_0$ . On the other hand, if the on-line error function is not 0-1,  $x_0$  makes “very large” mistakes on inputs in which it gives the wrong answer. Thus, although  $x_0$  makes few errors, the average error  $E_g(x_0)$  may be made arbitrarily large.

It is important to note that the on-line error function in Example 5.2 does not satisfy Assumption 2.2, thus Theorem 2.3 may not be applied to this case.

## 6 Concluding remarks

The main stumbling block regarding the process 2.1 is its local nature, which is due to the assumption that  $\nu(C_x)$  is bounded away from 0. We were not able to remove this restrictive assumption, but it may be possible to do so using a more sophisticated probabilistic argument.

Let us point out that the reason for the assumption  $\lambda_n \rightarrow \infty$  in Kim and Sompolinsky, (1998) was to overcome the possibility that the teacher makes a mistake. We did not treat this problem outside the one example presented in section 3 and it deserves additional consideration.

Note that an easy way to generalize part (a) of Theorem 2.2 is to formulate a stopping procedure which freezes the process once a state is close enough to a correct answer. One possibility is to count the number of consecutive correct responses at each state and stop the process once the number passes a given threshold. This gives an estimate on the measure  $\nu(C_x)$ . However, if the error function is not 0-1, the fact that  $\nu(C_x)$  is close to its global maximum does not imply that  $\mathbb{E}_g(x)$  is close to the global minimum (this is the idea behind Example 5.2). For that, one needs additional assumptions on the structure of the error function.

Our final remark concerns Theorem 3.2. There are several other probabilistic

schemes which may solve the relevant optimization problem, at least when the function  $f$  has a unique maximum. Particularly well known is the so-called *Stochastic Approximation* scheme (see Kushner and Yin, (1997)). We do not claim that the process 3.2 yields the best solution to the optimization problem. Our goal was to present one complete example in which the 0 temperature process with a bounded change converges to the minimum of  $\mathbb{E}_g$ , even in the presence of noise.

We were not able to formulate a more general theorem than the one presented here. Even when the error function is 0-1, the function  $\nu(C_x)$  does not determine the transition density from state to state. All we know is that when  $\nu(C_x) > \nu(C_y)$  there is a positive transition density from  $y$  to  $x$ . Unfortunately, it is possible to construct natural examples for which  $\nu(C_y) > \nu(C_x)$ , but still there is a positive transition density from  $y$  to  $x$  and it is possible that the analogous convergence theorem is not true.

## References

- Cheney, E.W. (1966)** Introduction to approximation theory, McGraw-Hill
- Haykin, S. (1994)** Neural Networks: A comprehensive foundation, MacMillan College Press
- Kim, J.W. Sompolinsky, H. (1996)** On-line Gibbs learning, Physical Review Letters, Vol 76, 16, 3021-3024
- Kim, J.W. Sompolinsky, H. (1998)** On-line Gibbs learning I, Physical Review E, Vol 58, 2, 2335-2347
- Kushner, H.J. Yin, G.G. (1997)** Stochastic Approximation Algorithms and Applications, Springer
- Loève, M. (1963)** Probability Theory, 3rd edition, D. Van Nostran
- Leshno, M. Lin, V.Y., Pinkus, A. Schocken, S. (1993)** Multilayer feed-forward networks..., Neural Networks, 6, 861-867
- Minsky, M. Papert, S. (1972)** Perceptrons, MIT Press
- Orey, S. (1971)** Limit theorems for Markov chain transition probabilities, Van Nostran Reinhold
- Ritter, H. Martinetz, T. Schulten, K. (1992)** Neural computation and self organizing maps, Addison-Wesley