Imperial College of Science,
Technology and Medicine

(University of London)

Department of Computing

The Storage Capacity of Forgetful Neural Networks

by

Peter John Potts

Submitted in partial fulfilment
of the requirements for the MSc
Degree in Engineering of the
University of London and for the
Diploma of Imperial College of
Science, Technology and Medicine.

September 1995

Abstract

In this report, we derive a two stage algorithm to evaluate the storage capacity of a forgetful neural network using any smooth learning scheme.

In the first stage, we evaluate the exponential decay rate of the embedding strengths of memorised patterns. We do this using a generalised form of Riemann integration taken from recent advances in Domain theory and the theory of chaotic dynamical systems.

In the second stage, we derive a simple formula to equate the evaluated decay rate with a corresponding neural network using the so called marginalist learning scheme. This enables us to use a solved Ising model taken from statistical mechanics to derive the storage capacity.

In order to compare the theoretical predictions with experiment, we take the parameterised hyperbolic tangent function and the parameterised error function as concrete examples of the smooth learning scheme.

In summary, we show that the highest attainable storage capacity for any smooth forgetful neural network is $0.0489585N$ where N is the total number of neurons in the system and we derive an algorithm to evaluate the optimal parameter to achieve this.

Acknowledgements

I would like to thank my supervisor Abbas Edalat for his help and guidance.

Contents

Chapter 1

Introduction

The greatest computer in the world is undoubtedly the human brain. Neural networks [18] are the study of idealised systems containing very large numbers of connected neurons deliberately constructed to make use of organisational principles found in the human brain. In this report, we are going to use the latest techniques in domain theory to evaluate the storage capacity of certain kinds of forgetful neural network and compare this with experiment.

As early as 1943, McCulloch and Pitts [25] showed that any binary logical operation can be represented by simplified neurons.

In 1949, Hebb [16] suggested that learning may take place by the modification of the synaptic couplings between neurons. In other words, memory resides between neurons and not in the neurons themselves. More specifically, he suggested that a synapse may be modified according to the temporal average of the correlated activity of the two neurons it connects. This led to a variety of models for associative memory and pattern recognition.

In 1954, Cragg and Temperley [7] introduced the analogy between neural networks and the Ising models of magnetic systems and speculated about whether biological equivalents could be found for the ideas of temperature and energy.

In 1974, Little [23] discovered that noise could be added to a neural network in such a way that it corresponded to temperature in the Ising model.

In 1982, Hopfield [20] fired the public imagination for the first time by explaining how neural networks could be used in practice. This heralded the modern era of neural networks. He suggested that neural networks learn patterns by developing a locally stable state in state space for each pattern. All other states flow into these stable states, called *attractors*. Errors in a state near to a stable state are corrected as it flows into the stable state.

The Hopfield model is deterministic. A more general model is the Boltzmann machine, which is the Hopfield model with noise. This makes the Boltzmann machine stochastic. Put another way, the Hopfield model can be seen as the zero temperature case of the Boltzmann machine.

Hopfield made the critical observation that the Boltzmann machine is isomorphic to the Ising model. The Ising model came from the physics of spin glass. This has allowed a deluge of physical theory describing spin glass models to transform the field of neural networks. The Ising model has opened the way for understanding a whole universe of systems consisting of large numbers of strongly interacting elements.

A spin glass is a special magnetic alloy that exhibits ferromagnetic and antiferromagnetic properties, such as Manganese Fluoride, or Chromium Bromide. These properties are conflicting in tendency. This is also an intrinsic feature of neural networks by which neurons interact synaptically via intense mixtures of excitatory and inhibitory synapses. This leads to a system that exhibits many diverse stable states. Such a combination is ideal for an associative memory and pattern recognition.

In 1987, Amit et al [1] solved the Hopfield model exactly using the replica method with the approximation of replica symmetry on the fully connected Ising model for the limiting case as the number of neurons tends to infinity.

The solution reveals that 0.14 is the critical ratio between the number of stored patterns and the total number of neurons. This value is known as the storage capacity of the neural network. The retrieval quality is good when the number of stored patterns is below the critical number, but deteriorates suddenly above it. This phenomenon is known as *catastrophic forgetting.*

Hopfield suggested alternate learning schemes in his original paper [20] that avoided this catastrophic forgetting. In these learning schemes, new patterns are learnt at the expense of gradually forgetting previously stored patterns. These models are called forgetful neural networks. However, there is a price to pay for this more desirable behaviour. Namely, the storage capacity is lower than that for the Hopfield model.

In 1986, Nadal et al [27] explored various learning schemes for forgetful neural networks including the *marginalist* scheme and the *smooth* scheme. The marginalist learning scheme specifies that the contribution of each pattern to the synaptic couplings decays exponentially with age. This is a very attractive learning scheme because it is easily accomplished by an iterative procedure that is biologically realistic.

In 1986, Mézard et al [26] formulated and solved a general learning scheme that incorporated both the Hopfield model and the marginalist learning scheme.

The smooth learning scheme specifies a general iterative procedure for the synaptic couplings using a restricted class of functions. This scheme is interesting because it suggests a level of independence between the function used and the overall properties of the system. This lends support to the idea that in real neurons the exact shape of the various electrochemical functions, namely the Hodgkin-Huxley equations [19], may not actually matter.

In 1988, van Hemmen et al [35] analyzed the *smooth* learning scheme by interpreting the evolution of the synaptic couplings as a Markov process, studying its asymptotics and deriving the embedding strength decay rate of the stored patterns.

In 1992, Behn et al [4] analyzed the invariant distributions of the synaptic couplings and showed that they exhibited fractal and multifractal properties. Interestingly, he showed that the nature of the invariant distribution undergoes a number of sharp transitions as the parameter associated with the smooth learning scheme changes, but that this does not effect the overall performance of the neural network.

The embedding strength of a stored pattern is a measure of how well represented it is in the synaptic couplings and should decrease with time in a forgetful neural network. However, we will derive computationally the embedding strength decay rate of the stored patterns using recent advances in domain theory by Edalat [13, 11, 10, 12].

This may seem like a strange marriage of ideas, since domain theory was introduced by Dana Scott [30] in 1970 as a mathematical theory of semantics of programming languages. The rise of domain theory over a quarter of a century was primarily motivated by the need to solve recursive procedures and data types in computer science.

In 1993, Edalat found the first new application for domain theory, outside of denotational semantics, in fractal image decompression [9]. Theoretically, he showed that some important areas of mathematics [13] have natural domain-theoretic computational models.

The key to understanding the leap between the use of domain theory in denotational semantics and other more esoteric mathematical applications is in the wider understanding that domain theory provides a means of representing infinite objects in a finite manner suitable for consumption by a computer in a sound theoretical way.

Domain theory has provided new ways to represent and manipulate fractals. In contrast to the simple objects manipulated in classical geometry, such as lines and circles, fractals are very complex objects. Fractals have a fine structure at every magnification. Many objects found in nature, such as ferns and clouds, can be thought of as fractals. Fractals have proved to be a much more appropriate vehicle for representing natural objects than traditional geometry.

A Dutch mathematician called Mandlebrot [24] was the first person to use the term fractal from the Latin word fractus, which means fractured or broken, to describe these new objects.

It is easy to see that the self similarity at different magnifications in fractals is a form of recursion and it is this feature that has made it amenable to a domain-theoretic representation.

In particular, Edalat devised an algorithm [11] to evaluate the expectation of continuous functions over fractal probability distributions. This is very pertinent to the problem at hand because it can be shown [12] that for random stored patterns, the probability distribution of the synaptic couplings is a fractal and evaluating the embedding strength decay rate of the stored patterns involves calculating the expectation of a certain continuous function over this probability distribution.

In order to understand how this algorithm works and get a full picture, it is necessary to start with some fundamental constructs in mathematics. In Chapter 2, we start by explaining the elementary ideas of domains, metric spaces and topological spaces. We then proceed to describe a variety of useful topologies and show how the three elementary ideas above can be related to each other. We then define the idea of a measure and show how it can be used to represent a probability distribution. We then proceed to describe the idea of a valuation [6], which is a restricted version of a measure. We then bring all this theory together to describe the normalised probabilistic power domain and show how any probability distribution can be represented by a sequence of simple valuations.

This leads us to the definition of the generalised Riemann integral, which will provide our means for evaluating the expectation that we require over the fractal probability distribution. This is an improvement over standard Riemann integration, which can only be used to evaluate the expectation of a continuous function over a continuous probability distribution.

We cannot use Lebesgue integration [36], because although it is extremely general, it loses the constructive nature of Riemann integration. However, we have the nice property that when the generalised Riemann integral does exist, it coincides with its Lebesgue integral [11].

In a nutshell, Riemann integration involves slicing a function vertically into an ever increasing number of thinner slices and summing them in an equally weighted manner. This is equivalent to a uniform probability distribution. Lebesgue integration involves slicing a function horizontally into an ever increasing number of fragments and summing them according to an arbitrarily complex probability distribution. Generalised Riemann integration involves finding a sequence of increasingly better approximations for the arbitrarily complex probability distribution consisting of an increasing number of simple elements and applying them to the function.

We continue by introducing the notion of an iterated function system with probabilities [17], which has proved to be a very useful way of representing an interesting class of fractals. In particular we are interested in the smaller class of weakly hyperbolic iterated function systems with probabilities, because it is sufficiently general for our purposes and Edalat [10] demonstrated a constructive technique to approximate the fractals that they generate.

We next introduce the notion of a non-deterministic dynamical system and demonstrate the vital step made possible be Elton [14], who showed that its limiting probability distribution is equal to the fractal probability distribution of its corresponding iterated function systems with probabilities provided certain criteria are satisfied.

In Chapter 3, we will show how the synaptic couplings of a forgetful neural network is an example of a non-deterministic dynamical system. It then follows that the synaptic couplings can be represented by a fractal probability distribution and so we derive a sequence of increasingly better approximations for the fractal probability distribution.

In Chapter 4, we outline how the embedding strength decay rate of the stored patterns is derived from the Lyapunov exponent [35, 4, 12] of the non-deterministic dynamical system that describes the synaptic couplings. The Lyapunov exponent of a dynamical system is the average exponential rate at which the resulting motion of the system starting from two slightly different initial positions depart from each other, assuming that it is exponential.

We then evaluate the embedding strength decay rate of the stored patterns using the generalised form of Riemann integration for various smooth learning schemes.

In Chapter 5, we give an overview of the fully connected Ising model and its relationship with neural networks. In particular, we run through the extremely relevant general learning scheme devised by Mézard et al [26], which incorporates both the Hopfield model and the marginalist learning scheme.

In Chapter 6, we derive the relationship between the smooth and the marginalist learning schemes using first order differential calculus and probability theory. We then use results established by Mézard et al [26] with respect to the marginalist learning scheme to formulate the storage capacity of the corresponding smooth learning scheme.

We then constructed a 1500 neuron computational model, stored 375 random patterns in the synaptic couplings using various smooth learning schemes and computed the retrieval quality for each of the last 90 patterns and showed that they were consistent with the theoretical storage capacity.

Finally, we explored some variations of the smooth learning scheme.

Chapter 2

Domain theoretic background

The aim of this chapter is to introduce a generalised form of Riemann integration as formulated by Edalat [11], which we need later to evaluate the embedding strength decay rate of the stored patterns in a forgetful neural network.

For completeness, we will start by reviewing the fundamental ideas behind domain theory, metric spaces and topological spaces. Then we consider various topologies and show how the three fundamental ideas above are linked together.

2.1 Domains

The term *domain* is often used in reference to partial orders. A *partial order* (or *poset*) $\mathbf{D} = (D, \sqsubseteq_D)$ is a set D with a binary relation $\sqsubseteq_D$ which is reflexive, transitive and anti-symmetric.

Partial orders provide a means of approximating a complex object by a sequence of simple objects, referred to more technically as a chain. In particular, they provide a means for representing recursion in computing.

So when we have two objects $x, y \in D$, if $x \sqsubseteq_D y$ then we can think of x as an approximation of y or that y contains more information than x. A subset A of D is a *chain* if every $x, y \in A$ satisfies $x \sqsubseteq_D y$ or $y \sqsubseteq_D x$. A is an ω-chain if its elements are countable. Clearly, an ω-chain can be numbered so that

$$x_0 \sqsubseteq_D x_1 \sqsubseteq_D x_2 \sqsubseteq_D x_3 \cdots$$

which we shall denote more concisely by $\langle x_i \rangle_{i \geq 0}$. If the chain has a least upper bound then in a sense it contains all the information in the chain and no more.

A subset A of D is *directed* if every finite subset of A has an upper bound in D. Clearly, a chain is directed. The directed subsets of a partial order provide a useful class of subsets.

This leads to the most important class of partial orders in domain theory, namely the directed complete partial order. A *directed complete partial order* (or dcpo) $\mathbf{D} = (D, \sqsubseteq_D)$ is a poset such that every directed subset A of D has a least upper bound, denoted $\bigsqcup A$.

The next important thing to consider is maps between directed complete partial orders. In order to be useful, we need to ensure that information order and least upper bounds are preserved so that the information analogy can be carried from one dcpo to another. A map $f : \mathbf{D} \to \mathbf{E}$ between the dcpo $\mathbf{D} = (D, \sqsubseteq_D)$ and the dcpo $\mathbf{E} = (E, \sqsubseteq_E)$ is *monotone* if every $x, y \in D$ satisfies $f(x) \sqsubseteq_E f(y)$ whenever $x \sqsubseteq_D y$. If every directed subset A of D satisfies $\bigsqcup f(A) = f(\bigsqcup A)$ then f is *continuous* and interestingly the *least fixed point* is given by $\bigsqcup_n f^n(\bot_D)$. This last property is crucial because in computer science recursively defined objects are in essence the least fixed point of the recursive definition and we have here a constructive technique to approximate the least fixed point to any required degree of accuracy.

It is also important to be able to distinguish between elements of a dcpo that contain a finite and an infinite amounts of information because it is in the nature of computers that they can only handle finite amounts. Given a dcpo $\mathbf{D} = (D, \sqsubseteq_D)$, an element a of D is *finite* if for every directed subset A of D whenever $a \sqsubseteq_D \bigsqcup A$ then there exists an element b of A such that $a \sqsubseteq_D b$. The set of all finite elements of $\mathbf{D}$ is denoted K_D.

A dcpo $\mathbf{D} = (D, \sqsubseteq_D)$ is *algebraic* if for every element x of D, the set $\{y \in K_D \mid y \sqsubseteq_D x\}$ is directed with least upper bound x. This property ensures that it is sufficient to work with the finite elements only. It is ω-*algebraic* if the finite elements are countable. It is *bounded complete* if every bounded subset has a least upper bound. A bounded complete ω-algebraic dcpo is also known as a Scott domain.

Given a dcpo $\mathbf{D} = (D, \sqsubseteq_D)$, an element x of D is *way below* an element y of D, denoted $x \ll y$, if whenever $y \sqsubseteq_D \bigsqcup A$, there exists an element z of A such that $x \sqsubseteq_D z$. The way below relation is a generalisation of finiteness since x is finite if $x \ll x$.

The set of all elements way below x is denoted $\downdownarrows x = \{y \in \mathbf{D} \mid y \ll x\}$. A dcpo $\mathbf{D} = (D, \sqsubseteq_D)$ is *continuous* if $\downdownarrows x$ is directed with least upper bound x for all $x \in D$. A subset B of D forms a *base* for $\mathbf{D}$ if $\downdownarrows x \cap B$ is directed with least upper bound x for every element x of D. The dcpo is ω-continuous if has a countable base.

Note that an (ω-)algebraic dcpo is an (ω-)continuous dcpo due to fact that the way below relation is a generalisation of finiteness.

2.2 Metric spaces

Metric spaces [32] are important because they allow the degree to which one elements approximates another to be quantified via a distance function.

A *metric space* $X = (X, d_X)$ consists of a non-empty set X together with a distance function $d_X : X \times X \to \mathbb{R}$ satisfying

- $d_X(x, y) \geq 0$

- $d_X(x, y) = 0 \iff x = y$

- $d_X(x, y) = d_X(y, x)$

- $d_X(x, y) + d_X(y, z) \geq d_X(x, z)$

We now define some terminology that will be crucial in making the link between metric spaces and topological spaces. Given a point $x \in X$ and a strictly positive real number ϵ, the *open ϵ-ball* of $x \in X$ is the set $B_\epsilon(x) = \{y \in X \mid d_X(x, y) < \epsilon\}$. A subset O of X is *open* if given any $x \in O$ there exists $\epsilon > 0$ such that $B_\epsilon(x) \subset O$.

A map $f : X \to X$ is *contracting* if there exists $\alpha < 1$ such that $d_X(f(x), f(y)) \leq \alpha\, d_X(x, y)$ for all $x, y \in X$. Contracting maps gives us the simplest tool to generate fractals for complete metric spaces as we shall see later in this chapter when we discuss iterated function systems. They also have unique fixed points given by the element of the singleton set $\bigcap_n f^n(X)$ if X is a compact metric space.

2.3 Topological spaces

In many ways, topological spaces [32] provide an alternate perspective of metric spaces. However, they are even more important than that because they are in fact a generalisation of metric spaces.

A *topological space* $\mathcal{X} = (X, \Omega(X))$ consists of a non-empty set X together with a collection $\Omega(X)$ of subsets of X satisfying

- $X, \emptyset \in \Omega(X)$

- the intersection of any two sets in $\Omega(X)$ is again in $\Omega(X)$

- the union of any collection of sets in $\Omega(X)$ is again in $\Omega(X)$

The collection $\Omega(X)$ is called a *topology* for X, and the members of $\Omega(X)$ are the *open* sets of $\mathcal{X}$. A subset C of X is *closed* if $X - C$ is open.

A topological space $\mathcal{X} = (X, \Omega(X))$ is *Hausdorff* if for every distinct $x, y \in X$ there exists $A, B \in \Omega(X)$ such that $x \in A$ and $y \in B$ and $A \cap B = \emptyset$.

A subcollection B of $\Omega(X)$ is a *basis* for $\Omega(X)$ if every set in $\Omega(X)$ is the union of sets from B. $\mathcal{X}$ is *second countable* if it has a countable basis.

