

Recurrence Methods in the Analysis of Learning Processes

S. Mendelson and I. Nelken

Shahar Mendelson, Dept. of Mathematics, Technion, and Institute of Computer Science, Hebrew University. E-mail: shahar@tx.technion.ac.il.

Israel Nelken (corresponding author), Dept. of Physiology, Hebrew University - Hadassah Medical School and the Interdisciplinary Center for Neural Computation, Hebrew University, Jerusalem. Email: israel@music.md.huji.ac.il. Telephone: (972)-2-6758381, Fax: (972)-2-6439736. Mailing Address: Dept. of Physiology, The Hebrew University - Hadassah Medical School, P.O.Box 12272, Jerusalem 91120, ISRAEL.

Abstract

The goal of most learning processes is to bring a machine into a set of “correct” states. In practice, however, it may be difficult to show that the process enters this target set. We present a condition which ensures that the process visits the target set infinitely often almost surely. This condition is easy to verify and is true for many well known learning rules. To demonstrate the utility of this method, we apply it to four types of learning processes: the Perceptron, learning rules governed by continuous energy functions, the Kohonen rule, and the Committee Machine.

Keywords: learning processes, recurrence, Perceptron, Kohonen algorithm, committee machine

1 Introduction

A central issue in the study of learning processes is that of their long term behavior; namely, if the process converges, what are its limit points, and if not, which sets are visited frequently by the orbits of the process (which are the sequences of states generated by the application of the learning rule). The purpose of this paper is to introduce a method which can be applied to a wide range of well known learning rules, and may reduce the difficulties in answering this question considerably.

Given a learning rule, one may define a “target set” into which the orbits of the learning process should enter often. For example, consider the set of correct states in a supervised learning process. In the supervised learning setup, the student is adapted until it agrees with the teacher on every possible example, which means that it enters the set of correct states. In this case, the target set is absorbing: once an orbit enters the set of correct states, it will never leave it. However, in the case of many unsupervised learning processes, there is no guarantee that the target set is absorbing.

A property of the target set, which is weaker than being absorbing, is recurrence. The target set is recurrent if the orbits enter it infinitely often with probability 1. Recurrence of the target set is also weaker than ergodicity: when a process is ergodic, every set of states with positive measure is recurrent. In many learning processes, we would like to eventually limit the set of states which are visited by the orbits of the process, and thus we would like only specific subsets of the state space (those containing the target set) to be recurrent.

Although proving that the target set is recurrent does not ensure con-

vergence of the process to this set, such a proof is useful. Recurrence proof can serve as part of a convergence proof – if the target set is not recurrent, convergence cannot be achieved. If the target set is recurrent, the other piece of information which is required in order to prove convergence is that once the process enters this target set, it will not leave it again easily (in the sense that escape probabilities from the target set are small enough). Therefore, the main result of this paper is a method which allows us to determine, under very general conditions, whether a target set is recurrent or not.

The method consists of two parts. First, we need to prove that the learning process satisfies a (rather weak) continuity assumption. We show that if this assumption holds, and if from any initial condition there is a positive probability of entering the target set, the orbits will enter the target set infinitely often with probability 1. However, there may be initial conditions with a 0 probability to enter the target set. The identification of this set of “undesirable” initial conditions, which we denote by A , is the second part of our method. We demonstrate that with probability 1, the orbits either converge to the set A or enter the target set infinitely often. In particular, if A is empty, the target set is recurrent.

This approach does not deal directly with the issue of how long it takes to reach a state in the target set starting from a generic initial state. We do not obtain quantitative results, nor do we attempt to estimate the time it takes to reach the target set. Rather, we focus on a “soft” approach, which makes it easy to check whether the target set is recurrent.

We present four examples of learning rules to which our method may be applied. The first is the Perceptron learning rule (Haykin, 1994). We prove

that even under conditions which are less restrictive than those required by the Perceptron convergence theorem, the orbits of the process come arbitrarily close to the teacher with probability 1, although convergence does not necessarily occur. The second example is the Kohonen learning rule (Kohonen, 1989; Ritter et al., 1992). Again, we show that the continuity assumption holds. We define a target set for a small 2-dimensional Kohonen process, and demonstrate that with probability 1, the orbits of the process must enter the target set infinitely often. Then, we investigate the 2-dimensional Kohonen process with respect to a lattice order. We introduce an algorithm which should allow us to construct a sequence of inputs which takes a generic initial state to the target set; i.e., to a set which has a lattice order. Thus, with probability 1, the orbits of the process become ordered. The third example is a general convergence proof for processes for which it is possible to build a smooth energy function, using the theory presented here. Finally, we examine a version of a “bounded change” learning algorithm suggested for the Committee Machine (Kim and Sompolinsky, 1996). We demonstrate that although the continuity assumption is fulfilled, if the bound on the step size is too tight, there is a non-trivial set of initial conditions from which it is impossible to enter the set of correct states.

2 General Theory

As stated in the Introduction, in many learning processes there is some set of states to which one hopes that the orbits of the learning process converge. The minimal requirement is that this set of states is recurrent, in the sense that the orbits of process enter it infinitely often almost surely. Our goal is

to formulate an easy to verify condition which ensures that the target set is indeed recurrent.

We start with some notation. For a set A , denote by χ_A the characteristic function of A , i.e., a function which is 1 on A and vanishes outside A . Put $d(x, A)$ the distance between x and the set A . Finally, $B_x(r)$ is the open ball with radius r centered in x .

We view the learning process T as a dynamical process on some compact state space X . During each learning step, the system is exposed to an input v selected from some set V and it adapts by the learning rule $x \rightarrow T(x, v)$. The inputs are selected by an i.i.d law on V which is given by a Borel measure ν . Thus, our process is a time homogeneous Markov process (X_n, V_n) defined by the transition density

$$P\left((X_1, V_1) = (a, b) | (X_0, V_0)\right) = P\left(T(X_0, V_0) = a\right)P\left(V_1 = b\right).$$

Hence, (V_n) are selected independently while (X_n) are adapted according to T . Equip X with a probability measure μ , let P_0 be the measure induced on the orbits (X_n) and set P_x to be the induced measure given that $X_0 = x$. For every k -tuple $(v_1, \dots, v_k) \in V^k$, set $T^k(x; v_1, \dots, v_k)$ to be the state of the initial condition x after it was exposed to the examples v_1, \dots, v_k .

The basis of our discussion is the following theorem:

Theorem 2.1. *Let (X_n) be a homogeneous Markov process and set $A, O \subset X$. Assume that $\inf_{x \in A} P_x(X_n \in O \text{ for some } n) > 0$. Then*

$$\{X_n \in A \text{ infinitely often}\} \subset \{X_n \in O \text{ infinitely often}\} \quad P_0\text{-almost surely}$$

Roughly speaking, the assertion of the theorem is that if the transition probability from every $x \in A$ to some set O is strictly positive (that is, larger

than some positive constant), an orbit which visits A infinitely often will also visit O infinitely often. The proof of Theorem 2.1 is deferred to the appendix.

This theorem is useful for proving that a set is recurrent, but given some $O \subset X$, it is usually difficult to prove the existence of a positive lower bound on $P_x(X_n \in O \text{ for some } n)$. We bypass this difficulty by showing that the function $h_O(x) = P_x(X_n \in O \text{ for some } n)$ is *lower semi continuous* (defined below) under mild assumptions on the process T and on the set O .

Definition 2.2. *A real valued function f is lower semi continuous if for every $x_n \rightarrow x$, $\liminf_{n \rightarrow \infty} f(x_n) \geq f(x)$.*

An important property of lower semi continuous functions which is stated in the following lemma, is that like continuous functions, they attain their minimum on compact sets.

