Stochastic Approximation of Artificial Neural Network-Type Learning Algorithms: A Dynamical Systems Approach

by

Israel Ncube

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Applied Mathematics

Waterloo, Ontario, Canada, 2001

© Israel Ncube 2001

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-60559-0

Canada

The University of Waterloo requires the signatures of all persons using or photocopying this thesis. Please sign below, and give address and date.

Abstract

Stochastic approximation is concerned with characterisation of the long term behaviour of recursive random algorithms. For example, *does the algorithm converge to a unique fixed point, for all initial points?* This problem is well-understood, via the Kushner-Clark theorem, only if the so-called associated ordinary differential equation (ODE) has exactly one locally asymptotically stable equilibrium point. In this case, it is known that, under some fairly reasonable assumptions, the random algorithm converges, with probability one, to the equilibrium point of the ODE. However, if the ODE has multiple locally asymptotically stable equilibria, not much is currently known about convergence of the algorithm to any *specific* one of these equilibria. The primary objective of the thesis is the investigation of this problem, both qualitatively and quantitatively. We study random fields generated by discrete algorithms, and then draw relationships between dynamics on the continuous (associated ODE) and discrete phase spaces. A novel computer algorithm, which estimates probabilities of convergence of a simple discrete system to particular stable equilibria of the ODE, is introduced. Simulation results suggest that the probabilities so estimated are almost independent of the initialisation of the discrete system. We reformulate the analysis of evolution of densities of algorithms, under the action of the Frobenius-Perron operator, on a new space, i.e. the space of normalised positive distributions. Endowed with a suitable metric, it is shown that the resulting metric space is complete.

Acknowledgements

The scope, depth, and direction of this research are all primarily a consequence of the exceptional supervisorship of Drs. Sue Ann Campbell and Edward R. Vrscay. I would like to thank both of them for their constant encouragement, patience, empathy, and understanding. Even during the times when I felt like I was at the end of my tether, they were always there with me, as instructors and as mentors. I would like to express my profound gratitude to them. Needless to say, I am forever indebted to these guys.

I would also like to thank my son, Mondli, for unconditionally standing by my side throughout the course of this work. Thanks, Mondli, for teaching me the value of patience, perseverance, listening, understanding, and *relaxing*. Your youthful innocence, boundless energy, and your insatiable curiosity about *everything* are an inspiration for me. Also, thank you for teaching me how to play like a kid, something I had almost forgotten over the years. I know you immensely enjoyed those simulated "horse rides" on my back! Above all, thanks for the priceless hugs and kisses, in bad times and in good times.

Finally, I would like to acknowledge the generous financial support that I received from the Department of Applied Mathematics, my supervisors, and the University of Waterloo.

Dedication

This thesis is dedicated to all the underpriviledged children of the world. especially those in rural sub-Saharan Africa. Their innocence, resilience, and courage in the face of adverse poverty and a brutal "global economy" should be an inspiration and a lesson for us all. Babylon, your walls will burn down soon.

Contents

Chapter 1

Preliminaries and Motivation

1.1 Introduction

Consider a recursive random algorithm of the form

$$w_{n+1} = w_n + \gamma_n h(w_n, x_n) \ , \tag{1.1}$$

where the w_n's and x_n's are in $\mathbb{R}$ (the generalisation to vectors in $\mathbb{R}^r$ is straightforward), $\{\gamma_n\}$ is a sequence of positive decreasing-to-zero real numbers such that $\sum_n \gamma_n = \infty$, $h : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a continuous function, and $\{x_n\}$ is a sequence of random variables that are distributed according to some given law. We may write

$$\begin{aligned} w_{n+1} &= w_n + \gamma_n \bar{h}(w_n) + \gamma_n [h(w_n, x_n) - \bar{h}(w_n)] \\ &= w_n + \gamma_n \bar{h}(w_n) + \gamma_n \xi_n \ , \end{aligned} \tag{1.2}$$

where

$$\bar{h}(z) = E[h(z, x_n) \mid x_0, \cdots, x_{n-1}] \,. \tag{1.3}$$

for $z \in \mathbb{R}$. The main objective of this thesis is the characterisation of the long term behaviour of the $\{w_n\}$ sequence, in the case when the so-called *associated ODE* (see [1] and [2] for details)