A *cover* C for a set A is a collection of sets such that $A \subset \bigcup_{B \in C} B$. The topological space $\mathcal{X}$ is *compact* if every open cover of X has a finite subcover. Again this is important because finiteness is an ever reoccurring concern with computers.

2.4 Various topologies

Any metric space $X = (X, d_X)$ gives rise to a topological space $\mathcal{X} = (X, \Omega(X))$, where $\Omega(X)$ is defined to be the collection of all those subsets which are open in the metric space. This is called the *usual topology*. A topological space which arises in this way from a metric space is called *metrizable*.

In contrast, the *discrete topology* of a set X is the collection of all subsets of X and the *indiscrete topology* is $\{\emptyset, X\}$.

The *Scott topology* of a dcpo $\mathbf{D} = (D, \sqsubseteq_D)$ consists of Scott open sets O satisfying

- $x \in O \wedge x \sqsubseteq_D y \Rightarrow y \in O$
- $(\forall \text{ directed } A \subseteq D) \bigsqcup A \in O \Rightarrow O \cap A \neq \emptyset$.

This shows how a dcpo can be used to generate a topological space.

Additionally, the *Scott topology* of an algebraic dcpo $\mathbf{D} = (D, \sqsubseteq_D)$ has a base of $\{\uparrow x \mid x \in K_D\}$ where $\uparrow x = \{y \in D \mid x \sqsubseteq_D y\}$ and the *Scott topology* of a continuous dcpo $\mathbf{D} = (D, \sqsubseteq_D)$ has a base of $\{\twoheaduparrow x \mid x \in K_D\}$ where $\twoheaduparrow x = \{y \in D \mid x \ll y\}$.

The *specialisation ordering* $\sqsubseteq_{\text{s.o.}}$ of a topological space $\mathcal{X} = (X, \Omega(X))$ is defined as $x \sqsubseteq_{\text{s.o.}} y$ if $x \in A$ implies $y \in A$ for all $A \in \Omega(X)$. It can be shown that $(X, \sqsubseteq_{\text{s.o.}})$ is a preorder and for the Scott topology on a dcpo $\mathbf{D} = (D, \sqsubseteq_D)$ that $\sqsubseteq_D = \sqsubseteq_{\text{s.o.}}$.

We now have a full circle going from dcpo to topological space and back again through the vehicles of the Scott topology and specialisation ordering respectively. This will prove very useful when applied to the upper space topology.

Given any topological space $\mathcal{X} = (X, \Omega(X))$, its *upper space* $\mathcal{UX} = (UX, \Omega(UX))$ is defined by

- UX is the set of all non-empty compact subsets of X
- $\{\Box x \mid x \in \Omega(X)\}$ is a base of pow UX where $\Box x = \{y \in UX \mid y \subseteq x\}$

If X is a compact metric space with the usual topology then UX with specialisation ordering is a bounded complete ω-continuous dcpo with bottom X. In fact, the specialisation ordering of UX is reverse inclusion.

Note that any compact subspace of a metric space is bounded and any compact subspace of a Hausdorff space is closed.

2.5 Normalised Borel measures

We are concerned here with normalised Borel measures on topological spaces because they provide a rigorous framework for studying probability distributions.

The class of *Borel sets* $\mathcal{B}(X)$ of the topological space $\mathcal{X} = (X, \Omega(X))$ is the smallest collection of subsets of X which contains the open sets $\Omega(X)$ and is closed under complements and countable unions. In plain language, a Borel set is basically any reasonably normal subset.

A *Normalised Borel measure* μ on a topological space $\mathcal{X} = (X, \Omega(X))$ is a mapping

$$\mu : \mathcal{B}(X) \to [0, 1]$$

satisfying

- $\mu(\emptyset) = 0$
- $\mu(X) = 1$
- $\mu(\bigcup_{i\geq 0} B_i) = \sum_{i\geq 0} \mu(B_i)$ where B_i are disjoint subsets of X.

We denote the set of all normalised Borel measures on X by M^1X.

It is plain to see that a normalised Borel measure μ represents a random variable ξ on the sample space X where given a Borel subset B of X, the probability that $\xi \in B$ is given by

$$\mathbf{P}\{\xi \in B\} = \mu(B). \tag{2.1}$$

Therefore, M^1X represents the set of all probability distributions on the sample space X.

2.6 Normalised valuations

A normalised valuation is basically a normalised Borel measure whose source has been restricted to open sets instead of Borel sets. Conversely, it can be seen that any normalised valuation extends uniquely to a normalised Borel measure on certain nice spaces such as compact metric spaces.

This difference affects the definition because the subtraction of one open set with another is not an open set.

Definition 2.6.1 (Normalised valuation)
A normalised valuation *ν on a topological space $\mathcal{X} = (X, \Omega(X))$ is a mapping*

$$\nu : \Omega(X) \to ([0,1], \leq)$$

satisfying

- $\nu(\emptyset) = 0$
- $\nu(X) = 1$

- $\nu(a) \leq \nu(b)$ *if* $a \subseteq b$

- $\nu(a) + \nu(b) = \nu(a \cup b) + \nu(a \cap b)$

Intuitively, in order for a valuation to be computable, it must be continuous. A valuation on a topological space $\mathcal{X} = (X, \Omega(X))$ is *continuous* if every directed subsets A of $\Omega(X)$ satisfies

$$\nu\left(\bigcup_{O \in A} O\right) = \sup_{O \in A} \nu(O).$$

We need to construct a sequence of approximations for a valuation. Let us start by considering the simplest valuation of all. For any $x \in X$, the *point valuation* δ_x at x is the mapping $\delta_x : \Omega(X) \to ([0,1], \leq)$ defined by

$$\delta_x(O) = \begin{cases} 1 & \text{if } x \in O \\ 0 & \text{otherwise} \end{cases} . \tag{2.2}$$

These point valuations can be combined to give simple valuations. A *simple valuation* is any finite linear combination

$$\sum_{i=0}^{n} r_i \delta_{x_i}$$

of point valuations where $r_i \in (0,1]$ and $\sum_{i=0}^{n} r_i = 1$.

We shall see in the next section that in certain situations, simple valuations are sufficiently complex to provide an approximation of any valuation to any desired degree of accuracy.

2.7 Normalised probabilistic power domains

Finally, before we actually tackle the generalised Riemann integral itself, we examine the most complex structure which underpins it, namely the normalised probabilistic power domain.

Definition 2.7.1 (Normalised probabilistic power domain)
Given a topological space $\mathcal{X} = (X, \Omega(X))$*, its* normalised probabilistic power domain $\mathbf{P^1X} = (P^1X, \sqsubseteq_{P^1X})$ *is defined by*

- P^1X *is the set of all normalised continuous valuations* ν *on* $\mathcal{X}$

- $\mu \sqsubseteq P^1X\nu$ *if* $\mu(O) \leq \nu(O)$ *for all* $O \in \Omega(X)$

Every normalised probabilistic power domain is a dcpo.

In addition, it has been shown by Edalat [11] that if the topological space $\mathcal{X}$ is derived from an ω-continuous dcpo with bottom $\perp$ using the Scott topology then $\mathbf{P^1X}$ is an ω-continuous dcpo with bottom $\delta_\perp$ and with a basis consisting of all the simple valuations on $\mathcal{X}$ [11].

If X is a compact metric space then $(UX, \supseteq)$ is an ω-continuous dcpo and so $\mathbf{P^1UX}$ is an ω-continuous dcpo.

The critical observation made by Edalat [13] was that the maximal elements of the normalised probabilistic power domain of $\mathcal{UX}$ where $\mathcal{X}$ is a compact metric space with the usual topology contains all the normalised Borel measures filtered through the singleton map. The *singleton map* $s : X \to UX$ embeds X onto the set of maximal elements of UX.

This is good news because $\mathbf{P^1UX}$ is an ω-continuous dcpo. Therefore given any normalised Borel measure, there exists a chain of simple valuations whose least upper bound filtered through the singleton map extend uniquely to it.

More technically speaking, let us consider the precise maximal elements of interest. A valuation $\mu \in P^1UX$ is *supported* in $s(X)$ if its unique extension to a normalised Borel measure on $\mathcal{UX}$ satisfies $\mu(s(X)) = 1$. The *support* of μ is the set of points $x \in s(X)$ such that $\mu(O) > 0$ for all $x \in O \in$ pow UX. Let S^1X be the set of all normalised valuations supported in $s(X)$. Then M^1X is isomorphic with S^1X with isomorphism

$$\begin{aligned} e : M^1X &\to S^1X \\ \mu &\to \mu \circ s^{-1}. \end{aligned}$$

It is not known at the time of writing whether S^1X includes all the maximal elements of the power domain.

2.8 Generalised Riemann Integration

The aim of the generalised form of Riemann integration is to provide a constructive technique to obtain the value of the integral of bounded real valued functions with respect to normalised Borel measures on compact metric spaces.

We observe from the last section that a normalised Borel measure can be obtained as the least upper bound of an ω-chain of simple valuations. This leads naturally to a sequence of increasingly better approximations to the value of an integral.

Let $g : X \to \mathbb{R}$ be a continuous real valued function on a compact metric space $X = (X, d_X)$.

Let μ be a normalised Borel measure. It corresponds to a unique valuation $\mu \circ s^{-1} \in S^1X \subseteq P^1UX$ [11, Theorem 2.15]. Since $\mathbf{P^1UX}$ is an ω-continuous dcpo, there exists a chain $\langle \nu_i \rangle_{i \geq 0}$ of simple valuations $\nu_n \in P^1UX$ such that $\mu \circ s^{-1} = \bigsqcup_{n \geq 0} \nu_n$.

Definition 2.8.1 (Generalised upper and lower Riemann sum) *For any simple valuation*

$$\nu = \sum_{a \in A} r_a \delta_a \in P^1UX \tag{2.3}$$

the generalised upper Riemann sum of g with respect to ν is

$$S_X^u(g, \nu) = \sum_{a \in A} r_a \sup\{g(x) \mid x \in a\} \tag{2.4}$$

and the generalised lower Riemann sum of g with respect to ν is

$$S_X^l(g, \nu) = \sum_{a \in A} r_a \inf\{g(x) \mid x \in a\}. \tag{2.5}$$

The full unexpedited definition of generalised Riemann integration [11, Definition 4.5] will not be described here. However, the generalised Riemann integral of any continuous function $g : X \to \mathbb{R}$ with respect to any normalised Borel measure on a compact metric space X exists [11, Theorem 6.1] and is given by [11, Proposition 4.9]

$$\mathbf{R}\int g \; \mathrm{d}\mu = \lim_{n\to\infty} S_X^u(g,\nu_n) = \lim_{n\to\infty} S_X^l(g,\nu_n). \tag{2.6}$$

In fact, it can also be shown that $S_X^u(g,\nu_n)$ and $S_X^l(g,\nu_n)$ are monotonically decreasing and increasing respectively [11, Corollary 4.10].

$$S_X^u(g,\nu_n) \searrow \mathbf{R}\int g \; \mathrm{d}\mu \nwarrow S_X^l(g,\nu_n) \tag{2.7}$$

as $n \to \infty$. Also, it is Lebesgue integrable and the two integrals coincide [11, Theorem 7.2]

$$\int g \; \mathrm{d}\mu = \mathbf{L}\int g \; \mathrm{d}\mu = \mathbf{R}\int g \; \mathrm{d}\mu. \tag{2.8}$$

Therefore, provided that g can be analyzed sufficiently, so that the supremums and infimums in equations 2.4 and 2.5 can be evaluated exactly, the upper and lower Riemann sums can be evaluated for increasing n according to equation 2.7 until the desired accuracy is achieved.

2.9 Iterated function systems with probabilities

In 1981, Hutchinson [21] showed how a useful class of fractals could be represented by so called iterated function systems (or IFS).

Definition 2.9.1 (Iterated function system with probabilities) *An* iterated function system *with probabilities* $\{X; f_1,\ldots,f_N; p_1,\ldots,p_N\}$ *is given by a finite number of continuous maps* $f_i : X \to X (1 \leq i \leq N)$ *on a compact metric space* $X = (X, d_X)$*, such that each* f_i *is assigned a probability* p_i *where* $0 < p_i < 1$ *and* $\sum_{i=0}^{N} p_i = 1$.

In order to make theoretical progress, it is necessary to consider restricted classes of iterated function systems. The most fruitful class over the last few years has proved to be the hyperbolic iterated function system.

Definition 2.9.2 (Hyperbolic IFS)

An iterated function system $\{X; f_1, \ldots, f_N\}$ *is* hyperbolic *if* $f_1, \ldots, f_N$ *are contracting maps.*

However, this is too restricting for our purposes. So we have to consider something more general, namely the weakly hyperbolic iterated function system, which was coined and studied by Edalat [10].

Definition 2.9.3 (Weakly hyperbolic IFS)

An iterated function system $\{X; f_1, \ldots, f_N\}$ *is* weakly hyperbolic *if every infinite sequence* $i_1, \ldots \in \{1, \ldots, N\}$ *satisfies*

$$\lim_{n\to\infty} |f_{i_1} \cdots f_{i_n} X| = 0. \tag{2.9}$$

Clearly, every hyperbolic IFS is a weakly hyperbolic IFS.

It has been shown [10] that every weakly hyperbolic IFS with probabilities has a unique invariant measure with fractal characteristics given by

$$\left(\bigsqcup_n H^n \delta_X \right) \circ s \tag{2.10}$$

where

$$H^n \delta_X = \sum_{i_1, \ldots, i_n = 1}^{N} p_{i_1} \cdots p_{i_n} \delta_{f_{i_1} \cdots f_{i_n} X} \tag{2.11}$$

whose support is the unique attractor of the IFS. This is exactly in the form we require for use in generalised Riemann integration.

2.10 Non-deterministic dynamical systems

We next introduce the non-deterministic dynamical system, which will form the theoretical bridge between iterated function systems and the synaptic couplings of a forgetful neural network using the smooth learning scheme.

Given an IFS with probabilities $\{X; f_1, \ldots, f_N; p_1, \ldots, p_N\}$ the corresponding non-deterministic dynamical system is the iterative orbit of a single point in X, in which at each iteration a map f_i is selected with probability p_i.

In 1986, Elton [14] made the important observation that the time average of a continuous function g for almost all initial points $x \in X$ and for almost all sequences $i_1, i_2, \ldots \in \{1, \ldots, N\}$ tends with probability one to its integral with respect to the unique invariant measure of the IFS

$$\lim_{n \to \infty} \frac{1}{n+1} \sum_{m=0}^{n} g(f_{i_m} \cdots f_{i_1}(x)) = \int g \ \mathrm{d}\mu \tag{2.12}$$

provided that there exists $r < 1$ such that

$$\prod_{i=1}^{N} d_X(f_i(x), f_i(y))^{p_i} \leq r d_X(x, y) \tag{2.13}$$

for all $x, y \in X$. This is known as Elton's Ergodic Theorem because it implies that the dynamics is *ergodic* meaning phase space averages are the same as time averages.

Chapter 3

Fractal probability distribution of synaptic couplings

In this chapter, we show that the synaptic couplings of a forgetful neural network using the smooth learning scheme can be represented by a fractal probability distribution and derive a sequence of increasingly better approximations for this distribution. In the next chapter we will use this sequence of approximations to evaluate the embedding strength decay rate of the stored patterns using the generalised form of Riemann integration described in the last chapter.

3.1 Forgetful neural networks using smooth learning scheme

A forgetful neural network consists of N fully connected neurons. Each neuron is a processing unit with one output x_i and $N-1$ inputs connected to the outputs of each of the other neurons. Each output x_i is either in the firing state ($x_i = 1$) or in the quiescent state ($x_i = -1$). Each connection from neuron j to neuron i is determined by the synaptic coupling parameter J_{ij}. If J_{ij} is negative, zero or positive then the connection is inhibitory, void or excitatory respectively. Each neuron updates its output asynchronously according to the following rule

$$x_i \text{ becomes } \begin{cases} 1 & \text{if } \sum_{j=1}^{N} J_{ij} x_j > 0 \\ -1 & \text{otherwise} \end{cases} . \tag{3.1}$$

The network should work as an associative memory. In other words, if the network is set to a stored pattern (or close to) then it should relax under the dynamics described above towards a close stationary state. Proximity is measured by the overlap between the stored pattern and the stationary state and is called the *retrieval quality.*

The *overlap* m between two patterns $\mathbf{x}$, $\mathbf{y}$ is the ratio of differing bits to the total number of bits and it is easy to see that it is given by

$$m = \frac{1}{N} \sum_{i=1}^{N} x_i y_i \tag{3.2}$$

and the fraction of errors is $\frac{1}{2}(1 - m)$.

Assuming that the synaptic couplings are symmetric, as suggested by Hopfield, the Hamiltonian of the network is the energy given by

$$H = -\frac{1}{2} \sum_{i,j=1}^{N} J_{ij} x_i x_j . \tag{3.3}$$

This means that the state of the network moves on the energy landscape to the local minima that should correspond to one of the stored patterns.

The simplest storage prescription for M patterns $\mathbf{X}^m (1 \leq m \leq M)$, which corresponds to the Hopfield model, is

$$J_{ij} = \frac{1}{N} \sum_{m=1}^{M} X_i^m X_j^m . \tag{3.4}$$

This prescription works very well for low storage levels, but suffers from catastrophic forgetting when $M > 0.14N$. Above this level, only a negligible number of patterns are remembered.

The forgetful storage prescription is given by the local iterative procedure

$$J_{ij}^m = \frac{1}{N}\phi(NJ_{ij}^{m-1} + \epsilon X_i^m X_j^m) \tag{3.5}$$

where $J_{ij}^0 = 0$ and $J_{ij} = J_{ij}^M$. Here J_{ij}^m represents the stored information up to and including pattern m. This prescription avoids catastrophic forgetting by gradually forgetting the oldest patterns.

The fact that the forgetful storage prescription is nonlinear [33, 34] in contrast to the Hopfield model makes the analysis more complex.