Lemma 2.3. *Let f be a real valued function which is lower semi continuous. Then on every compact set, f attains its minimum.*

The proof of the lemma is easy and we shall omit it.

From Lemma 2.3 it follows that in order to prove that a set O is recurrent, it is sufficient to demonstrate that h_O is lower semi continuous, and that for every $x \in X$, $h_O(x) > 0$. Indeed, since X is compact and h_O is lower semi continuous, it attains its minimum in X , implying that this minimum is positive. Hence

$$\inf_{x \in X} h_O(x) = \inf_{x \in X} P_x(X_n \in O \text{ for some } n) > 0.$$

By Theorem 2.1, $\{X_n \in X \text{ infinitely often}\} \subset \{X_n \in O \text{ infinitely often}\}$ almost surely and our claim follows.

Unfortunately, may one sometimes encounter a process in which there is a set A on which $h_O(x) = 0$, and $X \setminus A$ is not compact. Thus, the idea we used above does not apply directly to this case. However, only a small modification is required to prove the following:

Theorem 2.4. *Assume that $h_O(x)$ is lower semi continuous and that $A \subset X$ is a set such that for every $x \in X \setminus A$, $h_O(x) > 0$. Then*

$$P_0\left(\{d(X_n, A) \rightarrow 0\} \cup \{X_n \in O \text{ infinitely often}\}\right) = 1.$$

Sketch of Proof: Fix some integer m and set $R_m = \{x | d(x, A) \geq 1/m\}$. If an orbit (X_n) does not converge to A , then it belongs to some R_m infinitely often. Since R_m is compact and h_O is positive on this set, then there is some $\delta > 0$ such that $h_O(x) > \delta$ on R_m , which implies that the orbit enters O infinitely often. ■

Next, we shall formulate a sufficient condition on the dynamical process T which ensures that for every open set O , h_O is lower semi continuous.

Assumption 1. *For every $x \in X$ and every sequence (y_n) which converges to x , $T(y_n, -) \rightarrow T(x, -)$ almost surely, i.e., there is some set $\Omega \subset V$ (which may depend both on x and on (y_n)) with $\nu(\Omega) = 1$ such that for every $\omega \in \Omega$, $T(y_n, \omega) \rightarrow T(x, \omega)$.*

Note that this assumption is considerably weaker than a condition of ν -almost everywhere continuity in x since the set Ω may depend on the specific sequence (y_n) .

This weak condition suffices to ensure the stability of the stochastic process in the sense that for every integer k there is a neighborhood of x and

a “large” set of inputs such that for every y in the neighborhood and every such k -tuple of inputs v_1, \dots, v_k , $T^k(y; v_1, \dots, v_k)$ is close to $T^k(x; v_1, \dots, v_k)$. This stability condition is at the heart of the proof of the next Theorem:

Theorem 2.5. *Assume that Assumption 1 holds and that O is an open subset of X . Then $h_O(x) = P_x\{X_n \in O \text{ infinitely often}\}$ is lower semi continuous.*

The proof of this theorem is deferred to the appendix. However, the proof presented there uses a less direct approach. Therefore, we present a sketch of a direct proof of Theorem 2.5 in which we assume that T is continuous.

Sketch of Proof: Let $B_x^n = \{v_1, \dots, v_n \mid T^n(x; v_1, \dots, v_n) \in O\}$ and set C_x^n to be a compact subset of B_x^n whose measure is “almost” the measure of B_x^n . Since T^n is continuous, then $T^n(x; C_x^n)$ is a compact subset of O , implying that there is a positive distance between this set and $X \setminus O$. Therefore, there is a neighborhood U of x such that for every $y \in U$ and every $(\omega_1, \dots, \omega_n) \in C_x^n$, $T^n(y; \omega_1, \dots, \omega_n)$ is “close” to $T^n(x; C_x^n)$, hence $T^n(y; \omega_1, \dots, \omega_n)$ belongs to O .

By this argument, it follows that for every $\varepsilon > 0$ there is a neighborhood U_x such that for every $y \in U_x$,

$$\nu^n\left(\bigcup_{i=1}^n B_y^i\right) \geq \nu^n\left(\bigcup_{i=1}^n B_x^i\right) - \varepsilon.$$

Hence, for every $y \in U_x$,

$$P_y\{X_n \in O \text{ for some } i \leq n\} \geq P_x\{X_n \in O \text{ for some } i \leq n\} - \varepsilon. \quad (1)$$

Define a stopping time τ to be the time in which the process first enters O , and set $f_n(x) = P_x\{\tau \leq n\}$. Thus, by (1), f_n is an increasing sequence of lower semi continuous functions which converges pointwise to h_O . Therefore, h_O is lower semi continuous.

■

Corollary 2.6. *If Assumption 1 holds, O is open and A is a set such that for every $x \notin A$, $h_O(x) > 0$ then*

$$P_0\left(\{d(X_n, A) \rightarrow 0\} \cup \{X_n \in O \text{ infinitely often}\}\right) = 1. \quad (2)$$

3 Examples

In this section, we give some examples of learning processes to which our method can be applied. We chose on purpose well-known learning processes, in order to compare the theory presented here with previous treatments of the same processes. Each example has two parts: first, it is required to show that Assumption 1 holds. Second, given a target open set O , it is necessary to study the function $h_O(x)$, giving the probability of entering O starting at any state x . The second step is the more complicated one, requiring usually ad-hoc arguments adapted for each case individually.

3.1 The Perceptron

The Perceptron is a classical supervised learning model. Here, both the teacher T and the student X are positive sides of given maximal subspaces.

Let

$$\text{sign}(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}, \quad (3)$$

then for any Perceptron X $\chi_X(v) = \text{sign}(\langle x, v \rangle)$. Thus, $\chi_X(v)$ is the decision function of the Perceptron X . Let T be the teacher Perceptron and put X to

be the Perceptron we are training. Note that X and T agree on an example v if and only if $\chi_T(v) = \chi_X(v)$. Hence, \mathbb{R}^d is divided into two sets: on the first one χ_T and χ_X agree and on the other set χ_T and χ_X disagree (Fig. 1).

We can identify each maximal subspace X with its normal unit vector x . With this identification, $X = \{y | \langle x, y \rangle \geq 0\}$ and the boundary is $\partial X = \{y | \langle x, y \rangle = 0\}$. ∂X is called the decision boundary of X . Furthermore, as far as the Perceptron is concerned, an example y is classified in the same way as any example cy for $c > 0$. Thus, we can assume that the input space is the set of unit vectors, which is the d -dimensional unit sphere, $S^{d-1} = \{x | \|x\| = 1\}$.

Next, we have to select the probability distribution ν with which the examples for the training process are chosen. Assume that the measure is “smooth” in the sense that it does not assign positive probability to sets whose area is 0. In particular, it assigns zero probability to any specific input. Formally, we require that ν is absolutely continuous with respect to the Haar measure, which is the natural extension to the sphere of the notion of arc length on the circle.

Given a student X and a teacher T , set the function $f_{T,X}(v) : S^{d-1} \rightarrow \mathbb{R}$ by $f_{T,X} = \chi_T - \chi_X$. Thus, $f_{T,X}(v) = 0$ if and only if T and X agree on the input v . A Perceptron X is correct if it classifies essentially all inputs correctly; that is, if the set on which X and T disagree has probability 0. Hence, we say that the Perceptron X is correct if $\nu(\{v \in V | f_{T,X}(v) \neq 0\}) = 0$.

