Abstract

In this thesis we investigate various mathematical aspects of learning in neural networks.

During a learning process a network is exposed and responds to various inputs. Then, the system learns by changing parameters which influence its decision making process. For example, when the network has a teacher, it can inform the network if the response to an input is correct or not. If the network's response is wrong, the parameters change according to some learning rule such that the chance for the network to give a correct answer to the inputs increases.

We investigate several supervised learning models and focus on questions concerning the convergence of these processes. The tools we use in the analysis are martingale convergence theorems and recurrence theorems for Markov processes.

In chapter 2 we analyze linear learning processes (the Perceptron learning rule and the Projection model). We show that in certain cases the Perceptron learning algorithm does not converge, while the Projection model converges almost surely and the limit networks agrees with the teacher for almost every input.

In chapter 3 we introduce a general on–line supervised learning model. We show, for example, that if there are "many" states which are in complete agreement with the teacher then the process converges to one of those states.

There are many learning rules which are not aided by a teacher. A tool frequently used in the analysis of such models in the *Stochastic Approximation process*. Thus, in the first chapter we present a thorough investigation of this process in sufficiently nice Banach spaces. We are particularly interested in so–called stochastic Lyapunov systems. We show, for instance, that in finite dimensional spaces this process converges to a local minimizer of the system's Lyapunov function. Also, we introduce a new version of the stochastic approximation process in Hilbert spaces and use it to construct an adaptive process which converges to solutions of Reaction–Diffusion equations.

The unsupervised learning process we investigate in chapter 4 is the well known Kohonen learning rule. Here, we try to answer questions regarding the self–organization properties of this process. We show that the orbits of the Kohonen process frequently enter the set of "organized" states.

In the last section of chapter 4 we introduce a smooth version of the Kohonen rule which is connected to an optimization problem called the "lazy traveling salesman" problem. We use results derived in chapter 1 to offer an adaptive stochastic process which converges to the solution of this optimization problem.

List of symbols

X	normed space
V, Ω	sets containing the inputs
$\mathbb{E}f$	expectation of f
$\|f\|$	$\mathbb{E}\,\|f\|$
S^{d-1}	unit sphere in $\mathbb{R}^d$
$C^\alpha(Y, X)$	space of functions from $Y \subset \mathbb{R}$ to X which are α–Hölder
$C^k(X)$	space of functions from X to $\mathbb{R}$ whose k derivative is continuous
$C^{1,1}_{loc}(X)$	differentiable functions on X with a locally Lipschitz derivative
Df	derivative of f
$B_\lambda(x)$	open ball centered at x with radius λ
$\overline{A}$	closure of A
$A \Delta B$	$(A \cap B^c) \cup (A^c \cap B)$
χ_A	the characteristic function of A

What is a neural network

Roughly speaking, a neural network consists of a finite number of "black boxes" and an architecture of connections between the boxes. Each black box can receive an input and according to some rule it produces an output. The architecture determines which of boxes are connected and what is the strength of each connection.

During a learning process the network is exposed to inputs, yields outputs and changes its parameters according to some learning rule. The learning rule may change the strength of the connections between the boxes or even the decision making mechanism of each box.

We view a network as a point in some state space, hence the learning process is a stochastic process defined on that state space. In this thesis we investigate the behavior of limits of such stochastic learning processes.

We separate the discussion to two categories, the first of which is called supervised learning. In such learning rules the learning process is aided by a teacher in the following way: both the teacher and the student are exposed to inputs and the student changes its position when its response does not agree with that of the teacher. In cases of an incorrect answer, the learning rule should move the student closer to the teacher in some sense. In the second category, called unsupervised learning, the adaptive process is not aided by a teacher.

We wish to note that sometimes the network's architecture and its decision making mechanism are not described. Rather, we view the network as an element in a state space and the learning rule is given by the transition density for state to state.

Chapter 1 – Stochastic approximation of Lyapunov systems

0 – Introduction

The main object of this chapter is the study of stochastic approximation in infinite dimensional spaces. As an example we consider the scalar valued, semilinear parabolic equation on a domain $\Omega \times \mathbb{R}^+$ where $\Omega \subset \mathbb{R}^d$

$$\frac{\partial u}{\partial t} = \Delta u + \overline{f}(u) \quad ; \quad u(0) = u_0 \tag{0.1}$$

and $u \in \mathbb{H}_0^1(\Omega)$ for $t > 0$.

Let (V_j), $j = 0, 1, 2, \ldots$ be a sequence of independent identically distributed (i.i.d) random variables subjected to the probability law μ and consider a function $f = f(u, v)$ where $\overline{f}(u) = \int f(u, v)\mu(dv)$. Given a sequence $t_n \to \infty$ such that $t_{j+1} - t_j \to 0$, we define the random process V_t for $t \geq 0$ by $V_t = V_j$ if $t \in [t_j, t_{j+1})$. A stochastic approximation of (0.1) is obtained by

$$\frac{\partial U}{\partial t} = \Delta U + f(U, V_t) \quad ; \quad U(0) = u_0 \tag{0.1*}$$

In the present example, the system (0.1) is a gradient flow with respect to the functional

$$F(u) = \frac{1}{2}\int_\Omega |\nabla u|^2 - \int_\Omega g(u)$$

where g is the primitive function given by $g(u) = \int^u \overline{f}$. Hence

$$\frac{\partial u}{\partial t} = -D_u F(u) \tag{0.2}$$

Similarly, (0.1*) is a gradient flow as well:

$$\frac{\partial U}{\partial t} = -D_U G(U, V_t) \tag{0.2*}$$

where

$$G(U, V) = \frac{1}{2}\int_\Omega |\nabla U|^2 - \int_\Omega g(U, V)$$

and $g(U, V) = \int^U f(s, V)ds$.

In this case, we wish to investigate the relations between the critical points of F and the asymptotic limits of (0.2*).

In its original context, *Stochastic Approximation* is a process defined by a stochastic difference equation in $\mathbb{R}^d$. It was first investigated during the early 1950s by Kiefer

and Wolfowitz ([KW]) and by Robbins and Monro ([RM]) who considered the particular example

$$X_{n+1} = X_n - \frac{1}{n}\left(f(X_n) - \frac{1}{2}\right)$$

where $f(X_n)$ is a random function interpreted as the result of an experiment conducted at the state X_n. The object was to show that the process above converges almost surely to a deterministic value x which satisfies $h(x) = 1/2$, where $h(-\infty) = 0$ and $\frac{dh}{dx}$ is the probability distribution of the outcome of a single experiment.

After the pioneering work which appeared in [RM] and [KW], the process became the subject of a large number of papers and manuscripts. The definition of the stochastic approximation process was extended significantly to a more general form:

$$X_{n+1} = T_{\varepsilon_n}(X_n, V_n) \equiv X_n - \varepsilon_n H(X_n, V_n) \tag{0.3}$$

where $H(X_n, V_n)$ are random variables which represent samples of a given function on the state space, but unlike the example above may depend on other parameters and not just on X_n. In general, X_n and V_n are not assumed to be independent and even ε_n are sometimes assumed to be positive random variables instead of a fixed positive sequence. The questions asked in this context are under what assumptions does the process converge and do the limits give any additional information on the sampled function.

A detailed survey of the subject up to the mid 1960s can be found in [W]. Another source of information is Kushner and Clark's book from 1978 ([KC]). The most recent survey is due to Kushner and Yin ([KY]) which covers the latest developments concerning the stochastic approximation process in a finite dimensional space. There are fewer results concerning the process in an infinite dimensional setting, most of which are elementary extensions of the finite dimensional case. Some results concerning the process in Hilbert spaces may be found in [LPW].

In this chapter we offer two possible extensions to the process (0.3) in infinite dimensional spaces. In the first section we investigate an analog to (0.3) in Banach spaces with a sufficiently smooth norm and under the assumption that the system has a stochastic Lyapunov function. For example, if the space is a Hilbert space then the process (0.3) has a stochastic Lyapunov function if H is the derivative of some smooth function G. We also assume that (V_n) are i.i.d., that X_n and V_n are independent and that $\sum \varepsilon_n = \infty$, $\sum \varepsilon_n^p < \infty$ where $1 < p \leq 2$ is determined by the geometry of the space.

It is well known (see [KY]) that if G is a smooth function on $\mathbb{R}^d$ and if (X_n) are uniformly bounded then (X_n) converges almost surely to the set of critical points of F, where $F(x) = \int G(x,v)\mu(dv)$. It is also known that if x is a local minimum of F then there is a compact set K containing x, such that if a sample path (x_n) of the process enters K infinitely often then $x_n \to x$.

The proofs of those results are both due to Kushner and Clark ([KC]) and follow from the fact that orbits (x_n) can be approximated by the deterministic gradient flow $\dot{x} = -\nabla F(x)$, analogous to (0.2).

In the first section we present a simple proof of the convergence of (X_n) to the set of the critical points of F in sufficiently smooth Banach spaces. Our method, based on convergence theorems for Banach valued martingales ([P]) and classical martingale theory ([S]) enables us to obtain sharper results concerning the actual limits of X_n. In particular we obtain that under natural conditions in finite dimensional spaces the process X_n converges a.s. to a local minimizer of F. More generally, if all the solutions of $\dot{x} = -\nabla F(x)$, excluding a set of initial data of Lebesgue measure zero, converge to limit points in a set $C \subset K$, where K is the set of the critical points of F, then the corresponding stochastic approximations converge a.s. to limits in $\overline{C}$ as well. As an example, suppose that the stochastic approximation is given in $\mathbb{R}^2$ (i.e. $G = G((x,y),v)$) while the averaged $F = F(x)$. Then our results yield not only the convergence of the first coordinate sequence x_n to a critical point of F almost surely, but it implies that (x_n, y_n) converges as well.

In the second section we introduce a generalization of the process (0.3) in a Hilbert spaces. We may define T_{ε_n} as the compact operator given by the nonlinear semigroup generated by the flow (0.2*) for a time interval of length $\varepsilon_n = t_{n+1} - t_n$ where V is fixed on that interval of time. We show that, under some conceivable assumptions, a stochastic approximation of this type converges to a critical point of F, provided we have some a-priori local estimate of the type $\|T_\varepsilon(V) \circ X - X\| < C\varepsilon^{p-1}$ in the Hilbert space norm where $1 < p \leq 2$ and $(\varepsilon_n) \in l_p$.

In the third part we demonstrate that the conditions of section 2 hold for reaction diffusion equations of the type (0.1*) with $p = 3/2$ and in the fourth and final part we prove results concerning the weak convergence of stochastic Lyapunov systems.

1 – Stochastic gradient descent in Banach spaces – classical approach

The stochastic gradient descent process in a finite dimensional space is an example of the celebrated stochastic approximation process: define the process on $\mathbb{R}^d$ by $X_{n+1} = X_n - \varepsilon_n H(X_n, V_n)$, where $H : \mathbb{R}^d \times \Omega \to \mathbb{R}^d$ is continuous, $(V_n) \subset \Omega$ are i.i.d random variables with a common distribution which is induced by the measure μ on Ω and (ε_n) is a positive sequence which belongs to $l_2 \backslash l_1$, i.e., $\sum \varepsilon_i = \infty$, $\sum \varepsilon_i^2 < \infty$. It is well known ([KY]) that if X_n are uniformly bounded then the limit points of (X_n) are in the invariant set of the O.D.E $\dot{x} = \int H(x,v)\mu(dv)$ almost surely with respect to the measure induced by the process.

In this section we extend this result to a more general case under additional assumptions on the function H.

We begin with the following definitions and notations. For every Banach space $\mathbb{B}$ and $\tau > 0$ let $\rho_{\mathbb{B}}(\tau)$ be the modulus of smoothness of $\mathbb{B}$ defined by

$$\rho_{\mathbb{B}}(\tau) = sup\{(\|x+y\| + \|x-y\|)/2 - 1, \quad x, y \in \mathbb{B}, \|x\| = 1, \ \|y\| = \tau\}$$

We say that $\mathbb{B}$ is p–smooth if $\rho_{\mathbb{B}}(\tau) \leq C\tau^p$ for some $1 < p \leq 2$. $\mathbb{B}$ is called strictly convex if from the fact that $\|x\| = \|y\| = 1$ follows that $\|x+y\|/2 = 1$ if and only if $x = y$. Recall that if $\mathbb{B}$ is a strictly convex p–smooth Banach space then $\mathbb{B}$ is reflexive with a differentiable norm (see [D]).

For every Banach space E the normalized duality map $J : E \to E^*$ is a set valued function given by

$$J_E x = \{x^* \in E^* | \langle x^*, x \rangle = \|x\|^2, \ \|x\| = \|x^*\|\}$$

In our case $J_{\mathbb{B}}$ is a single valued one to one function and $J_{\mathbb{B}^*} J_{\mathbb{B}} = I_{\mathbb{B}}$. Also note that $J_{\mathbb{B}}$ is the derivative of $\frac{1}{2}\|x\|^2$. Let $H : \mathbb{B} \times \Omega \to \mathbb{B}$, put μ a probability measure on Ω, set V_n to be i.i.d random variables which are given by the measure μ and let $(\varepsilon_n) \in l_p \backslash l_1$. Assume that for every $x \in \mathbb{B}$, $\mathbb{E}_v H(x,v) = \int H(x,v)d\mu(v)$ exists, where the integration is in the Bochner sense. We define the stochastic approximation associated with H by

$$X_{n+1} = X_n - \varepsilon_n H(X_n, V_n) \tag{1.1}$$

Definition. *We say that the process (1.1) is a stochastic Lyapunov system if there exists a function $F : \mathbb{B} \to \mathbb{R}$ such that $F \in C^{1,1}_{loc}(\mathbb{B})$ and $DF = J_{\mathbb{B}}(\mathbb{E}_v H)$.*

If, for example, $\mathbb{B}$ is a Hilbert space then it has smoothness type $p = 2$ and $J_{\mathbb{B}}$ is the identity map. If H is the derivative of some function $G(x, v)$ with respect to the first variable, where $G(-, v) \in C^{1,1}_{loc}(\mathbb{B})$ uniformly μ–almost surely, i.e., for any bounded set $B \subset \mathbb{B}$ there exists a C such that for almost every $v \in \Omega$

$$\sup_{x \in B} \left\| D_x G_{(x,v)} \right\| < C, \quad \sup_{x \in B} \left\| D_x G_{(x_1,v)} - D_x G_{(x_2,v)} \right\| < C \left\| x_1 - x_2 \right\|$$

then the process (1.1) is a stochastic Lyapunov system. Indeed this follows by putting $F(x) = \int G(x, v) d\mu(v)$. The reason for the name "stochastic Lyapunov system" comes from the fact that F is a Lyapunov function for the deterministic gradient descent (z_n) defined by

$$z_{n+1} = z_n - \varepsilon_n J_{\mathbb{B}^*} DF(z_n) \tag{1.2}$$

which is in some sense the average of the process (1.1).
Throughout this section we assume that the process (1.1) is a stochastic Lyapunov system, that $\mathbb{B}$ is a p–smooth strictly convex Banach space, that $(\varepsilon_n) \in l_p \backslash l_1$ and:

A1. *$H(-, v)$ locally Lipschitz uniformly μ–almost surely, i.e., for any bounded set $B \subset \mathbb{B}$ there exists a C such that for almost every $v \in \Omega$*

$$\sup_{x \in B} \left\| H(x, v) \right\| < C, \quad \sup_{x \in B} \left\| H(x_1, v) - H(x_2, v) \right\| < C \left\| x_1 - x_2 \right\|$$

Let K be the set of the critical points of F (i.e. $DF(x) = 0$). Recall that λ is a critical value of F if there is a critical point x of F such that $F(x) = \lambda$ and denote by K_λ the set of such critical points. We assume that

A2. *K_λ is totally disconnected for any critical value of F.*

In certain cases we replace assumption A1 with the following:

A3. *$H(-, v)$ is both bounded and Lipschitz uniformly μ–a.s. and F is bounded from below on $\mathbb{B}$.*

We denote by τ the measure induced on the orbits of the process (1.1) and by $|\ |$ a Borel probability measure on $\mathbb{B}$.

We formulate our first claim:

Theorem 1.1. *Let X_n be defined by the process (1.1) and assume that (A1,A2) are satisfied. If X_n are uniformly bounded then $F(X_n)$ converges and $DF_{X_n} \to 0$ almost surely. Moreover, if F and DF are weakly continuous then for τ–almost every sample path (x_n) of the process (1.1) there exists a critical value λ of F such that (x_n) converges weakly and its limit belongs to the set K_λ. The same assertion holds even if (X_n) is not assumed to be uniformly bounded but conditions A2 and A3 are granted and F is coercive (i.e. $\lim_{\|x\|\to\infty} F(x) = \infty$).*

The idea behind the convergence theorem presented here is to divide the process into its stochastic part $H(x,v) - J_{\mathbb{B}^*}DF(x)$ and deterministic part $J_{\mathbb{B}^*}DF(x)$. We show that the stochastic element of the process (1.1) is well behaved in the sense that forms a converging series almost surely. Hence, the orbits of the process (1.1) are close to the orbits of the gradient descent process (1.2).

Lemma 1.2. *Put $Z_n = \varepsilon_n\big[\mathbb{E}\big(H(X_n,V_n)|X_n\big) - H(X_n,V_n)\big]$ and $Y_n = \big\langle DF(X_n), Z_n\big\rangle$, where X_n are given by the process (1.1). If $H(-,v) = DG_{-,v}$ is uniformly bounded τ–almost surely then $\sum_{n=1}^{\infty} Y_n$ and $\sum_{n=1}^{\infty} Z_n$ converge τ–almost surely.*

Proof: Let $\mathcal{F}_n$ be the σ-algebra generated by $X_1, V_1 ..., X_n, V_n$ and note that $\mathcal{F}_{n-1} = \sigma(\mathcal{F}_{n-1}, X_n)$. Clearly both Y_n and Z_n are $\mathcal{F}_n$–measurable uniformly bounded martingale difference sequences, i.e. $\mathbb{E}(Y_n|\mathcal{F}_{n-1}) = 0$ and $\mathbb{E}(Z_n|\mathcal{F}_{n-1}) = 0$. Indeed,

$$\mathbb{E}(Z_n|\mathcal{F}_{n-1}) = \varepsilon_n\mathbb{E}\Big(H(X_n,V_n)|\mathcal{F}_{n-1}\Big) - \varepsilon_n\mathbb{E}\Big(H(X_n,V_n)|\mathcal{F}_{n-1}\Big) = 0$$

and since $DF(X_n)$ is $\mathcal{F}_{\mathrm{n}-1}$ measurable then by the properties of the conditional expectation we see that

$$\mathbb{E}\big(\big\langle DF_{X_n}, Y_n\big\rangle|\mathcal{F}_{n-1}\big) = \big\langle DF_{X_n}, \mathbb{E}(Y_n|\mathcal{F}_{n-1})\big\rangle = 0$$

Since Y_n is scalar valued and $VarY_n \le C\varepsilon_n^2$ then $\sum VarY_n$ converges, hence by the martingale difference convergence theorem ([S]), $\sum Y_n$ converges almost surely.
Finally, set $S_n = \sum_1^n Z_i$. By Doob's maximal lemma (see [P1]), for every $t > 0$ and every $n \in \mathbb{N}$

$$\tau\Big(\big\{\sup_k \|S_{n+k} - S_n\| \ge t\big\}\Big) \le \frac{\sup_k \mathbb{E}\,\|S_{n+k} - S_n\|^p}{t^p}$$

Since $\mathbb{B}$ is p-smooth then by a result due to Pisier ([P2])

$$\frac{\sup_k \mathbb{E}\left\|S_{n+k}-S_n\right\|^p}{t^p} \leq \frac{1}{t^p}\sum_{k=1}^{\infty}\left\|Z_{n+k}\right\|^p \leq \frac{C}{t^p}\sum_{i=n}^{\infty}\varepsilon_i^p$$

Therefore $\sum Z_n$ converges almost surely.