We are going to concentrate on the *smooth* learning scheme [35] in which the function ϕ is assumed to be odd ($\phi(-x) = -\phi(x)$), monotonically increasing and strictly concave for $x > 0$ ($\phi''(x) < 0$) and $\phi'(0) = 1$.

In chapter 5, we will consider the *marginalist* learning scheme [26] where $\phi(x) = \exp\left(-\frac{\varepsilon^2}{2N}\right)x$.

If we set $x_m = NJ_{ij}^m$ and $h_m = X_i^m X_j^m$ [12] then equation 3.5 reduces to

$$x_{m+1} = \phi(x_m + \epsilon h_m) \tag{3.6}$$

with $x_0 = 0$.

Assuming that the stored patterns are random, this means that h_m is a random variable, which is equal to 1 with a probability of $\frac{1}{2}$ and is equal to -1 with a probability of $\frac{1}{2}$. Therefore, in the limit as $M \to \infty$, this is a non-deterministic dynamical system corresponding to the iterated function system with probabilities $\{[x_-, x_+]; \phi_+, \phi_-; \frac{1}{2}, \frac{1}{2}\}$ where $\phi_\pm(x) = \phi(x \pm \epsilon)$ and $x_\pm$ is the least fixed point of $\phi_\pm$.

For the sake of definiteness, when required, we are going to consider

$$\phi(x) = \tanh(x) \tag{3.7}$$

and

$$\phi(x) = \mathrm{erf}\left(\frac{\sqrt{\pi}}{2}x\right) \tag{3.8}$$

with

$$\epsilon = \frac{k}{N}. \tag{3.9}$$

We are now in a position to derive an approximating sequence for the fractal probability distribution μ that describes the synaptic couplings.

$$\mu = \left(\bigsqcup_n \nu_n \right) \circ s \tag{3.10}$$

where

$$\nu_n = \frac{1}{2^n} \sum_{i_1,\ldots,i_n \in \{+,-\}} \delta_{\phi_{i_1} \cdots \phi_{i_n} [x_-, x_+]}. \tag{3.11}$$

But, ϕ_+ and ϕ_- are monotonically increasing functions, therefore

$$\nu_n = \frac{1}{2_n} \sum_{i_1,\ldots,i_n \in \{+,-\}} \delta_{[\phi_{i_1} \cdots \phi_{i_n}(x_-), \phi_{i_1} \cdots \phi_{i_n}(x_+)]}. \tag{3.12}$$

This means that we only need to consider the orbits of the two points x_+ and x_-.

3.2 Existence of the fractal probability distribution

In this section, we are going to show that the fractal probability distribution associated with ϕ exists and is unique.

In order to demonstrate this, it is sufficient to show that the corresponding iterated function system with probabilities is weakly hyperbolic and then the result follows from section 2.9

Before we start, we will state and prove two useful propositions.

Proposition 3.2.1

Given an IFS with probabilities $\{X; f_1, \ldots, f_N; p_1, \ldots, p_N\}$ if there exists an integer $r \geq 1$ such that $f_{i_1} \circ \cdots \circ f_{i_r}$ is a contracting map for all sequences $i_1, \ldots, i_r \in \{1, \ldots, N\}$ then the IFS is weakly hyperbolic.

Proof

Since $X = (X, d_X)$ is a compact metric space and every compact metric space is bounded, there exists $K \geq 0$ such that $d_X(x,y) \leq K$ for all $x, y \in X$.

Since $f_{i_1} \circ \cdots \circ f_{i_r}$ is a contracting map for all sequences $i_1, \ldots, i_r \in \{1, \ldots, N\}$, there exists $\beta_{i_1 \cdots i_r} < 1$ such that

$$|f_{i_1} \cdots f_{i_r}(x) - f_{i_1} \cdots f_{i_r}(y)| \leq \beta_{i_1 \cdots i_r} |x - y|.$$

Let $\alpha = \max\{\beta_{i_1 \cdots i_r} \mid i_1, \ldots, i_r \in \{1, \ldots, N\}\}$. Clearly $\alpha < 1$ and

$$|f_{i_1} \cdots f_{i_r}(x) - f_{i_1} \cdots f_{i_r}(y)| \leq \alpha K.$$

So, given any $\epsilon > 0$, choose an integer

$$n > \frac{\log \epsilon - \log (2K+1)K}{\log \alpha}$$

and any $m \geq rn$. Then

$$\begin{aligned}
|f_{i_1} \cdots f_{i_m}(x) - f_{i_1} \cdots f_{i_m}(y)| &\leq |f_{i_1} \cdots f_{i_m}(x) - f_{i_1} \cdots f_{i_{rn}}(x)| + \\
&\quad\; |f_{i_1} \cdots f_{i_{rn}}(x) - f_{i_1} \cdots f_{i_{rn}}(y)| + \\
&\quad\; |f_{i_1} \cdots f_{i_{rn}}(y) - f_{i_1} \cdots f_{i_m}(y)| \\
&\leq (K + 1 + K)\alpha^n K \\
&< \epsilon
\end{aligned}$$

for all $x, y \in X$. □

Proposition 3.2.2

Given a map $f : [a,b] \to \mathbb{R}$ such that f and f' are continuous over $[a,b]$, if $|f'(x)| < 1$ for all $x \in [a,b]$ then f is contracting.

Proof

f' attains its supremum and infimum because $[a,b]$ is compact and f' is continuous. Therefore, there exists $\alpha < 1$ such that

$$|f'(x)| \leq \alpha.$$

Therefore, for all $x, y \in [a, b]$, there exists $z \in (a, b)$, such that

$$f(x) - f(y) = f'(z)(x - y)$$

by the mean value theorem. Therefore

$$|f(x) - f(y)| \leq \alpha|x - y|.$$

$$\square$$

Now we need to show that the iterated function system with probabilities given by $\{[x_-, x_+]; \phi_+, \phi_-; \frac{1}{2}, \frac{1}{2}\}$ where $\phi_\pm(x) = \phi(x \pm \epsilon)$, is weakly hyperbolic.

Using differential calculus, we have

$$\begin{aligned}
\phi'_{++}(x) &= \phi'(x+\epsilon)\phi'(\phi(x+\epsilon)+\epsilon) \\
\phi'_{+-}(x) &= \phi'(x+\epsilon)\phi'(\phi(x+\epsilon)-\epsilon) \\
\phi'_{-+}(x) &= \phi'(x-\epsilon)\phi'(\phi(x-\epsilon)+\epsilon) \\
\phi'_{--}(x) &= \phi'(x-\epsilon)\phi'(\phi(x-\epsilon)-\epsilon)
\end{aligned}$$

where

$$\begin{aligned}
\phi_{++}(x) &= \phi(\phi(x+\epsilon)+\epsilon) \\
\phi_{+-}(x) &= \phi(\phi(x+\epsilon)-\epsilon) \\
\phi_{-+}(x) &= \phi(\phi(x-\epsilon)+\epsilon) \\
\phi_{--}(x) &= \phi(\phi(x-\epsilon)-\epsilon).
\end{aligned}$$

Notice that $\phi'(x) > 0$ because ϕ is monotonically increasing, odd and strictly concave for $x > 0$. Also, $\phi'(x) < 1$ for $x \neq 0$ because ϕ is odd, strictly concave for $x > 0$ and $\phi'(0) = 1$. Therefore, $0 < \phi'_{\pm\pm}(x) \leq 1$.

Also, $\phi(0) = 0$ because ϕ is odd and continuous.

However, if $\phi'_{\pm\pm}(x) = 1$ then $x = \pm\epsilon$ and $\phi(x \mp \epsilon) = \phi(0) = 0$ and so $\epsilon = 0$. But $\epsilon > 0$, so $\phi'_{\pm\pm}(x) < 1$ by contradiction.

Therefore by proposition 3.2.2, ϕ_{++}, ϕ_{+-}, ϕ_{-+} and ϕ_{--} are contracting maps. Therefore by proposition 3.2.1, the iterated function system with probabilities $\{[x_-, x_+]; \phi_+, \phi_-; \frac{1}{2}, \frac{1}{2}\}$ is weakly hyperbolic. This proves that all smooth learning schemes give rise to a unique fractal probability distribution.

Chapter 4

Embedding strength decay rate of stored patterns

An important physical quantity in a forgetful neural network is the embedding strength decay rate of the stored patterns. It will be shown later how this is related to the storage capacity of the neural network.

4.1 Embedding strengths of stored patterns

The *embedding strength* of pattern $\mathbf{X}^m$ is defined as

$$e_m = \frac{1}{N} \sum_{i,j=1}^{N} J_{ij} X_i^m X_j^m. \tag{4.1}$$

We know that the embedding strengths of all stored patterns decay to zero as further patterns are subsequently stored by virtue of the properties of a forgetful neural network.

Also, it would seem reasonable that this decay is exponential on average when large numbers of patterns have been previously stored because all patterns are homogeneous in character. This leads effectively to investigating [35, 4, 12] the Lyapunov stability of the neural network, where the decay rate is called the *Lyapunov exponent.*

Therefore, in the limiting case as $n \to \infty$, we have

$$e_m \sim \exp(\gamma n) \tag{4.2}$$

where $n = M+1-m$ and γ is the average Lyapunov exponent. This average Lyapunov exponent is also called the embedding strength decay rate, which we expect to be a negative number.

4.2 Average Lyapunov exponent

Generally speaking, the Lyapunov exponent [22, 28] of a dynamical system is the average exponential rate at which the resulting motion of the system starting from two slightly different initial positions depart from each other, assuming that it is exponential.

In this particular case, the embedding strengths are asymptotically attracted to the same point, namely zero, regardless of the initial conditions, namely the stored pattern.

Recall that we are considering the non-deterministic dynamical system corresponding to the IFS with probabilities $\{[x_-, x_+]; \phi_+, \phi_-; \frac{1}{2}, \frac{1}{2}\}$ where $\phi_\pm(x) = \phi(x \pm \epsilon)$, $x_\pm$ is the least fixed point of $\phi_\pm$ and ϕ is odd ($\phi(-x) = -\phi(x)$), monotonically increasing and strictly concave for $x > 0$ ($\phi''(x) < 0$) and $\phi'(0) = 1$.

So, there is a single average Lyapunov exponent which is independent of x and the infinite sequence $i_1, i_2, \ldots \in \{+, -\}$ such that for large n and small $\mathrm{d}x$

$$|\phi_{i_n} \cdots \phi_{i_1}(x + \mathrm{d}x) - \phi_{i_n} \cdots \phi_{i_1}(x)| \approx \mathrm{d}x \exp(\gamma n). \tag{4.3}$$

Therefore

$$\begin{aligned} \gamma &= \lim_{n\to\infty} \lim_{\mathrm{d}x \to 0} \frac{1}{n} \log \left| \frac{\phi_{i_n} \cdots \phi_{i_1}(x+\mathrm{d}x) - \phi_{i_n} \cdots \phi_{i_1}(x)}{\mathrm{d}x} \right| \\ &= \lim_{n\to\infty} \frac{1}{n} \log \left| \frac{\mathrm{d}}{\mathrm{d}x} \phi_{i_n} \cdots \phi_{i_1}(x) \right| \end{aligned}$$

$$
\begin{aligned}
&= \lim_{n\to\infty} \frac{1}{n+1} \log \prod_{m=0}^{n} \left|\phi'_{i_{m+1}}\phi_{i_m}\cdots\phi_{i_1}(x)\right| \\
&= \lim_{n\to\infty} \frac{1}{n+1} \sum_{m=0}^{n} \log \left|\phi'_{i_{m+1}}\phi_{i_m}\cdots\phi_{i_1}(x)\right|
\end{aligned}
$$

by the definition of differentiation and the chain rule. So

$$
\begin{aligned}
\gamma &= \lim_{n\to\infty} \frac{1}{n+1} \sum_{m=0}^{n} \log \phi'_+ \begin{cases} \phi_{i_m}\cdots\phi_{i_1}(x) & \text{if } i_{m+1} = + \\ \phi_{-i_m}\cdots\phi_{-i_1}(-x) & \text{if } i_{m+1} = - \end{cases} \\
&= \lim_{n\to\infty} \frac{1}{n+1} \sum_{m=0}^{n} \log \phi'_+\phi_{i_m}\cdots\phi_{i_1}(x) \\
&= \lim_{n\to\infty} \frac{1}{n+1} \sum_{m=0}^{n} g(\phi_{i_m}\cdots\phi_{i_1}(x))
\end{aligned}
$$

where

$$g(x) = \log \phi'_+(x) \tag{4.4}$$

due to $\phi'_\pm \geq 0$, $\phi_+(-x) = -\phi_-(x)$, $\phi'_+(-x) = \phi'_-(x)$ and the independence property.

So, by equation 2.12 and 2.13 from Elton's ergodic theorem [14]

$$\gamma = \int g(x)\, \mathrm{d}\mu(x) \tag{4.5}$$

where μ is the unique invariant measure of the iterated function systems with probabilities $\{[x_-, x_+]; \phi_+, \phi_-; \frac{1}{2}, \frac{1}{2}\}$ provided that there exists $\alpha < 1$ such that

$$|\phi_+(x) - \phi_+(y)|^{\frac{1}{2}} |\phi_-(x) - \phi_-(y)|^{\frac{1}{2}} \leq \alpha |x - y| \tag{4.6}$$

for all $x \in [x_-, x_+]$.

Before we prove that this condition is satisfied, we will state and prove a proposition that we will need.

Proposition 4.2.1

Given an iterated function system with probabilities

$$\{[a, b]; f_1, \ldots, f_N; p_1, \ldots, p_N\}$$

where $f'_1, \ldots, f'_N$ are continuous over $[a, b]$, if

$$\prod_{i=1}^{N} |f'_i(x)|^{p_i} < 1 \tag{4.7}$$

for all $x \in [a, b]$ then there exists $\alpha < 1$ such that

$$\prod_{i=1}^{N} |f_i(x) - f_i(y)|^{p_i} \leq \alpha |x - y| \tag{4.8}$$

for all $x, y \in [a, b]$.

Proof

$\prod_{i=1}^{N} |f'_i(x)|^{p_i}$ attains its supremum because $[a, b]$ is compact and $f'_1, \ldots, f'_N$ are continuous. Therefore, there exists $\alpha < 1$ such that

$$\prod_{i=1}^{N} |f'_i(x)|^{p_i} \leq \alpha.$$

Therefore, for all $x, y \in [a, b]$ and $i \in \{1, \ldots, N\}$, there exists $z_i \in (a, b)$ such that

$$f_i(x) - f_i(y) = f'_i(z_i)(x - y)$$

by the mean value theorem. Therefore

$$\begin{aligned}
\prod_{i=1}^{N} |f_i(x) - f_i(y)|^{p_i} &= \prod_{i=1}^{N} |f'_i(z_i)(x - y)|^{p_i} \\
&= (\prod_{i=1}^{N} |f'_i(z_i)|^{p_i}) |x - y|^{\sum_{i=1}^{N} p_i} \\
&\leq \alpha |x - y|.
\end{aligned}$$

□

Returning to the problem in hand, we know that $0 < \phi'_+(x) < 1$ for $x \neq -\epsilon$ and $0 < \phi'_-(x) < 1$ for $x \neq \epsilon$. Therefore, $0 < \phi'_+(x)\phi'_-(x) < 1$ because $e > 0$ and so

$$|\phi'_+(x)|^{\frac{1}{2}} |\phi'_-(x)|^{\frac{1}{2}} < 1. \tag{4.9}$$

Hence, using proposition 4.2.1 on the interval $[x_-, x_+]$, condition 4.6 is satisfied. Therefore, equation 4.5 is valid for all smooth learning schemes.

4.3 Algorithm to compute the Lyapunov exponent

Using the theory so far, we can derive an algorithm for the Lyapunov exponent.

From equations 4.5, 2.7 and 2.8

$$S^l_{[x_-,x_+]}(g,\nu_n) \leq \gamma \leq S^u_{[x_-,x_+]}(g,\nu_n) \tag{4.10}$$

where from equations 2.4, 2.5 and 3.12 the generalised upper and lower Riemann sums are

$$S^u_{[x_-,x_+]}(g,\nu_n) = \frac{1}{2^n} \sum_{i_1,\ldots,i_n \in \{+,-\}} \sup\, g[\phi_{i_1}\cdots\phi_{i_n}(x_-), \phi_{i_1}\cdots\phi_{i_n}(x_+)] \tag{4.11}$$

$$S^l_{[x_-,x_+]}(g,\nu_n) = \frac{1}{2^n} \sum_{i_1,\ldots,i_n \in \{+,-\}} \inf\, g[\phi_{i_1}\cdots\phi_{i_n}(x_-), \phi_{i_1}\cdots\phi_{i_n}(x_+)] \tag{4.12}$$

where we recall that $g(x) = \log \phi'_+(x)$, $\phi_\pm(x) = \phi(x \pm \epsilon)$, $x_\pm$ is the least fixed point of $\phi_\pm$ and ϕ is odd ($\phi(-x) = -\phi(x)$), monotonically increasing and strictly concave for $x > 0$ ($\phi''(x) < 0$) and $\phi'(0) = 1$ and so

$$g'(x) = \frac{\phi''_+(x)}{\phi'_+(x)}.$$

Therefore, it follows that $g(x-\epsilon)$ is even with respect to x ($g(x-\epsilon) = g(-x-\epsilon)$) and $g'(x) < 0$ for $x > -\epsilon$ and $g'(x) > 0$ for $x < -\epsilon$ and $g'(x) = 0$ for $x = -\epsilon$.

Therefore

$$\sup\, g[a,b] = \begin{cases} g(a) & \text{if } a > -\epsilon \\ g(b) & \text{if } b < -\epsilon \\ g(-\epsilon) & \text{otherwise} \end{cases} \tag{4.13}$$

$$\inf g[a,b] = \begin{cases} g(a) & \text{if } |a+\epsilon| > |b+\epsilon| \\ g(b) & \text{otherwise} \end{cases}. \tag{4.14}$$

To summarise so far, the algorithm entails calculating the generalised upper and lower Riemann sums given by equations 4.11, 4.12, 4.13 and 4.14 for n large enough so that the difference between the sums is less than or equal to a given maximum allowed error θ, since we know that γ lies somewhere between them.