The standard Perceptron training algorithm can now be formulated in the following way: define a function T_p and a dynamical process $x_{n+1} = T_p(x_n, v_n)$ which takes the current Perceptron X_n represented by the unit vector x_n , an

input example v_n drawn independently by the probability distribution ν , which produces as an output a new Perceptron x_{n+1} . The function T_p is computed in two steps: first, the direction of the new Perceptron is chosen by the standard Perceptron update rule, and then it is normalized to a unit length:

$$\tilde{x}_{n+1} = x_n + \varepsilon f_{T, X_n}(v_n)v_n, \quad x_{n+1} = T_p(x_n, v_n) = \frac{\tilde{x}_{n+1}}{\|\tilde{x}_{n+1}\|}. \quad (4)$$

The standard Perceptron convergence theorem states that for a finite training set and for any sequence of examples taken from this set such that all the members of the training set appear an infinite number of times, the algorithm converges to a Perceptron which agrees with the teacher on the training set. A crucial part of the proof is that the smallest angle between a training example and the teacher is always greater than zero. This theorem can be generalized to an infinite set of training examples and a more general distribution on that set, but then the assumption that the angle between the training examples and the teacher is bounded from below by a positive number must be made explicitly. Here, we extend the Perceptron convergence theorem even further. Assume that the measure ν by which the examples are selected is absolutely continuous with respect to the Haar measure. We show that with probability 1, running the Perceptron training algorithm brings us arbitrarily close to the teacher infinitely often. The price paid for the weaker assumption is that although the orbits approach the teacher infinitely often, they may not converge and the student may oscillate around the position of the teacher.

To prove recurrence, first we have to show that the function T_p satisfies Assumption 1. Indeed, for every $x \in S^{d-1}$, the set on which $T_p(x, -)$ is not

continuous is $(\partial X \cup \partial T) \cap S^{d-1}$ and consists of the points of discontinuity of $f_{X,T}$. Fix $x \in S^{d-1}$, a sequence $y_n \rightarrow x$ and set $\Omega = V \setminus \partial X \cup \partial T \bigcup_{n=1}^{\infty} \partial Y_n$. Then, $\nu(\Omega) = 1$ and for every $\omega \in \Omega$, $T_p(y_n, \omega) \rightarrow T_p(x, \omega)$. Therefore, the conditions of Assumption 1 hold.

Next, we have to define the candidate recurrent set O and to study the function $h_O(x)$. We select $B_t(r)$, the open ball with radius r around the teacher, as the set O of Theorem 2.6. Thus, We show that with probability 1, the orbits enter $B_t(r)$ infinitely often.

Corollary 3.1. *Denote by (x_n) the orbits of the Perceptron Learning rule T_p . Then for every $r > 0$, $P_0 \left\{ x_n \in B_t(r) \text{ i.o.} \right\} = 1$.*

Proof: The idea behind the proof is to use the classical Perceptron convergence theorem to show that $h_O(x) > 0$ for $O = B_t(r)$. Let $V' = S^{d-1} \setminus \{d(T, x) \leq r\}$. Note that the set of correct states for the process (4) with V' as its input set and T as the teaching halfspace is $B_t(r)$, and that $d(\partial T, V') > 0$. Therefore, by the standard Perceptron convergence theorem, the orbits of the process (4) using V' and T as the input set and the teacher respectively, enter $B_t(r)$ in a finite number of steps almost surely. This implies that if $V = S^{d-1}$ then for every $x \in S^{d-1}$, $h_{B_t(r)}(x) > 0$. By the discussion above $h_{B_t(r)}$ is lower semi continuous, therefore, by Theorem 2.4 applied to $A = \phi$ it follows that $P_0 \left\{ x_n \in B_t(r) \text{ i.o.} \right\} = 1$ as claimed. ■

We wish to emphasize that the fact that the orbits are arbitrarily close to the teacher infinitely often does not imply that they converge to the teacher.

Remark 1. *Let $T = \{x \in \mathbb{R}^d | x_d \leq 0\}$ and assume that ν is the Haar measure restricted to $S^{d-1} \cap T$. Then, by corollary 3.1, (x_n) enters $B_r(t)$*

infinitely often almost surely. However, it can be shown (Mendelson, 1998) that $P_0\{x_n \rightarrow t\} = 0$. In other words, although X_n are arbitrarily close to T almost surely, the process (4) does not converge to the teacher with P_0 -probability 1.

3.2 The multi-dimensional Kohonen process

The Kohonen adaptive process is defined using two sets: one is the (finite) set we wish to adapt and the other one is the set of possible inputs. We assume that both these sets are contained in some finite dimensional normed space, and that the set of inputs is compact (that is, it is closed and bounded).

Again, denote the input set by V and the adapting set in the n -stage by x_n . Note that x_n is a set, consisting of the current configuration of the adapted set. Given $v \in V$, let $P_{x_n}(v)$ be the set of the nearest points to v in $x_n = \{x_n^1, \dots, x_n^m\}$ and put $i(y)$ the index of y in the set x . Our learning process T_k generates the next state x_{n+1} by moving some of the points in x_n in the direction of the current input. The points to be moved are determined by a function f which depends on the index of the nearest point to v , $i(P_{x_n}(v))$, and on the index of the point being considered, j .

$$x_{n+1}^j = [T_k(x_n, v_n)]^j = x_n^j + \varepsilon f(i(P_{x_n}(v)), j)(v_n - x_n^j), \quad (5)$$

where $x_n^j \in x_n$, $\varepsilon > 0$ and $f : \mathbb{N} \times \mathbb{N} \rightarrow [0, 1]$.

Thus, the function f implicitly defines a neighborhood relationships on the indices. Pairs for which $f(i, j)$ is large are “close” whereas pairs of points for which $f(i, j)$ is small or zero are “far”. At each step of the process, points which are close to $P_{x_n}(v)$ in this sense take a large step in the direction of the current input v , and in particular approach each other. Points which

are farther away take a relatively smaller step or do not move at all. The goal of the process is to cause points which are close in terms of indices (as determined by the function f) to be close in the sense of the metric distance. Each step of the Kohonen process improves this correspondence to some extent, but only for the nearest point $i(P_{x_n}(v))$ and its neighbors – because they move closer to each other. In so doing, however, the same step may worsen the situation for other points.

The main experimental observation about the Kohonen process is that in spite of the local character of each learning step, it is possible to generate global order. However, a proof of this fact is available only in the one dimensional case with the neighborhood function

$$f(i, j) = \begin{cases} 1 & |i - j| \leq 1 \\ 0 & \text{otherwise} \end{cases} .$$

Usually, it is assumed that the step size ε decreases to 0 in some way, but not too fast. This ensures that at the later stages of the process, the modification of the state at each moment becomes very small. However, if ε depends on the stage of the learning process (that is, ε is time dependent), the theory presented in this paper is not applicable. Thus, we analyze a “homogeneous” Kohonen process in which the step size ε is constant. One way to justify the analysis of this case is that in practical implementations of the Kohonen process, it is very often necessary to keep ε relatively large for a long time in order to reach approximate order, and only then ε is decreased in order to “freeze” this order. The analysis presented here is then relevant to the first stage of such implementations, and as we demonstrate below, the homogeneous case already ensures the emergence of order in the multi-dimensional Kohonen process, although only transiently (in the sense

that the set of ordered states is not absorbing). The homogenous Kohonen process was also analyzed by others (e.g. Fort and Pages, 1995).

To apply the theory presented here it is first necessary to check that the process has the required continuity property. Clearly, for every $x = \{x^1, \dots, x^m\}$ the set in which $T_k(x, -)$ is not continuous is a subset of

$$\bigcup_{i \neq j} \{d(x^i, v) = d(x^j, v)\}.$$

Thus, by the same method as in 3.1, one can show that the conditions of Assumption 1 hold, given that the measure ν assigns 0 probability to hyperplanes.