$$\dot{x} = \bar{h}(x)$$

corresponding to (1.1) has multiple locally asymptotically stable equilibria. For example, *does* $\{w_n\}_{n=1}^{\infty}$ *converge to a fixed point?* This is a classic *stochastic approximation* problem. Stochastic approximation finds application in a wide variety of problems arising in engineering, science, and economics, for example. The primary application of the type of problem addressed in this thesis is in the area of unsupervised artificial neural learning. Neural learning is a process of updating the free parameters of a network of neurons in order to achieve specific desired design objectives. Unsupervised learning is that paradigm of neural learning where there is no external critic to supervise the learning process. In the framework of neural networks, (1.1) is interpreted as follows. The $\{w_n\}_{n=1}^{\infty}$ sequence denotes iterates of a network weight, x_n is an input received at time n, which causes w_n to be updated to take account of new information. The randomness of inputs is a result of contamination by external noise signals. The γ_n serve as 'training parameters' that modulate the correction terms. The requirement that $\gamma_n \to 0$ as $n \to \infty$ reflects the desire to gradually "phase out" the corrections in order to ensure convergence of the algorithm. The condition that $\sum_n \gamma_n = \infty$ is to ensure that the algorithm does not converge prematurely, i.e. that it converges to the "right" point or set.

Over the years, the area of stochastic approximation has received a lot of attention from a myriad of researchers. Arguably, the seminal publication in this area was authored by Robbins and Monro, in 1951 (see [7] for details). The essence of the so-called Robbins-Monro procedure for dosage is outlined below. For each dosage θ of some chemical product, you make an experiment that delivers some effect X depending on θ. In other words, X is a random variable whose distribution depends on θ. The issue is to determine the value of θ for which the mean effect is α, viz. $E[X \mid \theta] = \alpha$. The Robbins-Monro procedure consists of conducting a series of experiments with changing values of θ according to

$$\theta_{n+1} = \theta_n + \gamma_n(\alpha - x_{n+1}) \,, \tag{1.4}$$

where x_{n+1} is the result of the experiment made with θ_n. One can prove that, under some reasonable assumptions, θ_n converges to the desired solution (see [1] for details).

In 1968, Fabian [18] proved the *asymptotic normality* of stochastic approximation algorithms of the form

$$U_{n+1} = (I - n^{-\alpha}\Gamma_n)U_n + n^{-0.5(\alpha+\beta)}\Phi_n V_n + n^{-\alpha-0.5\beta}T_n \,,$$

where $U_n,\ V_n,\ T_n \in \mathbb{R}^k$, $\Gamma_n,\ \Phi_n \in \mathbb{R}^{k\times k}$ and $\alpha,\ \beta \in \mathbb{R}$. The V_n's are random vectors, and hence so are the U_n's. Fabian showed that, under some conditions, the asymptotic distribution of $n^{\frac{\beta}{2}}U_n$ is normal with a mean μ and a covariance matrix M. To this end, he characterised the asymptotics of $\{\tilde{U}_n\}$ by describing the

asymptotics of its mean $E[\tilde{U}_n]$ and covariance matrix $E[\tilde{U}_n\tilde{U}_n']$, where

$$\tilde{U}_{n+1} = (I - n^{-\alpha}\Lambda)\tilde{U}_n + n^{-\frac{\alpha}{2}}V_n \ ,$$

$\tilde{U}_n \stackrel{def}{=} (n-1)^{\frac{\beta}{2}}U_n$, and Λ is a constant $k \times k$ diagonal matrix.

Another landmark paper in this area was published in 1977, by Ljung [2]. He considered a class of algorithms which includes stochastic approximation algorithms, recursive identification algorithms, and algorithms for adaptive control of linear systems. This paper was the first to attempt to characterise the limits of sequences $\{w_n\}_{n=1}^{\infty}$ in equation (1.1) by studying ODEs which were approximately satisfied by the asymptotic part of a natural continuous parameter interpolation of $\{w_n\}_{n=1}^{\infty}$. The method used is very technical and complicated. In 1978, Kushner and Clark [1] achieved what [2] had achieved earlier, using a relatively less cumbersome method. The technique pioneered by Ljung is nowadays referred to as *the ODE method*, and constitutes a central part of this thesis. The work of [1] and [2] led to the birth of a celebrated theorem on the convergence of $\{w_n\}_{n=1}^{\infty}$, the so-called Kushner-Clark theorem. The general idea behind this theorem is that, under some conditions, it is possible to derive an autonomous deterministic ODE that is associated with a given learning algorithm (which is, usually, in the form of a non-autonomous random difference equation). This association is in the sense that the asymptotic paths of the ODE and the algorithm are close with a large probability and, eventually, the solution of the algorithm tends, *with probability one* (wp1 in short), to a uniformly asymptotically stable solution of the ODE. This ability to associate a deterministic autonomous ODE to the asymptotic part of the

difference equation, even when the latter has additive (or other) noise, is subtle and very useful in applications.