So, it only remains to determine a value for n such that

$$S^u_{[x_-,x_+]}(g,\nu_n) - S^l_{[x_-,x_+]}(g,\nu_n) \leq \theta. \tag{4.15}$$

Note that evaluating the generalised upper and lower Riemann sums involves calculating $O(2^n)$ floating point expressions.

Therefore, evaluating the generalised upper and lower Riemann sums starting from $n = 1$ and incrementing by one until the difference drops to the desired accuracy θ would require approximately twice as much time as it would if the required n was already known.

There are various heuristic possibilities for improving on this, but they tend to be problem dependent. The one that we employed involves evaluating the generalised upper and lower Riemann sums for three values of n, two fixed and one variable. It is based on the assumption that the gap decreases exponentially on "average".

Choose n_1 and n_2 low enough so that the evaluation of the generalised upper and lower Riemann sums is quick, but not so low that the word "average" in the above assumption becomes meaningless. Define e_i as

$$e_i = \log\left(S^u_{[x_-,x_+]}(g,\nu_{n_i}) - S^l_{[x_-,x_+]}(g,\nu_{n_i})\right).$$

Then, solving the appropriate simultaneous equations

$$n_3 = \left\lceil \frac{(n_1 - n_2)\log\theta + e_1 n_2 - e_2 n_1}{e_1 - e_2} \right\rceil \tag{4.16}$$

where $\lceil a \rceil$ denotes the least integer greater than or equal to a.

Also, define l_i and u_i as

$$l_i \quad = \quad S^l_{[x_-,x_+]}(g,\nu_{n_i})$$

$$u_i = S^u_{[x_-,x_+]}(g, \nu_{n_i})$$

As an illustrative example for the remainder of this section, we will take $\phi(x) = \tanh(x)$, $\epsilon = 1.0$ and $\theta = 0.000001$.

So choosing $n_1 = 6$ and $n_2 = 7$, we have

$$\begin{aligned} e_1 &= -1.4963053 \\ e_2 &= -1.8143737 \\ n_3 &= 21 \\ l_3 &= -1.09399808 \\ u_3 &= -1.09399753. \end{aligned}$$

We will discuss here two other methods for estimating n, but unless it is equal to or one greater than the optimal n it is of purely academic value for this problem.

The first method relies on the fact that ϕ_{++}, ϕ_{+-}, ϕ_{-+} and ϕ_{--} as defined and shown in section 3.2 are contracting maps and therefore, the IFS

$$\{[x_-, x_+]; \phi_{++}, \phi_{+-}, \phi_{-+}, \phi_{--}; \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\}$$

is hyperbolic. So, if g satisfies a Lipschitz condition, then we can obtain a finite algorithm to calculate the integral to any given accuracy [10]. Suppose there exists $k > 0$ and $c > 0$ such that

$$|g(x) - g(y)| \leq c|x - y|^k$$

for all $x, y \in [x_-, x_+]$. Then equation 4.15 is satisfied for

$$n = \left\lceil \frac{\frac{1}{k}\log\left(\frac{\theta}{c}\right) - \log|X|}{\log s} \right\rceil \tag{4.17}$$

where s is the square root of the contractivity of the above hyperbolic IFS.

In fact, g is an analytic function. Therefore

$$|g(x) - g(y)| \leq c|x - y|$$

for $c = \max_{x_- \le x \le x_+} |g'(x)|$. So

$$\begin{aligned} c &= 2.57527 \\ k &= 1 \\ |X| &= 1.92236 \\ s &= 0.773724 \\ n &= 61 \end{aligned}$$

Clearly, this upper bound for n is too generous to be useful.

The second method is based on a proposition by Edalat [12, Proposition 2.2]. In summary, there exists $n \ge 0$ such that

$$\frac{k}{2^n} < \frac{\theta}{2}$$

where k is the number of sequences $i_1, \ldots, i_n \in \{+, -\}$ of length n such that the diameter of the set $\phi_{i_1} \cdots \phi_{i_n}[x_-, x_+]$ is at least $\frac{\theta}{2c}$. In which case, equation 4.15 holds. However, for $n = 21$, $k = 201448$ and therefore $\theta > 0.192116$. This is a long way from 0.000001 and so again, this method falls short of being practical for this application.

4.4 Asymptotes of the smooth learning scheme

Let us consider analytically the Lyapunov exponent for small ϵ. Maclaurin's expansion for $\phi(x)$ is

$$\phi(x) = \sum_{n=0}^{\infty} \frac{1}{n!} \phi^{(n)}(0) x^n$$

where

$$\phi^{(n)} = \frac{\mathrm{d}^n}{\mathrm{d}x^n} \phi(x).$$

However, ϕ is odd, therefore $\phi^{(n)}(0) = 0$ for n even. For n odd, let r be the least odd number greater than 1 such that $\phi^{(r)}(0) \ne 0$ and since ϕ is strictly concave for $x > 0$ this means that $\phi^{(r)}(0) < 0$. So, that gives

$$\phi(x) = x + \frac{1}{r!} \phi^{(r)}(0) x^r + O(x^{r+2}).$$

We know that x_+ satisfies

$$x_+ = \phi(x_+ + \epsilon)$$

and the inverse of ϕ is well defined because ϕ is monotonically increasing, therefore

$$\begin{aligned}\epsilon &= \phi^{-1}x_+ - x_+ \\ &\approx -\frac{\phi^{(r)}(0)}{r!}x_+^r + O(x_+^{r+2}).\end{aligned}$$

Therefore, it can be shown that

$$x_+ \approx -\left(\frac{r!\,\epsilon}{\phi^{(r)}(0)}\right)^{\frac{1}{r}} + O(\epsilon^{\frac{3}{r}}).$$

Therefore

$$\begin{aligned}S^u_{[x_-,x_+]}(g,\nu_0) &= g(-\epsilon) \\ &= \log\phi'(0) \\ &= 0 \\ S^l_{[x_-,x_+]}(g,\nu_0) &= g(x_+) \\ &\approx r\left(\frac{\phi^{(r)}(0)}{r!}\right)^{\frac{1}{r}}\epsilon^{\frac{r-1}{r}} + O(\epsilon^{\frac{r+1}{r}}).\end{aligned}$$

Therefore, for sufficiently small ϵ

$$(r+1)\left(\frac{\phi^{(r)}(0)}{r!}\right)^{\frac{1}{r}}\epsilon^{\frac{r-1}{r}} < \gamma(\epsilon) \le 0.$$

In other words, $\gamma \to 0$ as $\epsilon \to 0$.

4.5 Asymptotes of the hyperbolic tangent learning scheme

For the hyperbolic tangent learning scheme

$$\phi(x) = \tanh(x) \tag{4.18}$$

$$g(x) = -2\log\cosh(x+\epsilon). \tag{4.19}$$

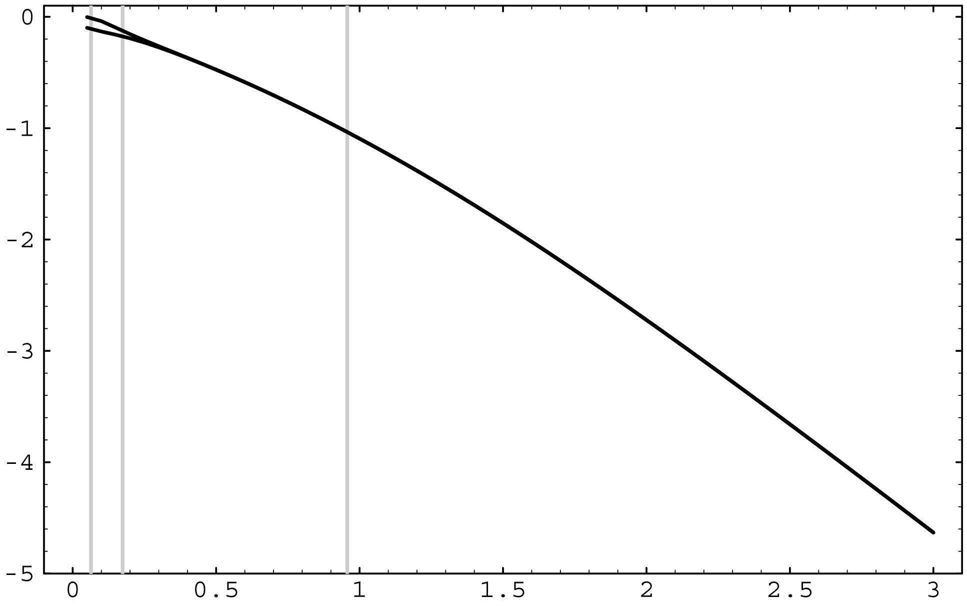

Figure 4.1: Plot of $S^l_{[-1,1]}(g,\nu_n)$ and $S^u_{[-1,1]}(g,\nu_n)$ against ϵ for the hyperbolic tangent learning scheme where the two curves are nearly indistinguishable

Let us consider analytically the Lyapunov exponent for large ϵ. The support of the invariant measure is in $[x_-, x_+]$, therefore

$$\int g(x)\ \mathrm{d}\mu(x) = \int h(x)\ \mathrm{d}\mu(x)$$

where

$$h(x) = \begin{cases} g(x) & \text{if } x \in [x_-, x_+] \\ 0 & \text{otherwise} \end{cases} .$$

But, for $x \in [x_-, x_+]$ and ϵ large

$$\begin{aligned} g(x) &\approx -2\log\left(\frac{\exp\epsilon}{2}\right) \\ &= \log 4 - 2\epsilon. \end{aligned}$$

Therefore $\gamma(\epsilon) \approx \log 4 - 2\epsilon$ for large ϵ.

Some of the computed results for various ϵ and n are listed in table 4.1 with the associated graph shown in figure 4.1. Notice that n has to be increased as ϵ decreases in order to maintain accuracy. However, due to

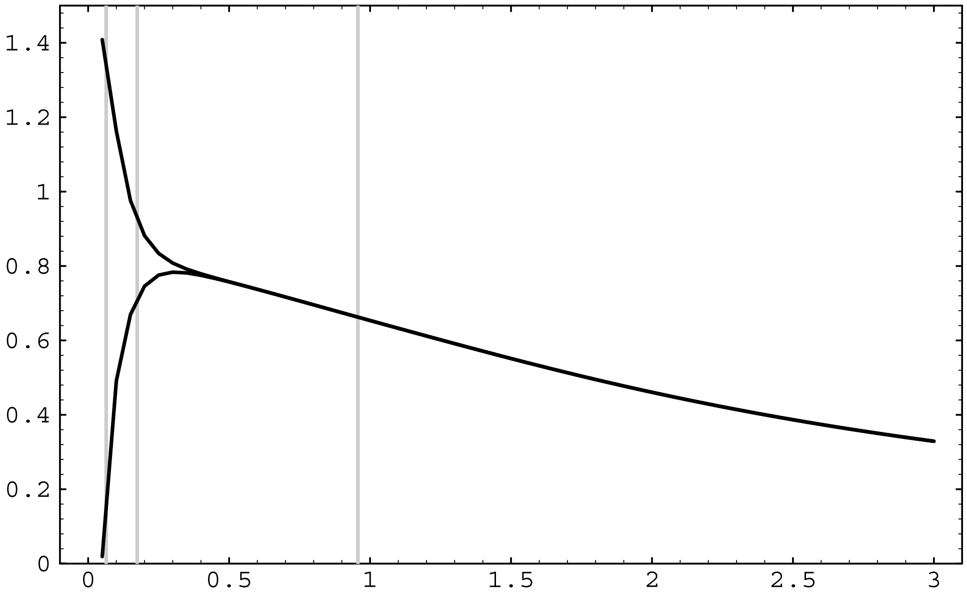

Figure 4.2: Plot of $h(S^l_{[-1,1]}(g, \nu_n))$ and $h(S^u_{[-1,1]}(g, \nu_n))$ against ϵ for the hyperbolic tangent learning scheme

limited computational power n had to be restricted to the maximum value of 26.

The Lyapunov exponent is between $S^l_{[-1,1]}(g, \nu_n)$ and $S^u_{[-1,1]}(g, \nu_n)$. Figure 4.2 shows the graph of $h(S^l_{[-1,1]}(g, \nu_n))$ and $h(S^u_{[-1,1]}(g, \nu_n))$ against ϵ where

$$h(x) = \frac{x + 2\epsilon}{\epsilon \log 4}. \tag{4.20}$$

The 4^{th}-degree polynomial least-squares fit to the data between $\epsilon = 0.4$ and $\epsilon = 3.0$ is

$$0.846879 - 0.151116\,\epsilon - 0.0705225\,\epsilon^2 + 0.0320542\,\epsilon^3 - 0.00364559\,\epsilon^4. \tag{4.21}$$

Figure 4.3 shows the graph of $h(S^l_{[-1,1]}(g, \nu_n))$, $h(S^u_{[-1,1]}(g, \nu_n))$ and equation 4.21, against ϵ in the critical region where the data diverges.

Therefore

$$\gamma(\epsilon) \quad \approx \quad (0.846879 \log 4 - 2)\epsilon$$

n	ϵ	$S^l_{[-1,1]}(g,\nu_n)$	$S^u_{[-1,1]}(g,\nu_n)$
26	0.1	-0.131913	-0.038778
26	0.2	-0.193185	-0.155600
26	0.3	-0.274155	-0.263971
26	0.4	-0.370240	-0.367999
26	0.5	-0.475128	-0.474708
26	0.6	-0.586711	-0.586643
26	0.7	-0.704538	-0.704528
26	0.8	-0.828478	-0.828477
23	0.9	-0.958360	-0.958359
20	1.0	-1.093998	-1.093997
18	1.1	-1.235260	-1.235259
16	1.2	-1.382036	-1.382035
14	1.3	-1.534179	-1.534177
13	1.4	-1.691454	-1.691452
12	1.5	-1.853548	-1.853547
11	1.6	-2.020089	-2.020087
10	1.7	-2.190682	-2.190680
10	1.8	-2.364935	-2.364934
10	1.9	-2.542478	-2.542478
10	2.0	-2.722967	-2.722967
10	2.1	-2.906085	-2.906085
10	2.2	-3.091543	-3.091543
10	2.3	-3.279073	-3.279073
10	2.4	-3.468428	-3.468428
10	2.5	-3.659384	-3.659384

Table 4.1: Computed values of $S^l_{[-1,1]}(g,\nu_n)$ and $S^u_{[-1,1]}(g,\nu_n)$ for various ϵ and n for the hyperbolic tangent learning scheme

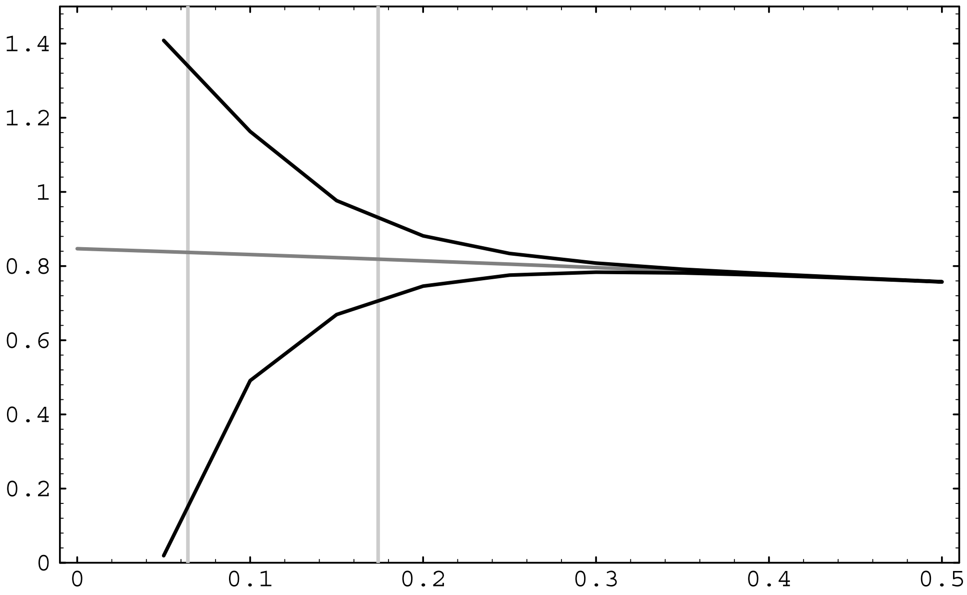

Figure 4.3: Plot of equation 4.21, $h(S^l_{[-1,1]}(g,\nu_n))$ and $h(S^u_{[-1,1]}(g,\nu_n))$ against ϵ in the critical region for the hyperbolic tangent learning scheme

$$\begin{aligned} &\approx -0.825976\,\epsilon && (4.22) \\ e_m &\sim \exp(-0.826 n\epsilon) && \\ &\sim \exp\left(\frac{-\theta(\tanh)kn}{N}\right) && (4.23) \end{aligned}$$

where

$$\theta(\tanh) = 0.826. \qquad (4.24)$$

Note that the grey vertical lines in figures 4.1, 4.2 and 4.3 correspond to the following critical values for ϵ evaluated by Behn et al [4]:

$$\begin{aligned} \epsilon_{\mathrm{C}}^{(1)} &= 0.957 \\ \epsilon_{\mathrm{C}}^{(2)} &= 0.174 \\ \epsilon_{\mathrm{C}}^{(3)} &= 0.064 \end{aligned}$$

For $\epsilon > \epsilon_{\mathrm{C}}^{(1)}$, the support of the invariant distribution is a fractal, while for $\epsilon < \epsilon_{\mathrm{C}}^{(1)}$, the support of the invariant distribution is the whole interval $[x_-, x_+]$.

For $\epsilon > \epsilon_{\rm C}^{(2)}$, the invariant distribution is infinite at the boundaries of it's support, while for $\epsilon < \epsilon_{\rm C}^{(2)}$, the invariant distribution is zero at the boundaries of it's support,

For $\epsilon > \epsilon_{\rm C}^{(3)}$, the invariant distribution has an infinite gradient at the boundaries of it's support, while for $\epsilon < \epsilon_{\rm C}^{(3)}$, the invariant distribution has a zero gradient at the boundaries of it's support,

So, although the invariant distribution has sudden changes in character at these three critical values, it does not appear to manifest itself as kinks or discontinuities in the evaluation of the average Lyapunov exponent.