In the multi-dimensional case, there is no known natural definition of order for which the set of ordered states is absorbing (see also Fort and Pages, 1996), nor do we supply one here. Instead, we use the theory developed here to find recurrent sets of states that capture some intuitive features of order. Thus, although we cannot demonstrate that the Kohonen process converges to an ordered state, we do demonstrate that order emerges as a result of the application of the Kohonen process even in the multi-dimensional case in the sense that the process visits such ordered states infinitely often.

To illustrate the essential features of this application in a toy example, consider the two dimensional setup, where the states of the Kohonen process are composed of five points denoted by $\{1 \ 2 \ 3 \ 4 \ 5\}$ with the natural neighborhood relationships as presented in Figure 2A. The relative weights for moving the points are chosen so that each time an input v is presented, the point x closest to v moves the largest distance. The neighbors of x move in the direction of v a distance which is proportional to their interaction with x – determined by the function f . For example, in the notation

above, we take $f(3,3)=1$, $f(3,j)=0.5$ for other values of j , $f(1,1)=1$, $f(1,3)=0.5$, $f(1,2)=f(1,5)=0.25$ and $f(1,4)=0$.

To emphasize the fact that it is the function f which implicitly defines the order, we can define an ordered state as a state in which the distance between the points is compatible with the neighborhood structure as imposed by the function f . Thus, a state is ordered if the distance between a point x and its neighbors (as imposed by f) is smaller than the distance between x and points which are not its neighbors. In addition, its distance to a closer neighbor (with a larger value of f) should be smaller than the distance to a farther neighbor (with a smaller value of f). The state in Fig. 2B is ordered according to this definition. Denote the set of all ordered states according to this definition by O .

Some of the problems in the definition of order in the multi-dimensional Kohonen process are apparent even in this toy problem. First, the definition of order is not unique: for example, the state in Fig. 2C is not in O , although it still captures some aspects of the ordering of the five points. Furthermore, the set O is not absorbing: even if the process enters O , it is easy to see that there is always a positive probability of leaving it in a finite number of steps. We tested a hierarchy of such definitions of order, all of which suffered from the same two problems. We chose to work with the set O as defined above, since the constraints it imposes are less stringent than the requirement of keeping an exact square arrangement of the points. On the other hand, it still imposes some metric conditions on the target set. Such metric conditions are absent from a more “topological” definitions of order, which would have included the state in Fig. 2C.

To demonstrate that the set O is visited infinitely often, we present an algorithm which takes any non-degenerate initial state containing five points to a state in O by generating a finite, deterministic sequence of input examples, none of which is on a discontinuity line of the process. This algorithm is described in the appendix. Once such a deterministic sequence is generated, we can (using continuity) replace each point in this sequence by a small neighborhood, and any sequence of points selected from the neighborhoods still takes the initial condition into O . If the measure on the set of inputs is equivalent to the Lebesgue measure, the probability of this set of sequences is greater than 0. This argument proves that $h_O(x) > 0$ for every non-degenerate state. Applying Corollary 2.6 we conclude that with probability 1, an orbit either visits O infinitely often or it converges to a degenerate state.

As mentioned above, O is not absorbing, so the process also leaves O with probability 1. In fact, there are open sets in the complement of O which have the same property, i.e., the orbits enter them with positive probability from any initial condition. Thus, the process does not converge to the set O . To illustrate this, define a function $e(x_n)$ on the set of states, which measures how many points are locally unordered, in the sense that their distances from the other points are not compatible with the neighborhood structure imposed by f . Thus, $x \in O$ if and only if $e(x) = 0$. Fig. 2D shows $e(x_n)$ for a typical run of the Kohonen process. Initially, e fluctuates between 2 and 4, but it decreases to smaller values rather rapidly. Eventually it reaches 0, meaning that the orbit entered O , but from time to time it increases again, showing the process left O . This is the typical behavior described by Corollary 2.6:

although the orbits may leave the set of ordered states, they will eventually enter this set again.

Next, we turn to the example of the two-dimensional lattice. Here, each state consists of $n \times n$ points in $[0, n]$. Each point is indexed by a pair (i, j) , such that $0 \leq i, j \leq n$. The neighbors of a point (i, j) are all the points (i', j') such that $|i - i'| + |j - j'| \leq 1$. Define $f((i, j), (i', j')) = 0.5$ if the two points are neighbors, and 0 otherwise. A state $(x_{i,j})$ is said to be ordered if the points $(1, j), \dots, (n, j)$ are linearly ordered along the x axis for each j (either increasing for all j or decreasing for all j), while the points $(j, 1), \dots, (j, n)$ are linearly ordered along the y axis for every j (again, either increasing for all j or decreasing for all j), or if the reverse situation occurs, that is, points $(1, j) \dots (n, j)$ are linearly ordered along the y axis whereas points $(j, 1) \dots (j, n)$ are linearly ordered along the x axis.

We would like to show that this set is recurrent, which implies that theoretically, the Kohonen process can be used to order the 2-dimensional lattice. In practice, the waiting time may be very long. In fact, in simulations, the process often seems to be trapped in a set of states in which there are several domains, each of which is ordered, but the domains are incompatible. Fig. 3A shows an initial condition in which each of the upper and lower halves of a 20×20 is ordered, but the two halves are reversed with respect to each other. After 800,000 iterations, the initial mismatch between the upper and the lower part remains (Fig. 3B).

In spite of this negative experimental result, we conjecture that with probability 1 the mismatch will eventually be resolved. As usual, the difficulty lies in the attempt to show that the set of ordered states can be reached

with positive probability from any non-degenerate initial state. We present an algorithm which should ensure this.

Begin the process by finding the quarter of the square in which the point $(1, 1)$ lies. Then, pull all other points to a small square near the opposite corner. This can be done (with positive probability), because it requires choosing examples in the small square, making sure that the points $(1, 1)$, $(1, 2)$, $(2, 1)$ and $(2, 2)$ are never the closest points to the chosen example. Within a finite number of steps, all points (except $(1, 1)$) are pulled to that corner.

Next, choose points which are in the quarter square in which $(1, 1)$ lies. At first, point $(1, 1)$ is the closest to any example we choose, hence, it pulls its nearest neighbors. As its neighbors are pulled towards $(1, 1)$, they enter the area from which examples are chosen. Therefore they become the nearest point to some of the examples, and they pull their own neighbors in. Since the points are pulled into the quarter square in the right order, we conjecture that they will always end up in the correct lattice order.

Simulations show that the suggested algorithm is useful for ordering non trivial states. Unfortunately, we were not able to formally prove this conjecture. If the conjecture is true, then $h_{\mathcal{O}}(x) > 0$ for any non-degenerate state and it follows from 2.6 that with probability 1, topologically ordered states will be visited infinitely often.

In Fig. 3C we show the state of the orbit the first time we detected entry into the set of ordered states. All the points are concentrated in one corner, but they are ordered according to our definition. Fig. 3D shows a plot of the number of disordered rows and columns during the operation of the

algorithm. Since the initial state of the algorithm is almost ordered, at first the amount of disorder actually increases. All points except $(1, 1)$ arrived to their corner after 29510 steps, marked by the left arrow in the plot. Next, the points are pulled to the opposite corner. At first, this increases the disorder, but eventually an ordered state is reached, as indicated by the right arrow in Fig. 3D. The simulation was continued afterwards, allowing examples to be chosen from the full 20×20 square without limitations. Since the points are already in a nearly correct order, ordered states are often hit.

Note that the number of steps required for the sequence in the simulation to take the initial condition to O is very large (approximately 10^5 steps). This indicates why the Kohonen process was not able to order the initial state in figure 3 in 800,000 steps. If the simulation is indeed a “typical” example, then the probability to enter O in approximately 10^5 steps is $(1/9)^{(10^5)}$ (since the area of the small squares we used in the application of the algorithm had $1/9$ the area of the full square). Thus the expected value of the time needed to enter O is approximately $10^5 \times 9^{(10^5)}$ which is huge compared with 800,000. Although this estimate is an upper bound on the time to enter O , it may be safe to say that it will be greater than 800,000. Thus, although the Kohonen rule may be used to order a lattice, it may not be very useful for this purpose.