Essentially, the Kushner-Clark theorem establishes that if the associated ODE

$$\dot{x} = \bar{h}(x) \tag{1.5}$$

corresponding to the random algorithm (1.1) has exactly one asymptotically stable equilibrium point $\hat{x}$, then, under some fairly reasonable assumptions (see Section 1.4), $w_n \overset{wpl}{\to} \hat{x}$ as $n \to \infty$. However, if (1.5) possesses multiple locally asymptotically stable equilibria, nothing is currently known about the convergence of $\{w_n\}$ to a *specific* element of the set $\{x | \bar{h}(x) = 0\}$. In fact, this is still an open problem in this area. Since the pioneering work of Ljung, Kushner, and Clark in the late 1970's, there has been minimal progress made to date in tackling the above problem. This, in part, is due to the almost intractability of the problem. In addition, because this problem commonly arises in engineering applications, many researchers have been more focussed on the "practicality" of stochastic approximation algorithms rather than in a more mathematically rigorous treatment of the problem.

Since the emergence of the Kushner-Clark theorem, only a few papers have appeared, specifically targetting the problem mentioned above. In 1996, Fort and Pagès [8] established and proved a theorem that allowed them to transfer the convergence of solutions of the associated ODE to $\{w_n\}_{n=1}^{\infty}$, if the ODE has no *pseudo-cycle*. The latter include bona fide periodic orbits as well as isolated equilibria. Their approach is primarily a development of the original proof of the Kushner-Clark theorem. They made further assumptions on the ODE: continuity of $\bar{h}$,

convergence of every bounded maximal solution toward some zero of $\bar{h}$, and finally, a simple dynamics for the ODE (i.e. no pseudocycle). The proof of their theorem amounts to proving that $\{w_n\}_{n=1}^{\infty}$ has only one limiting point in $\{x|\ \bar{h}(x) = 0\}$. Furthermore, they performed numerical simulations when the ODE *does* have pseudocycles. Finally, if one of the elements of $\{x|\ \bar{h}(x) = 0\}$ is a saddle point $\hat{x}$, Fort and Pagès have shown that the sequence $\{w_n\}_{n=1}^{\infty}$ will *not* converge to $\hat{x}$.

The main thrust of this thesis is the investigation of convergence of $\{w_n\}_{n=1}^{\infty}$ if the associated ODE has multiple locally asymptotically stable equilibria. We attempt to characterise this convergence to *specific* elements of the set $\{x|\ \bar{h}(x) = 0\}$. The formal mathematical analysis will hinge on principles drawn mainly from the areas of dynamical systems theory, probability theory, and stochastic processes. To this end, both analytical and numerical approaches will be utilised. The organisation of the thesis is outlined below.

1.2 Thesis Organisation

In Section 1.3, we outline some key concepts and definitions that will be encountered throughout the thesis. Then, in the rest of this chapter, we introduce the theory behind the notion of an associated ODE. To this end, we start off by stating the Kushner-Clark theorem. Then we elaborate the technical details behind the derivation of the ODE. Next, we present some examples to illustrate how this theory works and also to highlight the difficulty of characterising convergence of $\{w_n\}_{n=1}^{\infty}$ to a *specific* element of the set $\{x|\ \bar{h}(x) = 0\}$, particularly if the ODE has more than one locally asymptotically stable equilibrium. We close the chapter

with an introduction to the rudiments of artificial neural networks, particularly the notion of *learning* and how it relates to the subject matter of this thesis.