4.6 Asymptotes of the error function learning scheme

For the error function learning scheme

$$\phi(x) = \mathrm{erf}\left(\frac{\sqrt{\pi}}{2}x\right) \tag{4.25}$$

$$g(x) = -\frac{\pi}{4}(x+\epsilon)^2. \tag{4.26}$$

Let us consider analytically the Lyapunov exponent for large ϵ. The support of the invariant measure is in $[x_-, x_+]$, therefore

$$\int g(x)\ \mathrm{d}\mu(x) = \int h(x)\ \mathrm{d}\mu(x)$$

where

$$h(x) = \begin{cases} g(x) & \text{if } x \in [x_-, x_+] \\ 0 & \text{otherwise} \end{cases} .$$

But, for $x \in [x_-, x_+]$ and ϵ large

$$g(x) \approx -\frac{\pi}{4}\epsilon^2.$$

Therefore $\gamma(\epsilon) \approx -\frac{\pi}{4}\epsilon^2$ for large ϵ.

Some of the computed results for various ϵ and n are listed in table 4.2 with the associated graph shown in figure 4.4. Notice that n has to be

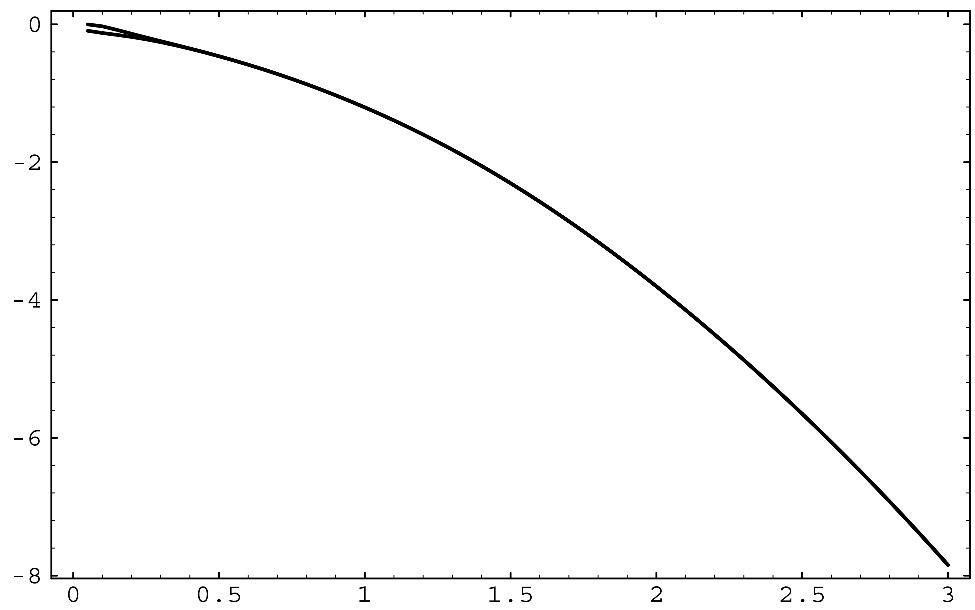

Figure 4.4: Plot of $S^{l}_{[-1,1]}(g,\nu_n)$ and $S^{u}_{[-1,1]}(g,\nu_n)$ against ϵ for the error function learning scheme where the two curves are nearly indistinguishable

increased as ϵ decreases in order to maintain accuracy. However, due to limited computational power n had to be restricted to the maximum value of 26.

The Lyapunov exponent is between $S^{l}_{[-1,1]}(g,\nu_n)$ and $S^{u}_{[-1,1]}(g,\nu_n)$. Figure 4.5 shows the graph of $h(S^{l}_{[-1,1]}(g,\nu_n))$ and $h(S^{u}_{[-1,1]}(g,\nu_n))$ against ϵ where

$$h(x) = \frac{\gamma(\epsilon)}{\epsilon} + \frac{\pi}{4}\epsilon. \tag{4.27}$$

The 4^{th}-degree polynomial least-squares fit to the data between $\epsilon = 0.3$ and $\epsilon = 3.0$ is

$$-0.742225 + 2.89879\,\epsilon - 0.281287\,\epsilon^2 + 0.0673135\,\epsilon^3 - 0.00527905\,\epsilon^4. \tag{4.28}$$

Figure 4.6 shows the graph of $h(S^{l}_{[-1,1]}(g,\nu_n))$, $h(S^{u}_{[-1,1]}(g,\nu_n))$ and equation 4.28, against ϵ in the critical region where the data diverges.

n	ϵ	$S^l_{[-1,1]}(g,\nu_n)$	$S^u_{[-1,1]}(g,\nu_n)$
26	0.1	-0.125764	-0.029083
26	0.2	-0.182791	-0.136472
26	0.3	-0.258522	-0.243816
26	0.4	-0.353779	-0.350056
26	0.5	-0.463458	-0.462669
26	0.6	-0.585668	-0.585527
26	0.7	-0.720405	-0.720384
26	0.8	-0.868216	-0.868213
24	0.9	-1.029603	-1.029602
21	1.0	-1.204878	-1.204877
18	1.1	-1.394365	-1.394363
16	1.2	-1.598595	-1.598592
15	1.3	-1.818244	-1.818242
14	1.4	-2.053908	-2.053907
13	1.5	-2.305894	-2.305893
12	1.6	-2.574156	-2.574155
11	1.7	-2.858359	-2.858358
10	1.8	-3.158009	-3.158008
10	1.9	-3.472560	-3.472560
10	2.0	-3.801508	-3.801508
10	2.1	-4.144436	-4.144436
10	2.2	-4.501038	-4.501038
10	2.3	-4.871132	-4.871132
10	2.4	-5.254652	-5.254652
10	2.5	-5.651635	-5.651635

Table 4.2: Computed values of $S^l_{[-1,1]}(g,\nu_n)$ and $S^u_{[-1,1]}(g,\nu_n)$ for various ϵ and n for the error function learning scheme

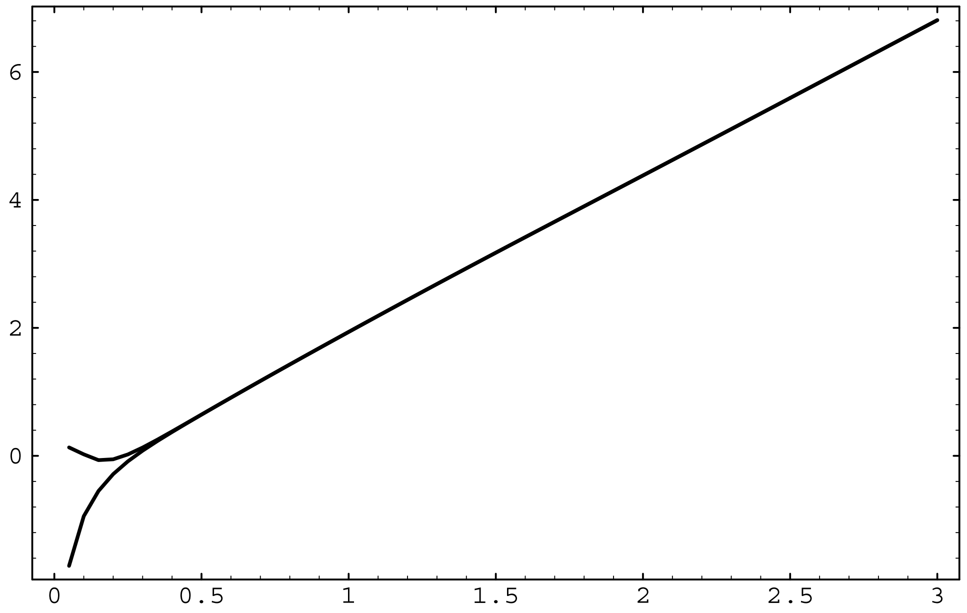

Figure 4.5: Plot of $h(S^l_{[-1,1]}(g, \nu_n))$ and $h(S^u_{[-1,1]}(g, \nu_n))$ against ϵ for the error function learning scheme

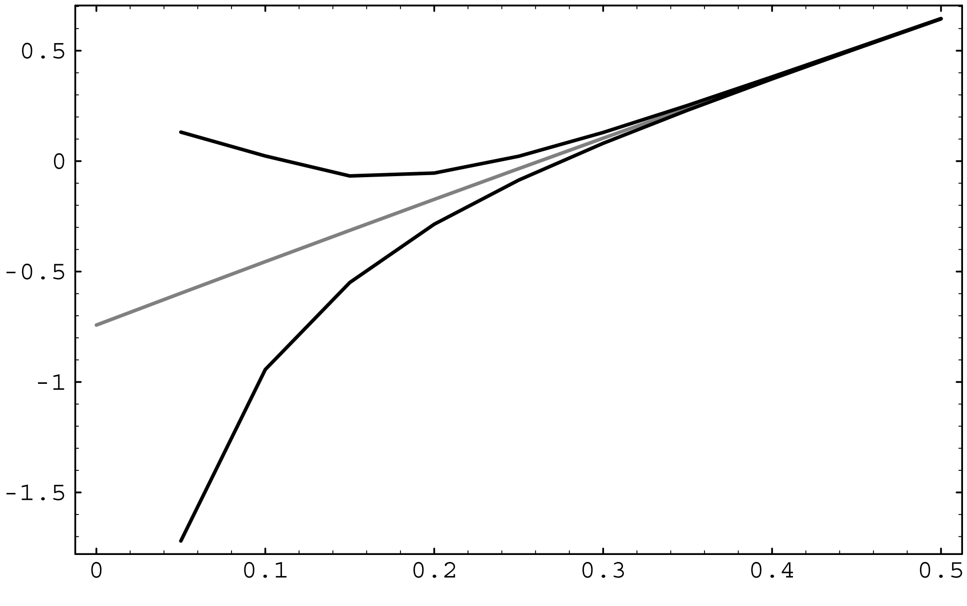

Figure 4.6: Plot of equation 4.28, $h(S^l_{[-1,1]}(g, \nu_n))$ and $h(S^u_{[-1,1]}(g, \nu_n))$ against ϵ in the critical region for the error function learning scheme

Therefore

$$\gamma(\epsilon) \approx -0.742225\,\epsilon \tag{4.29}$$

$$\begin{aligned} e_m &\sim \exp(-0.742 n\epsilon) \\ &\sim \exp\left(\frac{-\theta(\mathrm{erf})kn}{N}\right) \end{aligned} \tag{4.30}$$

where

$$\theta(\mathrm{erf}) = 0.742. \tag{4.31}$$

Chapter 5

Ising model for the marginalist learning scheme

In order to make analytical progress, it is useful to consider the addition of noise to the network and then take the limit as the noise tends to zero. The most fruitful way to do this is to model the network as a simplified description of magnetism, namely the Ising model. The noise level is usually called the temperature T.

5.1 General principles of statistical mechanics

The basic quantity characterising a system in statistical mechanics is the *energy* $H(\mathbf{x})$, which is defined as some function of the microscopic system state $\mathbf{x} = (x_1, \ldots, x_N)$. In general, the behaviour of the system is defined so that its energy tends to its minimum value. However, no observable system can be perfectly isolated from its surroundings and the effect of interaction with it manifests itself in the form of thermal noise.

Given an observable quantity $O(\mathbf{x})$, the quantity of interest in statistical mechanics is the time averaged value, namely

$$\langle O\rangle = \lim_{t\to\infty}\frac{1}{t}\int_0^t O(\mathbf{x}(\tau))\,\mathrm{d}\tau. \tag{5.1}$$

An *order parameter* is an observable quantity whose time average plays an important role in describing the macroscopic characteristics of the model.

The fundamental hypothesis of statistical mechanics is that if we know the energy $H(\mathbf{x})$ for every state $\mathbf{x}$ of the system, then the properties of the system, in equilibrium at temperature T, can be computed as if the probability of finding the system in a particular state is proportional to $\exp\left(\frac{-H(\mathbf{x})}{kT}\right)$, where k is Boltzmann's constant.

Therefore, the time average of $O(\mathbf{x})$ is given by

$$\langle O\rangle = \sum_{x_1,\ldots,x_N\in\{-1,1\}} O(\mathbf{x})\rho(\mathbf{x}) \tag{5.2}$$

where

$$\begin{aligned}
\rho(\mathbf{x}) &= \frac{1}{Z(\mathbf{x})}\exp(-\beta H(\mathbf{x})) &(5.3)\\
Z(\mathbf{x}) &= \sum_{x_1,\ldots,x_N\in\{-1,1\}} \exp(-\beta H(\mathbf{x})) &(5.4)\\
\beta &= \frac{1}{kT}. &(5.5)
\end{aligned}$$

Here $\rho(\mathbf{x})$ is the *probability distribution function* of the system, $Z(\mathbf{x})$ is called the *partition function* and β is called the *inverse temperature.*

In statistical mechanics, it is the *entropy* S that characterises the probability distribution. It is defined as the average of the logarithm of the distribution function.

$$S = -\langle\log\rho\rangle. \tag{5.6}$$

In general, entropy is a measure of the degree of disorder in a system. We can illustrate this with a simple example. Imagine a system whose states have a probability distribution such that L states have equal probability $\frac{1}{L}$

and all other states have probability zero. According to the definition of entropy

$$S = -\sum_{l=1}^{L} \frac{1}{L} \log\left(\frac{1}{L}\right) = \log L.$$

Therefore, the more broad the distribution, the larger the entropy. On the other hand, the more narrow the distribution, the smaller the entropy. In the extreme case with only one state, the entropy is zero. In Nature, systems change so as to increase their entropy in order to achieve maximum disorder. In a sense, noise is natural.

According to the basic hypothesis, the time average of the energy is

$$E \equiv \langle H \rangle = \sum_{x_1,\ldots,x_N \in \{-1,1\}} H(\mathbf{x})\rho(\mathbf{x}). \tag{5.7}$$

The *free energy* F of the system is defined as

$$F = E - TS \tag{5.8}$$

and is the natural extension of the energy landscape description to situations with noise. In other words, the system seeks to minimise its free energy.

The partition function defined in equation 5.4 plays a crucial role because all the observables described above can be derived from it using the following equations [8]

$$\begin{aligned} F &= -\frac{1}{\beta}\log(Z) && (5.9) \\ E &= -\frac{\partial}{\partial\beta}\log(Z) && (5.10) \\ S &= = \beta^2 \frac{\partial F}{\partial\beta}. && (5.11) \end{aligned}$$

5.2 Magnetic Ising spin system

In magnetic materials, the microscopic state of the system is determined by the spin orientations of the component electrons.

In the Ising spin system, we model this by the microscopic system state $\mathbf{x} = (x_1, \ldots, x_N)$, where x_i is the spin state of electron i, which takes the value 1 for "up" and -1 for "down".

Traditionally, the microscopic energy (also called the Hamiltonian) has the form

$$H = -\sum_{\langle i,j \rangle} J_{ij} x_i x_j - h \sum_{i=1}^{N} x_i. \tag{5.12}$$

Here the notation $\langle i, j \rangle$ means nearest neighbours only, J_{ij} are the spin-spin interactions and h is the external magnetic field.

5.3 Mean-field approximation

Although, this model seems very simple, it is a fact that an exact solution [5, 3] (which involves calculating the partition function) has only been found for the one- and two-dimensional cases with zero external magnetic field.

The only way progress has been made with other cases has been by making some form of approximation. One of the simplest is called the mean-field approximation. This involves assuming that $x_1, \ldots, x_N$ are independent and identically distributed variables in the equilibrium state.

$$\rho(\mathbf{x}) = \prod_{i=1}^{N} \rho(x_i) \tag{5.13}$$

5.4 The replica method

Sherrington and Kirkpatric [31] introduced the model of a spin glass described by the Hamiltonian

$$H = -\frac{1}{2} \sum_{i \neq j}^{N} J_{ij} x_i x_j \tag{5.14}$$

where J_{ij} is symmetric. This model has recently acquired new significance because it appears to provide a fruitful model for neural networks with noise.

The *replica method* was invented by Emery [15] and provides a way to make analytical progress on this model. It was intended to deal with cases where averaging $\rho(\mathbf{x})$ was easy, but averaging $\log(\rho(\mathbf{x}))$, in order to evaluate the free energy, was difficult. The idea is based on the following limiting form.

$$\lim_{n\to 0} \frac{Z^n - 1}{n} = \log Z \tag{5.15}$$

Although, no one has proved whether this is sound, it does seem to work.

Note that Z^n is the partition function of n non-interacting identical copies of the original system, known as *replicas*.

5.5 Replica symmetric solution

We are particularly interested in the model studied by Mézard, Nadal and Toulouse [26] with the general storage prescription

$$J_{ij} = \frac{1}{N} \sum_{n=1}^{M} \Lambda\left(\frac{M+1-n}{N}\right) X_i^n X_j^n \tag{5.16}$$

where $\Lambda(x)$ is any positive function such that

$$\int_0^\infty \Lambda^2(x) \ \mathrm{dx} = 1 \tag{5.17}$$

and for the marginalist storage prescription in particular

$$\Lambda(x) = \varepsilon \exp\left(-\frac{1}{2} x \varepsilon^2\right). \tag{5.18}$$

The free energy can be evaluated using the replica method. The resulting equations are very complex [2] and involve various two dimensional matrices. To make further progress, we assume that the matrices contain identical elements. This is known as the approximation of replica symmetry.