$\diamond$

Next, we show that the process (1.1) converges to the set $\{x|DF(x)=0\}$. The key part in the proof is the following deterministic lemma, which is a modification of a result appearing in [LPW].

Lemma 1.3. *Let $(x_n),(u_n)\subset\mathbb{B}$, $(\varepsilon_n)\in l_p\backslash l_1$, where $x_{n+1}=x_n-\varepsilon_n J_{\mathbb{B}^*}DF_{x_n}+\varepsilon_n u_n$. Assume that $(F(x_n))$ is bounded from below and that DF is a Lipschitz function. Also assume that both (u_n) and (DF_{x_n}) are bounded sequences and that $\sum\varepsilon_n\langle DF_{x_n},u_n\rangle<\infty$, $\sum\varepsilon_n u_n<\infty$. Then $F(x_n)$ converges, $\sum\varepsilon_n\left\|DF_{x_n}\right\|^2<\infty$ and $DF_{x_n}\to 0$.*

Proof: By Taylor's formula, there is a ρ_n such that

$$F(x_{n+1})=F(x_n)+\varepsilon_n\langle DF_{\rho_n},-J_{\mathbb{B}^*}DF_{x_n}+u_n\rangle=$$
$$\varepsilon_n\left(\langle DF_{x_n},u_n\rangle-\left\|DF_{x_n}\right\|^2\right)+\varepsilon_n\langle DF_{\rho_n}-DF_{x_n},u_n-J_{\mathbb{B}^*}DF_{x_n}\rangle$$

Since DF is Lipschitz and both u_n and $J_{\mathbb{B}^*}DF_{x_n}=\int H(x_n,v)d\mu(v)$ are bounded then

$$\left|\varepsilon_n\langle DF_{\rho_n}-DF_{x_n},u_n-J_{\mathbb{B}^*}DF_{x_n}\rangle\right|\leq\varepsilon_n C\left\|\rho_n-x_n\right\|\leq\varepsilon_n C\left\|x_{n+1}-x_n\right\|\leq C'\varepsilon_n^2$$

Iterating the expansion for $F(x_{n+1})$ we see that

$$F(x_{n+1})=F(x_1)+\sum_{i=1}^{n}\varepsilon_i\langle DF_{x_i},u_i\rangle-\sum_{i=1}^{n}\varepsilon_i\left\|DF_{x_i}\right\|^2+\sum_{i=1}^{n}\beta_i$$

where $\sum_1^\infty\beta_i$ converges absolutely.
Since $\sum_1^n\varepsilon_i\langle DF_{x_i},u_i\rangle$ converges and $F(x_n)$ is bounded from below, $\sum\varepsilon_n\left\|DF_{x_n}\right\|^2$ is a positive bounded series, thus it converges, which implies that $F(x_n)$ converges as well.
Next, assume that $\left\|DF_{x_n}\right\|\geq\delta$ infinitely often. Hence, we can find an N such that $\left\|DF_{x_N}\right\|\geq\delta$, $\sum_N^\infty\varepsilon_k\left\|DF_{x_k}\right\|^2<\delta^2/8C$ and $\left\|\sum_N^\infty\varepsilon_k u_k\right\|<\delta/4C$, where C is the Lipschitz constant of DF_x. We show using induction that $\left\|DF_{x_n}\right\|\geq\delta/2$ for every $n>N$. Indeed,

by the definition of the process and using the induction hypothesis

$$\|x_{n+1} - x_N\| \leq \sum_{k=N}^{n} \varepsilon_k \|DF_{x_k}\| + \left\| \sum_{k=N}^{n} \varepsilon_k u_k \right\| \leq$$
$$\leq \frac{2}{\delta} \sum_{k=N}^{n} \varepsilon_k \|DF_{x_k}\|^2 + \left\| \sum_{k=N}^{n} \varepsilon_k u_k \right\| < \delta/2C$$

On the other hand,

$$\left\|DF_{x_{n+1}}\right\| \geq \|DF_{x_N}\| - \left\|DF_{x_{n+1}} - DF_{x_N}\right\| \geq$$
$$\geq \|DF_{x_N}\| - C\|x_{n+1} - x_N\| \geq \delta - \delta/2 = \delta/2$$

Therefore, since $(\varepsilon_n) \notin l_1$, $\sum \varepsilon_n \|DF_{x_n}\|^2$ diverges, which is a contradiction.

$\diamond$

Proof of Theorem 1.1: Set $U_n = \Big(\mathbb{E}\big(H(X_n, V_n)|X_n\big) - H(X_n, V_n)\Big)$. If (X_n) are uniformly bounded or if A3 is satisfied then both $Z_n = \varepsilon_n U_n$ and $Y_n = \langle DF_{X_n}, Z_n \rangle$ are uniformly bounded, thus by Lemma 1.2 both $\sum \varepsilon_n U_n$ and $\sum \varepsilon_n \langle DF_{X_n}, U_n \rangle$ converge $\tau-$almost surely. Therefore, in both cases the assumptions of lemma 1.3 hold for $\tau-$almost every orbit. It follows that for $\tau-$almost every orbit (x_n), $F(x_n)$ converges and $DF(x_n) \to 0$. Hence, since F and DF are weakly continuous, there is a critical value λ of F such that

$$\Omega^w_{(x_n)} = \bigcap_{k>1} \overline{\bigcup_{j>k}(x_j)} \subset K_\lambda$$

where the closure is with respect to the weak topology. We claim that $\Omega^w_{(x_n)}$ is connected and nonempty. Indeed, note that in both cases (x_n) is bounded – in the first case by the assumption that X_n are uniformly bounded and in the second, since by lemma 1.2 $F(x_n)$ converges and since F is coercive then x_n must be bounded. Thus, $\overline{\bigcup_{j>k}(x_j)}$ is a compact set and $\Omega^w_{(x_n)}$ is nonempty as an intersection of nested compact sets. To show that $\Omega^w_{(x_n)}$ is connected, note that if it is not the case, there are disjoint weakly open and closed (compact) sets $C_1,\ C_2 \subset \Omega^w_{(x_n)}$ such that $C_1 \cup C_2 = \Omega^w_{(x_n)}$. Hence, since the weak topology on Ω is metrizable and by the properties of that metric, there are weakly open sets U_1 and U_2 with disjoint weak closures, such that $C_i \subset U_i$. Both C_1 and C_2 contain weak limit points of (x_n) and since $\|x_{n+1} - x_n\| \to 0$, there is a subsequence $(x_{n_j}) \subset (U_1 \cup U_2)^c$.

Therefore, there is a weak limit point of (x_n) outside $C_1 \cup C_2$ which is a contradiction. Since K_λ is totally disconnected, $\Omega^w_{(x_n)}$ consists of a single point, thus (x_n) itself must converge.

$\diamond$

The next step is to prove that the behavior of the orbits of the process (1.1) is determined in some sense by the O.D.E (1.3) below. From this follows that in the case $\mathbb{B} = \mathbb{R}^d$, if X_1 has a density which is equivalent to the Lebesgue measure and under mild assumptions on F, X_n converges almost surely to a local minimum of F.
We impose the additional assumption:

A4. *For any measurable set $A \subset \mathbb{B}$, every $n > 0$ and almost every $v \in \Omega$,*

$$|\{y | y = x - \varepsilon_n H(x,v);\ x \in A\}| = 0 \quad \text{if and only if } \ |A| = 0$$

Note that if $\mathbb{B} = \mathbb{R}^d$, $H(x,v) = \nabla_x G(x,v)$ and G satisfies A3, then if ε_n is sufficiently small A4 holds – since the map $x \to x - \varepsilon_n \nabla_x G(x,v)$ is invertible. Also note that from A4 follows that the distributions of X_1 and X_n are equivalent.

The next result deals with the connection of the O.D.E

$$\frac{dx}{dt} = -J_{\mathbb{B}^*} DF_x \tag{1.3}$$

and the process (1.1).

Theorem 1.4. *Assume that A3 and A4 hold, that X_1 has a distribution which is equivalent to $|\ |$ and that there is a set $C \subset K$ such that for $|\ |$–almost every $x \in \mathbb{B}$ the solution $x(t)$ of (1.3) for which $x(0) = X_1$ converges to some $k \in C$ as $t \to \infty$. Then for τ–almost every orbit (x_n) of the process (1.1), either $\Omega_{(x_n)}$ is empty or $\lim_{n\to\infty} x_n$ exists and belongs to $\overline{C}$.*

Proof: For every $x \in \mathbb{B}$, let $\overline{x}_x(t)$ be the solution of (1.3) such that $\overline{x}_x(0) = x$. Define $N_0 \subset \mathbb{B}$ in the following way: $x \notin N_0$ if $\lim_{t\to\infty} \overline{x}_x(t)$ exists and belongs to C. Since $|N_0| = 0$ and since X_1 has a distribution which is equivalent to $|\ |$ then $\tau\big(\{(X_n) | X_i \in N_0 \ \text{ for some i}\}\big) = 0$. Hence, for τ–almost every orbit (x_n), $\sum \varepsilon_n \Big(\mathbb{E}\big(H(X_n, V_n) | X_n\big) - H(X_n, V_n)\Big)$ converges, and for every n the solution $\overline{x}(t)$ of (1.3) such that $x(0) = x_n$ converges to some point in C as $t \to \infty$. Let (x_n) be such an orbit.

Clearly

$$x_{n+1} = x_n - \varepsilon_n J_{\mathbb{B}^*} DF_{x_n} + \varepsilon_n \Big(\mathbb{E}\big(H(x_n, v_n)|X_n = x_n\big) - H(x_n, v_n) \Big)$$

Denote $\mathbb{E}\big(H(X_n, v_n)|X_n = x_n\big) - H(x_n, v_n)$ by u_n. Put $t_n = \sum_1^n \varepsilon_i$ and let $U(t)$ be the linear interpolation defined by $U(0) = X_1$, $U(t_n) = \sum_1^n \varepsilon_i u_i$.
Set $x^*(t) = \sum \chi_{[t_n, t_{n+1}]} x_{n+1}$, $\hat{x}(t) = x_0 - \int_0^t J_{\mathbb{B}^*} DF_{x^*(s)} ds + U(t)$ and $\hat{x}_n(t) = \hat{x}(t + t_n)$. Note that $\hat{x}(t_n) = \hat{x}_n(0) = x_n$. Moreover,

$$\hat{x}_n(t) = \Big(\hat{x}_n(0) - \int_0^t J_{\mathbb{B}^*} DF_{\hat{x}_n(s)} ds \Big) + \Big(U(t + t_n) - U(t_n) \Big) -$$
$$- \Big(\int_0^t J_{\mathbb{B}^*} DF_{x^*(s+t_n)} ds - \int_0^t J_{\mathbb{B}^*} DF_{\hat{x}_n(s)\, ds} \Big) = \overline{x}_n(t) + A_n(t) + B_n(t)$$

Since $\overline{x}_n(t)$ is a solution of (1.3) then $\lim_{t \to \infty} \overline{x}_n(t)$ exists, belongs to C and we denote it by k_n.
By the definitions of $\hat{x}(t)$ and $x^*(t)$, we see that

$$\int_{t_n}^{t_{n+1}} \|\hat{x}(t) - x^*(t)\|\, dt \le \|x_{n+2} - x_{n+1}\|\, \varepsilon_{n+1}/2 \le C' \varepsilon_n^2$$

and since $J_{\mathbb{B}^*} DF = \int H(x, v) d\mu(v)$ is a Lipschitz function then

$$\left\| \int_0^t J_{\mathbb{B}^*} DF_{\hat{x}(s+t_n)} - J_{\mathbb{B}^*} DF_{x^*(s+t_n)} ds \right\| \le C \int_{t_n}^{t+t_n} \|\hat{x}(s) - x^*(s)\|\, ds \le C' \sum_n^\infty \varepsilon_n^2$$

therefore, for every n, $\|B_n(t)\| \le C' \sum_n^\infty \varepsilon_i^2$ and since $U(t)$ is the linear interpolation of $U(t_n) = \sum_1^n \varepsilon_i u_i$ then $\lim_{n \to \infty} U(t + t_n) - U(t_n) = 0$ uniformly. Fix $\delta > 0$. There is an $N(\delta)$ such that for every $t > 0$, $\|A_N(t) + B_N(t)\| \le \delta$, therefore, $\|\hat{x}_N(t) - \overline{x}_N(t)\| \le \delta$ for every $t > 0$. Since $\lim_{t \to \infty} x_N(t) = k_N$ then $\Omega_{(x_n)} \subset B_{k_N}(\delta_N)$. Hence, $\Omega_{(x_n)} \subset \bigcap_{\delta > 0} B_{k_{N(\delta)}}(\delta)$ which implies that if $\Omega_{(x_n)}$ is not empty, then $\lim_{\delta \to 0} k_{N(\delta)}$ exists and that $\Omega_{(x_n)} = \lim_{\delta \to 0} k_{N(\delta)}$. Thus (x_n) converges to some $k \in \overline{C}$.

$\diamond$

Note that the theorem will still be true if we assume A1,A4 and impose that X_n are uniformly bounded.

From here on, we assume that $\mathbb{B}$ is a finite dimensional space. Theorem 1.4 implies that the behavior of the process (1.1) is determined by the O.D.E (1.3) under the assumed

conditions. Also note that $\Omega_{(x_n)}$ is not empty, since by the proof of theorem 1.4 (x_n) is bounded. We present in Corollary 1.7 below an example in which the conditions of theorem 1.4 are fulfilled and prove that the process (1.1) converges to a local minimum of F τ–almost surely. For examples in which weaker conditions than the ones in corollary 1.7 are imposed but the limit of (1.3) still exists for Lebesgue almost every initial condition, we refer the reader to [A].
We add the additional assumption:

A5. *All the critical points of F are non degenerate. In particular, the set K of the critical points of F is (at most) countable.*

The following result concerning the O.D.E (1.3) in $\mathbb{R}^d$ is well known and its proof is omitted.

Proposition 1.5. Let $\mathbb{B} = (\mathbb{R}^d, \| \ \|_2)$ and let $|\ |$ be a probability measure equivalent to the Lebesgue measure.

a. *If the solution $x(t)$ of (1.3) is bounded on $t \geq 0$ then $\lim_{t\to\infty} x(t)$ exists and is a critical point of F. In particular, if F is coercive then the above limit exists for any $x(0) \in \mathbb{R}^d$.*

b. *For any non-degenerate critical point x_c of F the stable $W^s(x_c)$ and unstable $W^u(x_c)$ manifolds of (1.3) are embedded in $\mathbb{R}^d$.*

Since an embedded manifold of non zero co–dimension is of Lebesgue measure 0 we obtain:

Proposition 1.6. *If F is coercive and all its critical points are non degenerate then there exists a set $N_0 \subset \mathbb{R}^d$, $|N_0| = 0$ such that $\lim_{t\to\infty} x(t)$ is a local minimum of F provided that $x(0) \in N_0^c$.*

Proof: Let $K^u \subset K$ ($K^s \subset K$) be the set of unstable (stable) critical points of F and denote $W^s(K^u) \equiv \bigcup_{x\in K^u} W^s(x)$ (res. $W^s(K^s) \equiv \bigcup_{x\in K^s} W^s(x)$). Since $W^s(x)$ is an embedded manifold of dimension $k < d$ for any $x \in K^u$ and there are a countable number of such points, then $W^s(K^u) = 0$. Since $\mathbb{R}^d = W^s(K^u) \cup W^s(K^s)$ we obtain that $N_0 = W^s(K^u)$ as required.

$\diamond$

Corollary 1.7. *Let* $(\mathbb{B} = \mathbb{R}^d, \| \ \|_2)$ *and set* $H(x,v) = \nabla_x G(x,v)$. *Assume that A3 and A5 hold, that* X_1 *has a distribution which is equivalent to the Lebesgue measure and that* F *is coercive. Then* τ*–almost surely the process (1.1) converges and its limits are local minima of* F.

Proof: This follows immediately from proposition 1.6 and theorem 1.4, by setting C to be the set of local minima of F.

$\diamond$

2 – The infinite dimensional case – compact operator approach

In this section, we present a generalization of the process (1.1) to a Hilbert space. The essential difference between this and the previous case is that the gradient (Fréchet derivative) of G may be in general, a random *unbounded* operator on the underlying space. In actual applications (c.f. next section) the gradient of G generates a nonlinear random continuous semigroup which is a compact perturbation of a linear (deterministic) one. This example motivates the assumptions below:

Let $\mathbb{H}_1 \subset \mathbb{H}_0$ be Hilbert spaces equipped with the norms $\| \ \|_{\mathbb{H}_i}$ $i = 0, 1$. Assume that on $\mathbb{H}_1$, $\|x\|_{\mathbb{H}_0} \leq C \|x\|_{\mathbb{H}_1}$ and that $\mathbb{H}_1$ is compactly embedded in $\mathbb{H}_0$. Denote by $\langle \ , \ \rangle$ the inner product in $\mathbb{H}_0$ and let (Ω, μ) be a Borel probability space. Suppose that $G : \mathbb{H}_1 \times \Omega \to \mathbb{R}$ is a Fréchet differentiable function on $\mathbb{H}_1$ for $\mu-$almost every $v \in \Omega$. For every fixed v, denote by $D_x G$ the derivative of G with respect to the first variable and assume that for every $x \in \mathbb{H}_1$, $D_x G_{x,-}$ is μ measurable and that $\operatorname{ess\,sup}_{v \in V} \|D_x G_{x,v}\|_{(\mathbb{H}_1, \| \ \|_0)^*} \leq C(x)$. We denote $(\mathbb{H}_1, \| \ \|_0)^*$ by $\mathbb{H}_{-1}$ and let $C^\alpha([0,T]; \mathbb{H}_1)$ be the space of functions from $[0,T]$ to $\mathbb{H}_1$ which are α–Hölder.

The following observation, which was noted in the first section, is standard and its proof is omitted.

Lemma 2.1. *Put* $F(x) = \int G(x,v) d\mu(v)$. *Then* F *is Fréchet differentiable with respect to the norm* $\| \ \|_{\mathbb{H}_0}$ *on* $\mathbb{H}_1$ *and* $DF = \int D_x G_{(x,v)} d\mu(v)$, *where the integration is in the Bochner sense.*

Fix $0 < \alpha < 1$, let $(\varepsilon_n) \in l_p \backslash l_1$ where $p = \alpha + 1$ and set $t_n = \sum_1^{n-1} \varepsilon_i$, $t_1 = 0$.