3.3 Processes with energy functionals

3.3.1 General considerations

In the context of dynamical systems, a Lyapunov functional is an “energy function” which decreases along the orbits of the process. For the non-

deterministic systems studied here, the requirement that for every given input and every state the energy of $T(x, v)$ will be smaller than the energy of x may be too restrictive. A less strict requirement is that instead of decreasing along the orbit, the energy may increase from time to time, but that averaged on all examples, the energy decreases. Even this definition is too strict. Roughly speaking, near minima of the energy function, the stochastic noise is expected to increase, rather than decrease, the energy on average. Therefore, we first treat a general case in which the functional is allowed to both increase and decrease, and later we show that under additional assumptions on the structure of the process, neighborhoods of the critical points are recurrent.

Let F be a continuous functional for T on the space X . Set $G(x) = \mathbb{E}_\nu F(T(x, v)) - F(x)$. Since X is compact and F is continuous, then F and G must be bounded. For every $\delta > 0$ let $O_\delta = \{x | G(x) > -\delta\}$; O_δ is the set of states on which the average “energy” either increases, or decreases by less than δ .

Theorem 3.2. *Let T be a process which has a continuous functional and satisfies Assumption 1. Then, for every $\delta > 0$,*

$$P_0 \left(X_n \in O_\delta \text{ infinitely often} \right) = 1.$$

Proof: Intuitively, since the energy function is bounded, it cannot decrease by more than δ an infinite number of times. Formally, we shall show that the conditions of Corollary 2.6 hold. Using the notation of Corollary 2.6 let A be an empty set. To use the Corollary, we have to show that O_δ is open and that for every $x \in X$, $h_{O_\delta}(x) > 0$.

First, note that G is a continuous function. Indeed, let $(x_n) \subset X$ be a sequence which converges to x . By Assumption 1, there is a set $\Omega \subset V$, such

that $\nu(\Omega) = 1$ and for every $v \in \Omega$, $T(x_n, v)$ converges to $T(x, v)$. Since F is continuous then $F(T(x_n, v)) - F(x_n)$ tends almost surely to $F(T(x, v)) - F(x)$. Because F is bounded, then by the bounded convergence theorem, $G(x_n) \rightarrow G(x)$. Thus, O_δ is open as an inverse image of an open set by a continuous function.

Next, fix some $x \in X$ and recall that

$$h_{O_\delta}(x) = P_x\left(X_n \in O_\delta \text{ for some } n\right).$$

If $h_{O_\delta}(x) = 0$, then for every integer n the conditional expectation

$$\mathbb{E}\left(F(X_{n+1}) - F(X_n) | X_n\right) \leq -\delta.$$

Therefore, for every integer n ,

$$\begin{aligned} \mathbb{E}\left(F(X_{n+1}) - F(X_1)\right) &= \sum_{i=1}^n \mathbb{E}\left(F(X_{i+1}) - \mathbb{E}F(X_i)\right) = \\ &= \sum_{i=1}^n \mathbb{E}\left(\mathbb{E}\left(F(X_{i+1}) - F(X_i) | X_i\right)\right) \leq -n\delta, \end{aligned}$$

which is impossible because F is bounded on X . ■

In the special case when there is a continuous and bounded energy function F which decreases along orbits, the sets O_δ are neighborhoods of the minima of F , and Theorem 3.2 demonstrates that the process visits such neighborhoods infinitely often. However, as mentioned above, there are situations in which a natural definition for energy function exists, but the energy does not necessarily decrease along orbits. Even in these cases, it will be shown now that with some additional assumptions, neighborhoods of the critical points of the energy function are recurrent.

3.3.2 Stochastic Gradient

To apply these results in a more concrete setup, let $H : X \times V \rightarrow \mathbb{R}$ be a twice differentiable function. Set

$$T(x, v) = x - \varepsilon D_x H(x, v), \quad (6)$$

where $D_x H(x, v)$ is the derivative of H with respect to the first variable and ε is some positive parameter. If $F(x) = \mathbb{E}_\nu H(x, v)$ then one can show that for every $x \in X$, the derivative of F at x is $DF_x = \mathbb{E}_\nu D_x H(x, v)$. Put $G(x) = \mathbb{E}_\nu F(T(x, v)) - F(x)$. Note that since $T(x, v)$ is continuous with respect to both variables then G is continuous. Next, we show that the set O_δ , which is the candidate recurrent set by Theorem 3.2, consists of states for which the expected value of the gradient of H is small.

Lemma 3.3. *There is a constant C such that for every $\varepsilon, \delta > 0$*

$$O_\delta \subset \left\{ \|DF_x\|^2 \leq \frac{\delta}{\varepsilon} + C\varepsilon \right\}.$$

Proof: For every $v \in V$, apply Taylor's expansion for F . Since F has a bounded second derivative, then

$$F(T(x, v)) - F(x) = \langle DF_x, T(x, v) - x \rangle + O(\|T(x, v) - x\|^2).$$

By the definition of T , it follows that there is a constant C such that $\|T(x, v) - x\| = \varepsilon \|D_x H(x, v)\| \leq C\varepsilon$. Hence, if $G(x) = \mathbb{E}_\nu F(T(x, v)) - F(x) \geq -\delta$, then (for another constant C),

$$\begin{aligned} -\delta &\leq \mathbb{E}_\nu F(T(x, v)) - F(x) \leq -\varepsilon \mathbb{E}_\nu \langle DF_x, D_x H(x, v) \rangle + C\varepsilon^2 = \\ &= -\varepsilon \|DF_x\|^2 + C\varepsilon^2. \end{aligned}$$

Therefore, $\|DF_x\|^2 \leq \frac{\delta}{\varepsilon} + C\varepsilon$, as claimed. ■

Corollary 3.4. *There is some constant C such that for every $0 < \varepsilon < 1$ the orbits X_n of the process \mathfrak{G} enter the set $\{\|DF_x\|^2 \leq C\varepsilon\}$ infinitely often almost surely.*

Proof: Clearly, T satisfies Assumption 1. If $\delta = O(\varepsilon^3)$, then by Lemma 3.3,

$$O_\delta \subset \left\{ \|DF_x\|^2 \leq C\varepsilon \right\}.$$

Hence, by Theorem 3.2 our claim follows. ■

Roughly speaking, when the process is at a state x in which $\|DF_x\|^2 = 0$, the change in the state of the process \mathfrak{G} is 0 on average. Therefore, these states are the candidate states for convergence of the process. Thus, the corollary implies that the process will visit neighborhoods of these critical states infinitely often, as is expected intuitively.

3.4 The Committee Machine with Bounded Change

The Committee Machine is an extension of the Perceptron. It consists of an odd number of Perceptrons and the decision is taken by a majority vote: if an example belongs to the majority of the closed halfspaces determined by the Perceptrons, the Committee Machine assigns the value 1 to that example, otherwise, the result will be 0.

We analyze below two variants of a learning algorithm for the committee machine. In both cases, there is a non-trivial set A from which the target set of states is not accessible. Thus, the learning algorithm may never reach a “good” target state. We prove that A is large, in the sense that it has a non-empty interior.

Formally, let

$$\text{sgn}(x) = \begin{cases} 1 & x \geq 0 \\ -1 & x < 0 \end{cases}, \quad (7)$$

and put X^i to be a Perceptron in \mathbb{R}^d . For every $X = \{X^1, \dots, X^n\}$ and $v \in S^{d-1}$ let the decision function of the Committee Machine X be $g_X(v) = \text{sgn}\left(\sum_{i=1}^n \text{sgn}\langle x^i, v \rangle\right)$. Just like in the previous case, we define the error function which enables us to determine whether two states agree or disagree on a given input. For every Committee Machine X , let the on-line error function be

$$f_{X,T} = \begin{cases} 0 & g_X(v) = g_T(v) \\ 1 & \text{otherwise} \end{cases},$$

where $T = \{T^1, \dots, T^n\}$ is the teaching Committee Machine.