In the limit as $\beta \to \infty$, the following equations for the $N + 2$ order parameters $\{m^n \mid 1 \leq n \leq N\} \cup \{q, r\}$ crystallise out [26]

$$\begin{aligned} m^n &= \begin{cases} m & \text{if } n = \alpha N \\ 0 & \text{otherwise} \end{cases} && (5.19) \\ q &= 1 - \frac{C}{\beta} && (5.20) \\ r &= \int_0^\infty \left(\frac{\Lambda(x)}{1 - C\Lambda(x)} \right)^2 \, \mathrm{dx} && (5.21) \end{aligned}$$

where

$$\begin{aligned} m &= \mathrm{erf}(m\eta) && (5.22) \\ C &= \sqrt{\frac{2}{\pi r}} \exp\left(-m^2\eta^2\right) && (5.23) \\ \eta &= \frac{\Lambda(\alpha)}{\sqrt{2r}} && (5.24) \\ \mathrm{erf}(x) &= \frac{2}{\sqrt{\pi}} \int_0^x \exp\left(-y^2\right) \, \mathrm{dy}. && (5.25) \end{aligned}$$

5.6 Phase space and the order parameters

Broadly speaking a neural network has three states in phase space.

- The ferromagnetic state occurs at low temperature and the system behaves like an associative memory
- The spin glass state also occurs at low temperature, but is characterised by the system being frozen into a random state
- The paramagnetic state occurs at high temperature where the system converges to a state distribution which is independent of the initial state

It is worth considering what the order parameters represent.

$$\begin{aligned} m^n &= \frac{1}{N} \sum_{i=1}^{N} X_i^n \langle x_i \rangle \\ q &= \frac{1}{N} \sum_{i=1}^{N} \langle x_i \rangle^2 \end{aligned}$$

The order parameter m^n is the time averaged overlap between the pattern $\mathbf{X}^n$ and the state of the system. So when this value is close to one, it means that that pattern is being retrieved with good quality and the system is in a ferromagnetic state. However, when all the time averaged overlaps are zero, it may indicate the spin glass state or the paramagnetic state.

The Edwards-Anderson order parameter q allows these states to be distinguished. In the spin glass state, x_i is frozen and so $\langle x_i \rangle = \pm 1$ implying $q = 1$. In the paramagnetic state x_i is randomly fluctuating and so $\langle x_i \rangle = 0$ implying $q = 0$.

Finally, the order parameter r represents the noise due to the overlap of unwanted patterns, so r must be low for retrieval to be good.

5.7 Critical values in the phase space

Numerical analysis [26] of equations 5.18, 5.19, 5.21, 5.22, 5.23, 5.24 and 5.25 of the kind shown below in Mathematica

```
In[1] :=
  rrrr[x_] :=
    -2(x + (1 - x) Log[1 - x]) / (x^2 (x - 1));
  rrr[c_, e_] := rrrr[c e];
  rr[m_, n_, e_] :=
    x /. FindRoot[
    x == rrr[Sqrt[2 / (Pi x)] Exp[- n^2 m^2], e],
    {x, 1}];
  nn[m_] := n /. FindRoot[Erf[m n] == m, {n, 1}];
  aaa[m_, n_, e_] := Log[2 rr[m, n, e] n^2 / e^2] / e^2;
  aa[m_, e_] := aaa[m, nn[m], e];
In[7] :=
  FindMinimum[aa[m, e], {m, 0.97, 0.98}, {e, 4, 4.1}]
```

```
Out[7] :=

  {-0.0489585, {m -> 0.971971, e -> 4.10812}}

In[8] :=

  FindMinimum[aa[m, 2.464805], {m, 0.97, 0.98}]

Out[8] :=

  {5.29038 10^-8  , {m -> 0.933347}}

In[9] :=

  FindMinimum[aa[m, 2.464815], {m, 0.97, 0.98}]

Out[9] :=

  {-1.00144 10^-6  , {m -> 0.933347}}
```

reveals that they have no solution with $m \neq 0$ for $\varepsilon < \varepsilon_{\mathrm{C}}$ where

$$\varepsilon_{\mathrm{C}} = 2.46481. \tag{5.26}$$

For $\varepsilon = \varepsilon_{\mathrm{C}}$, there is one stable solution with $\alpha = 0$ and $m = m_{\mathrm{C}}$ where

$$m_{\mathrm{C}} = 0.933347. \tag{5.27}$$

For $\varepsilon > \varepsilon_{\mathrm{C}}$, there are two solutions with $m \neq 0$, but only the highest value is stable and $m > m_{\mathrm{C}}$. The maximum value for α is attained at

$$\varepsilon_{\mathrm{opt}} = 4.10812 \tag{5.28}$$

with a capacity of

$$\alpha_{\mathrm{opt}} = 0.0489585. \tag{5.29}$$

Figure 5.1 shows the graph of storage capacity α against ϵ for the marginalist learning scheme plotted using the following additional commands in Mathematica.

```
In[10] :=

  maa[e_] := -FindMinimum[aa[m, e], {m, 0.97, 0.98}][[1]];
```

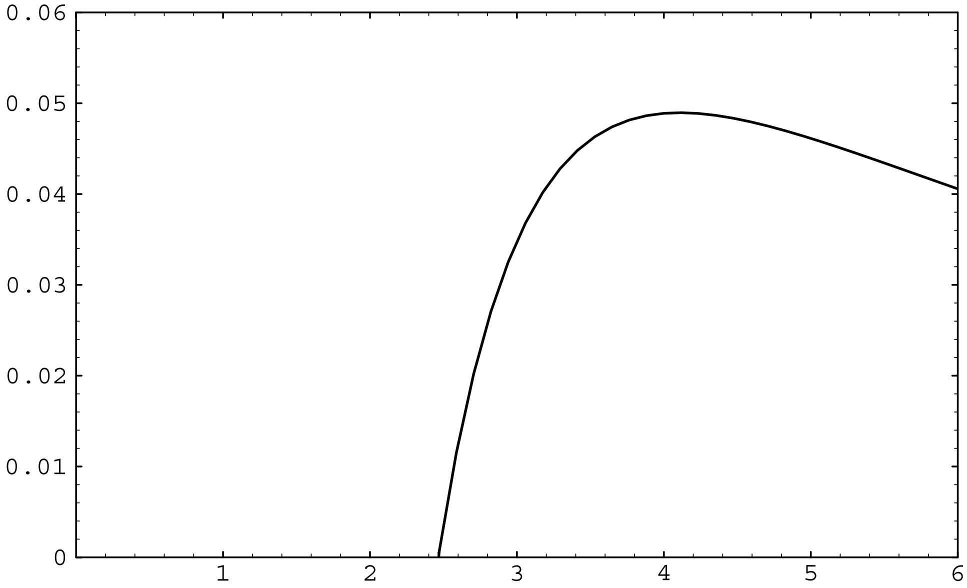

Figure 5.1: Plot of storage capacity α against ε for the marginalist learning scheme

```
Plot[
   maa[e],
   {e, 2.47, 6},
   Frame -> True,
   PlotRange -> {{0, 6}, {0, 0.06}}];
```

5.8 Crude upper bound of the storage capacity

Just as an aside, we can find an upper bound analytically for the storage capacity α. Equations 5.21 and 5.18 give

$$r = \frac{-2(c\varepsilon + (1 - c\varepsilon)\log(1 - c\varepsilon))}{c^2\varepsilon^2(c\varepsilon - 1)} \tag{5.30}$$

and from equations 5.24 and 5.18, we have

$$\alpha = \frac{1}{\varepsilon^2}\log\left(\frac{\varepsilon^2}{2r\eta^2}\right). \tag{5.31}$$

For constant r and η, α is maximised by setting

$$\varepsilon = \eta\sqrt{2re}. \tag{5.32}$$

Therefore

$$\alpha \leq \frac{1}{2re\eta^2}. \tag{5.33}$$

But from equation 5.30 it is can be shown that $r \geq 1$ and from equation 5.22 that $\eta \geq \frac{\sqrt{\pi}}{2}$, therefore

$$\alpha \leq \frac{2}{\pi e} \approx 0.234. \tag{5.34}$$

This implies that patterns greater than $0.234N$ are forgotten.

Chapter 6

Storage capacity for the smooth learning scheme

In this chapter, we will derive the critical values for k in the hyperbolic tangent and error function learning schemes.

6.1 Embedding strength to storage capacity

We need to show the connection between the embedding strengths computed in chapter 4 and the marginalist storage prescription described in chapter 5.

From equations 4.1 and 5.16, we have

$$\begin{aligned} e_m &= \frac{1}{N}\sum_{i,j=1}^{N}\left(\frac{1}{N}\sum_{k=1}^{M}\Lambda\left(\frac{M+1-k}{N}\right)X_i^kX_j^k\right)X_i^mX_j^m \\ &= \frac{1}{N^2}\sum_{k=1}^{M}\Lambda\left(\frac{M+1-k}{N}\right)\left(\sum_{i=1}^{N}X_i^kX_i^m\right)^2 . \end{aligned}$$

Pulling out the term $k = m$ gives

$$e_m = \Lambda\left(\frac{n}{N}\right) + \frac{1}{N^2}\sum_{k=1,k\neq m}^{M}\Lambda\left(\frac{M+1-k}{N}\right)\left(\sum_{i=1}^{N}X_i^kX_i^m\right)^2 \qquad (6.1)$$

where $n = M + 1 - m$.

Recall that $\{\mathbf{X}^m \mid 1 \leq m \leq M\}$ are random patterns. Therefore, by the central limit theorem $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} X_i^k X_i^m$ tends to a gaussian random variable with zero mean and variance one as $N \to \infty$. Therefore, the second term in equation 6.1 tends to a gaussian random variable with zero mean and a variance of

$$\begin{aligned}
\sigma^2 &\to \frac{1}{N^2} \sum_{k=1, k \neq m}^{M} \Lambda^2 \left(\frac{M+1-k}{N} \right) \\
&\to \frac{1}{N} \int_0^\infty \Lambda^2(x) \; \mathrm{dx} \\
&= \frac{1}{N} \\
&\to 0.
\end{aligned}$$

So for large N

$$e_m \approx \Lambda \left(\frac{n}{N} \right). \tag{6.2}$$

But

$$e_m \sim \exp \left(-\frac{\theta k n}{N} \right) \tag{6.3}$$

where θ is a function of the function ϕ used in the smooth learning scheme.

Therefore, using equation 5.18

$$k = \frac{\varepsilon^2}{2\theta}. \tag{6.4}$$

6.2 Critical values for the hyperbolic tangent scheme

Figure 6.1 shows the graph of storage capacity α against the parameter k for the hyperbolic tangent learning scheme.

No patterns are stored for $k < k_{\mathrm{C}}$ where

$$k_{\mathrm{C}} = \frac{\varepsilon_{\mathrm{C}}^2}{2\theta(\tanh)} = 3.68 \tag{6.5}$$

and the optimal value for k is

$$k_{\mathrm{opt}} = \frac{\varepsilon_{\mathrm{opt}}^2}{2\theta(\tanh)} = 10.2. \tag{6.6}$$

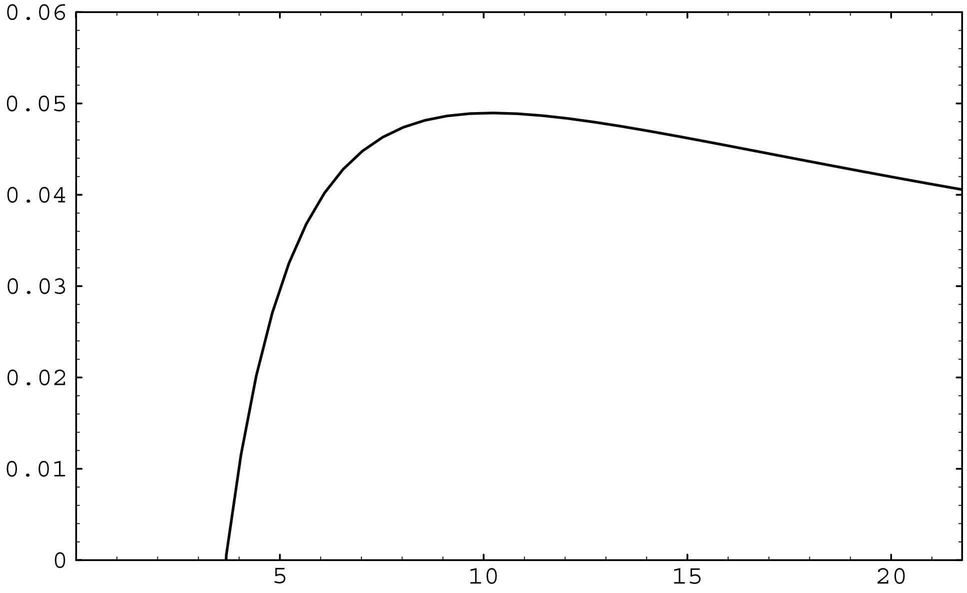

Figure 6.1: Plot of storage capacity α against the parameter k for the hyperbolic tangent learning scheme

6.3 Critical values for the error function scheme

Figure 6.2 shows the graph of storage capacity α against the parameter k for the error function learning scheme.

No patterns are stored for $k < k_{\rm c}$ where

$$k_{\rm c} = \frac{\varepsilon_{\rm c}^2}{2\theta(\rm erf)} = 4.09 \tag{6.7}$$

and the optimal value for k is

$$k_{\rm opt} = \frac{\varepsilon_{\rm opt}^2}{2\theta(\rm erf)} = 11.4. \tag{6.8}$$

6.4 Theoretical storage capacity

The striking result is that the optimal storage capacity 0.0489585 is independent of the function ϕ chosen in the smooth learning scheme. The function ϕ only affects the actual value of k that gives rise to this optimal storage

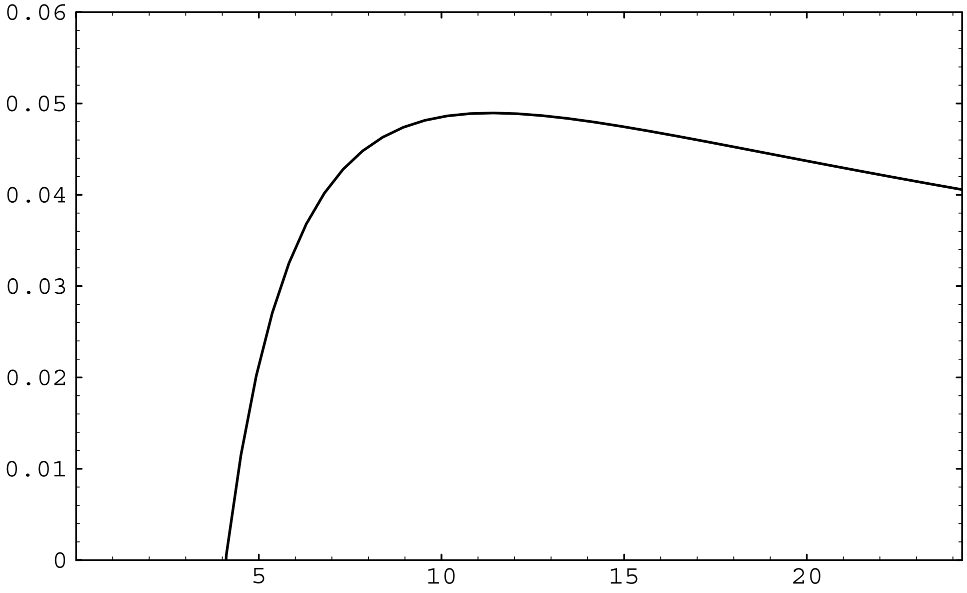

Figure 6.2: Plot of storage capacity α against the parameter k for the error function learning scheme

capacity.

$$\alpha_{\text{opt}} = 0.0489585 \tag{6.9}$$

$$k_{\text{opt}} = \frac{\varepsilon^2_{\text{opt}}}{2\theta(\phi)} \tag{6.10}$$

6.5 Experimental storage capacity

In order to confirm, that the theoretically derived storage capacities as depicted in figures 6.1 and 6.2 correspond to reality, we constructed a forgetful neural network with 1500 neurons utilising the smooth learning scheme and performed a number of experiments.

For the hyperbolic tangent learning scheme, we stored 375 randomly generated patterns using 5 different values of k and then evaluated the retrieval quality m of the 90 most recently stored patterns. The results are shown postscriptally in figures 6.3, 6.4, 6.5, 6.6 and 6.7.