The process is defined as follows: Let $X_0 \equiv u_0 \in \mathbb{H}_1$. Then $X_{n+1} = u(\varepsilon_n)$ where u is

the solution of

$$\frac{du}{dt} = -D_x G(u, V_n) \tag{2.1}$$

and $u(0) = X_n$. We make the following assumptions:

A1. *For any $u(0) \in \mathbb{H}_1$ there exists $T > 0$ and C, depending only on $\|u_0\|_{\mathbb{H}_1}$, such that the equation (2.1) is solvable in the interval $[0,T]$, the solution $u \in C^\alpha([0,T];\mathbb{H}_1)$ is unique in this interval and $\|u\|_{C^\alpha([0,T];\mathbb{H}_1)} < C$.*

A2. *For every bounded set $B \subset \mathbb{H}_1$ exists a C such that for every $x, y \in B$, $z \in \mathbb{H}_1$ and $v \in \Omega$, $\left|\left\langle D_x G_{x,v}, z\right\rangle\right| \le C\|z\|_{\mathbb{H}_1}$ and $\left|\left\langle D_x G_{x,v} - D_x G_{y,v}, z\right\rangle\right| \le C\|x-y\|_{\mathbb{H}_1}\|z\|_{\mathbb{H}_1}$. Clearly, the same estimates also hold for F, i.e., $\left|\left\langle DF_x, z\right\rangle\right| \le C\|z\|_{\mathbb{H}_1}$ and $\left|\left\langle DF_x - DF_y, z\right\rangle\right| \le C\|x-y\|_{\mathbb{H}_1}\|z\|_{\mathbb{H}_1}$.*

A3. *For almost every $v \in \Omega$, $Q_v = D_x G_{(-,v)} - DF$ maps $\mathbb{H}_1$ to $\mathbb{H}_1$. The family (Q_v) is uniformly bounded on bounded sets in $\mathbb{H}_1$ and $\operatorname{ess\,sup}_{v\in\Omega} \|Q_v(x) - Q_v(y)\|_{\mathbb{H}_1} \le C\|x-y\|_{\mathbb{H}_1}$*

Assumption A3 implies that although $D_x G$ maps $\mathbb{H}_1$ into $\mathbb{H}_{-1}$, the "stochastic part" of $D_x G$ is a bounded random operator into $\mathbb{H}_1$, namely $D_x G - DF$ maps $\mathbb{H}_1$ to $\mathbb{H}_1$.

Evidently, by A1, the process is well defined if we have an a–priori bound on $\|X_n\|_{\mathbb{H}_1}$ for ε_n which are sufficiently small. We shall refer to the natural extension of (X_n) into a continuous orbit $X(t)$, $t \ge 0$ by $X(t) = u(t - t_n)$ for $t \in (t_n, t_{n+1})$ where u is the solution of (2.1) subjected to $u(0) = X_n$. Therefore, if $X_1 = x$ then every sequence (v_n) induces a time continuous sample path.

We make the following additional assumption:

A4. *There is a $K \subset \mathbb{H}_1$ and C such that if $X_0 \in K$ a.s., the process (2.1) is well defined and $\sup_{\mathbb{R}^+} \|X(t)\|_{\mathbb{H}_1} < C$ holds for almost every time continuous sample path.*

We limit the discussion to the case where the initial conditions are selected from K.

Recall the definition of a version of the Palais–Smale condition (P.S.) (see [MW]):

Definition 2.2. *Let $F \in C^1(\mathbb{H}_1, \| \ \|_{\mathbb{H}_0})$. We say that F satisfies the P.S. condition if from the fact that $F(x_n) \to \lambda$ and $DF(x_n) \to 0$ in $\mathbb{H}_{-1}$ follows that λ is a critical value of F and (x_n) contains a subsequence which converges weakly to a critical point in K_λ.*

Our main result is:

Theorem 2.3. *For almost every sample path of the process (2.1), $\big(F(x_n)\big)$ converges and there exists a subsequence n_k along which*

$$\lim_{k \to \infty} D_x F(x_{n_k}) = 0$$

holds in $\mathbb{H}_{-1}$. Also, if F satisfies the P.S. condition, then for almost every sample path (x_n) there exists a critical value λ of F such that all the limit points of (x_n) are contained in K_λ.

Corollary 2.4. *Under the condition of Theorem 2.3, if F satisfies the P.S. condition and the set K_λ is totally disconnected in $\mathbb{H}_0$ for every critical value λ then for almost every sample path (x_n) there is a critical value λ and $x \in K_\lambda$ so that $x_n \to x$ in $\mathbb{H}_0$.*

The proof of this Corollary is identical to the proof of the corresponding part in Theorem 1.1, by setting

$$\Omega_{(x_n)} = \bigcap_{n \geq 0} \overline{\bigcup_{k \geq n} (x_k)}$$

where the closure is in $\mathbb{H}_0$. Since $\mathbb{H}_1$ is compactly embedded in $\mathbb{H}_0$ then $\Omega_{(x_n)}$ is compact and connected in $\mathbb{H}_0$. Hence $\Omega_{(x_n)} \subset K_\lambda$, therefore it consists of a single point.

To prove Theorem 2.3 we introduce:

Lemma 2.5. *For almost every continuous time sample path $x(t)$ the series*

$$\sum_{n=1}^{\infty} \int_{t_n}^{t_{n+1}} \big\langle DF_{x(t)}, DG_{x(t),v_n} - DF_{x(t)} \big\rangle dt$$

converges.

Proof: Let $x(t)$ be induced by the sequence of inputs $(v_1, v_2, ...)$. Note that

$$\beta_n = \int_{t_n}^{t_{n+1}} \big\langle DF_{x(t)}, DG_{x(t),v_n} - DF_{x(t)} \big\rangle dt = \int_{t_n}^{t_{n+1}} \big\langle DF_{x_n}, DG_{x(t),v_n} - DF_{x(t)} \big\rangle dt +$$
$$+ \int_{t_n}^{t_{n+1}} \big\langle DF_{x(t)} - DF_{x_n}, DG_{x(t),v_n} - DF_{x(t)} \big\rangle dt = (1) + (2)$$

To estimate (2), note that since $x(t)$ is bounded in $\mathbb{H}_1$ then so is $Q_v(x(t)) = DG_{x(t),v_n} - DF_{x(t)}$. Thus, by A2, $|(2)| \leq C \int_{t_n}^{t_{n+1}} \|x(t) - x_n\|_{\mathbb{H}_1}$, hence by A1

$$|(2)| \leq C' \varepsilon_n \sup_{t \in [t_n, t_{n+1}]} \|x(t) - x(t_n)\|_{\mathbb{H}_1} \leq C'' \varepsilon_n^p$$

Next, note that

$$(1) = \varepsilon_n \Big\langle DF_{x_n}, Q_{v_n}(x_n) \Big\rangle + \Big\langle DF_{x_n}, \int_{t_n}^{t_{n+1}} \big(Q_{v_n}(x(t)) - Q_{v_n}(x_n)\big) dt \Big\rangle$$

Again, since by A4 $x(t)$ is bounded in $\mathbb{H}_1$ then by A2, A3 and A1

$$\begin{aligned} &\left| \Big\langle DF_{x_n}, \int_{t_n}^{t_{n+1}} \big(Q_{v_n}(x(t)) - Q_{v_n}(x_n)\big) dt \Big\rangle \right| \leq \\ &\leq \varepsilon_n C' \sup_{t \in [t_n, t_{n+1}]} \|Q_{v_n}(x(t)) - Q_{v_n}(x_n)\|_{\mathbb{H}_1} \leq C \varepsilon_n^p \end{aligned}$$

Since $\sum \varepsilon^p < \infty$ it is enough to show that $\sum_{n=1}^{\infty} \varepsilon_n \big\langle DF_{x_n}, Q_{v_n}(x_n) \big\rangle$ converges for almost every orbit (x_n).

Let $\mathcal{F}_{\mathrm{n}}$ be the σ–algebra generated by $X_1, V_1, ...X_n, V_n$ and put $Y_n = \varepsilon_n \big\langle DF_{X_n}, Q_v(X_n) \big\rangle$. Note that Y_n is $\mathcal{F}_{\mathrm{n}}$ measurable and X_n is $\mathcal{F}_{\mathrm{n}-1}$ measurable. moreover

$$\mathbb{E}\big(Q_v(X_n)|\mathcal{F}_{\mathrm{n}-1}\big) = \mathbb{E}\big(\mathrm{D_x G_{X_n,V_n}} - \mathrm{DF_{X_n}}|\mathcal{F}_{\mathrm{n}-1}\big) = \int_{\Omega} \mathrm{D_x G_{X_n,v}} \mathrm{d}\mu(\mathrm{v}) - \mathrm{DF_{X_n}} = 0$$

Hence,

$$\mathbb{E}(Y_n|\mathcal{F}_{\mathrm{n}-1}) = \varepsilon_{\mathrm{n}} \mathbb{E}\big(\big\langle \mathrm{DF_{x_n}}, \mathrm{Q_v(x_n)} \big\rangle | \mathcal{F}_{\mathrm{n}-1}\big) = \varepsilon_{\mathrm{n}} \big\langle \mathrm{DF_{x_n}}, \mathbb{E}(\mathrm{Q_v(x_n)}|\mathcal{F}_{\mathrm{n}-1}) \big\rangle = 0$$

thus, Y_n forms a martingale difference sequence. By A4, X_n are uniformly bounded in $\mathbb{H}_1$, thus by A2,

$$\big|\big\langle DF_{X_n}, Q_v(X_n) \big\rangle\big| \leq C \|Q_v(X_n)\|_{\mathbb{H}_1} \leq C'$$

therefore, Y_n are uniformly bounded and $\sum_1^{\infty} Var Y_n$ converges absolutely almost surely, implying, just as in section 1, that $\sum_{n=1}^{\infty} Y_n$ converges almost surely.

$\diamond$

Proof of Theorem 2.3: Note that

$$\begin{aligned} F(x_m) - F(x_n) &= \int_{t_n}^{t_m} \frac{\partial F(x(t))}{\partial t} dt = -\sum_{i=n}^{m-1} \int_{t_i}^{t_{i+1}} \big\langle DF_{x(t)}, DG_{x(t),v_i} \big\rangle = \\ &= -\int_{t_n}^{t_m} \big\|DF_{x(t)}\big\|^2 - \sum_{i=n}^{m-1} \int_{t_i}^{t_{i+1}} \big\langle DF_{x(t)}, DG_{x(t),v_i} - DF_{x(t)} \big\rangle \end{aligned}$$

By lemma 2.5, the second term is a converging series and since $(F(x_n))$ is a bounded sequence, it follows that $F(x_n)$ converges and $\int_0^\infty \left\|DF_{x(t)}\right\|^2 < \infty$. In particular, there is a subsequence x_{n_k} such that $\left\|DF_{x_{n_k}}\right\| \to 0$.

To prove the second part of the theorem, assume that there exists an $\mathbb{H}_0$ limit point of (x_n), denoted by x, such that $x \in \mathbb{H}_1$ but is not a critical point of F. Then there is a $\delta > 0$ and a radius R such that $\|DF_y\| \geq \delta$ on $N = \{y| \, \|y - x\|_{\mathbb{H}_0} \leq R\} \cap \mathbb{H}_1$. Indeed, if this is not the case, there is a sequence $(y_n) \subset \mathbb{H}_1$ such that $\|y_n - x\|_{\mathbb{H}_0} \to 0$ and $DF_{y_n} \to 0$. By the P.S. condition there is a subsequence (y_{n_j}) converging weakly to a critical point of F – thus x must be a critical point.

Assume that $\|x_n - x\|_{\mathbb{H}_0} < \rho < R$ and let $T^* = \inf\limits_{\tau > 0} \left\{ \|x(t_n + \tau) - x\|_{\mathbb{H}_0} = R \right\}$, i.e., $T^* \leq \infty$ is the minimal time required to leave N given that $x(t_n) = x_n$. We now use the uniform $\alpha-$Hölder estimate on the solutions (assumption A1) to show that T^* cannot be too small. If T^* is smaller than the interval of existence T of (2.1) (c.f. assumption A1), then $\|x(t_n + T^*) - x_n\| < C(T^*)^\alpha$ by A1 and

$$R = \|x(t_n + T^*) - x\|_{\mathbb{H}_0} \leq \|x(t_n + T^*) - x_n\|_{\mathbb{H}_1} + \rho \leq C(T^*)^\alpha + \rho$$

Therefore, $T^* \geq \left(\frac{R - \rho}{C}\right)^{\frac{1}{\alpha}}$. This implies that

$$\int_{t_n}^{\infty} \left\|DF_{(x(t))}\right\|^2 dt > \int_{t_n}^{t_n + T^*} \left\|DF_{(x(t))}\right\|^2 dt > \delta \left(\frac{R - \rho}{C}\right)^{\frac{2}{\alpha}}$$

which contradicts the convergence of $\int_0^\infty \left\|DF_{x(t)}\right\|^2 dt$.

$\diamond$

3 – Applications for reaction-diffusion equations

In this section we demonstrate an application of the results from section 2 to a class of reaction-diffusion equations of the form

$$\frac{\partial U}{\partial t} = AU + f(U, V_t) \quad , \quad u(0) = u_0 \in \mathbb{H}_0 \tag{3.1}$$

where A is an unbounded, self-adjoint negative operator in a Hilbert space $\mathbb{H}_0$ which generates the continuous semigroup $e^{tA} : \mathbb{R}^+ \times \mathbb{H}_0 \to \mathbb{H}_1$.

For any $i > 0$, define $\mathbb{H}_i$ in terms of the spectral family of projection valued measures $\mathbb{P}_\lambda$ of the operator A (see [RS]). Let $\sigma(A)$ be the spectrum of A, then $\phi \in \mathbb{H}_i$ if and only if

$$\|\phi\|^2_{\mathbb{H}_i} \equiv \int_{\sigma(A)} (1+|\lambda|^{2i})^{1/2} \langle \phi, d\mathbb{P}_\lambda(\phi) \rangle < \infty$$

Assume that $\mathbb{H}_0 = \mathbb{L}^2(\Omega)$ where $\Omega \subset \mathbb{R}^d$ and A is a uniformly elliptic operator in Ω with appropriate boundary conditions (One may keep in mind, for example, $A = \Delta$ is the Laplacian with Dirichlet boundary conditions and $\mathbb{H}_1 = W_0^{1,2}(\Omega)$).

As in the previous section the parameter V is a random function of time give by $V_t = v_n$ if $t \in [t_n, t_{n+1})$, where

$$t_0 = 0 \quad , \quad \lim_{n\to\infty} t_n = \infty \quad , \quad \sum_{n=0}^{\infty} (t_{n+1} - t_n)^{3/2} < \infty$$

while $V_n \in \Omega$ are distributed according to a probability measure μ.

Assume that for almost every v, $f(s,v)$ is a continuous function with respect to the parameter s. Let $g(s,v)$ be the primitive function $g(s,v) = \int^s f(\rho, v) d\rho$ and put

$$G(U,v) = -\frac{1}{2}\langle AU, U \rangle - \int_\Omega g(U,v) dx$$

then $D_u G(U,v) = -AU - f(U,v)$, hence (3.1) is an example of the process (2.1). Also, we assume that the Nimitzky operator $f(-,v)$ maps $\mathbb{H}_i \cap \mathbb{L}_\infty$, for $i = 0, 1$ into itself uniformly with respect to v. In the case where Ω is bounded, this follows if, for example, $f(-,v) \in C^1_{loc}(\mathbb{R})$ uniformly with respect to v, and when Ω is unbounded, if $f = f(-,x,v) \in C^1_{loc}(\mathbb{R}, \Omega)$ uniformly in v and satisfies certain decay properties where $x \to \infty$. Moreover, assume that there exists a function $M > 0$, independent of v, such that $Uf(U,-) < 0$ whenever $|U| > M$.

Denote by $\overline{f} = \int f(u,v) d\mu(v)$ and set $F(u) = \int G(u,v) d\mu(v)$. Clearly $u\overline{f}(u) < 0$ for $|u| > M$ and $\overline{f}$ maps $\mathbb{H}_1 \cap \mathbb{L}_\infty$ into itself as well. Thus, F is defined on $\mathbb{H}_1 \cap B$ for any bounded set $B \subset \mathbb{L}_\infty(\Omega)$. Moreover, by the maximum principle [PW] the solution of

$$\frac{\partial u}{\partial t} = Au + \overline{f}(u) \quad , \quad u(0) = x_0 \tag{3.2}$$

and the solution U of (3.1) both satisfy

$$\|U(-,t)\|_\infty < M' \quad ; \quad \|u(-,t)\|_\infty < M' \tag{3.3}$$

for any $t \geq 0$ provided $\|x_0\|_\infty < M'$.

Note that the steady states of (3.2) are the critical points of F and let $\mathcal{K}^{(1)}_{M'} = \mathbb{H}_1 \cap B_{M'}$, where $B_{M'}$ is the ball of radius M' in $\mathbb{L}_\infty(\Omega)$. By the maximum principle, all the steady states of (3.2) are in the set $\mathcal{K}^{(1)}_M$, which implies that the set of the critical points of F are in $\mathcal{K}^{(1)}_M$ as well.

We shall show that under our assumptions, the process (2.1) converge a.s. to a critical point of F, i.e., to a solution $w \in \mathbb{H}_1$ of $Aw + \overline{f}(w) = 0$ provided that the set of its solutions is totally disconnected (for example, if the number of solutions is countable). To that end we show that the assumptions A1–A4 are fulfilled, hence the assertions of theorem 2.3 and corollary 2.4 hold.

Let the set of initial conditions K (appearing in assumption A4) be $\mathcal{K}^{(1)}_{M'}$ for some $M' > M$. For this K, assumption A4 is verified. Indeed, the solution U of (3.1) is bounded by M' in $\mathbb{L}_\infty$ by the maximum principle. By the assumed properties of f we obtain that $f(U(-, V_t))$ is bounded uniformly in time with respect to the $\mathbb{H}_0$ norm.

Consider the *linear* equation

$$\frac{dU}{dt} = AU + g(-,t) \quad , \qquad U(0) = x_0 \in \mathbb{H}_1 \tag{3.4}$$

where $g(-,t) = f(U, V_t)$. Since $g \in \mathbb{L}_\infty(\mathbb{R}^+; \mathbb{H}_0)$ it follows that $U \in \mathbb{L}_\infty(\mathbb{R}^+; \mathbb{H}_1)$. Together with the uniform $\mathbb{L}_\infty$ estimate we have on U and the assumptions on f we obtain that $g \in \mathbb{L}_\infty(\mathbb{R}^+; \mathbb{H}_1)$.

Next we show that assumption A1 holds in an abstract Hilbert space setting. To verify A1 it is enough to show that $U \in C^{1/2}(\mathbb{R}^+; \mathbb{H}_1)$. This is obtained by a uniform estimate of the solution in $\mathbb{H}_2$:

Proposition 3.1. *The solution U of (3.1) satisfies $U \in \mathbb{L}_\infty(\mathbb{R}^+, \mathbb{H}_2)$.*

The proof of the Hölder continuity of U in the $\mathbb{H}_1$ norm follows from proposition 3.1 in two steps. In the first step, we use the $\mathbb{H}_2$ estimate on U to show that U is Lipschitz in the $\mathbb{H}_0$ norm. Indeed, Let $z_\tau(t) = U(t+\tau) - U(t)$. Since $\|U\|_{\mathbb{H}_2}$ and $\|AU\|_{\mathbb{H}_0}$ are equivalent then

$$\begin{aligned} \frac{1}{2}\frac{\partial}{\partial \tau}\|z_\tau(t)\|^2_{\mathbb{H}_0} &= \left\langle z_\tau(t), \frac{\partial U(t+\tau)}{\partial \tau} \right\rangle = \left\langle z_\tau, AU(t+\tau) + g(t+\tau) \right\rangle \leq \\ &\leq \|z_\tau\|_{\mathbb{H}_0} \left(\|g(t+\tau)\|_{\mathbb{H}_0} + \|U(t+\tau)\|_{\mathbb{H}_2} \right) \end{aligned}$$

Thus, since $g = f(U, v_t)$ is uniformly bounded in $\mathbb{H}_1$ and by the proposition, $U \in Lip(\mathbb{R}^+; \mathbb{H}_0)$. From the interpolation $\|z_\tau\|_{\mathbb{H}_1} \leq C \|z_\tau\|^{1/2}_{\mathbb{H}_0} \|z_\tau\|^{1/2}_{\mathbb{H}_2}$ follows that $U \in C^{1/2}(\mathbb{R}^+; \mathbb{H}_1)$, as required. The proofs of the remaining assumptions A2 and A3 are self-evident.