There are several possible ways to define a learning process for Committee Machines and we shall focus on two of them. Both learning rules are based on the same idea: given an example in which the student disagrees with the teacher, we wish to move several Perceptrons from the (wrong) majority vote to the correct minority one.

In the first learning rule, given such an example, we find the Perceptron whose decision boundary is nearest to the example and then change it so that it agrees with the teaching Committee Machine. The problem with this learning rule is that even now, the student may be wrong – because the wrong opinion may still be in the majority. Thus, the second learning rule is to take a large enough set of Perceptrons and change their vote, just as we did with the single Perceptron. We shall take the smallest set possible and the Perceptrons selected are those with decision boundaries closest to the example.

We define the learning process T_c in the following way: if $f_{X,T}(v) = 1$, let $W = W(x, v)$ be the set of Perceptrons (X^i $i = 1, \dots, n$) such that the response of the Perceptron X^i does not agree with that of the teaching Committee Machine T . From W select X^j such that ∂X^j is the nearest to v . If v and ∂X^j are “close enough” in the sense that $d(\partial X^j, v) \leq \delta_n$, then X^j is adapted so that it agrees with the teacher on v with a minimal change in its location, where δ_n may be fixed, or even a sequence decreasing to 0. Otherwise, if the boundary of X^j is too far away from v , the student remains stationary. The Assumption that no learning takes place on “far” examples, is called bounded change. This assumption is made to help in cases where the teacher itself may randomly make mistakes (for example, Kim and Sompolinsky, 1996).

Note that when both the teacher and the student are single Perceptrons ($n = 1$), then this learning process guarantees that the $d(x_n, t) \rightarrow 0$ almost surely (Mendelson, 1998).

Now we apply the theory of this paper to the committee machine. Using a similar method to the one presented in 3.1, it is easy to see that the conditions of Corollary 2.6 are fulfilled in this case. However, as the following example shows, the set A of states from which the target set is inaccessible may be large and (X_n) may remain far away from the set of the system’s correct states, even when $n = 3$ (see Fig. 4).

Example 3.5. *The only set in which the student and the teacher do not agree is the domain bounded by (1) and (3). For every input selected from that set, we assume that the only ∂X^i which is near enough to allow the adaptation process is (2). Hence, (2) converges to (3) P_0 -almost surely and*

the limit state is an absorbing state of the process which is not a correct state.

Using this idea, one can demonstrate that for every odd n the Committee Machine containing n Perceptrons can behave in the same manner as in the example above. It is possible to construct such examples in which both the teacher and the student can not be realized by a single Perceptron.

Clearly, only a slight modification in the examples above is needed to show that the set A has a non-empty interior. Indeed, a small perturbation of the student in figure 4 yields a state which, by the same argument as above, converges to the same limit state as the student in figure 4.

These examples indicate that one should be careful when selecting the bounded change parameter δ . A wise choice of δ is necessary, otherwise, one may stumble on such “bad” states. Of course, when the geometry of the problem is not known, it may be difficult to make this selection. Also, even if the bounded change parameter works for the initial state, there may be a positive probability to move to a state for which δ is too small.

Modifying the learning algorithm to move sets of Perceptrons instead a single Perceptron does not solve this problem, and in fact the same example is a counter-example to such learning algorithms as well.

4 Concluding Remarks

We presented a theory which makes it possible to check if a target set of a learning algorithm is recurrent. The main disadvantage of this method is that it does not yield a quantitative estimate on the time required to enter the target set. However, it should be noted that the first step in obtaining any such estimate is to study the properties of h_O . Our method takes at least

a step in this direction, in that it gives conditions under which a positive lower bound exists. The advantage of this method is its simplicity, and as we showed in the examples, it applies in many standard settings, simplifying, unifying and extending existing results.

References

- Fort JC, Pages G (1995) On the a.s. convergence of the Kohonen algorithm with a general neighborhood function, *The Annals of Applied Probability*, Vol 5, 4 1177–1216
- Fort JC, Pages G (1996) About the Kohonen Algorithm: Strong or Weak Self Organization, *Neural Networks*, Vol 9, 5 773–785
- Haykin SS (1994) *Neural Networks: A Comprehensive foundation*, MacMillan College Press
- Kim JW, Sompolinsky H (1996) On-line Gibbs learning, *Physical Review Letters*, Vol 76, 16 3021–3024
- Kohonen T (1989) *Self-organization and associative memory*, Springer-Verlag
- Loève M (1963) *Probability Theory*, 3rd edition, D. Van Nostrand
- Mendelson, S (1998) *Mathematical Aspects of Learning in Neural Networks*, Ph.D Thesis, Technion – I.I.T
- Orey S (1971) *Limit theorems for Markov chain transition probabilities*, Van Nostrand Reinhold
- Ritter H, Martinetz T, Schulten K (1992) *Neural computation and self organizing maps*, Addison-Wesley

5 Figure Legends

Fig. 1: The geometry of Perceptron learning. In two dimensions, hyperplanes are lines dividing the plane into a positive and a negative halves, denoted here by the + and - signs. The teaching hyperplane T is in this case a vertical line, and its positive half plane is to its right. The student hyperplane X gives identical answers to the teacher in the regions marked Correct, but in the other two regions, the student gives opposite answers to those of the teacher. As long as there is a positive probability for training examples to be chosen from those regions, the student Perceptron will continue to adapt. x and t are the unit vectors orthogonal to X and T, respectively.

Fig. 2: A. The canonical ordered set. B. An example of a state in which the distances between the points is in accordance with the neighborhood function f . This state belongs to O . C. An example of a state which preserves some of the features of the order in A, but does not belong to O . For example, point 4 is closer to point 2 than to point 1, although $f(4,2)=0$ whereas $f(4,1)=0.25$. D. The number of unordered points during a typical run of the learning process.

Fig. 3: Ordering the lattice with the Kohonen process. A. The initial state of the simulations. To improve visibility, only one out of every four points is shown. B. The state of the process after 800,000 iterations. There are roughly three domains. The top half of the square keeps roughly its initial order; the bottom half is divided into two domains, rotated by 90° relative to each other. In this simulation run, the three domains were already apparent after 200,000 iterations ($\varepsilon = 1$). C. The state of the process at the end of the 2nd stage of our algorithm, when an ordered state is formed at the bottom

left part of the set of examples. Note the change in scale of the x and y axes. D. The number of disordered rows and columns during the operation of the algorithm. Since the simulations were run on a 20×20 lattice, the maximal number of disordered rows and columns is 40. The arrows show transitions between states of the algorithm (further details are in the text).

Fig. 4: The teaching committee machine is shown in A, while the student appears in B. The "+" signs near each hyperplane represent the positive side of the hyperplane on which the Perceptron votes "+". The numbers represent the final vote of the Committee Machine on examples selected from that area. For example, +1 means that the example is on the positive side of two Perceptrons and on the negative side of the third. In C we see the teacher and student together. The area bounded by Perceptrons 1 and 3 is the only area in which the teacher and the student disagree. If the bounded change parameter is too small, then 2 is the only Perceptron that can adapt, and it will converge to 3, to form an absorbing state which is not correct.