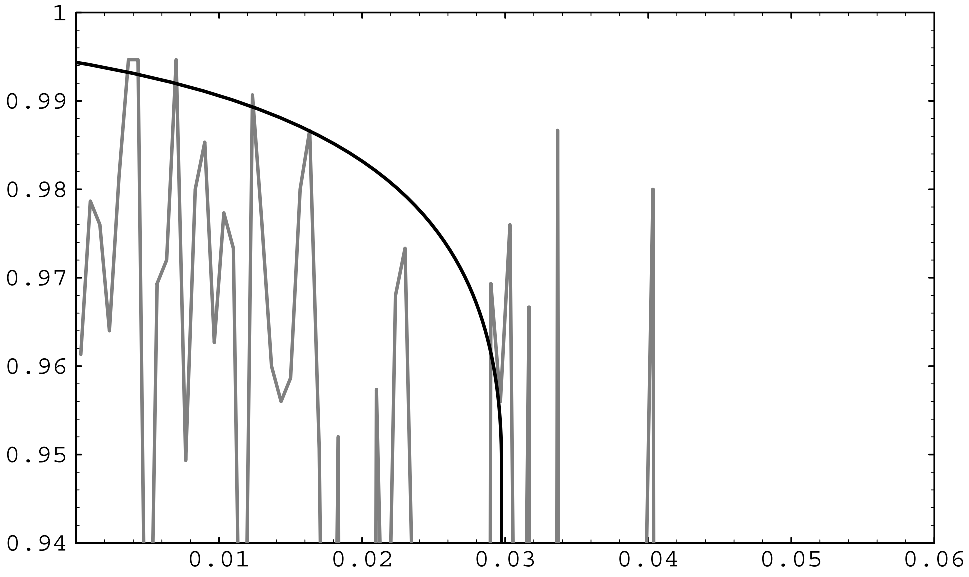

Figure 6.3: Plot of retrieval quality using the hyperbolic tangent learning scheme with $k = 5$ against the ratio $\frac{n}{N}$

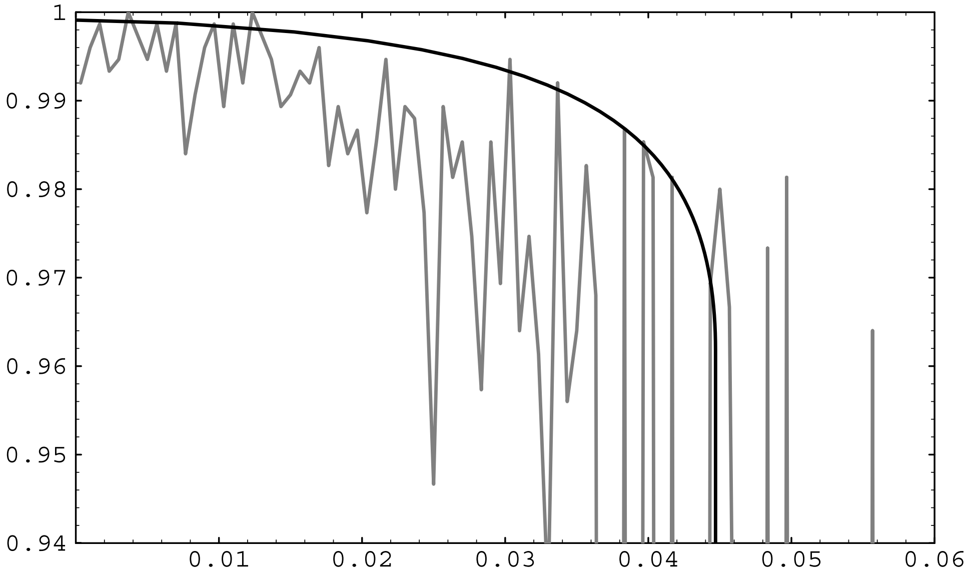

Figure 6.4: Plot of retrieval quality using the hyperbolic tangent learning scheme with $k = 7$ against the ratio $\frac{n}{N}$

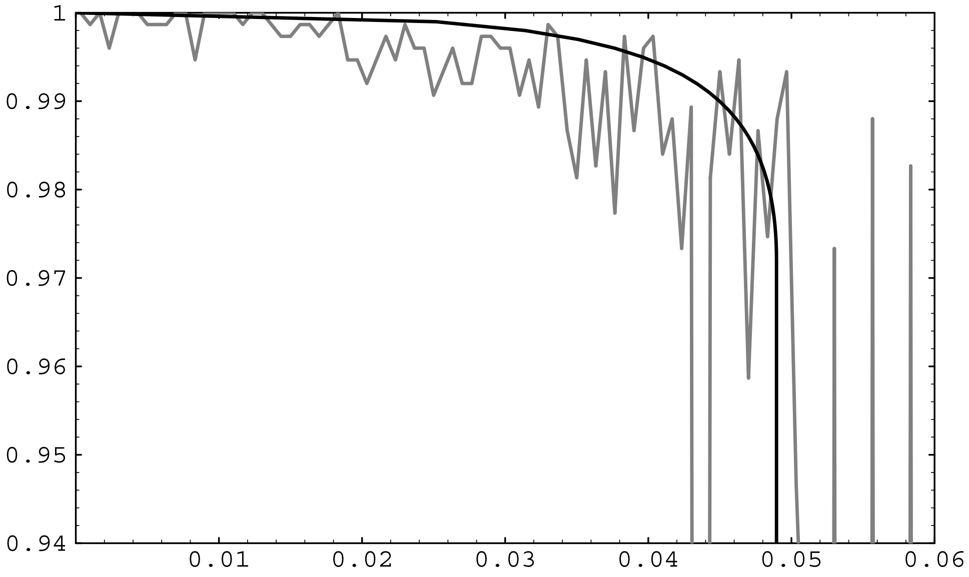

Figure 6.5: Plot of retrieval quality using the hyperbolic tangent learning scheme with $k = 10.2$ against the ratio $\frac{n}{N}$

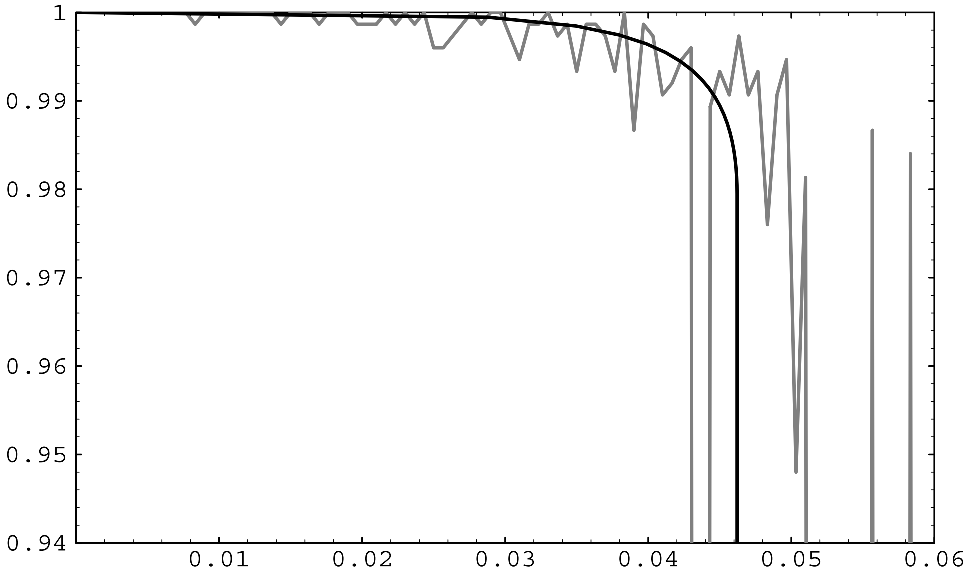

Figure 6.6: Plot of retrieval quality using the hyperbolic tangent learning scheme with $k = 15$ against the ratio $\frac{n}{N}$

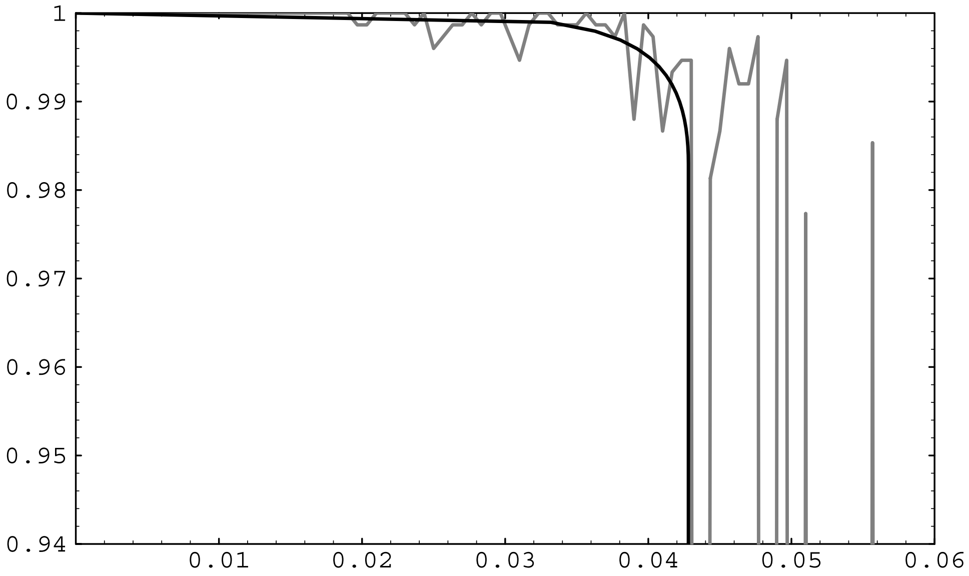

Figure 6.7: Plot of retrieval quality using the hyperbolic tangent learning scheme with $k = 19$ against the ratio $\frac{n}{N}$

The black lines show the theoretical retrieval quality m of pattern n against the ratio $\frac{n}{N}$ in the limit as $N \to \infty$. The theoretical storage capacity α corresponds to the value of $\frac{n}{N}$ where the black line drops to zero. The grey lines show the actual retrieval quality m of pattern n against the ratio $\frac{n}{1500}$.

A similar experiment was conducted with the error function learning scheme and the corresponding graphs are shown in figures 6.8, 6.9, 6.10, 6.11 and 6.12.

The correlation between the black and grey lines is not overwhelming, but then again the grey lines only correspond to single runs.

So, we ran the hyperbolic tangent learning scheme with $k = 10.2$ and the error function learning scheme with $k = 11.4$ ten times to give figures 6.13 and 6.14. These show a much better correlation, particularly with regard to the storage capacity.

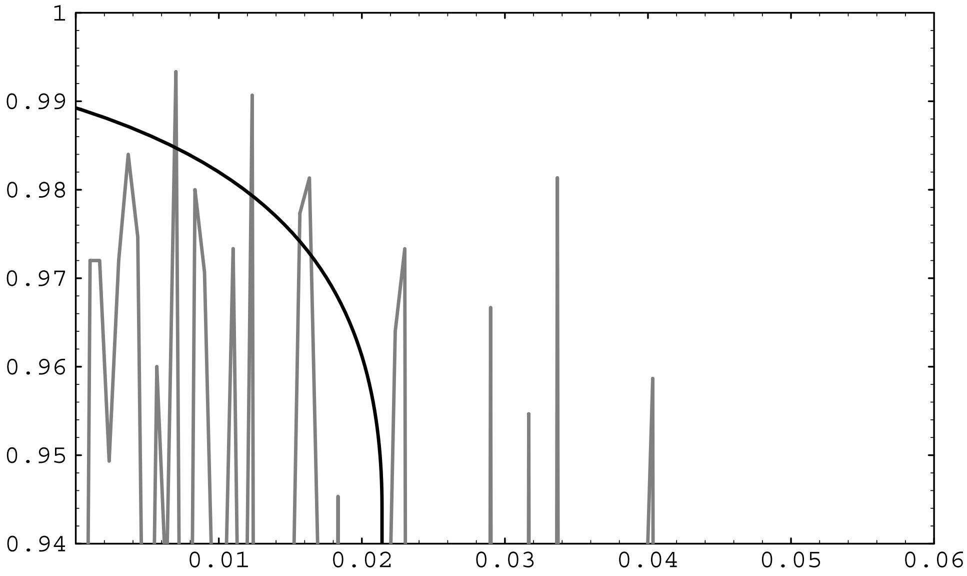

Figure 6.8: Plot of retrieval quality using the error tangent learning scheme with $k = 5$ against the ratio $\frac{n}{N}$ where N is the total number of neurons

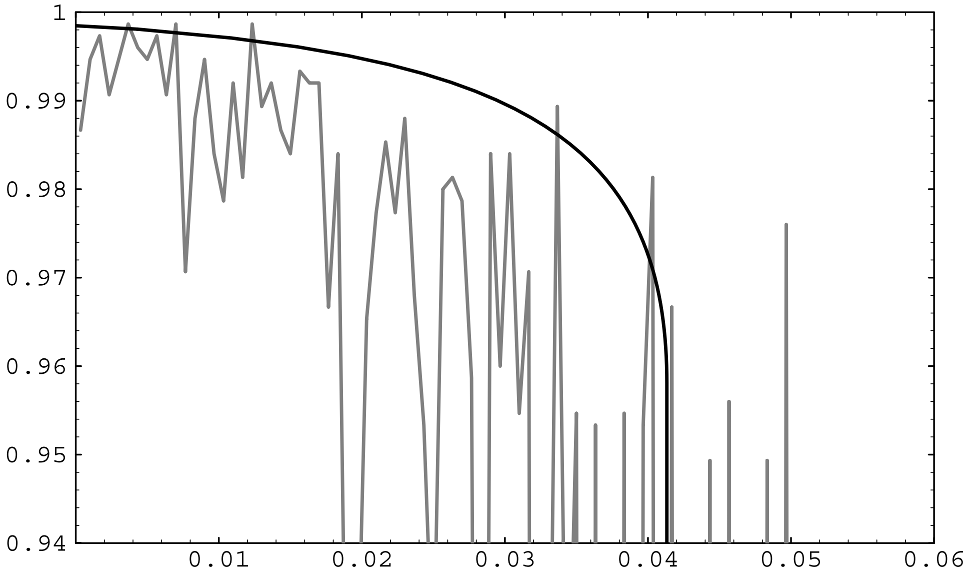

Figure 6.9: Plot of retrieval quality using the error function learning scheme with $k = 7$ against the ratio $\frac{n}{N}$

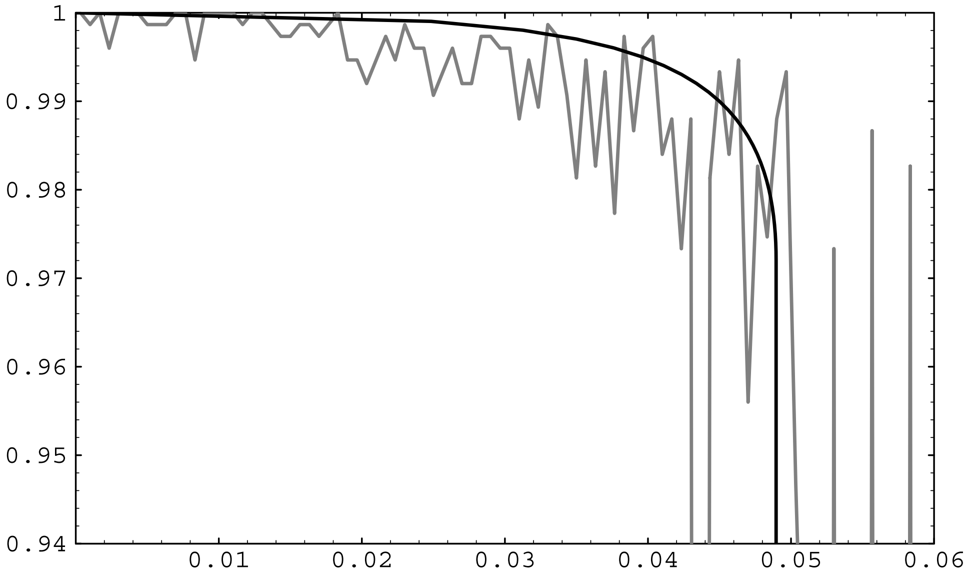

Figure 6.10: Plot of retrieval quality using the error function learning scheme with $k = 11.4$ against the ratio $\frac{n}{N}$

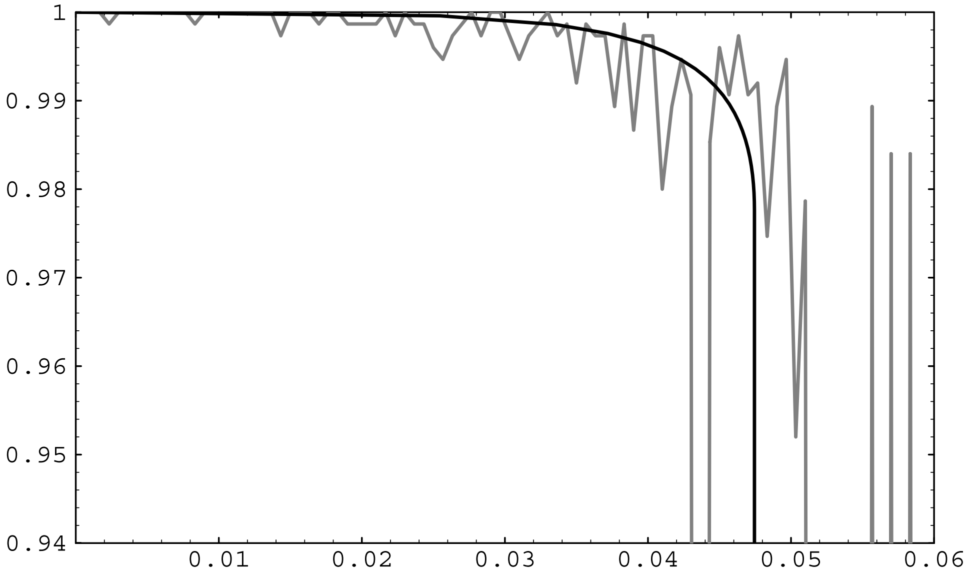

Figure 6.11: Plot of retrieval quality using the error function learning scheme with $k = 15$ against the ratio $\frac{n}{N}$

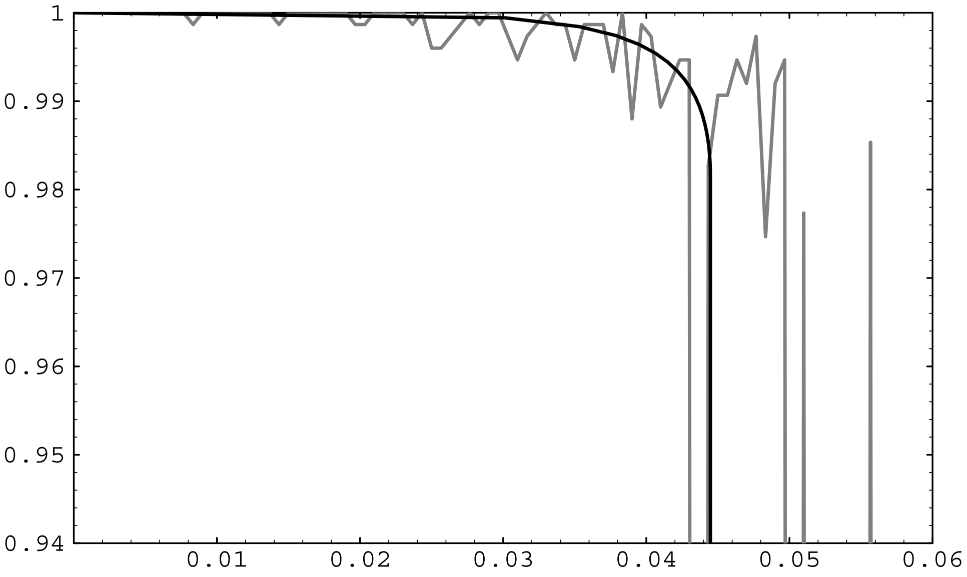

Figure 6.12: Plot of retrieval quality using the error function learning scheme with $k = 19$ against the ratio $\frac{n}{N}$