Proof of Proposition 3.1: We may assume $x_0 = 0$ because it differs from the actual solution by a solution of the homogeneous equation, which belongs to $\mathbb{H}_k$ for every $k > 0$. Using the presentation (3.4) with $g(t) \equiv f(U, V_t)$ and since $g \in \mathbb{L}_\infty(\mathbb{R}^+; \mathbb{H}_1)$ then $U(t) = \int_0^t e^{(t-s)A} g(s) ds$. Denote by $\mu_\phi(d\lambda) = \langle \phi, d\mathbb{P}_\lambda \phi \rangle$ the spectral measures associated with $\phi \in \mathbb{H}_0$. Hence

$$d\mathbb{P}_\lambda U(t) = \int_0^t e^{(t-s)\lambda} d\mathbb{P}_\lambda g(s) ds \tag{3.5}$$

Taking the inner product of $U(t)$ with (3.5) and since $d\mathbb{P}_\lambda$ are orthogonal projections, it follows that

$$\begin{aligned} \mu_{U(t)}(d\lambda) &= \Big\langle \int_0^t e^{(t-s')A} g(s') ds', \int_0^t e^{(t-s)\lambda} d\mathbb{P}_\lambda g(s) ds \Big\rangle = \\ &= \int_0^t \int_0^t e^{(t-s)\lambda} e^{(t-s')\lambda} \big\langle g(s'), d\mathbb{P}_\lambda g(s) \big\rangle ds ds' \end{aligned}$$

By the Cauchy-Schwartz inequality,

$$\big\langle g(s'), d\mathbb{P}_\lambda g(s) \big\rangle = \big\langle d\mathbb{P}_\lambda g(s'), d\mathbb{P}_\lambda g(s) \big\rangle \leq \big\langle g(s'), d\mathbb{P}_\lambda g(s') \big\rangle^{1/2} \big\langle g(s), d\mathbb{P}_\lambda g(s) \big\rangle^{1/2}$$

and since the spectrum of A is negative then

$$\begin{aligned} \mu_{U(t)}(d\lambda) &\leq \Big(\int_0^t e^{(t-s)\lambda} \mu^{1/2}_{g(s)}(d\lambda) \Big)^2 \leq \Big(\int_0^t e^{(t-s)\lambda} ds \Big) \Big(\int_0^t e^{(t-s)\lambda} \mu_{g(s)}(d\lambda) ds \Big) \leq \\ &\leq -\lambda^{-1} \int_0^t e^{(t-s)\lambda} \mu_{g(s)}(d\lambda) ds \end{aligned}$$

Therefore, if $\lambda_0 = \sup \sigma(A) < 0$ then

$$\begin{aligned} &\|U(t)\|^2_{\mathbb{H}_2} = \int_{\sigma(A)} \lambda^2 \mu_{U(t)}(d\lambda) \leq - \int_{\sigma(A)} \int_0^t \lambda e^{(t-s)\lambda} \mu_{g(s)}(d\lambda) ds \leq \\ &- \int_0^t e^{(t-s)\lambda_0} \int_{\sigma(A)} \lambda \mu_{g(s)} d\lambda ds \leq - \int_0^t e^{(t-s)\lambda_0} \|g(s)\|^2_{\mathbb{H}_1} ds \leq -\lambda_0^{-1} \sup_{s \in \mathbb{R}^+} \|g(s)\|^2_{\mathbb{H}_1} \end{aligned}$$

$\diamond$

4 – Weak convergence of Lyapunov systems

In this section we make a few observations on the question of weak convergence of Lyapunov systems defined by (1.1) in p–smooth strictly convex Banach spaces. Throughout this section we assume that X_n are uniformly bounded almost surely, that H is a Lipschitz function and that $\varepsilon_n < 1$ for every n. We derive our results from the following maximal lemma:

Lemma 4.1. *Let τ be the measure induced by the process (1.1). As in proposition 1.4, denote by $X(t)$ the linear interpolation of X_n, put $\hat{X}_n = X(t_n + t)$ and set $\overline{X}_n(t)$ to be the solution of (1.3) given that $\overline{X}_n(0) = X_n$. Then $\tau\Big(\{\sup_{1\le i\le m}\Big\|\overline{X}_n(t_i) - \hat{X}_n(t_i)\Big\| \ge \lambda\}\Big) \le C(\lambda)S_{m,n}$ where $S_{m,n} = \sum_n^{n+m}\varepsilon_i^p$.*

Proof: Recall that $\hat{X}_n(t) - \overline{X}_n(t) = U(t_n + t) - U(t_n) + B_n(t)$. Fix some t_m, then for every $1 \le i \le m$

$$\tau\Big(\{\max_i\Big\|\overline{X}_n - \hat{X}_n\Big\|(t_i) \ge \lambda\}\Big) \le$$
$$\le \tau\Big(\{\max_i\|U(t_i + t_n) - U(t_n)\| \ge \lambda/2\}\Big) + \tau\Big(\{\max_i\|B_n(t_i)\| \ge \lambda/2\}\Big)$$

As in the proof of lemma 1.1, since X_n are uniformly bounded, H is Lipschitz and $\mathbb{B}$ is p–smooth ([P1], [P2]) then by Doob's maximal inequality follows that $\tau\Big(\{\max_i\|U(t_i + t_n) - U(t_n)\| \ge \lambda/2\}\Big) \le C(\lambda)S_{n,m}$. Also, by Chebyshev's inequality $\tau\Big(\{\max_i\|B_n(t_i)\| \ge \lambda/2\}\Big) \le 2\mathbb{E}(\max_i\|B_n(t_i)\|)/\lambda$. Our claim follows since $\|B_n(t_i)\| \le C\sum_n^{n+i}\varepsilon_j^2$

$\diamond$

Theorem 4.2.

a. *If $(\varepsilon_n) \in l_p$ then $\Big\|\overline{X_n} - \hat{X}\Big\|_{L_\infty(X)} \to 0$ uniformly τ-almost surely.*

b. *If $\varepsilon_n \to 0$ then for every $\lambda > 0$ the fraction of time in which $\Big\|X(t_i) - \overline{X}(t_i)\Big\| \ge \lambda$ tends to 0 in probability.*

c. *If $\varepsilon_n \to 0$ then there are subsequences n_k and m_k such that $\sup_{1\le i\le m_k}\Big\|\overline{X}_{n_k} - \hat{X}_{m_k}\Big\| \to 0$ as $k \to \infty$ almost surely. In particular, for every $T > 0$ there are an infinite number of functions $\overline{X}_{n_k}$ and $\hat{X}_{n_k}$ which are arbitrarily close on an interval of time whose length is at least T.*

d. *Assume that $\varepsilon_n = \varepsilon$ for every n and let X_n^ε be the position of the process at the n-th stage. Then $X_n^\varepsilon \to \hat{X}_1(t_n)$ in probability as $\varepsilon \to 0$.*

Proof: (a) follows immediately from the fact that

$$\tau_{n,\lambda} = \tau\Big(\big\{\big\|\overline{X}_n - \hat{X}_n\big\|_{L_\infty(X)} \geq \lambda\big\}\Big) \leq \tau\Big(\big\{\sup_i \big\|\overline{X}_n - \hat{X}_n\big\|(t_i) \geq \lambda\big\}\Big) \leq C(\lambda)\sum_{i=n}^{\infty} \varepsilon_i^p$$

Since $(\varepsilon_n) \in l_p$ then for every $\lambda > 0$, $\tau_{n,\lambda} \to 0$, hence $\big\|\overline{X}_n - \hat{X}_n\big\|_{L_\infty(X)} \to 0$ almost surely. To prove (b), put $Z(t) = \big\|X(t) - \overline{X}(t)\big\|$, fix $\lambda > 0$ and let $g_m = \frac{1}{m}\sum_1^m \chi_{\{Z(t_i)\geq\lambda\}}$. Then $\mathbb{E}g_m = \frac{1}{m}\sum_1^m \tau\Big(\big\{Z(t_i) \geq \lambda\big\}\Big) \leq \frac{C(\lambda)}{m}\sum_1^{1+m} \varepsilon_i^p$ and since $\varepsilon_i \to 0$ then $\mathbb{E}g_m \to 0$ for every $\lambda > 0$. Hence, since g_m are nonnegative they converges in probability to 0.

As for (c), since $\varepsilon_n \to 0$ there are sequences n_k and m_k such that $\sum_{k=1}^\infty \sum_{n_k}^{n_k+m_k} \varepsilon_i^p < \infty$. Set $Z_{n_k}(t) = \big\|\overline{X}_{n_k}(t) - \hat{X}_{n_k}(t)\big\|$.
By the lemma $\sum \tau\Big(\big\{\max_{1\leq i\leq m_k} \big\|\overline{X}_{n_k} - \hat{X}_{n_k}\big\|(t_i) \geq \lambda\big\}\Big) < \infty$. Thus, by the Borel–Cantelli lemma, $\max_{1\leq i\leq t_m} Z_{n_k}(t_i)$ converges a.s. to 0.

Finally, since $\tau\Big(\big\{\big\|X_n - \hat{X}_1(t_n)\big\| \geq \lambda\big\}\Big) = \tau\Big(\big\{\big\|X(t_n) - \hat{X}_1(t_n)\big\| \geq \lambda\big\}\Big) \leq C(\lambda)n\varepsilon$, then X_n converges in probability to $\hat{X}_1(t_n)$.

$\diamond$

Corollary 4.3. *If $\varepsilon_n \to 0$ and $\lim_{t\to\infty} \overline{X}(t)$ exists and belongs to a set C for almost every initial condition then the fraction of time in which $d(X(t), C) \geq \lambda$ tends to 0 in probability.*

Introduction to supervised learning models

The following two chapters are devoted to supervised learning models. 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. The first supervised learning model thoroughly investigated was the Perceptron (see [MP],[H]). In this model, both the student S and the teacher T are halfspaces in $\mathbb{R}^d$. Given an input $v \in \mathbb{R}^d$, 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. Note that 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.

In the following chapters, we examine several learning processes. We formulate the well known Perceptron convergence theorem and give a counterexample which demonstrates that the Perceptron may not converge if the distance between the teacher and the set of inputs is 0. From this follows that the Perceptron may not be a proper learning process – if one wishes the student to converge to the teacher. Therefore, we suggest an alternative learning model, in which (again) both the teacher and the student are halfspaces and discuss its convergence properties. The approach we take in both cases in mostly geometric and using that approach we show that the second model is useful in cases where the Perceptron fails. Finally, we turn to a general on-line learning model. In this case the response of the system is determined by an on-line error function, denoted by $\mathcal{E}(x, v)$. It is a nonnegative function such that for every state x (which represents a student) and every input v, $\mathcal{E}(x, v) = 0$ if and only if x gives a correct response to the input v, i.e., $\mathcal{E}(x, v) = 0$ if and only if x agrees with the teacher on the input v.

An example of such a model is the on-line Gibbs learning, first suggested in [KS]. The idea behind the on-line Gibbs learning is to describe the conditional transition density from state to state – given the input at the n-th stage using 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. We introduce a new on-line learning process, investigate its convergence properties and compare it to the on-line Gibbs learning.

Chapter 2 – Linear separation models

0 – Introduction

The most elementary model for a two layer network consists of an input layer and an output layer with a single neuron. The output neuron has two possible responses: it can either fire or not fire, and it fires if and only if the total input it receives surpasses a given threshold. Denote by s the threshold of the output neuron and let w_i be the synaptic weight representing the strength of the interaction between the i-th neuron in the input layer and the output neuron, which fires if and only if

$$\sum_i w_i v_i > s \tag{0.1}$$

Let $U = (w_1, ..., w_n)$ be the "synaptic weights" vector. The neuron fires when $U(v) = \langle U, v \rangle > s$ where $v = (v_1, .., v_n)$ is the input, which implies that the "firing zone" of each neuron consists of vectors which are in the positive halfspace $\{v | U(v) > s\}$. If we assume, for example, that the threshold of every cell is 0 and that $\mathbb{R}^d$ is the set of all the possible inputs, the response of each cell in the output layer is 1 on an open halfspace of $\mathbb{R}^d$ and 0 on its complement, when the boundary of the "firing zone" is a maximal subspace. In the learning process the synaptic weights vector of each neuron is adapted by some pre–determined rule, hence the relevant halfspace changes its location. Given a teacher T and a student S which are positive sides of a given maximal subspaces, $\mathbb{R}^d$ is divided to two sets: on the first one T and S agree – which means that for every input selected from that set, S gives the same response as the T and on the other set T and S disagree.

Left: the Perceptron fires when $\sum v_i w_i > 0$. *Right: On the shaded area* S *and* T *disagree*

figure 0.1

Our goal is to ensure that the student converges to the teacher. To that end, we examine two learning models: the Perceptron learning rule and the projection model.

Although the Perceptron convergence theorem implies that the learning process converges to a correct state, a crucial assumption used in the proof is that the distance between the set of inputs and the boundary of the teacher is positive. This assumption implies that there are many halfspaces which agree with the teacher on every input, thus the limit of the process can be any one of the correct states and not just the teacher.
If we impose that there is a unique correct state, the distance between the boundary of the teacher and the set of inputs must be 0 and the Perceptron convergence theorem fails. Indeed, in the first section we present a counterexample which demonstrates that with probability 1 the Perceptron learning rule does not converge.

To ensure the convergence of the learning process to the teacher, we formulate an alternative linear learning model – the projection model. We show that if one uses this learning rule the student converges to a correct state almost surely without further assumptions on the set of inputs. Thus, if the set of inputs is rich enough to ensure a unique correct state, (which is, of course, the teacher) the process converges to the teacher almost surely.

Finally, we use the fact that we can control the limit of the learning process to construct three layer networks with several neurons in its second layer and a single neuron in the third layer, which can approximate any pre-determined number of bounded convex sets.

Note that the number of cells in the second layer determines the number of halfspaces the network teaches. Of course, each student halfspace may have a different teacher and we assume that each student learns independently of the other.

To simplify our notations, we identify our inner product space $\mathbb{R}^d$ with its dual space and each open halfspace X with the norm 1 functional x such that $X = \{y | \langle x, y \rangle > 0\}$. Clearly, we can assume that both the input space V and the state space are subsets of the sphere S^{d-1}, and we equip V with a probability measure ν which is assumed to be absolutely continuous with respect to the Haar measure on S^{d-1}. For every two halfspaces T and S, let $f_{T,S}(v) : S^{d-1} \to \mathbb{R}$ be defined by $f_{T,S} = \chi_T - \chi_S$, i.e., $f_{T,S}(v) = 0$ if and only if T and S agree on the input v. We say that a state is correct if $\nu\big(\{v \in V | f_{T,S}(v) \neq 0\}\big) = 0$.

1 – The Perceptron

The first learning process we examine is the Perceptron learning rule, which is defined by

$$\tilde{s}_{n+1} = s_n + \varepsilon f_{T,S_n}(v_n)v_n, \quad s_{n+1} = \frac{\tilde{s}_{n+1}}{\|\tilde{s}_{n+1}\|} \tag{1.1}$$

where v_n are i.i.d. random variables representing the input to the system and are distributed according to the probability measure ν, S_n is the student halfspace in the n-th stage, s_n is its representation in S^{d-1} and ε is some positive constant.

All the analysis presented here is in the case $\varepsilon_n = \varepsilon$ for every n, but note that in the case $(\varepsilon_n) \in l_2 \backslash l_1$ and if we put

$$H(s,v) = f_{T,S}(v)v = \big(\chi_T(v) - \chi_S(v)\big)v$$

our process is a stochastic approximation process with the constraint $\|s_n\| = 1$.

Let $\tilde{T}$ be the boundary hyperplane of the teacher T. It is known (see [H]) that if $d(\tilde{T}, V) > 0$ (i.e., if the infimum of the distances between the boundary of the teacher and the set of all the inputs is positive) then the process (1.1) converges to a correct answer in a finite number of learning steps almost surely, and there in an estimate on the number of steps required which depends only on ε and $d(\tilde{T}, V)$. This is the well know Perceptron convergence theorem.

Theorem 1.1. *If $d = d(\tilde{T}, V) > 0$ there exists a random variable $N = N(\varepsilon, d)$ such that for every $n > N$ and every initial condition s, s_n is a correct state. Also, the number of times in which the system changes its position is uniformly bounded.*

However, in this case a limit of the process (1.1) may not be T but some other halfspace which is "far away" from the teacher, yet they agree on almost every input. The following counterexample shows that theorem 1.1 does not hold in cases when the teacher is the unique correct state of the system: with probability 1 the learning rule does not converge at all.

Example 1.2. *Let both the state space and the input set be subsets of S^1. Put $T = \{(x,y)|x < 0\}$ to be the teaching halfspace, hence $t = (-1,0)$ is its representation in S^1, and assume that $V = T \cap S^1$ with the probability measure induced by the Haar*

measure on S^1 (see figure 1.1). The idea behind the construction is as follows: since the set of correct states has 0 measure, then with probability 1 the process does not enter that set in a finite number of steps. As for an infinite number of learning steps, the orbits of the process oscillate around t. Thus, for almost every orbit the process does not converge to t due to overshooting. Formally, note that for every $s \in S^1$ and $v \in V$, $f_{T,S}(v) \geq 0$. Let $IC_S = \{v | f_{T,S}(v) > 0\}$, i.e., IC_S is the set of inputs on which S and T disagree. The process converges to a correct state if and only if $s_n \to t$, which implies that $d(IC_{S_n}, t) \to \sqrt{2}$. Assume that (v_n) is a sequence of inputs such that $s_n \to t$ and w.l.o.g. $v_n \in IC_{S_n}$ for every n, hence $d(v_n, t) \to \sqrt{2}$. By taking a converging subsequence of inputs with a limit u and passing to a limit in (1.1) we obtain that $t = \frac{t+\varepsilon u}{\|t+\varepsilon u\|}$. Since $\|u\| = 1$ and since t and u are linearly dependent then either $t = u$ or $t = -u$, which is impossible since $\|t - u\| = \sqrt{2}$.

$\diamond$

figure 1.1

Example 1.2 shows that in some sense the Perceptron learning rule with a constant learning step is not useful when the task at hand is to approximate a given halfspace. It also demonstrates that the important property used in the Perceptron convergence theorem is that all the possible inputs are "far away" from the boundary of the teacher, which implies that the set of correct states has a positive measure. In the following section we formulate another learning model in which both the student and the teacher are halfspaces, however, in this model, the student converges to the teacher when the system has a unique correct state. This learning rule previously appeared in [KS] as an example to the on-line Gibbs learning process. Although their treatment deals with a more general case, it contains several gaps and differs from the approach presented here.