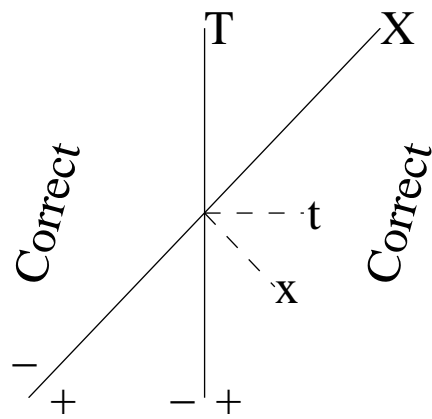


Fig. 1 (Mendelson, single column)

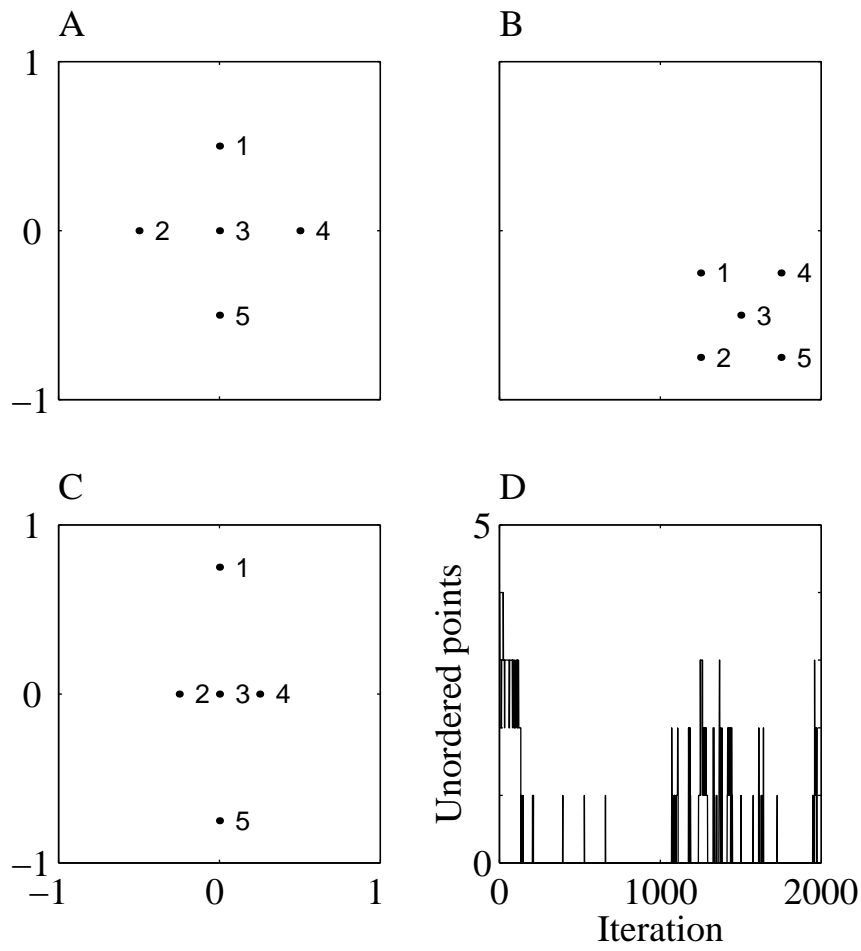


Fig. 2 (Mendelson, double column)

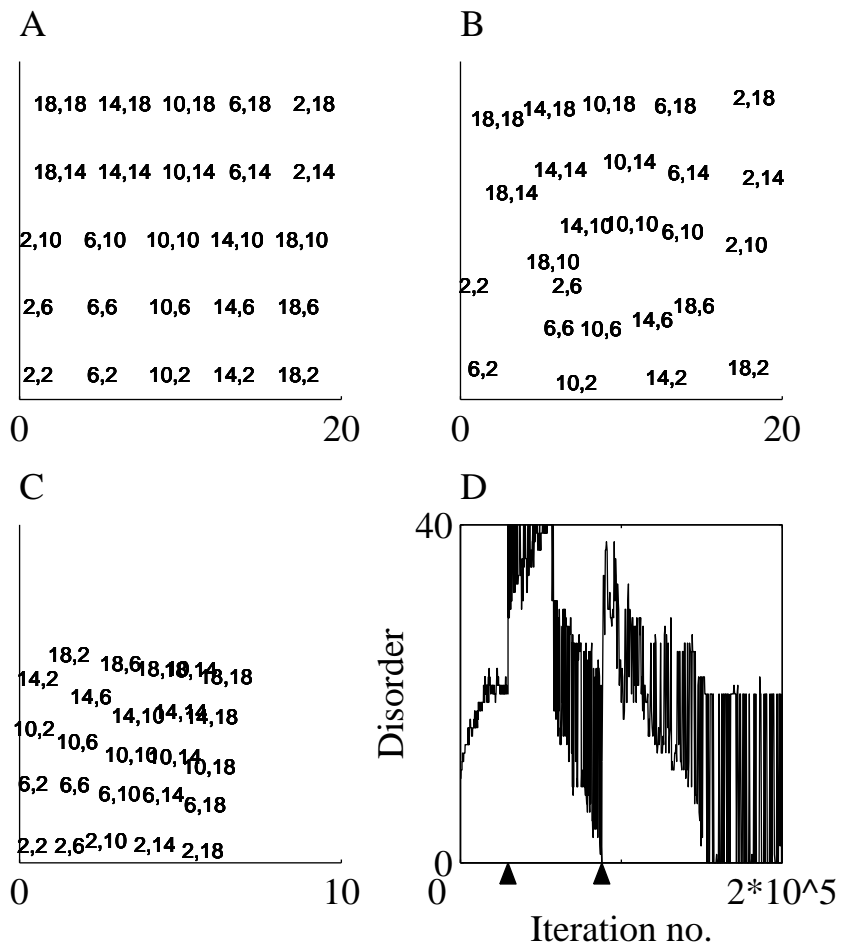


Fig. 3 (Mendelson, double column)

Fig. 4 (M

A Appendix

A.1 Proofs of the Theorems from section 2

The appendix is devoted to the proofs of the main theorems in section 2. We begin with the proof of the well known recurrence result, which may be found, for example, in Orey (1971). We present the proof for the sake of completeness.

Proof of Theorem 2.1: The first part of the proof is a version of a 0-1 law which is due to P. Lévy (Loève, 1963): Let Y_1, Y_2, \dots , be a sequence of random variables and let Y be a random variable defined on Y_1, Y_2, \dots such that $\mathbb{E}|Y| < \infty$. Note that $Z_n = \mathbb{E}(Y|Y_1, \dots, Y_n)$ forms a martingale, thus, by the martingale convergence theorem (Loève, 1963, p. 393), Z_n converges almost surely to Y . In particular, if we set $U_i = \{X_i \in O\}$, $U = \{X_n \in O \text{ i.o.}\}$, $Y_n = X_n$ and $Y = \chi_U$, then $P_0(U|Y_1, \dots, Y_n) = \mathbb{E}(Y|Y_1, \dots, Y_n)$ converges a.s. to χ_U and $P_0(\cup_k^\infty U_i|Y_1, \dots, Y_n)$ tends to $\chi_{\cup_k^\infty U_i}$ for every fixed k .

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

$$P_0(\cup_k^\infty U_i|Y_1, \dots, Y_n) \geq P_0(\cup_n^\infty U_i|Y_1, \dots, Y_n) \geq P_0(U|Y_1, \dots, Y_n).$$

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

$$\chi_{\cup_k^\infty U_i} \geq \limsup_{n \rightarrow \infty} P_0(\cup_n^\infty U_i|Y_1, \dots, Y_n) \geq \liminf_{n \rightarrow \infty} P_0(\cup_n^\infty U_i|Y_1, \dots, Y_n) \geq \chi_U.$$

Again, taking $k \rightarrow \infty$, the left side converges almost surely to χ_U , hence $P_0(\cup_n^\infty U_i|Y_1, \dots, Y_n)$ tends to χ_U .