Chapter 2 is a systematic study of the dynamical behaviour of Ljung's automatic classifier [2], which is essentially a two-dimensional discrete random algorithm. This algorithm seeks to classify a one-dimensional set of data, and thus addresses a very basic classification problem. We perform qualitative and quantitative analyses of both the algorithm and its associated ODE, in the case when the latter has two locally asymptotically stable equilibria. The underlying motivation is the desire to characterise the relationship between dynamics on the discrete and continuous phase spaces. In chapter 3, we introduce and investigate the notion of a stationary density of the *Frobenius-Perron operator* associated with an algorithm, and how this may be exploited to help understand the long term behaviour of $\{w_n\}_{n=1}^{\infty}$ if the associated ODE has multiple locally asymptotically stable equilibria. We introduce an *appropriate space* for the analysis of densities under iteration of the Frobenius-Perron operator. We find stationary densities for both linear and nonlinear random algorithms, the stationary densities for the latter being "determined" numerically. We emphasize that this chapter is motivated and strongly influenced by the work of [20], [3], and [5].

Chapter 4 investigates some generalisations of Ljung's doubly-triangular *pdf*. Essentially, we consider *pdf*s whose underlying components are *overlapping*. We then derive and analyse the associated system of ODEs. We close the chapter by performing the usual (akin to Chapter 2) simulations. Finally, Chapter 5 is a summary and conclusion of our work. In addition, we give a brief description of

possible future research directions.

1.3 Some Basic Definitions

In this section, we give definitions of some essential concepts that will be encountered throughout the thesis.

Definition 1.1 *A sequence of continuous functions* $\{f_n\}$, $f_n : I \to \mathbb{R}^k$, *is said to be equicontinuous on* I *if, for each* $\epsilon > 0$, *there is a* $\delta > 0$ *such that for all* $t_1, t_2 \in I$

$$\|f_n(t_1) - f_n(t_2)\| \leq \epsilon \text{ whenever } |t_1 - t_2| \leq \delta ,$$

for all n, *and where* $\|\cdot\|$ *denotes the Euclidean norm.*

Definition 1.2 *A sequence* $\{f_n\}$, $f_n : I \to \mathbb{R}^k$, *is said to be uniformly bounded if there exists an* $M > 0$ *such that*

$$\|f_n(t)\| \leq M , \text{ for all } t \in I \text{ and all } n .$$

The following result is the well-known Arzelã-Ascoli theorem [25]:

Theorem 1.1 *If* $\{g_n\}$, $g_n : I \to \mathbb{R}^k$, *is a uniformly bounded and equicontinuous sequence of functions, then there exists a subsequence which converges uniformly on* I.

Definition 1.3 *Consider a sequence of random variables,* $\{X_n\}$, *defined on a probability space* $(\Omega, \mathcal{F}, \mathcal{P})$. *We say that* X_n *converges to* X *with probability one (w.p.1) if* $P(B) = 0$, *where* $B = \{\omega |\ X_n(\omega) \overset{n\to\infty}{\not\to} X(\omega)\}$.

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a finite probability space and $\mathcal{D}_1 \leq \mathcal{D}_2 \leq \cdots \leq \mathcal{D}_n$ a sequence of decompositions.

Definition 1.4 *A sequence of random variables $\xi_1, \cdots, \xi_n$ is called a martingale (with respect to the decomposition $\mathcal{D}_1 \leq \mathcal{D}_2 \leq \cdots \leq \mathcal{D}_n$) if*

1. *ξ_k is $\mathcal{D}_k$-measurable,*

2. *$E(\xi_{k+1}|\mathcal{D}_k) = \xi_k$, $1 \leq k \leq n-1$.*

Consider a system of ODEs given by

$$\dot{x} = f(t,x)\ ,\ (t,x) \in \mathbb{R}^1 \times \mathbb{R}^n\ ,\ x(t_0) = x_0\ ,\ (t_0,x_0) \in U\ , \qquad (E)$$

where $f : U \to \mathbb{R}^n$, with U an open set in $\mathbb{R}^1 \times \mathbb{R}^n$. Assume that (E) possesses an isolated equilibrium at the origin. In other words, $f(t,0) = 0$ for all $t \geq 0$. Furthermore, suppose that $\phi(t,t_0,x_0)$ is a unique solution of (E) which depends continuously on the initial data (t_0,x_0).