2 – The projection model

The model we suggest is as follows: let S_n, T be halfspaces and select $v \in S^{d-1}$ such that $f_{T,S_n}(v) \neq 0$. We define

$$s_{n+1} = \frac{s_n - \langle s_n, v \rangle v}{\left\| s_n - \langle s_n, v \rangle v \right\|} = \frac{P_{v^\perp}(s_n)}{\|P_{v^\perp}(s_n)\|} \tag{2.1}$$

where $P_{v^\perp}$ is the orthogonal projection on the space orthogonal to v. If $f_{T,S_n}(v) = 0$ then $s_{n+1} = s_n$.

Our main result concerning the convergence of the process (2.1) is the following:

Theorem 2.1. *Let C be the set of the correct states and assume that the probability measure ν by which the inputs are selected is absolutely continuous with respect to the Haar measure on S^{d-1}. Then for almost every orbit the process (2.1) converges and its limit belongs to C. In particular, if there is a unique correct answer, the process converges almost surely to the teacher.*

As an example (see figure 2.1 below), assume that the process (2.1) takes place in $\mathbb{R}^2$ and that $V = S^1$. In this case, as shown in figure 2.1 below, the angle between s_n and t is a decreasing function of n. It is also clear that by selecting inputs close to T, s_n moves arbitrarily close to t, which implies that the process 2.1 converges to t almost surely.

figure 2.1

The proof of theorem 2.1 goes along the same lines as the example: first we show that $\|s_n - t\|$ is monotone decreasing and then we prove that the distance between s_n and t becomes arbitrarily small with probability 1.

Proof: Clearly, C is a closed set and for every $x, y \in C$, $f_{X,S} = f_{Y,S}$ almost surely. Put $t \in C$ and let $h_t(s_n, v) = \|s_n - t\| - \|s_{n+1} - t\|$. Note that if $f_{T,S}(v) \neq 0$ then

$\langle s,v\rangle\langle t,v\rangle \le 0$. Since $\|P_{v^\perp}(s_n)\| \le \|s_n\| = 1$ then

$$\langle s_{n+1},t\rangle = \langle s_n,t\rangle / \|P_{v^\perp}(s_n)\| - \langle s_n,v\rangle\langle v,t\rangle / \|P_{v^\perp}(s_n)\| \ge \langle s_n,t\rangle$$

hence

$$\|s_n - t\| - \|s_{n+1} - t\| = h_t(s_n,v) \ge 0 \tag{2.2}$$

Therefore, the sequence $(\|s_n - t\|)$ is decreasing and bounded, which implies that for every $\varepsilon > 0$ and every $t \in C$, the ball $B_\varepsilon(t)$ is an absorbing set. Thus, to prove our claim it is enough to show that for almost every sequence (s_n) and every $\varepsilon > 0$, there is some $T \in C$ such that $s_n \in B_\varepsilon(t)$ for some n, i.e., that $s_n \in \bigcup_{x\in C} B_\varepsilon(x)$.

Put $A_s = \{v | f_{T,S}(v) \ne 0\}$ and $A_{n,s} = \{v | \, |f_{T,S}(v)|\, h_t(s,v) \ge 1/n\}$, let $g(s) = \nu(A_s)$ and $g_n(s) = \nu(A_{n,s})$. Clearly, for every s, $A_{n,s}$ is an increasing sequence and $\bigcup_n A_{n,s} = A_s$, thus g_n is a monotone sequence which tends to g pointwise. Note that both g_n and g are continuous functions. Indeed, if $s_m \to s$ then $S^{d-1} \cap S_m \to S^{d-1} \cap S$ in the Hausdorff metric, hence f_{T,S_m} converges to $f_{T,S}$ almost surely – which implies that g is continuous. By the same argument, $|f_{T,S_m}(v)|\, h_t(s_m,v)$ converges almost surely to $|f_{T,S}(v)|\, h_t(s,v)$, therefore, up to a set of zero measure, $\limsup_{m\to\infty} A_{n,s_m} \subset A_{n,s}$. Also, since $\nu\big(\{v | \, |f_{T,S}(v)|\, h_t(s,v) = 1/n\}\big) = 0$ then $A_{n,s} \subset \liminf_{m\to\infty} A_{n,s_m}$ up to a set of zero measure. Thus

$$\limsup_{m\to\infty} \nu\big(A_{n,s_m}\big) \le \nu\big(\limsup_{m\to\infty} A_{n,s_m}\big) \le \nu\big(A_{n,s}\big) \le \nu\big(\liminf_{m\to\infty} A_{n,s_m}\big) \le \liminf_{m\to\infty} \nu\big(A_{n,s_m}\big)$$

hence for every $t \in C$ $\lim_{m\to\infty} g_n(s_m) = g_n(s)$. In particular, by Dini's theorem $g_n \to g$ uniformly.

Fix $\varepsilon > 0$ and $t \in C$. Since g is positive on the compact set $S^{d-1} \setminus \bigcup_{x\in C} B_\varepsilon(x)$ and since $g_n \to g$ uniformly, then there are $\delta > 0$ and N such that $g_n(s) \ge \delta$ for every $n > N$ and every $s \in S^{d-1} \setminus \bigcup_{x\in C} B_\varepsilon(x)$. Thus, if $s_m \notin \bigcup_{x\in C} B_\varepsilon(x)$ there is a set $I_{s_m} \subset V$ such that $\nu(I_{s_m}) \ge \delta(\varepsilon)$, and for every $v \in I_{s_m}$ $\|s_{m+1} - t\| \le \|s_m - t\| - 1/n$. From this follows that almost every orbit (s_m) must enter $\bigcup_{x\in C} B_\varepsilon(x)$.

$\diamond$

So far, we formulated a linear learning process with the feature that the student S converges to the teacher T if we assume that the set of inputs is rich enough to allow a unique correct state. Next, we show how in that case the process (2.1) may be used to learn complex geometric shapes. This feature is important since one of the disadvantages found in the Perceptron learning rule was that it could be used only in systems which are linearly separable (see [MP],[H]). Here, we show that it is possible for a network to learn the most general shape possible, which is in this case, a pre–determined number of convex sets. We shall approximate every convex set from the inside by a polytope. Since a polytope is the intersection of halfspaces we can use the process 2.1 to ensure convergence to the halfspaces determining the polytope.

First, note that theorem 2.1 may be used even when the boundary of either the teacher's or the student's "firing zone" is not a maximal subspace but some hyperplane. In a neural network that boundary is a maximal subspace if and only if the threshold of the output unit is 0. Therefore, given a Perceptron for which the output unit has a threshold s, we add a unit to the input layer which receives the input -1, its synaptic interaction with the output unit is s and the output unit is assumed to have a threshold 0.

From a geometric point of view, we identify our copy of $\mathbb{R}^d$ with the set $X = \{x \in \mathbb{R}^{d+1} | x_{d+1} = -1\}$, extend each hyperplane in X to a maximal subspace in $\mathbb{R}^{d+1}$ and continue the analysis in $\mathbb{R}^{d+1}$. Now, both the teacher and the student have boundaries which are maximal subspaces, thus, both the process (2.1) and theorem 2.1 can be extended to include the case where the boundaries are hyperplanes. Our next goal is to show how this may be used to "teach" complex shapes to a given network. To this end we demonstrate how a convex set may be approximated by an intersection of a fixed number of halfspaces.

Assume that $K \subset \mathbb{R}^d$ is a bounded convex set, denote by $|\ |$ the Lebesgue measure on $\mathbb{R}^d$ and let $\varepsilon > 0$. Suppose that we can construct a polytope P which has $\alpha(\varepsilon, d)$ facets such that $K \subset P$ and $|P \backslash K| < \varepsilon$. Therefore $P = \bigcap_1^\alpha T_i$ where T_i are halfspaces. Let N be a network which contains α units in its second layer, such that every Perceptron determines one of the halfspaces. The network has a single neuron in the third layer with a threshold $\alpha - 1 < s < \alpha$ and its synaptic interaction with every cell in the second layer is 1. Thus, a vector v belongs to P if and only if every cell in the second layer of this network fires in response to v and this occurs if and only if the neuron in the output layer fires. Given

$S_1, ..., S_\alpha$ halfspaces and since S_i and S_j learn independently, then by theorem 2.1 if we adapt S_i using T_i then $\bigcap_1^\alpha S_i \to \bigcap_1^\alpha T_i$, i.e., the limit network N determines the polytope P which approximates K.

The following theorem allows us to construct the set of teaching halfspaces which determine P and gives a bound on the number of vertices required for this construction. From this follows that such a bound also exists on the number of facets needed, which is the number of neurons required in the second layer of the adapting network.

Theorem 2.2 ([GMR]). *For every bounded convex set $K \subset \mathbb{R}^d$ and a given number of vertices m, one can construct a polytope P with m vertices contained in K such that the volume ratio $\frac{|K \setminus P|}{|K|} \leq \frac{f(d)}{m^{\frac{2}{d-1}}}$.*

In particular, for every $K \subset B_1(0)$ and $\varepsilon > 0$, one can construct a polytope P with $m = m(\varepsilon, d)$ vertices at the most, such that $P \subset K$ and $|K \setminus P| < \varepsilon$.

Corollary 2.3. *Let $K_1, ...K_l \subset B_1(0)$ be convex sets. Then for every $\varepsilon > 0$ it is possible to construct $T_1, ...T_n$, $n \leq l\alpha(\varepsilon, d)$ teachers which determine $P_1, ...P_l$ such that $\sum |P_i \setminus K_i| < \varepsilon$. A network with at most $l\alpha(\varepsilon, d)$ neurons in its second layer adapted by the process (2.1) using the teachers T_i will converge, and for every $v \in V$ and $1 \leq i \leq l$ the limit network can determine if $v \in P_i$.*

3 – Conclusions

We demonstrated that the Perceptron learning rule may be problematic when the only correct state is the teacher, while the projection model converges to that correct state. Thus, the projection model is useful when one wishes to ensure that the student converges to the teacher. Then, we used the fact that every bounded convex set may be approximated by a polytope with a pre–determined number of facets to construct a set of teaching halfspaces which determine the approximating polytope. Using the projection model for each halfspace separately, it is possible for the network to "learn" the approximating polytope. Hence, a 3–layer network with a pre–determined number of units in its second layer and with a fixed synaptic interactions between the second and third layers can adapt and approximate every bounded convex set.

Chapter 3 – General on-line learning models

0 – Introduction

In this chapter we discuss an on-line learning process. The behavior of such a process is determined by an on-line error function $\mathcal{E}(x,v)$ which is a nonnegative function whose domain is the product space of the state space and the input set. $\mathcal{E}(x,v)=0$ if and only if the student x agrees with the teacher on the input v. In such a process x is a correct state if it agrees with the teacher on almost every input, i.e., $\mathcal{E}(x,v)=0$ almost surely. Clearly, this is the same as having $\mathbb{E}_g(x)=\int_V \mathcal{E}(x,v)d\nu(v)=0$. The function $\mathbb{E}_g$ is called the global error function.

In the general case, there may not be a correct state. Therefore, the goal of the learning process is to converge to the global minimum of $\mathbb{E}_g$. To that end, we formulate a process for which, on average, every learning step decreases $\mathbb{E}_g$. Unfortunately, as shown in later sections, even those processes might not converges to the global minimum of $\mathbb{E}_g$.

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 response 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, hence, it is easy to guarantee that the student converges at least to a local minimum of $\mathbb{E}_g$. In [KS] the authors claimed that the on-line learning model they introduced, called "On–line Gibbs learning" had the capabilities of the non on-line models: it converges to a minimum of $\mathbb{E}_g$. Moreover, they claimed that the convergence is to a global minimum of $\mathbb{E}_g$, both when the on-line error function is smooth and when it is a $0-1$ function. In section 4 we show that there are several difficulties with the results stated in [KS] and that some of them are not true.

The on-line model we introduce 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 in this chapter is on the process with a constant learning step, but we also give an example of a convergence theorem for the non homogeneous case under some additional assumptions on the on-line error function.

This chapter is divided to five sections. In the first one we define our model and state most of the convergence results concerning the homogeneous process. We also present

several examples in which the process may be used, two of which are the well known Perceptron and the multi layer Perceptron learning rules (see [H],[RMS]). In the second section, we give an example of a convergence theorem in the non-homogeneous case. 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 section we prove the results stated in the first section and the fourth contains an analysis of the main results from [KS]. We show that in some cases, there are counterexamples to claims concerning the convergence of the on-line Gibbs learning. We end the chapter with some concluding remarks.

Let us turn to some definitions and notations: throughout the section (X, μ) is a compact metric probability space. X^n is the product space of n copies of X with the induced topology and measure. (V, ν) is the compact metric space of all possible inputs where ν is a probability measure on V. Denote by τ the product measure $\mu \times \nu$ on $X \times V$. For a random variable X, $\mathbb{E}X$ is the expectation of X, and $\|X\|_1 = \mathbb{E}|X|$. $B_\lambda(x)$ is the closed ball of radius λ centered at x, and for a set A, $\overline{A}$ denotes its closure. Finally, we say that $(a_n) \in l_p$, if $\sum |a_n|^p < \infty$.

1 – 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. We begin with the following notations:

For every $x \in X$, put $C_x = \{v | \mathcal{E}(x, v) = 0\}$ and denote by $\mathcal{O} \subset X$ the set of local minima of the energy function $\mathbb{E}_g(x)$.

Our process is A Markov process (X_n, V_n), where (V_n) are i.i.d. which are distributed according to ν and independent of (X_n), while (X_n) are adapted using the conditional transition density from x to x' given the input v which is

$$P_n(x'|x, v) = \frac{1}{c_n(x)} \begin{cases} \mathcal{E}(x, v) e^{-\mathcal{E}(x', v)/T_n} & \mathcal{E}(x, v) > 0 \ , d(x, x') \leq \lambda \\ \delta_{x,x'} & otherwise \end{cases} \tag{1.1}$$

$c_n(x) = \nu(C_x) + \int_{B_\lambda(x)} \int_V \mathcal{E}(x, v) e^{-\mathcal{E}(x', v)/T_n} d\nu(v) d\mu(x)$ is a normalizing constant, T_n is a sequence decreasing to 0, and λ is the size of the maximal learning step.

The definition implies that the state x does not move if x responds in a correct way to the input v and it never moves to a point for which $d(x, x') > \lambda$.

The 0-temperature process is the process for which $P(x'|x,v) = \lim_{n\to\infty} P_n(x'|x,v)$ and the limit is in the pointwise sense. A key part in the analysis of the model 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 our process by the 0-temperature process. Throughout we Assume that $\nu(C_x \Delta C_{x'})$ is continuous in each variable separately, where $A\Delta B = (A\cap B^c)\cup(B\cap A^c)$. We also assume that there is a $\delta > 0$ such that $\nu(C_x) \geq \delta$ for every x.

We formulate the main results concerning the process (1.1):

Theorem 1.1. Assume that the error function is a 0-1 function. Then:

a) For every $\lambda > 0$ and every sequence $(T_n) \to 0$ the process (1.1) will enter every neighborhood of $\mathcal{O}$ infinitely often almost surely.

b) Put $Q = \{x|\mathbb{E}_g(x) = 0\}$ and assume that Q has a μ–positive measure. Then for $\lambda = diam(X)$ and for every sequence $(T_n) \to 0$ the process (1.1) converges almost surely to Q.

Note that Q is an absorbing set, i.e., 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:

Theorem 1.2. *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 also that for every $x \notin \mathcal{O}$ and for every $\varepsilon > 0$ there exists a $y \in B_\varepsilon(x)$ such that $C_y \supset C_x$. Then the assertions of theorem 1.1.a remain true. The assertion of theorem 1.1.b holds with no additional assumptions.*

The reason for the additional assumption is simple. When we deal with a 0-1 error function, $\mathbb{E}_g(x) = 1 - \nu(C_x)$. Therefore, if we increase $\nu(C_x)$, we come closer to a minimal point of $\mathbb{E}_g(x)$. On the other hand, for a general error function, we do not know if $C_y \backslash C_x$ and $\mathbb{E}_g(y) - \mathbb{E}_g(x)$ are correlated. In the final section, we present a counterexample (see example 4.2), in which the error function is a smooth nonnegative bounded function, but the orbits can not leave the global maximum of $\mathbb{E}_g(x)$.

Next we present three examples in which model (1.1) may be used. In all three cases, the error function is a 0-1 function.

Example 1.3: The simplest case in which we can apply the process (1.1) is the Perceptron

learning rule (see chapter 2). In this process we adapt a halfspace in $\mathbb{R}^d$ using a teacher which is also a halfspace. Denote the teacher by $T = \{x \in \mathbb{R}^d | y^*(x) > 0\}$ where y^* is a linear functional on $\mathbb{R}^d$, and the student by $S = \{x \in \mathbb{R}^d | z^*(x) > 0\}$, $\|z^*\| = \|y^*\| = 1$. In this case, the error function is 0 for x which are in $T \cap S$ or in $T^c \cap S^c$ and 1 otherwise. Assume that the input set V is a finite union of balls disjoint from the boundary of T and that the probability measure is given by a continuous density function supported on V. (see figure 1.1)

figure 1.1

Clearly, $\nu(C_x \Delta C_y)$ is continuous and the conditions of theorem 1.1.b hold. Therefore, the process (1.1) converges a.s. to $\{x | \mathbb{E}_g(x) = 0\}$.

Example 1.4: Here we present two examples which deal with the uniform approximation of a continuous function $f : V \to \mathbb{R}$ by a function selected from a family of continuous functions $\{g_x : V \to \mathbb{R} | x \in X\}$. Here, X and V are compact sets in $\mathbb{R}^d$ and $\mathbb{R}^k$. Fix $\eta > 0$, and define

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

Of course, if we want to use theorem 1.1, we have to find some kind of continuity condition on the family $\{g_x\}$, which is the object of the following lemma:

Lemma 1.5. *Assume that* $\lim_{r \to s} \|g_r - g_s\|_1 = 0$. *Then* $\nu(C_x \Delta C_y)$ *is continuous with respect to each variable separately.*

Proof: First, we show that for every s, $\nu(C_s \backslash C_r)$ is continuous with respect to r. Since ν is regular, there exist a compact set $K \subset C_s$ such that $\nu(C_s \backslash K) < \varepsilon/2$. Therefore, there is a $\delta > 0$ such that $\sup_{t \in K} |g_s(v) - f(v)| \leq \eta - \delta$. By Chebyshev's inequality, $\nu\{|g_s(v) - g_r(v)| \geq \delta/2\} \leq \frac{\|g_s - g_r\|_1}{\delta/2}$, hence, if s and r are close enough, $\nu(K \cap C_r^c) \leq \varepsilon/2$,

implying that $\nu(C_s \backslash C_r) = \nu(K \cap C_r^c) + \nu\big((C_s \backslash K) \cap C_r^c\big) \leq \varepsilon$. A similar argument shows that $\lim_{r \to s} \nu(C_r \backslash C_s) = 0$.