Denote by $L_n(X_n, O) = P_0(\cup_n^\infty U_i|X_1, \dots, X_n)$ then by the 0-1 law $L_n(X_n, O)$ tends to the characteristic function of the set $\{X_n \in O \text{ i.o.}\}$. Since $L_n(X_n, O)$ is the probability of an orbit to visit O for some $m > n$ given its history

X_1, \dots, X_n , then by our assumption $L_n(-, O)$ is strictly positive on the set $\{X_n \in A \text{ i.o.}\}$ implying that $L_n(-, O) \rightarrow 1$ on that set. Hence, $\{X_n \in O \text{ i.o.}\} \supset \{X_n \in A \text{ i.o.}\}$ P_0 -almost surely. ■

Next we shall prove that under Assumption 1, h_O is lower semi continuous. We begin the proof with the following lemma:

Lemma A.1. *If Assumption 1 holds, then for every $x \in X$, every sequence $y_n \rightarrow x$ and every integer k , there is a set $\Omega \subset V^k$, for which $\nu^k(\Omega) = 1$, and for every k -tuple $(\omega_1, \dots, \omega_k) \in \Omega$, $T^k(y_n; \omega_1, \dots, \omega_k) \rightarrow T^k(x; \omega_1, \dots, \omega_k)$*

Proof: We will prove our claim for $k = 2$. The general case follows by induction. Fix $x \in X$, and a sequence (y_n) such that $y_n \rightarrow x$. Set $z_n(v) = T(y_n, v)$ and $z(v) = T(x, v)$. Hence, by Assumption 1, $z_n \rightarrow z$ ν -almost surely. Again by Assumption 1, for every such v , $T(z_n(v), \omega) \rightarrow T(z(v), \omega)$ ν -almost surely. Note that $T^2(y_n; v, \omega) = T(z_n(v); \omega)$ and that $T^2(x; v, \omega) = T(z(v); \omega)$. Thus our claim follows from Fubini's theorem. ■

Proof of Theorem 2.5: Define stopping times τ_x by the first time in which X_n enters O given that $X_0 = x$. If the orbit does not enter O we set $\tau_x = \infty$. Let $f_k(x) = P_x\{\tau_x \leq k\}$. Clearly, f_k in an increasing sequence which converges to h_O . Thus, to show that h_O is lower semi continuous, it is enough to show that each f_k is lower semi continuous.

Fix $x \in X$ and a sequence (y_n) converging to x . By Lemma A.1, for every integer k there is a set $\Omega \subset V^k$ such that $\nu^k(\Omega) = 1$, and $T^k(y_n; -) \rightarrow T^k(x; -)$ on Ω . Note that if O is open and if $T^k(x; \omega_1, \dots, \omega_k) \in O$ then for n

large enough so is $T^k(y_n; \omega_1, \dots, \omega_k)$. Thus

$$\chi_{\{\tau_x \leq k\}} \leq \liminf_{n \rightarrow \infty} \chi_{\{\tau_{y_n} \leq k\}}. \quad (8)$$

Taking the expectation and by Fatou's lemma it follows that

$$\mathbb{E}\chi_{\{\tau_x \leq k\}} \leq \mathbb{E}\liminf_{n \rightarrow \infty} \{\tau_{y_n} \leq k\} \leq \liminf_{n \rightarrow \infty} \mathbb{E}\chi_{\{\tau_{y_n} \leq k\}}, \quad (9)$$

hence, $f_k(x) \leq \liminf_{n \rightarrow \infty} f_k(y_n)$. ■

A.2 Algorithm for the simple 2-d Kohonen process

1. Given the random initial state, denote by d the minimum distance between point 3 and any of the four corners of the unit square. Assume that $d \leq 3\varepsilon$. Otherwise, it is necessary to “pull” point 3 away from the border so that this condition is fulfilled. Alternatively one may decrease the size of the step of the process ε to be less than $d/3$.

2. Choose point 3. All other four points will approach point 3, so that the maximum distance between point 3 and any of the other four points decreases (by $1 - \varepsilon/2$). Repeat this step until the maximum distance between point 3 and the other points is small enough (e.g., smaller than $\varepsilon/10^6$).

At this point, we do the following rescaling. Since point 3 did not move, it is still at a distance of 3ε from the border. Build a square around point 3 with sides of length 2ε . The whole square will be inside the borders of the original region. To simplify notations, rescale the coordinates so that 1 corresponds to ε . With this rescaling, the small square now has corners $[\pm 1, \pm 1]$ and the new ε is equal 0.5. For the rest of the discussion, all the

coordinates refer to this new coordinate system. The coordinates of all 5 points are $[0, 0]$ (up to perturbations of size 10^{-6}).

3. At least one corner of the unit square is closer to one of points 1, 2, 4, 5 than to point 3. Choose one such corner (up to a perturbation of size 10^{-6}) as the next example. We assume without loss of generality that it is corner $[1, 1]$ and that it is closest to point 1. The coordinates after this step are given in the table below, up to perturbations of size 2×10^{-6} .

4. Consider the point $[-0.874, 1]$. This point is (up to the small perturbations) close to points 2 and 4, and farther away from all other points. Clearly, we may assume that it is actually closest to point 2. Also, it can be slightly perturbed so that it does not have two nearest neighbors. The coordinates are as given in the table up to perturbations of size 3×10^{-6} .

5. The next example is point $[-1, 1]$ which is closest to point 4. It is easy to check that all points whose distance from $[-1, 1]$ is smaller than 10^{-6} will also be closest to point 4. The coordinates are as given in the table up to perturbations of size 4×10^{-6} .

6. The last example is point $[-1, -1]$, which is closest to point 5 (together with a neighborhood of size 10^{-6}). The coordinates are as given in the table up to perturbations of size 5×10^{-6} .

At this point, the relative distances between the points are consistent with the function f . For example, the distance between point 1 and the other points is 0.7481, 0.5542, 0.7677, 1.5712. It is thus closest to point 3, farther from points 2 and 4, and farthest from point 5, as required.

After:	step 3		step 4		step 5		step 6	
<i>point</i>	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>	<i>x</i>	<i>y</i>
1	0.5000	0.5000	0.3282	0.5625	0.4122	0.3672	0.4122	0.3672
2	0.1250	0.1250	-0.3745	0.5625	-0.3745	0.5625	-0.4527	0.3672
3	0.2500	0.2500	-0.0310	0.4375	0.2268	0.0781	-0.0799	-0.1914
4	0.1250	0.1250	0.1250	0.1250	0.5625	-0.4375	0.3672	-0.5078
5	0	0	-0.1092	0.1250	0.0294	-0.0156	-0.4853	-0.5078

References

- Fort JC, Pages G On the a.s. convergence of the Kohonen algorithm with a general neighborhood function *The Annals of Applied Probability*, Vol 5, 4 1177–1216
- Fort JC, Pages G About the Kohonen Algorithm: Strong or Weak Self Organization *Neural Networks*, Vol 9, 5 773–785
- Haykin SS *Neural Networks: A Comprehensive foundation*. MacMillan College Press MacMillan College Press
- Kim JW, Sompolinsky H On-line Gibbs learning. *Physical Review Letters*, Vol 76, 16 3021–3024
- Kohonen T *Self-organization and associative memory*. Springer-Verlag
- Loève M *Probability Theory*. 3rd edition, D. Van Nostran
- Mendelson S *Mathematical Aspects of Learning in Neural Networks*. PhD thesis, Technion – I.I.T.
- Orey S *Limit theorems for Markov chain transition probabilities*. Van Nostran Reinhold
- Ritter H, Martinetz T, Schulten K *Neural computation and self organizing maps*. Addison-Wesley