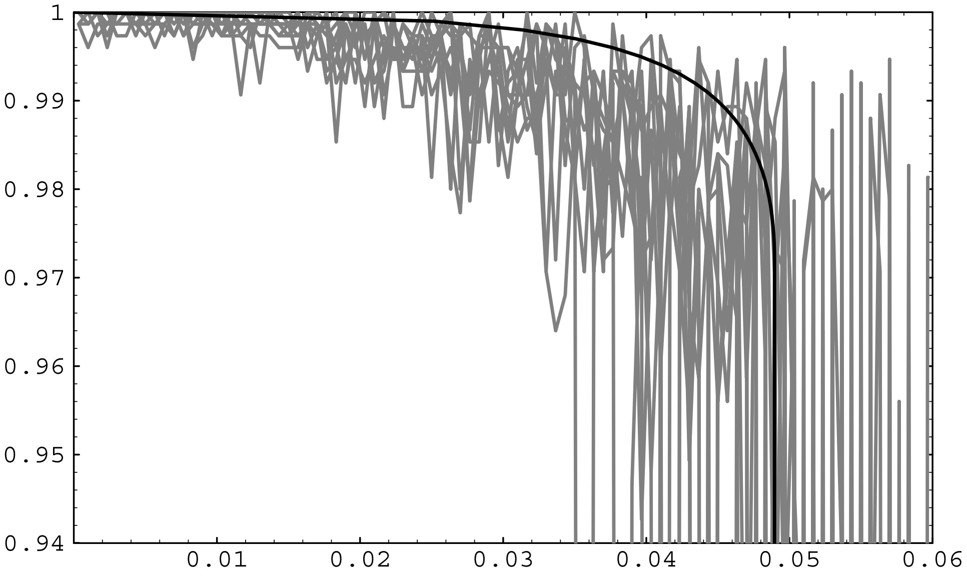

Figure 6.13: Plot of retrieval quality using the hyperbolic tangent learning scheme with $k = 10.2$ against the ratio $\frac{n}{N}$

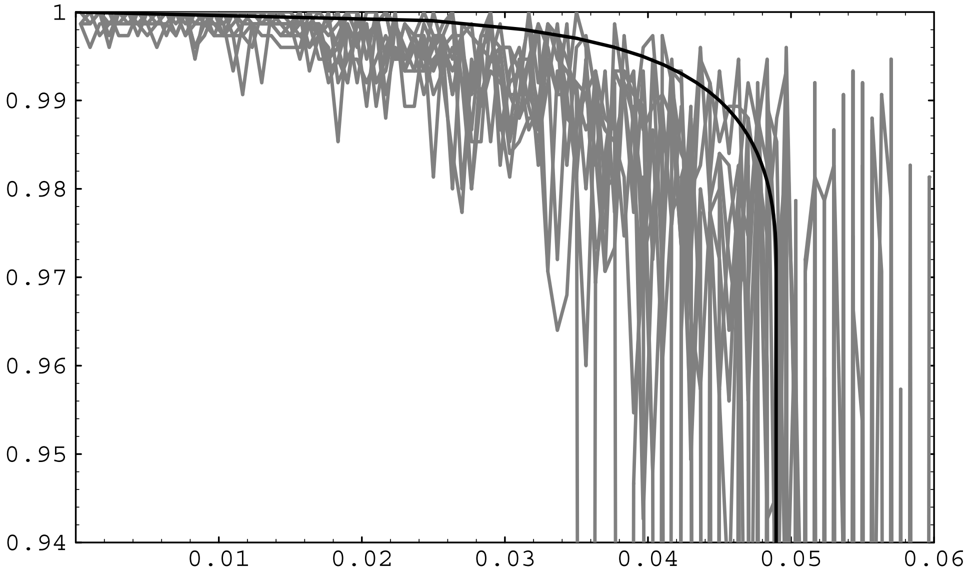

Figure 6.14: Plot of retrieval quality using the error function learning scheme with $k = 11.4$ against the ratio $\frac{n}{N}$

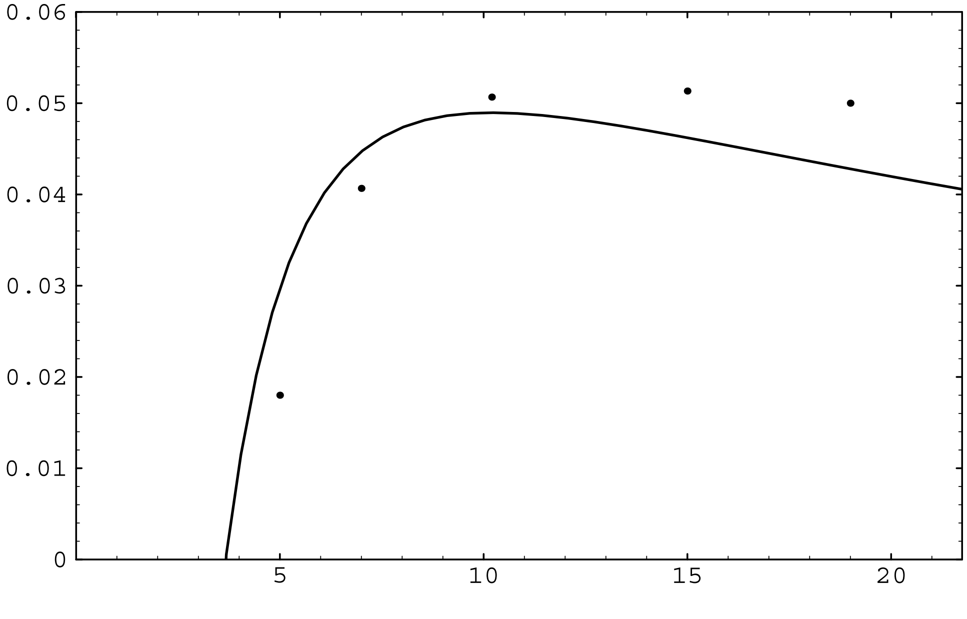

Figure 6.15: Plot of remembered patterns ratio and theoretical storage capacity against ϵ for the hyperbolic tangent learning scheme

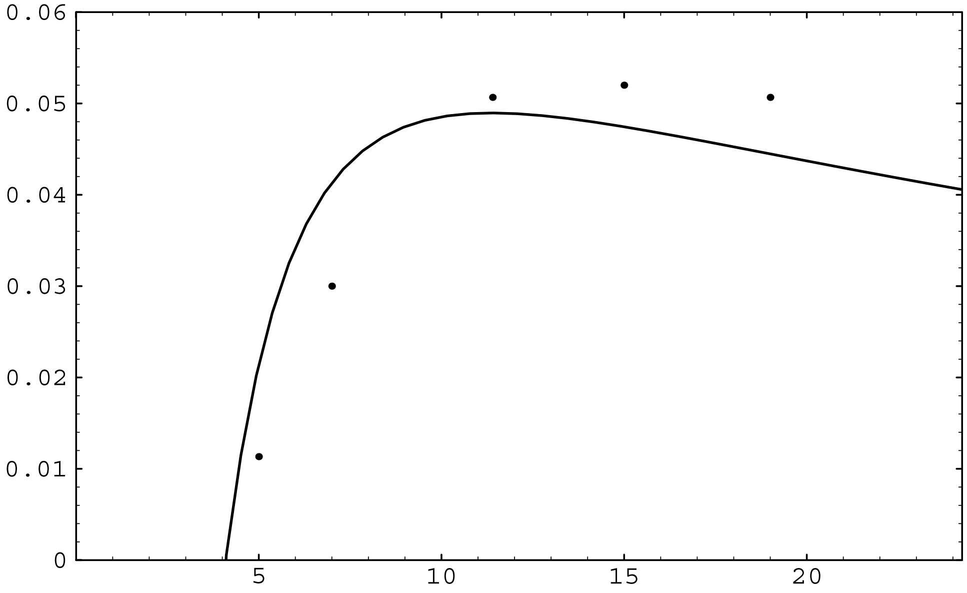

Figure 6.16: Plot of remembered patterns ratio and theoretical storage capacity against ϵ for the error function learning scheme

Alternatively, figures 6.15 and 6.16 show the number of patterns that are remembered by the neural network as a ratio of the total number of neurons, where by definition [29] the pattern is remembered if no more than 2% of the neurons in the final configuration differ from the stored pattern.

6.6 Exploration of other schemes

What happens if we drop the restriction that $\phi'(0) = 1$ for the smooth learning scheme?

Let

$$a = \phi'(0)$$

and suppose $a < 1$, then it is can be seen that the associated fractal probability distribution exists and is unique by following the same argument as in section 3.2. It then follows that the embedding strength decay rate is given by equation 4.5 using the same argument as in section 4.2.

Let us consider analytically the Lyapunov exponent for small ϵ. As before, using Maclaurin's expansion for $\phi(x)$, we have

$$\phi(x) = ax + bx^r + O(x^{r+2})$$

where for the sake of brevity

$$b = \frac{\phi^{(r)}(0)}{r!}$$

and we know that $b < 0$.

We know that x_+ satisfies

$$x_+ = \phi(x_+ + \epsilon)$$

and the inverse of ϕ is well defined because ϕ is monotonically increasing, therefore

$$\begin{aligned} \epsilon &= \phi^{-1}x_+ - x_+ \\ &\approx \frac{1-a}{a}x_+ - \frac{b}{a^{r+1}}x_+^r + O(x_+^{r+2}). \end{aligned}$$

Therefore, it can be shown that

$$x_+ \approx \frac{a}{1-a}\epsilon + \frac{b}{(1-a)^{r+1}}\epsilon^r + O(\epsilon^{r+2}).$$

Therefore

$$\begin{aligned} S^u_{[x_-,x_+]}(g,\nu_0) &= g(-\epsilon) \\ &= \log\phi'(0) \\ &= \log a \\ S^l_{[x_-,x_+]}(g,\nu_0) &= g(x_+) \\ &\approx \log a + \frac{r\,b}{(1-a)^{r-1}a}\epsilon^{r-1} + O(\epsilon^{r+1}). \end{aligned}$$

Therefore, for sufficiently small ϵ

$$\frac{(r+1)\,b}{(1-a)^{r-1}a}\epsilon^{r-1} < \gamma(\epsilon) - \log a \leq 0.$$

In other words, $\gamma \to \log a$ as $\epsilon \to 0$.

However, we need $\gamma \approx -\theta\epsilon$. This is satisfied by

$$a = 1 - \theta\epsilon. \tag{6.11}$$

Using equations 3.9 and 6.4, we have

$$a = 1 - \frac{\varepsilon^2}{2N}. \tag{6.12}$$

So, no patterns are stored for $\varepsilon < \varepsilon_{\mathrm{c}}$ and the optimal value is $\varepsilon = \varepsilon_{\mathrm{opt}}$.

Suppose $a > 1$, then for sufficiently small ϵ, $\phi_+ \circ \phi_-$ has three fixed points. Therefore the corresponding IFS is not weakly hyperbolic by proposition 3.2.1 and so we cannot make any theoretical progress.

Chapter 7

Conclusion

In this report, we have illustrated in a striking manner the union of three seemingly unrelated subjects, namely that of domain theory, statistical physics and neural networks.

We showed that the stored patterns in a forgetful neural network using the smooth learning scheme can be represented by a non-deterministic dynamical system with a unique fractal probability distribution. More specifically, we derived an approximating sequence for the distribution using a domain theoretic approach.

This approximating sequence allows us to compute the expectation of any continuous function over a fractal probability distribution using a generalised form of Riemann integration to any desired accuracy. In particular, we were interested in the embedding strength decay rate of the stored patterns, which just so happens to correspond to the Lyapunov exponent of the identified dynamical system.

By numerically analysing the asymptotics, we were able to estimate the decay rate for a large neural network. Specifically, we evaluated the decay rate γ for the hyperbolic tangent and error function learning schemes.

$$\gamma = \frac{-\theta(\phi)k}{N}$$

$$\begin{aligned} \theta(\tanh) &= 0.826 \\ \theta(\mathrm{erf}) &= 0.742 \end{aligned}$$

We then showed that a neural network using the smooth learning scheme is equivalent to a neural network using the marginalist learning scheme through the simple equation

$$k = \frac{\varepsilon^2}{2\theta(\phi)}$$

where k and ϕ parameterises the smooth learning scheme and ε parameterises the marginalist learning scheme.

This was a useful step because we were then able to use the solution of an Ising model from statistical physics that generalised the marginalist learning scheme. The solution gives a storage capacity for each value of the parameter ε and therefore for k and ϕ as well and showed that the maximum attainable storage capacity is $0.0489585N$, where N is the total number of neurons.

Finally, we constructed a neural network with 1500 neurons and showed that the experimental storage capacities were consistent with those derived theoretically. However, more work needs to be done here in order to prove a connection beyond doubt. Much larger neural networks and averaging over many runs is required. Current technology would require massive parallelism to achieve this.

Bibliography

[1] D. J. Amit, H. Gutfreund, and H. Sompolinsky. Information storage in neural networks with low levels of activity. *Phys. Rev. A*, 35:2293–2303, 1987.

[2] Daniel J. Amit. *Modelling Brain Function.* Cambridge University Press, 1992.

[3] R. J. Baxter. *Exactly Solved Models in Statistical Mechanics.* Academic Press, 1989.

[4] U. Behn, J. L. van Hemmen, R. Kühn, A. Lange, and V. A. Zagrebnov. Multifractality in forgetful memories. *Physica D*, 68:401–415, 1993.

[5] G. M. Bell and D. A. Lavis. *Statistical Mechanics of Lattice Models*, volume 1: Closed Form and Exact Theories of Co-operative Phenomena. Ellis Horwood, 1989.

[6] G. Birkhoff. *Lattice Theory.* American Mathematical Society, 1967.

[7] B. G. Cragg and H. N. V. Temperley. The organization of neurons: a cooperative analogy. *Electroenceph. Clin. Neurophysiol.*, 6:85, 1954.

[8] Viktor Dotsenko. *An Introduction to the Theory of Spin Glasses and Neural Networks*, volume 54 of *World Scientific Lecture Notes in Physics.* World Scientific, 1994.

[9] A. Edalat. Power domains algorithms for fractal image decompression. Technical Report Doc 93/44, Department of Computing, Imperial College, 1993.

[10] A. Edalat. Power domains and iterated function systems. Technical Report Doc 94/13, Department of Computing, Imperial College, 1994. To appear in *Information and Computation.*

[11] A. Edalat. Domain theory and integration. *Theoretical Computer Science*, 151:163–193, 1995.

[12] A. Edalat. Domain theory in learning processes. In S. Brooks, M. Main, A. Melton, and M. Mislove, editors, *Proceedings of the Eleventh Annual Conference on Mathematical Foundations of Programming Semantics*, volume 1 of *Electronic Notes in Theoretical Computer Science.* Elsevier, 1995. `http://www1.elsevier.nl/mcs/tcs/pc/covvol1.htm`.

[13] A. Edalat. Dynamical systems, measures and fractals via domain theory. *Information and Computation*, 120(1):32–48, July 1995.

[14] John H. Elton. Ergodic theorem for iterated maps. *Ergod. Th. & Dynam. Sys.*, 7:481–488, 1987.

[15] V. J. Emery. Critical properties of many component systems. *Phys. Rev. B*, 11:239–247, 1975.

[16] Donald O. Hebb. *The Organization of Behavior.* Wiley, New York, 1949.

[17] Donald O. Hebb. *Fractal Geometry.* John Wiley & Sons, 1989.

[18] John Hertz, Anders Krogh, and Richard G. Palmer. *Introduction to the Theory of Neural Computation.* Addison-Wesley Publishing Company, 1993.

[19] A. L. Hodgkin. The ionic basis of electrical activity in nerve and muscle. *Biol. Rev.*, 26:339–409, 1951.

[20] J. J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc. Natl. Acad. Sci. USA*, 79:2554–2558, 1982.

[21] J. E. Hutchinson. Fractal and self-similarity. *Indiana University Mathematics Journal*, 30:713–747, 1981.

[22] E. Atlee Jackson. *Perspectives of nonlinear dynamics*, volume 1: Nonlinear dynamical systems. Cambridge University Press, 1981.

[23] W. A. Little. The existence of persistent states in the brain. *Math. Biosci.*, 19:101, 1974.

[24] B. Mandelbrot. *The Fractal Geometry of Nature.* W. H. Freeman and co., San Francisco, 1982.

[25] W. S. McCulloch and W. A. Pitts. A logical calculus of the ideas immanent in neural networks. *Bull. Math. Biophys.*, 5:115, 1943.

[26] M. Mézard, J. P. Nadal, and G. Toulouse. Solvable models of working memories. *J. Phys.*, 47:1457–1462, 1986.

[27] J. P. Nadal, G. Toulouse, J. P. Changeux, and S. Dehaene. Networks of formal neurons and memory palimpsests. *Europhys. Lett.*, 1:535, 1986.

[28] Edward Ott. *Chaos in dynamical systems.* Cambridge University Press, 1993.

[29] Giorgio Parisi. A memory which forgets. *J. Phys. A*, 19:L617–L620, 1986.

[30] D. S. Scott. Outline of a mathematical theory of computation. In *Fourth Annual Princeton Conference on Information Sciences and Systems*, pages 169–176, 1970.

[31] D. Sherrington and S. Kirkpatrick. Unknown title. *Phys. Rev. Lett.*, 35:1972, 1975.

[32] W. A. Sutherland. *Introduction to Metric and Topological Spaces.* Oxford University Press, 1993.

[33] J. L. van Hemmen, D. Grensing, A. Huber, and Kühn R. Nonlinear neural networks. i. general theory. *J. Stat. Phys.*, 50(1):231–257, 1988.

[34] J. L. van Hemmen, D. Grensing, A. Huber, and Kühn R. Nonlinear neural networks. ii. information processing. *J. Stat. Phys.*, 50(1):259–293, 1988.

[35] J. L. van Hemmen, G. Keller, and R. Kühn. Forgetful memories. *Europhys. Lett.*, 5:663–668, 1988.

[36] Alan J. Weir. *Lebesgue Integration & Measure.* Cambridge University Press, 1994.