$\diamond$

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 h. It is formed by using the output of each layer of Perceptrons as an input to the next one. Usually, the output function h of each Perceptron in the MLP is assumed to be smooth. We assume that h Lipschitz and that the network has three layers. Let X be the closed unit ball in $\mathbb{R}^d$, in which case the MLP may be viewed as a function $M : X^n \times X \times \mathbb{R}^d \to \mathbb{R}$. For every $v \in \mathbb{R}^d$ the response of the MLP is $M_{x,y}(v) = h\big(\sum_{j=1}^{n}(y_j h(\sum_{i=1}^{d} x_{ij} v_i)\big)$, where $x \in X^n$ and $y \in X$. If we view the MLP as a neural network, d is the number of cells in the input layer, n is the number of units in the second layer and $(x_{ij})_{i=1}^n$ represents the synaptic weights between the first layer and the j-th cell in the second layer. Hence $h(\sum_{i=1}^n x_{ij} v_i)$ 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 $M_{x,y}(v)$. (see figure 1.2)

figure 1.2

We assume that all the inputs v are selected from some compact set $K \subset \mathbb{R}^d$. It is known that if h is not an algebraic polynomial, the set of all possible MLPs is dense in $C(K)$ (see [LLPS]). Thus, for every $\varepsilon > 0$ there are an n and $\{x_0, y_0\}$ such that $\|M_{x_0,y_0} - f\|_\infty \leq \eta - \varepsilon$, i.e., it is possible to construct an MLP which can $(\eta - \varepsilon)$ approximate f.

We show that the conditions of theorem 1.1.b hold, implying that the MLP converges to a correct state almost surely.

First, to prove that the set of correct states Q has a positive measure, it is enough to show that the set $\{x, y\}$ for which $M_{x,y}$ η–approximates f has a non empty interior. Indeed, if $M_{x_0,y_0} \in Q$ then for some $\varepsilon > 0$, M_{x_0,y_0} $\eta - \varepsilon$ approximates f. Since h is Lipschitz then $\sup_{v \in K} |M_{x_0,y_0}(v) - M_{x,y}(v)| \leq C\left(\|x - x_0\| + \|y - y_0\|\right)$, where C is some absolute constant. Therefore, a small perturbation in $\{x_0, y_0\}$ gives an η–approximation of f. By the same argument, if $\{x_n, y_n\}$ converges to $\{x, y\}$, then M_n converges to M uniformly which implies convergence in the L_1 norm. Therefore, since the conditions of lemma 1.6 hold, $\nu(C_x \Delta C_y)$ is continuous. Thus the process (1.1) converges to a function which η-approximates f.

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

The idea is similar to the one presented above, so most of the details are omitted. There is a 1-1 correspondence between the polynomials of degree $\leq n$ and $\mathbb{R}^n$. Note that it is possible to η–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 λ. For a bound on the degree of the approximating polynomial, we estimate $E_n(f) = \inf_{a_0,\ldots,a_n \perp 1} \sup_{v \in [-1,1]} \left| \sum_{i=0}^{n-1} a_i v^i - f(v) \right|$. By Jackson's theorem ([C]), if $E_n(f) = \eta$ then $n \leq C\frac{\lambda}{\eta}$, 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 $\|p_n'\|_\infty \leq n \|p_n\|_\infty$ where p_n is a polynomial of degree n. Thus, $|a_0| \leq \|p_n\|_\infty$, $|a_1| \leq \left|p_n'(0)\right| \leq \left\|p_n'\right\|_\infty$, and so on. Therefore, for every η, there are an integer n and a vector $(a_1, ..., a_n)$ such that $\sup_{t \in [-1,1]} \left| \sum_{1}^{n} a_i v^i - f(v) \right| < \eta$, where both n and $\|(a_1, ..., a_n)\|_\infty$ depend only on λ and η. A similar argument to the one used in example (a) shows that the conditions of theorem 1.1.b hold. Thus, the process (1.1) converges to a correct state almost surely.

2 – An example of a 0 temperature process

Here, we present an example of a 0 temperature process. In this example we replace the constant learning step λ by a positive sequence λ_n which decreases to 0. We prove that

in the case presented here, the process converges a.s. to a correct state even if at every step there is a probability $p < 1/2$ for the teacher to make a mistake.

Since the learning steps decrease to 0, the Markov process is not homogeneous. To deal with that we use a similar idea to the one Blum used in his proof of the convergence of the n-dimensional Robbins-Monro process (see [W]).

Example: Let $X \subset \mathbb{R}^d$ be a compact set and put $V = [0,1]$, both equipped with the normalized Lebesgue measure. 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:

$$\mathcal{E}(x,v) = \begin{cases} 1 & v > f(x) \\ 0 & v \leq f(x) \end{cases} \tag{2.1}$$

If we assume that in every step the teacher has a probability $p < 1/2$ to make a mistake, then for $y \in B_{\lambda_n}(x)$ the conditional transition density from x to y is

$$G_n(y|x,v) = \frac{1}{d_n}\Big((1-p)\mathcal{E}(x,v)e^{-\mathcal{E}(y,v)/T_n} + p\big(1-\mathcal{E}(x,v)\big)\,e^{-\big(1-\mathcal{E}(y,v)\big)/T_n}\Big) \tag{2.2}$$

where d_n is a normalizing constant. Therefore, in our example the transition operators for the 0 temperature process are:

(2.3)

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

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

Theorem 2.1. *Denote by λ_n the sequence of the learning step and assume that $(\lambda_n) \in l_{d+3}\backslash l_{d+2}$. If f has a unique critical point in X and if that point is a global maximum, the process defined by (2.3) converges in probability to that point. If $p = 0$ the convergence is almost surely.*

Let X_n be the position of our process in the n-th stage. 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 $U_n(x) = \mathbb{E}\big(\langle \nabla f(x), Y_n(x)\rangle | X_n = x\big)$. Let o be the unique point for which $\nabla f(x) = 0$ and assume that o is a global maximum of f.

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

Lemma 2.2. *For every compact set K for which $d(K,o)>0$ there are N and $L>0$ such that for every $n>N$, $\inf_{x\in K} U_n(x)\geq L$. Also, for n large enough $\mathbb{E}U_n\geq 0$. Next, put $a_n=\sup_x \rho_n^2\mathbb{E}\left\|Y_n(x)\right\|^2$. Then $(a_n)\in l_1$.*

Proof: Let S be the unit sphere in $\mathbb{R}^d$ and denote by $|\ |$ 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 $U_n(x)=\frac{1}{\rho}\langle\nabla f(x),\mathbb{E}(X_{n+1}-X_n|X_n=x)\rangle$ it is enough to prove the first claim for the function $U_n(x)c_n(x)$

$$\frac{1}{\rho_n}\Big\langle(1-p)\int_{A_n(x)}(y-x)(f(y)-f(x))dy+p\int_{B_n(x)}(y-x)(f(x)-f(y))dy,\nabla f(x)\Big\rangle=(*)$$

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 $f(y)-f(x)=\langle\nabla f(x),y-x\rangle+O(\left\|y-x\right\|^2)$, hence

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

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

$$\frac{1}{\rho_n}\Big((1-p)\int_{A_n(x)}\langle\nabla f(x),y-x\rangle^2dy-p\int_{B_n(x)}\langle\nabla f(x),y-x\rangle^2dy\Big)=$$
$$\frac{1}{\rho_n}\int_0^{\lambda_n}\Big((1-p)\int_{S^+_{x,r}}\langle\nabla f(x),rs\rangle^2r^{d-1}h(s)ds-p\int_{S^\perp_{x,r}}\langle\nabla f(x),rs\rangle^2r^{d-1}h(s)ds\Big)dr=$$
$$\frac{1}{\rho_n}\int_0^{\lambda_n}r^{d+1}\Big((1-p)\int_{S^+_{x,r}}\langle\nabla f(x),s\rangle^2h(s)ds-p\int_{S^\perp_{x,r}}\langle\nabla f(x),s\rangle^2h(s)ds\Big)dr$$

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

$$g(x,r)=\int_{S^+_{x,r}}\langle\nabla f(x),s\rangle^2h(s)ds-p\int_S\langle\nabla f(x),s\rangle^2h(s)ds=g_1(x,r)-pg_2(x)$$

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, note that in this case if $\lambda_n<R$, then $\frac{1}{\rho_n}\int_0^{\lambda_n}r^{d+1}g(x,r)dr\geq\frac{1}{\rho_n}\int_0^{\lambda_n}r^{d+1}l\geq l/(d+2)$.

Denote $D_{x,v} = \{s \in S | \langle \nabla f(x), s \rangle > v\}$. Clearly, $D_{x,v}$ increases to S_x as v tends to 0 and since $|D_{x,v}|$ and $|S_x|$ are both continuous functions of x, $|D_{x,v}|$ converges uniformly to $|S_x|$ on K by Dini's theorem. Fix $\varepsilon > 0$. There is some v such that for every x, $|S_x \backslash D_{x,v}| = |S_x| - |D_{x,v}| < \varepsilon/2M$, where $M = sup_x \|\nabla f(x)\|$. Since ∇f is continuous, there is a $\delta_x > 0$ such that for every $y \in B_{\delta_x}(x)$, $D_{x,v} \subset D_{y,v/2}$ and $\nu(C_x \Delta C_y) < \varepsilon/2M$. Note that if $\|y - x\| < \delta_x$, $s \in D_{x,v}$ and $z = y + rs$ then $f(z) - f(y) = \langle \nabla f(y), s \rangle r + o(r^2) \geq vr/2 - Cr^2$, where C is some uniform constant. Hence, there is an $R > 0$ such that if $r < R$ and $y \in B_{\delta_x}(x)$, then $D_{x,v} \subset S^+_{y,r} \cap D_{y,v/2} \subset S_y$.

For such r and y we see that

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

on the other hand for every y

$$p g_2(y) = p \int_S \langle \nabla f(y), s \rangle^2 h(s) ds = 2p \int_{S_y} \langle \nabla f(y), s \rangle^2 h(s) ds$$

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, so $g(y,r) > (1-2p)C - \varepsilon$. The rest follows from a standard compactness argument.

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

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

Therefore, by Fatou's lemma

$$\liminf_{n \to \infty} U_n(x) c(x) = \liminf_{n \to \infty} \frac{1}{\rho_n} \int_0^{\lambda_n} r^{d+1} g(x,r) dr \geq \frac{1}{\rho_n} \int_0^{\lambda_n} r^{d+1} \eta(x) dr = \eta(x)/(d+2)$$

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

Turning to the final claim, note that

$$\rho_n^2 \mathbb{E}\left\|Y_n(x)\right\|^2 \leq \frac{1}{(1-p)f(x)+p(1-f(x))} \int_{A_n(x)\cup B_n(x)} \left\|y-x\right\|^2 \left|f(y)-f(x)\right| dy \leq$$
$$\leq C \int_{A_n(x)\cup B_n(x)} \left\|y-x\right\|^2 \left|\left\langle \nabla f(x), y-x\right\rangle + O(\left\|y-x\right\|^2)\right| dy = (*)$$

For every $n > N$ and every x, $(*) \leq C' \int_{B_{\lambda_n}(x)} \left\|y-x\right\|^3 dy = C''\lambda_n^{d+3}$, where C,C' and C'' are some absolute constants. Thus, $\sum_N a_n \leq C'' \sum_N \lambda_n^{d+3} < \infty$.

$\diamond$

Corollary 2.3. *If $\mathbb{E}U_n$ converges to 0, $\mathbb{E}\left|U_n\right|$ converges to 0 too.*

Proof: For every $\varepsilon > 0$ fix a compact set $K \subset X$, such that $\mu(X\backslash K) < \varepsilon$ and $d(K,o) > 0$. By the lemma, for n large enough, 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. Thus:

$$\mathbb{E}\left|U_n\right| = \int_{X\backslash K} \left|U_n\right| d\mu + \int_K U_n d\mu \leq$$
$$= \mathbb{E}U_n + 2\int_{X\backslash K} \left|U_n\right| d\mu < \mathbb{E}U_n + 2\varepsilon$$

$\diamond$

Proof of Theorem 2.1: 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

$$f(X_{n+1}) = f(X_n + \rho_n Y_n) = f(X_n) + \rho_n \left\langle \nabla f(X_n), Y_n\right\rangle + 1/2\rho_n^2 \left\langle Y_n, H(X_n + \theta\rho_n Y_n)Y_n\right\rangle$$

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

$$\mathbb{E}\left(f(X_{n+1})|X_n\right) = f(X_n) + \rho_n U_n(x) + 1/2\rho_n^2 V_n(x) \tag{2.4}$$

Taking expectations on both sides and iterating, we see that

$$\mathbb{E}\left(f(X_{n+1})\right) = \mathbb{E}f(X_1) + \sum_1^n \rho_i \mathbb{E}U_i + \sum_1^n 1/2\rho_i^2 \mathbb{E}V_i \tag{2.5}$$

For every sequence (x_n), $f(x_n)$ is bounded by 1, thus $\mathbb{E}\left(f(X_{n+1})\right)$ is bounded. By the Cauchy-Schwarz inequality $\left|V_i\right| \leq C \sup_x \mathbb{E}\left\|Y_i(x)\right\|^2$, which implies that $\rho_i^2 \mathbb{E}\left|V_i\right| \leq$

$C\rho_i^2 \sup \mathbb{E}\left\|Y_i(x)\right\|^2 = Ca_i$. By Lemma (2.2) $(a_i) \in l_1$, so $\sum_1^n 1/2\rho_i^2 \mathbb{E}V_i$ converges, therefore $\sum_1^n \rho_i \mathbb{E}U_i$ is bounded. Again, by the lemma, $\sum \rho_i \mathbb{E}U_i$ is a nonnegative series for i large enough so it must converge – which implies that $\mathbb{E}\big(f(X_n)\big)$ converges too. Since $(\rho_n) \notin l_1$, there is a subsequence $\mathbb{E}U_{n_j}$ which tends to 0 and by corollary 2.3 $\mathbb{E}\left|U_{n_j}\right|$ tends to 0 too. Using Chebyshev's inequality U_{n_j} converges in probability to 0, thus there is a subsequence of U_{n_j}, also denoted by U_{n_j}, which converges a.s to 0. According to lemma (2.2) U_n are uniformly bounded away from 0 on every compact set not containing o, therefore X_{n_j} must converge a.s. to o. Since $\mathbb{E}\big(f(X_n)\big)$ converges and since it is a continuous function of X_n, it must converge to $\mathbb{E}f(o) = f(o)$. Note that o is a unique maximum of f, hence for every $\varepsilon > 0$ there is a $\delta > 0$ 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 $p = 0$, $f(X_n)$ is increasing a.s. – therefore, it converges almost surely. Since $\mathbb{E}\big(f(X_n)\big)$ converges to $f(o)$, $f(X_n)$ must converge to $f(o)$ a.s., thus, since o is a unique maximum, X_n must converge to o almost surely.

$\diamond$

3 – Proofs of the results from part 1

Our next goal is to proof the results stated in section 1. Recall the following notations: 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

$$P_n(A|x) = \frac{1}{c_n(x)} \int_{A \cap B_\lambda(x)} \int_V \mathcal{E}(x,v) e^{-\mathcal{E}(x',v)/T_n} d\mu(x') d\nu(v) = \frac{f_n^A(x)}{c_n(x)} \tag{3.1}$$

converges pointwise to

$$P(A|x) = \frac{\int_{A \cap B_\lambda(x)} \int_V \mathcal{E}(x,v) \chi_{C_{x'}} d\nu(v) d\mu(x')}{\nu(C_x) + \int_{B_\lambda(x)} \int_V \mathcal{E}(x,v) \chi_{C_{x'}} d\nu(v) d\mu(x')} = \frac{f^A(x)}{c(x)} \tag{3.2}$$

Let $\mathcal{P}^0$ be the probability measure induced by the orbits (X_n) of the 0-temperature process (3.2) and $\mathcal{P}$ is the induced measure by the orbits of the process (3.1).

We will show that the convergence of P_n to P is uniform in both x and A. First, we prove that $f_n^A(x)$ converges uniformly in both x and A to $f^A(x)$. With a similar argument one shows 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. Indeed: w.l.o.g. assume that $\mathcal{E}(x',v)$ is bounded by 1 and set $E = \{(x',s)|\mathcal{E}(x',s) > 0\}$. 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 $K \subset E$ such that $\tau(E \backslash K) < \varepsilon$ and $\mathcal{E}(x',v) > \beta$ on K. Then

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

$$= \sup_{x \in X} \int_E \mathcal{E}(x,v) \left| e^{-\mathcal{E}(x',v)/T_n} - \chi_{C_{x'}}(v) \right| d\tau = (*)$$

Since the integrands are uniformly bounded by 2 and $\tau(E \backslash K) < \varepsilon$ then:

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

Clearly the estimates above are uniform in the set A, which proves our claim.
Another observation which follows using a similar computation is that for every A the function $P(A|x)$ is continuous. Moreover, for every measurable $A_2, A_3, ..., A_n$ the function $P(X_n \in A_n, ..., X_2 \in A_2 | X_1 = x)$ is continuous in x. To prove this fact, we use the assumption that for every x, $\nu(C_x \Delta C_{x'})$ is a continuous function of x'.

Next, our aim is to use information concerning the 0-temperature process (3.2) to derive similar results about process (3.1). We begin with some additional notations. If (X_n) denotes an orbit then for every set $O \subset X$ put $O_i = \{(X_n)|X_i \in O \ for \ n = i\}$, $L_n^0(x,O) = \mathcal{P}^0\{X_i \in O \ for \ some \ i \geq n | X_n = x\}$, $L_n(x,O) = \mathcal{P}\{X_i \in O \ for \ some \ i \geq n | X_n = x\}$, and $L^0(x,O)$ is the $\mathcal{P}^0$ probability to enter O infinitely often given that $X_1 = x$. Assume that O has the following property: there are $\alpha > 0$ and N, such that for every $x \in X$, $\mathcal{P}^0\{\cup_1^N O_i | X_1 = x) > \alpha$. Since the 0–temperature is a homogeneous Markov process, $\mathcal{P}^0\{\cup_m^{m+N} O_i | X_m = x) > \alpha$ for every m and every x, and since P_n converges uniformly to P then for m large enough and for every x, $\mathcal{P}\{\cup_m^{m+N} O_i | X_m = x\} > \alpha/2$. Hence for m large enough and every x, $L_n(x,O) > \alpha/2$.

Lemma 3.1. *If there are N and α such that for every $n > N$ and every x $L_n(x,O) > \alpha$ then the orbits of the process (3.1) enter O i.o. $\mathcal{P}$–almost surely.*

This result appears in [O] in a slightly weaker form. The proof uses the same idea as the one presented in [O] 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 [L]: Let $Y_1, Y_2, ...,$ be a sequence of random variables and let Y be a random variable defined

on $Y_1, Y_2, ...$ such that $\mathbb{E}|Y| < \infty$. Note that $Z_n = \mathbb{E}(Y|Y_1, ..., Y_n)$ forms a martingale, thus, by the martingale convergence theorem ([L], pg. 393), Z_n converges a.s. to Y. In particular, if we set $B_i = \{X_i \in O\}$, $B = \{X_n \in O \ i.o.\}$, $Y_n = X_n$ and $Y = \chi_B$, then $\mathcal{P}(B|Y_1, ..., Y_n) = \mathbb{E}(Y|Y_1, ..., Y_n)$ converges a.s. to χ_B and $\mathcal{P}(\cup_k^\infty B_i|Y_1, ..., Y_n)$ tends to $\chi_{\cup_k^\infty B_i}$ for every fixed k.
On the other hand, for every $k \leq n$, note that

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

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

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

Again, taking $k \to \infty$, the left side converges a.s. to χ_B, hence $\mathcal{P}(\cup_n^\infty B_i|Y_1, ..., Y_n)$ tends to χ_B.