Definition 1.5 *The equilibrium $x = 0$ of (E) is stable if, for every $\epsilon > 0$ and any $t_0 \in \mathbb{R}^+$, there exists a $\delta(\epsilon,t_0) > 0$ such that*

$$|\phi(t,t_0,x_0)| < \epsilon \text{ for all } t \geq t_0$$

whenever

$$|x_0| < \delta(\epsilon,t_0)\ ..$$

Definition 1.6 *The equilibrium $x = 0$ of (E) is locally asymptotically stable if*

1. *it is stable. and*

2. *for every* $t_0 \geq 0$, *there exists an* $\eta(t_0) > 0$ *such that*

$$\lim_{t \to \infty} \phi(t, t_0, x_0) = 0$$

whenever $|x_0| < \eta$.

Definition 1.7 *The set of all* $x_0 \in \mathbb{R}^n$ *such that* $\phi(t, t_0, x_0) \to 0$ *as* $t \to \infty$ *for some* $t_0 \geq 0$ *is called the domain of attraction of the equilibrium* $x = 0$ *of (E).*

1.4 The Main Theorem

Theorem 1.2 *Let* $\{\mathbf{w}(n)\}$ *be given by*

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \gamma(n) h[\mathbf{w}(n), \mathbf{x}(n)], \quad n = 0, 1, \cdots \tag{1.6}$$

where n *denotes the iteration number,* $\{\mathbf{w}(n)\}$ *is a sequence of vectors in* $\mathbb{R}^r$ *that are the object of interest, and* $\mathbf{x}(n) \in \mathbb{R}^r$ *is an input vector received at time* n *, which causes* $\mathbf{w}(n)$ *to be updated to take account of new information. Assume that (1.6) satisfies the following set of conditions:*

i. $\{\gamma(n)\}$ *is a decreasing sequence of positive real numbers, such that*

$$\begin{aligned} &(a) \quad \sum_{n=1}^{\infty} \gamma(n) = \infty \text{ , and} \\ &(b) \quad \gamma(n) \to 0 \text{ as } n \to \infty \text{ .} \end{aligned} \tag{1.7}$$

ii. The sequence of vectors $\{\mathbf{w}(n)\}$ *is bounded wp1.*

iii. $h(\cdot,\cdot)$ *is a continuous* $\mathbb{R}^r$*-valued function on* $\mathbb{R}^r \times \mathbb{R}^r$.

iv. The conditional expectation (see [1]) defined by

$$\bar{h}(\mathbf{z}) = E[h(\mathbf{z}, \mathbf{x}(n)) \mid \mathbf{x}(0), \cdots, \mathbf{x}(n-1)], \tag{1.8}$$

where $\mathbf{z} \in \mathbb{R}^r$ *is a regular (i.e. non-random) variable. exists. The statistical expectation is taken with respect to* $\mathbf{x}(n)$*, under the assumption that the probability density function (pdf) of the random variable* $\mathbf{x}(n)$ *is known.*

v. Assume that

$$\lim_{n\to\infty} P\{\sup_{m\geq n} \mid \sum_{i=n}^{m} \gamma(i)\xi(i) \mid \geq \epsilon\} = 0 \ , \tag{1.9}$$

for each $\epsilon > 0$*, and where* $\{\xi(i)\}$ *is a sequence of* $\mathbb{R}^r$*-valued random variables defined by*

$$\begin{aligned} \mathbf{w}(n+1) &= \mathbf{w}(n) + \gamma(n)\bar{h}[\mathbf{w}(n)] + \gamma(n)\{h[\mathbf{w}(n), \mathbf{x}(n)] - \bar{h}[\mathbf{w}(n)]\} \\ &= \mathbf{w}(n) + \gamma(n)\bar{h}[\mathbf{w}(n)] + \gamma(n)\xi(n) \ . \end{aligned} \tag{1.10}$$

Then, $\{\mathbf{w}(n)\}$ *converges to a solution of the system of ODEs*

$$\frac{d\mathbf{z}(t)}{dt} = \bar{h}(\mathbf{z}(t)), \tag{1.11}$$

where t *denotes continuous time, and* $\mathbf{z} \in \mathbb{R}^r$.