Denote by X_∞ the set of all the orbits of the process.
Since $L_n(X_n, O) = \mathcal{P}(\cup_n^\infty B_i|Y_1, ..., Y_n)$ then by the 0-1 law $L_n(X_n, O)$ tends to the characteristic function of the set $\{X_n \in O \ i.o.\}$. By our assumption for n large enough and every x, $L_n(x, O) > \alpha$, thus for such n $L_n(X_n, O) > \alpha$ almost surely. Therefore,

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

Hence, almost every orbit enters O infinitely often almost surely.

$\diamond$

Remark 3.2: 1. By a similar method one shows that if there are N and $\alpha > 0$ such that for $n > N$ and every x $L_n(x, B) > \alpha$, then $\mathcal{P}$–almost surely the orbits which visits A infinitely often also visit B infinitely often.
2. Lemma 3.1 implies that in order to prove theorem 1.1.a, it is enough to show that for every neighborhood A of $\mathcal{O}$ there are N and $\alpha > 0$ such that for every x, $\mathcal{P}^0(\cup_1^N A_i|x) > \delta$.

Proof of Theorem 1.1: We begin with the proof of (b). Let A be an open set containing Q. Note that since $\lambda = diam X$ and since Q has a μ–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 $x \notin A$, $P(Q|x) \geq \delta$. Therefore,

for n large enough $\inf_{x \notin A} P_n(Q,x) \geq \delta/2$. Hence, by remark 3.2, 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 process (3.1) converges $\mathcal{P}3$–a.s. to Q.

To prove (a) we use the second part of remark 3.2. Assume that we limit the size of the learning step to λ and let A be a neighborhood of $\mathcal{O}$. Recall that $A_i = \{(X_n)|X_i \in A \text{ for } n = i\}$ and that $\mathcal{P}^0(A^i \backslash \cup_1^{i-1} A_j | X_1 = x) = \mathcal{P}^0(X_i \in A, X_{i-1} \in A^c, ..., X_2 \in A^c | X_1 = x)$ is a continuous function of x. Since $\mathcal{P}^0(\cup_1^n A_i | X_1 = x) = \sum_1^n \mathcal{P}^0(A_i \backslash \cup_1^i A_j | X_1 = x)$ then for every n, $h_n(x) = \mathcal{P}^0(\cup_1^n A_i | X_1 = x)$ are continuous functions too.

Note that if we show that for every x there is some n 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 $h_n(x) > \varepsilon/2$. Since X is compact and $\big(h_n(x)\big)$ is a monotone increasing sequence, we can take a finite sub-cover (U_{x_i}) and find $\alpha > 0$ and N such that for every x, $h_N(x) > \alpha$. Hence, it is enough to show that $L^0(x,A) > 0$ for every x, since this implies that for every x there is an n such that $h_n(x) > 0$.

Indeed, for every x, let y_x be a point in $B_\lambda(x)$ in which the maximum of $\nu(C_x)$ is attained. 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 if $\nu(C_x) < \nu(C_y)$ there is a positive transition density from x to y. By the continuity of the transition density, x has a positive probability to enter A which implies that $L^0(x,A) > 0$.

$\diamond$

Theorem 1.2 – Sketch of proof: The proof of theorem 1.2 goes along the same lines as the proof of theorem 1.1. The only difference is in proving that $L^0(x,A) > 0$. 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 $B_\lambda(x_0)$. According to our assumption, the only elements which are maximal in their neighborhood are elements of $\mathcal{O}$.

The proof of 1.2.b is identical to that of 1.1.b and does not require any additional assump-

tions.

◇

4 – Discussion

This investigation was motivated by the work of Kim and Sompolinsky [KS]. We begin this section with a summary of their main results.

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

$$P_{n,m}(x'|x,v) = \frac{1}{c} e^{\frac{\perp E_n(x',x,v)}{2T_m}} \tag{4.1}$$

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.

The main results presented is [KS] are as follows:
1. In the limit $\lambda_n \to \infty$ and $T_m \to 0$ the process (4.1) converges in distribution to a measure supported on the set of global minima of $\mathbb{E}_g$. The rates by which the limits are taken is not stated.
2. In the limit $T_m \to 0$ and for a large enough λ, process (4.1) behaves like the 0 temperature process both when $\mathcal{E}(x,v)$ is a C^2 function and when it is a 0-1 function.
3. Using simulations, the authors analyzed several well known learning models (for example, the Perceptron and the committee machine) even in cases where at every stage the teacher has a probability $p < 1/2$ to make a mistake. It was claimed that the 0 temperature process converges to an optimal solution and estimates on the optimal convergence rates were given.

There are several difficulties with the results stated in [KS]. To demonstrate this, we will construct two on-line error functions. The 0–temperature process induced by the first on-line error function will be a counterexample to (1), since it does not converge to a global minimum of the energy function. The 0–temperature process induced by the second error function implies that (1) and (2) may contradict each other in the case where the

error function is C^2, since in this case the global maximum of the energy function is an absorbing state.

Example 4.1: 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 4.1:

figure 4.1

Here, $\mathcal{E}(x, v) = 0$ on the shaded area and 1 otherwise. Note that he global minimum of $\mathbb{E}_g$ in X is attained at $x = 2$ and at 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 is supported in [-1/2,-1/4] and $\lambda_n > 32$, the 0 temperature process converges 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 0-temperature does not converge in distribution to the global minimum of $\mathbb{E}_g(x)$. Moreover, for some initial distributions and for every sequence λ_n such that each $\lambda_i > M$, the process (4.1) converges almost surely to a local minimum of $\mathbb{E}_g$ which is not the global minimum in contradiction to (1).

Example 4.2: 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. Therefore, it is impossible that both (1) and (2) hold for this process.

Define the error function on D by $\mathcal{E}(x, v) = \begin{cases} \frac{2(1-x^2)(v+x^2-1)}{x^2+1} & v > 1 - x^2 \\ 0 & v \leq 1 - x^2 \end{cases}$. 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 $0 \leq v \leq 1$, $\mathcal{E}(0,v) = 0$.

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.

5 – Concluding remarks

Note that an easy way to generalize part (a) of theorem 1.1 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 4.2). For that, one needs additional assumptions on the structure of the error function.

Let us point out that the reason for the assumption $\lambda_n \to \infty$ in [KS] was to overcome the possibility that the teacher makes a mistake. We did not treat this problem outside the case presented in section 2 and it deserves additional consideration. Our final remark concerns theorem 2.2. 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 may not be true.

Chapter 4 – The Kohonen learning rule

0 – Introduction

For several years researchers have been puzzled by the self organizing property of the adaptive model suggested by Kohonen in [K].

The basic idea behind the Kohonen model is simple: given a set of elements and an input, the input influences a neighborhood of the nearest point to it in the given set. If the inputs are selected at random, the given set is adapted to form an approximation of the set of inputs.

Formally, let $K_n = \{K_n^1, ...K_n^i, ..\}$ be the given set, $C_{k_n^i}$ is the neighborhood of $K_n^i \in K_n$, $\tilde{v}_n$ is the input and $P_{K_n}(\tilde{v})$ is the nearest element to $\tilde{v}$ in K_n. Denote by C_n the neighborhood of $P_{K_n}(\tilde{v}_n)$. Then the Kohonen adaptive process is:

$$K_{n+1}^i = K_n^i + \varepsilon_n \chi_{c_n}(\tilde{v}_n)(\tilde{v}_n - K_n^i) \tag{0.1}$$

where χ_{c_n} is the characteristic function of the set $C_{P_{K_n}(\tilde{v})}$ and (ε_n) is a positive sequence representing the size of the learning step.

Usually, the process is divided into two parts. In the first one, the learning step ε_n is assumed to be constant – which means that the process is a homogeneous Markov process. After the adapting set becomes organized in some sense, (ε_n) decreases to 0 fast enough to ensure the convergence of the process.

Let us give a simple example of this process. Assume that the given set is some finite set $\left\{K_0^1, \ldots, K_0^m\right\} \subset [-1, 1]$ with the metric $d(x, y) = |x - y|$. For every index $1 < i < m$, the neighborhood of K_t^i is $\left\{K_t^{i-1}, K_t^i, K_t^{i+1}\right\}$ while $C_{K_t^1} = \left\{K_t^1, K_t^2\right\}$, $C_{K_t^m} = \left\{K_t^{m-1}, K_t^m\right\}$ and the set from which inputs are selected is [-1,1] equipped with the Lebesgue measure. It can be shown (see [K]) that if (ε_n) is a constant sequence the set $\left\{K_0^1, \ldots, K_0^m\right\}$ adapts to a set $\left\{K^1, \ldots, K^m\right\}$ organized in a monotone order almost surely. The proof of this claim is very difficult and the idea behind it can not be applied to other examples.

An example with some biological significance may be found in [RMS]. It describes a monotone ordering of cells within a one–dimensional layer of the bat's auditory cortex.

In this model, five one-dimensional layers simulate an area in the auditory cortex. A neighborhood of a cell contains cells which are close enough to it, i.e., that the distance between them and the cell is smaller than some given $r > 0$.

Each cell has an initial value between 20 and 100 assigned to it which describes the best frequency in Khz to which the cell is tuned. The probability measure on [20,100] is concentrated around a small neighborhood of 61Khz.

A computerized simulation of the Kohonen process shows that each layer becomes linearly ordered and most of the cells are tuned to a best frequency near 61Khz.

In both examples the set enters an "organized" state – which in this case is a linear order. In general, two major questions are open: the convergence of the process, and its self organizing abilities. In this chapter we investigate the second question. We prove a general recurrence theorem which guarantees that the process always enters an organized state, and from this theorem we easily derive the previously mentioned result in the one–dimensional case.

Note that there is no clear way to define "organization" in a multi–dimensional case, at least, not as obvious interpretation as linear order on the interval $[-1,1]$. We give this term a very broad interpretation which may be applied in many cases. Furthermore, we introduce a simple $\mathbb{R}^n$ analog to the concept of linear ordering in $\mathbb{R}$. This chapter is divided to four sections. The first section consists of our generalization to Kohonen's model and several notations. In the second and third sections we prove a recurrence theorem and then give examples for its use in proving self–organization results.

1 – The basic model

The Kohonen adaptive process in made up of two sets: one is the set we wish to adapt and the other one is the set of possible inputs. Our first assumption is that both these sets are contained in some finite dimensional ormed space and that the set of inputs is compact.

Denote the input set by $\tilde{V}$, the state space (i.e. - all the possible positions our adaptive sets can arrive at) by S and the adapting set in the n-stage by K_n. Given $\tilde{v} \in \tilde{V}$, let $P_{K_n}(\tilde{v})$ be the set of nearest points to $\tilde{v}$ in K_n and put $i(x)$ the index of x in K_n. Our learning process is:

$$x_{n+1} = x_n + \varepsilon_n f_n(i(P_{K_n}(\tilde{v}))(\tilde{v}_n - x_n) \tag{1.1}$$

where $x_n \in K_n$, $\varepsilon_n > 0$ and $f_n : \mathbb{N} \to [0,1]$.

this implies that the value of f_n depends only on the index of the nearest point in K_n to the selected input $\tilde{v}$. We influence our adaptive process through the functions f_n, for example, define f_n to be 1 for a nearest point to the input and zero otherwise. A more interesting possibility is to change not only the nearest point, but also points which are near in some topological sense to elements of $P_{K_n}(\tilde{v})$.

figure 1.1

Note that if $\varepsilon_n \not\to 0$ the process (1.1) does not converge pointwise. In sections 1–3 we deal only with the question of organization so throughout those sections (ε_n) is assumed to be a constant sequence.

Clearly, if all the functions f_n are equal then the Kohonen learning process is a stochastic approximation process. Since the nearest point map is not smooth the methods developed in chapter 1 can not be applied in this case. However, most of the results concerning the convergence of the process (1.1) are derived using general results in stochastic approximation.

Let $\tilde{\mu}$ be a regular Borel probability measure on $\tilde{V}$ through which the inputs are selected. Let V be the space of input vectors, (i.e. $V = \prod \tilde{V}$ with the product topology) equipped with the measure μ which is defined by i.i.d. copies of μ. Since $\tilde{V}$ is compact, so is V.

Given $v \in V$ and $x \in \mathbb{R}^d$, denote by x^v the orbit of $x_1 = x$ when $v = (\tilde{v}_1, \ldots)$ is the sequence of inputs to the process (1.1). Thus, x_n^v is the position x arrived at after n steps.

Assume that the adapting sets are finite and that each has N elements. Hence, the state space is $S = \{(x_1, \ldots, x_N) \,|\, \|x_i\| \leq M\}$ and the metric on S is $\| \ \|_\infty$, implying that S is compact. Let $\mathcal{B}$ be a Borel σ-algebra and ν a measure on S. Denote by $\tau = \mu \times \nu$ the probability measure on $(S^\infty, \mathcal{B}^\infty)$, which are the infinite product of S with the induced σ-algebra on the product space. Hence, τ is the probability measure induced by the Markov

process on the set of possible orbits.

Finally, we assume that $\tilde{\mu}\left\{\tilde{v}\in\tilde{V}|\,\|x-\tilde{v}\|=\|y-\tilde{v}\|\right\}=0$ for all $x\neq y\in\mathbb{R}^d$. The reason for this final assumption is that the set of points in which the n-th stage of the process in not continuous is contained in $\bigcup_{x,y\in K_n}\left\{\tilde{v}\in\tilde{V}|\,\|x-\tilde{v}\|=\|y-\tilde{v}\|\right\}$, which is assumed to be of zero measure.

2 – The recurrence theorem

In this section our aim is to prove a general recurrence theorem which shows that the adapting set visits a desired set in the state space infinitely often almost surely.

Lemma 2.1. *Let $O\in\mathcal{B}$ which is open in S. Set $h_O:S\to\mathbb{R}$ by:*

$$h_O(s)=\mu\left(\{s \text{ enters } O \text{ after a finite number of steps}\}\right)$$

Then h is a lower semicontinuous function.

Proof: Let $B_s=\left\{\tilde{v}\in\tilde{V}|s \text{ enters } O \text{ after a finite number of steps}\right\}$. We separate B_s into disjoint sets $B_{s,i}=\{s \text{ first enters } O \text{ in the i}-\text{th step}\}$. The main observation in the proof is that for every $\varepsilon>0$ and every i, there exists a neighborhood U_i of s and $C_{s,i}\subset B_{s,i}$ for which $\mu\left(B_{s,i}\backslash C_{s,i}\right)<\varepsilon$, such that if $r\in U_i$ and $v\in C_{s,i}$, then $v\in B_{r,i}$. In other words, it is possible to find a large subset of $B_{s,i}$ such that if the states r and s are close enough, the elements of $C_{s,i}$ bring r into O after i steps. For the sake of simplicity, we shall prove this claim for $i=1$, but the general case follows in a similar fashion.
Fix $\varepsilon>0$. For every set $A\subset B_{s,1}$, let $\delta_A=inf\left\{v\in A, i\neq j\ |\ \left|\ \left\|\tilde{v}-s^i\right\|-\left\|\tilde{v}-s^j\right\|\ \right|\right\}$, where $s=\{s_1,\ldots,s_N\}$. Since $\mu\left\{\tilde{v}|\,\|\tilde{v}-s_i\|-\|\tilde{v}-s_j\|=0, i\neq j\right\}=0$, there is a compact set $C_{s,1}\subset B_{s,1}$ such that $\mu\left(B_{s,i}\backslash C_{s,i}\right)<\varepsilon$ and $\delta_{C_{s,1}}>0$. Clearly, for every r in the set $\left\{r\in S|d(s,r)<\frac{\delta_{C_{s,1}}}{2}\right\}$ and for every $v\in C_{s,1}$, $i(P_s(\tilde{v}_1))=i(P_r(\tilde{v}_1))$. Hence $f(s,x,\tilde{v}_1)=f(r,y,\tilde{v}_1)$ when $x\in s$, $y\in r$ with the same index, implying that $d\left(s_2^v,r_2^v\right)\leq d(s,r)$. Let $\delta_2=inf\left\{d(s_2^v,\partial O),\ \text{when } v\in C_{s,1}\right\}$. Since $C_{s,1}$ is compact, O is open and the map $v\to s_2^v$ is continuous on $C_{s,1}$, it is clear that $\delta_2>0$. Therefore, if $d(s,r)<\delta_2$ then $r_2^v\in O$. Put $\eta=min\left\{\delta_2,\frac{\delta_{C_{s,1}}}{2}\right\}$ and let $U_1=\{r\in S|d(r,s)<\eta\}$. Clearly, if $v\in\ C_{s,1}$ and $r\in U_1$, then $r_2^v\in O$ as claimed.

Next, in order to end the proof, fix $\rho>0$. Since $\sum_{i=1}^{\infty}\mu\left(B_{s,i}\right)<\infty$, there is some n_0 for which $\mu\left(\bigcup_{i=n_0}^{\infty}B_{s,i}\right)<\frac{\rho}{2}$. Choose $\rho_1,\ldots,\rho_{n_0-1}$ such that $\sum_{i=1}^{n_0-1}\rho_i<\frac{\rho}{2}$, let U_i and

$C_{s,i}$ be the sets constructed for $\rho = \rho_i$ and define $U = \bigcap_{i=1}^{n_0-1} U_i$, $C_s = \bigcup_{i=1}^{n_0-1} C_{s,i}$. It is easy to verify that $\mu(B_s \backslash C_s) < \rho$ and that if $r \in U$, $v \in C_s$ then r^v enters O after a finite number of steps. Thus $C_s \subset B_r$, which implies that $\mu(B_s \backslash B_r) < \rho$. Hence, if $r_n \to s$, $\liminf_{n\to\infty} \mu(B_{r_n}) \geq \mu(B_s)$.

$\diamond$

Theorem 2.2. *Let O be an open set in S and $A \subset S$ such that for every $s \in S \backslash A$, $h_O(s) > 0$. Then almost every orbit (s_n^v) either enters O infinitely often, or it converges to A.*

Proof: Put $R_m = \{v | d\big((s_n^v), A\big) \geq \frac{1}{m} \text{ i.o.}\}$ and denote $O_n = \{v | s_n^v \in O\}$. Our claim will follow if we show that for every m, $R_m \subset limsup O_n$. Note that the set $\{s \in S | d(s, A) \geq \frac{1}{m}\}$ is compact, thus by lemma 2.1 h_O attains a positive minimum on that set. From the proof of lemma 3.3.1 (see also [O], pg. 22, proposition 5.1) follows that in this case $R_m \subset limsup\, O_n$ as claimed.

$\diamond$

Remark 2.3: Note that the proof of lemma 2.1 does not use the fact that our process is homogeneous. However, for a general Markov process, even the claim "if every state has a positive probability (which is larger than some δ) to enter O, the process enters O a.s." is wrong. The reason for the failure of the non-homogeneous claim is that since the probability to enter O from a state s depends on the step in which the orbit visits s. Put $\eta_i = \varepsilon_{n+i}$ and assume that one uses the learning steps (η_i). Define $\lambda_n = \inf_{s\in S} \mu\big(\{s \text{ enters } O \text{ in a finite number of steps}\}\big)$. Since lemma 2.1 remains true, $\lambda_n > 0$ for every n. Thus to ensure the validity of theorem 2.2 it is enough to find positive lower bound for (λ_n).

3 – A few examples of self organization

We begin with a re–statement of the example from the introduction.

Example 3.1: Let $(x_0^1, \ldots, x_0^n) \in [-1, 1]^n$. Our object is to use Kohonen's process to order the points according to their index – in either ascending or descending order. Define $\tilde{V} = [-1, 1]$, let $\tilde{\mu}$ be the normalized Lebesgue measure and $S = [-1, 1]^n$. A neighborhood of x^i is the set $C_{x^i} = \{x^{i-1}, x^i, x^{i+1}\}$ for $1 < i < n$, and $\{x^1, x^2\}$, $\{x^{n-1}, x^n\}$ for $i = 1$,

$i = n$. Let $f_n\left(s_n, \left(x^i\right)_n, \tilde{v}_n\right) = \begin{cases} 1 & x_n^i \in P_{s_n}(\tilde{v}_n) \\ 0 & \text{otherwise} \end{cases}$.

During the process, a neighborhood of each of the nearest points to the input is adapted by the learning step and the rest of the points remain stationary. Note that the set of "organized" states (i.e. - $x^1 < x^2 < \ldots < x^n$ or $x^1 > x^2 > \ldots > x^n$) is an open set in S and let $A = \{(x_1, ...x_n) | x_i = x_j \text{ for some } i \neq j\}$. We show that for every $s \notin A$, $h_O(s) > 0$, hence the assertion of theorem 2.2 holds. The key point in our proof is the fact that from any given state s and for every $x \in s$, the set $\{\tilde{v} | P_s(v) = x\}$ has positive measure. We do not give a complete proof to the claim, rather, we demonstrate how a simple disorder may be corrected.

Suppose that $s = (x^1, \ldots, x^4)$ and $x^1 < x^4 < x^2 < x^3$. First, we move x^2 to its rightful position – by selecting any input $\tilde{v}$ for which $P_s(\tilde{v}) = x^1$. The new state is

$$s_1 = (x^1 + \varepsilon(\tilde{v} - x^1), x^2 + \varepsilon(\tilde{v} - x_2), x^3, x^4) \tag{3.1}$$

An easy calculation shows that:

$$d(x^1_{new}, x^2_{new}) = (1 - \varepsilon) d(x^1, x^2) \tag{3.2}$$

Using (3.2) repeatedly, after a finite number of steps, "x^2" takes its place between "x^1" and "x^4". We continue with such a $\tilde{v}$ for which the nearest point is "x^2", which brings "x^3" to its place, giving us an organized state.

Remark 3.2: In example 3.1 the set $\overline{O}$ is made up of all the organized states, and "degenerate" organized states, i.e., sets with the correct ordering, but instead of strict inequality, we may find that $x^i = x^{i+1}$. Note that O is an absorbing set: once an adapting set enters it, the set remains inside O, hence, by theorem 2.2, all the limit points of the process belong to $\overline{O}$ almost surely.

It is easy to see that every state has a positive probability to enter O if and only if every interval has $\tilde{\mu}$ positive probability. One direction of the proof follows in a similar fashion to the argument presented in example 3.1. On the other hand, if $\tilde{\mu}([a, b]) = 0$, the state $(x^2, x^1, \ldots, x^{n-1}, x^n) \subset [a, b]^n$ never becomes organized, since the only way to correct the disorder is to select an input for which $P_K(\tilde{v}) = x_3$ and that event has zero probability.

Example 3.1 may be extended to $\mathbb{R}^d$: let $\tilde{V}$ be a compact set in $\mathbb{R}^d$ with a non-empty interior and set $S = \tilde{V}^N$. Note that an alternative way to describe a linearly ordered set

in $\mathbb{R}$ is to say that the two nearest points to k^i are k^{i-1} and k^{i+1}. Having that in mind, we suggest the following definition:

Definition 3.3. *A state s is called organized, if the two nearest points to k^i are elements in C_{k_i} – defined in example 3.1.*

Let f_n be as in example 3.1, $\tilde{\mu}$ the normalized Lebesgue measure on $\tilde{V}$ and assume that (ε_n) is a constant sequence. Using the same idea as in the example above, it is clear that one can reach O from any initial state $s \notin A$ with a positive probability.
Again, let us emphasize that the key point is that for every $x \in s$, $\tilde{\mu}\{\tilde{v}|P_s(\tilde{v}) = x\} > 0$. Using theorem 2.2, almost every orbit either enters O i.o. or converges to A.
The attempt to generalize this example to the case where S is not necessarily contained in $\tilde{V}$ causes a difficulty, since we have no prior knowledge of the function $P_{s_n}^{-1}$. To use the algorithm mentioned above, it is vital that for every $x \in s_n$, $\tilde{\mu}\left(P_{s_n}^{-1}(x)\right) > 0$. Since we know nothing about the norm on $\mathbb{R}^d$, there is a possibility that $\tilde{\mu}\left(P_{s_n}^{-1}(x)\right) = 0$ (see figure 3.1).

figure 3.1

To solve this problem, we introduce a new selection of $\tilde{\mu}$ and f_n, in what we call "the ace up the sleeve" routine.

Example 3.4: Let $\tilde{V}$,$\tilde{\mu}$ be as in the extension of example 3.1 to $\mathbb{R}^d$. Fix m points $\{y^1, \ldots, y^m\}$ in $\mathbb{R}^d$, let $\tilde{\nu}$ be an atomic measure supported on $\{y^1, \ldots, y^m\}$ such that $\tilde{\nu}(y^i) = \frac{\delta}{m}$, where $0 < \delta < 1$ and set $\tilde{\beta} = \frac{\tilde{\mu}+\tilde{\nu}}{1+\delta}$.

Say that $\mu\left(P_s^{-1}(x^i)\right) = 0$. We define the process to respond to the input y^i as if x^i itself was selected as the nearest point (even though it may not be the case) and the points in the neighborhood of x^i move in some random direction. It is easy to see that for every $s \in S$ and $x^i \in s$, $\tilde{\beta}\left(P_s^{-1}(x^i)\right) \geq \frac{\delta}{n(1+\delta)}$.

Thus, it is possible to use the ordering algorithm and for every initial state, the process arrives to an organized state with β-positive probability. Next, we have to show that lemma 3.1 is still valid. Recall that one of the conditions in the proof was that $\tilde{\tau}\left\{\tilde{v} \in \tilde{V} | \, \|x - \tilde{v}\| = \|y - \tilde{v}\|\right\} = 0$, but since $\tilde{\tau}$ is atomic this assumption does not hold. However, examining the proof closely reveals that only minor technical adjustments are necessary to extend it so that it includes this case too. For example, divide $B_{s,1}$ to two sets: $\left\{v|\tilde{v}_1 = y^i\right\}$ and $\left\{v|\tilde{v}_1 \neq \ y^i\right\}$. For the second set, proceed just as in the original proof and find the set $C_{s,1}$. The required set will be $C_{s,1} \cup \left\{v \in B_{s,1}|\tilde{v}_1 = y^i\right\}$. To conclude, the assertion of theorem 2.2 holds in this case too.

4 – Stochastic Approximation and its application to a smooth Kohonen process

The Kohonen learning rule has two main features. The first one is that the process is a "winner takes all" process, in the sense that the nearest point to the input is selected. The other feature is that only an "index topology" neighborhood of the nearest point is adapted. Both these facts indicate that this process is not smooth, which makes its analysis difficult. Our aim is to formulate a smooth process which has similar properties to those presented above. Let

$$X_{n+1}^i = X_n^i + \varepsilon_n \left(\sigma(X_n^{i-1} + X_n^{i+1} - 2X_n^i) + \frac{e^{\frac{\perp \left\| X_n^i \perp v_n \right\|^2}{T}}}{\sum_{j=1}^m e^{\frac{\perp \left\| X_n^j \perp v_n \right\|^2}{T}}}(v_n - X_n^i)\right) \tag{4.1}$$

where X_n^i is the i-th element in the set X_n, $V \subset \mathbb{R}^d$ is the compact set of inputs with the probability measure μ and $(\varepsilon_n) \in l_2 \backslash l_1$.

Note that X_{n+1}^i is influenced by the neighborhood of X_n^i in the index topology using a diffusion type interaction.

This process is smooth because it is derived from the smooth stochastic Lyapunov function

$$G_T(X, v) = \frac{\sigma}{2} \sum_{i=1}^m \left\| X^{i+1} - X^i \right\|^2 - \frac{T}{2} log \sum_{i=1}^m e^{\frac{\perp \left\| X^i \perp v \right\|^2}{T}} \tag{4.2}$$

and the second term in (4.1) is a smooth approximation to the "winner takes all" condition in the Kohonen learning rule.

The process (4.1) has a continuous version when $X_n \in \mathbb{H}_1(S^1, \mathbb{R}^d)$, i.e., X_n is a periodic function on $[0, 2\pi]$. Then the process (4.1) becomes:

$$X_{n+1}(s) = X_n(s) + \varepsilon_n \Big(\sigma \Delta X_n(s) + \frac{e^{\frac{\perp \| X_n(s) \perp v_n \|^2}{T}}}{\int_{S^1} e^{\frac{\perp \| X_n(u) \perp v_n \|^2}{T}} du} (v_n - X_n(s)) \Big) \tag{4.3}$$

which is induced by the stochastic Lyapunov functional

$$G_T(X, v) = \frac{\sigma}{2} \|\nabla X(s)\|^2_{L_2(S^1)} - \frac{T}{2} log \int_{S^1} e^{\frac{\perp \| X(u) \perp v \|^2}{T}} du \tag{4.4}$$

Next, Consider the following problem: a traveling salesman has to visit the cities $v_1, ..v_n$. If he fails to visit a city he loses income which depends on the minimal distance to that city. On the other hand, he wishes the trip to be as short as possible. We wish to help the salesman find the optimal path in two cases. In the first case we assume that he must make m stops and that he can work only in those stops. Thus, the money he loses is a function of the distance between each city and the nearest stop to that city. In the second (continuous) case, he can do business at any time, in which case his loses depend on the minimal distance between each city and his path. First, we turn our attention to the discrete case. Let

$$F_T(X) = \frac{\sigma}{2} \sum_{i=1}^{m} \left\| X^{i+1} - X^i \right\|^2 - \frac{T}{2} \sum_{k=1}^{n} log \sum_{i=1}^{m} e^{\frac{\perp \left\| X^i \perp v_k \right\|^2}{T}}$$

then, when $T \to 0$, $F_T(X)$ tends pointwise to

$$F(X) = \frac{\sigma}{2} \sum_{i=1}^{m} \left\| X^{i+1} - X^i \right\|^2 + \frac{T}{2} \sum_{k=1}^{n} \min_i \left\| X^i - v_k \right\|^2$$

We wish to find stops X^i such that $X = (X^1, ...X^m)$ is the global minimum of F. Since F in not smooth, we relax the problem and seek the minimum of F_T. Note that $F_T(X) = \mathbb{E}\big(G_{T'}(X, V)|X\big)$ when $T' = nT$, $G_{T'}(X, v)$ is given by (4.2) and $\mu(v_i) = \frac{1}{n}$. Hence, to find local minima of F_T we can use the stochastic approximation process

$$X_{n+1} = X_n - \varepsilon_n \nabla_x G_T(X_n, V_n) \tag{4.5}$$

We may assume that all the critical points of F_T are non degenerate because the set of functions with non degenerate critical points is of the second category.

Lemma 4.1. *For every $T > 0$ the process (4.5) is uniformly bounded.*

Proof: Let $\lambda_n^i = \frac{e^{\perp \left\| X_n^i \perp v_n \right\|^2 / T}}{\sum_{j=1}^m e^{\perp \left\| X_n^j \perp v_n \right\|^2 / T}}$. It is easy to see that

$$X_{n+1}^i = \varepsilon_n \sigma X_n^{i-1} + \varepsilon_n \sigma X_n^{i+1} + \varepsilon_n \lambda_n^i v + (1 - \lambda_n^i - 2\varepsilon_n \sigma) X_n^i$$

for ε_n small enough, all the coefficients are nonnegative and their sum is 1. Hence, X_{n+1}^i is a convex combination of X_n^{i+1}, X_n^i, X_n^{i-1} and v_n.

◇

Corollary 4.2. *The process (4.5) converges almost surely to a local minimum of F_T.*

Proof: The proof follows immediately from corollary 1.1.7.

◇

Remark 4.3. A straight forward computation shows that for T large enough F_T is convex, thus it has a unique local minimum. By corollary 4.2 the process (4.5) converges to that minimum almost surely.

Next, we turn to the continuous case. Here, the relaxed cost function we wish to minimize is

$$F_T(X) = \frac{\sigma}{2} \left\| \nabla X(s) \right\|_{L_2(S^1)}^2 - \frac{T}{2} \int_V log \int_{S^1} e^{\frac{\perp \left\| X(u) \perp v \right\|^2}{T}} du d\mu(v)$$

which is the average of the stochastic Lyapunov functional (4.4). Note that $D_x G(X, v) = \sigma \Delta X(s) + \frac{e^{\perp \left\| X(s) \perp v \right\|^2 / T}}{\int_{S^1} e^{\perp \left\| X(u) \perp v \right\|^2 / T} du} (v - X(s))$. Thus, we can use the process (1.2.1), i.e., stochastic approximation of the form $\frac{\partial u}{\partial t} = -DG_x(U, v)$. Using the notations of section 1.3, set $A = -\sigma\Delta$ and $f(u, v) = \frac{e^{\perp \left\| u(s) \perp v \right\|^2 / T}}{\int_{S^1} e^{\perp \left\| u(s) \perp v \right\|^2 / T} ds} (v - X(s))$. Then, for i=0,1 $f(u, v)$ maps $\mathbb{H}_i \cap L_\infty$ into itself uniformly with respect to v and since the input set V is bounded, there exists some function M such that $uf(u, -) < 0$ whenever $|u| > M$. It is also clear that F_T satisfies the P.S. condition on bounded sets in $\mathbb{H}_1$. Indeed, since $(DF_T)_{X_n} = \Delta X_n + h(X_n)$ then if $(DF_T)_{X_n} \to 0$ in $\mathbb{H}_{-1}$ and (X_n) is bounded in $\mathbb{H}_1$, there is a subsequence X_{n_j} which converges weakly in $\mathbb{H}_1$. Since weak convergence in $\mathbb{H}_1$ implies uniform convergence, then for every $g \in \mathbb{H}_1$, $\left\langle \nabla(X_{n_g} - X), g \right\rangle \to 0$ and $\left\langle h(X_{n_j}), g \right\rangle \to \left\langle h(X), g \right\rangle$. Hence, for every $g \in \mathbb{H}_1$

$$\left\langle (DF_T)_X, g \right\rangle = \left\langle \nabla(X_{n_j} - X), g \right\rangle + \left\langle h(X_{n_j}) - h(X), g \right\rangle + \left\langle DF_{X_n}, g \right\rangle \to 0$$

Therefore, the assertions of theorem 1.2.3 and corollary 1.2.4 hold. In particular, for every sample path (X_n) there is a critical value λ such that all the limit points of (X_n) are contained in K_λ.

5– Concluding remarks

Clearly, there are major difficulties when one passes from the one-dimensional model to the multi-dimensional one. The key stumbling block is the fact that the set of organized states in the multi-dimensional case is not an absorbing set. To this day, there is no definition of multi-dimensional organization for which the set of organized states is absorbing (see, for example, [F],[FP] which show that with positive probability, the exit time from the set O is finite).

Another question concerns the process in the non–homogeneous case. One possible course of action is to find an analogous result to the recurrence theorem 2.2. A different approach it to apply methods of stochastic approximation. Unfortunately, we where unable to obtain sharp results similar to those presented in chapter 1 for non–smooth processes.

References

[**A**] B. Aulbach: Continuous and discrete dynamics near manifolds of equilibria, LNM 1058, Springer–Verlag, 1984

[**C**] E.W. Cheney: Introduction to approximation theory, 1966, McGraw–Hill

[**D**] J. Diestel: Geometry of Banach spaces, selected topics, LNM 485, Springer–Verlag, 1970

[**GMR**] Y. Gordon, M. Meyer, S. Reisner: Volume approximation of convex bodies by polytopes, Studia math. 111 (3), 1994, 81–95

[**F**] J.A. Flanagan: Self organization in Kohonen's SOM, Neural Networks Vol 6 (7), 1996, 1185–1197

[**FP**] J.C. Fort, G. Pagès: About the Kohonen algorithm, Neural Networks, Vol 9 (5), 1995, 773–785

[**H**] S.S. Haykin: Neural Networks: A Comprehensive Foundation, 1994, MacMillan College Press

[**K**] T. Kohonen: Self-organization and associative memory, 1989, Springer-Verlag

[**KC**] H.J. Kushner, D.S. Clark: Stochastic Approximation for constrained and unconstrained systems, Springer–Verlag, 1978

[**KS**] J.W. Kim, H. Sompolinsky: On-line Gibbs learning, Physical Review Letters, Vol 76, 16 (1996) 3021–3024

[**KW**] J. Kiefer, J. Wolfowitz: Stochastic estimation of the maximum of a regression function, Ann. math. stat. 23 (1952), 462–466

[**KY**] H.J. Kushner, G.G. Yin: Stochastic approximation algorithms and applications, Application of mathematics – Stochastic modelling and applied probability 35, Springer, 1997

[**L**] M. Loève: Probability Theory, 3rd edition, 1963, D. Van Nostran

[**LLPS**] M. Leshno, V.Y. Lin, A. Pinkus, S. Schocken: Multilayer feedforward networks..., Neural Networks, 6 (1993) 861–867

[**LPW**] L. Ljung, G. Pflug, H. Walk: Stochastic approximation and optimization of random systems, DMV seminar, BD. 17 Birkhauser Pub., 1992

[**MP**] M. Minsky, S. Papert: Perceptrons, 1972, MIT Press

[**MW**] J. Mawhin, M. Willem: Critical point theory and Hamiltonian systems, Applied math. sciences 74, Springer–Verlag, 1989

[**O**] S. Orey: Limit theorems for Markov chain transition probabilities, 1971, Van Nostran Reinhold

[**P1**] G. Pisier: Probabilistic Methods in the geometry of Banach spaces, Probability and Analysis, 167–241, LNM 1206, Springer–Verlag 1986

[**P2**] G. Pisier: Martingales with values in uniformly convex spaces, Israel J. of Math. 20 (1975) 326–350

[**PW**] M.H. Protter, H.F Weinberger: Maximum Principles in Differential Equations, Prentice – Hall, Inc, 1967

[**RM**] H. Robbins, S. Monro: A stochastic approximation method, Ann. math. stat. 22 (1951), 400–407

[**RMS**] H. Ritter, T. Martinetz, K. Schulten: Neural computation and self organizing maps, 1992, Addison-Wesley

[**RS**] M. Reed, B. Simon: Methods in modern mathematical physics – Vol. 1, Academic Press, 1980

[**RW**] J. Rubinstein, G. Wolansky: Pattern formation in neural networks, 1993, Technion report

[**S**] A.N. Shiryayev: Probability, Graduate texts in mathematics 95, Springer–Verlag, 1979

[**W**] M.T. Wasan: Stochastic Approximation, 1969 Cambridge University Press