The following extra condition specifies criteria which ensure the convergence of $\{\mathbf{w}(n)\}$ *to a particular solution of (1.11).*

vi. Let $\mathbf{z}_0$ *be a locally asymptotically stable equilibrium solution to (1.11). with domain of attraction* $DA(\mathbf{z}_0)$. *If there is a compact set* $A \subset DA(\mathbf{z}_0)$ *such that* $\mathbf{w}(n) \in A$ *infinitely often, then*

$$\mathbf{w}(n) \to \mathbf{z}_0 \ \ wp1 \ , \ \ as \ \ n \to \infty.$$

Note: By "infinitely often", we mean that there exists an infinite sequence of integers $\{n_i\}_{i=1}^{\infty}$, $n_i < n_{i+1}$, $i \geq 1$, *such that* $\mathbf{w}(n) \in A$ *for all* i.

Remarks on the noise condition (v)

Equation (1.10) may be rewritten as

$$\frac{\mathbf{w}(n+1) - \mathbf{w}(n)}{\gamma(n)} = \bar{h}[\mathbf{w}(n)] + \xi(n) \ , \ \ n \geq 0 \ ,$$

which resembles the well-known Euler difference scheme for approximating solutions of ODEs. Thus, if $\xi(n)$ is *suitably constrained*, we expect $\{\mathbf{w}(n)\}$ to converge (wp1) to a particular solution of the system of ODEs given by (1.11). The condition given by (1.9) is not usually straightforward to verify in most practical applications. However, it may be shown to hold under the following conditions [1], viz. $\{\sum_{i=0}^{n} \gamma(i)\xi(i)\}$ is a martingale sequence, and

$$\sum_{i=0}^{\infty} \gamma^p(i) < \infty \ \text{ for some } \ p > 1. \tag{1.12}$$

Remarks on conditional expectation

The conditional expectation in (1.8) is taken with respect to $\mathbf{x}(n)$, since the values $\{\mathbf{x}(0), \cdots, \mathbf{x}(n-1)\}$ are given. From (1.6), this implies that $\{\mathbf{w}(1), \cdots, \mathbf{w}(n)\}$ is known. Replacing $\mathbf{z}$ by $\mathbf{w}(n)$ leads to

$$\bar{h}(\mathbf{w}(n)) = E[h(\mathbf{w}(n), \mathbf{x}(n)) \mid \mathbf{x}(0), \cdots, \mathbf{x}(n-1)] \,,$$

where the conditional expectation is still with respect to $\mathbf{x}(n)$. Note that $\mathbf{w}(0)$ is given in most practical implementations of (1.6). Usually, it is randomly picked (for symmetry-breaking purposes) from a predetermined set. Finally, note that if $\{x(n)\}$ is a sequence of *independent, identically distributed* (*i.i.d.* for short) random variables, then the conditioning is redundant, i.e. we have that

$$\bar{h}(\mathbf{w}(n)) = E[h(\mathbf{w}(n), \mathbf{x}(n))] \,.$$

1.5 Derivation of the associated ODE

The basic goal of stochastic approximation, using ODE and compactness methods, is to investigate the asymptotic properties of the sequence $\{\mathbf{w}(n)\}$, and to relate them to properties of the associated ODE. In order to do this, we need to interpolate $\{\mathbf{w}(n)\}$ and $\{\sum_{i=0}^{n} \gamma(i)\xi(i)\}$ into continuous time processes. Then, we define sequences of left shifts which bring the asymptotic parts of these two sequences to a neighbourhood of the time origin [1].

Define the *piecewise linear* interpolations, $\mathbf{w}^0(.)$ and $M^0(.)$, of $\{\mathbf{w}(n)\}$ and

$\{\sum_{i=0}^{n-1}\gamma(i)\xi(i)\}$. respectively:

$$\begin{aligned}\mathbf{w}^0(t_n) &= \mathbf{w}(n)\ , \text{ and}\\ \mathbf{w}^0(t) &= \frac{(t_{n+1}-t)}{\gamma(n)}\mathbf{w}(n)+\frac{(t-t_n)}{\gamma(n)}\mathbf{w}(n+1),\ \ t\in[t_n,t_{n+1})\ ,\end{aligned}\tag{1.13}$$

where $t_n \triangleq \sum_{i=0}^{n-1}\gamma(i)$, $t_0=0$.

$$\begin{aligned}M^0(t_n) &= \sum_{i=0}^{n-1}\gamma(i)\xi(i)\ , \text{ and}\\ M^0(t) &= \frac{(t_{n+1}-t)}{\gamma(n)}M^0(t_n)+\frac{(t-t_n)}{\gamma(n)}M^0(t_{n+1}),\ \ t\in[t_n,t_{n+1})\ .\end{aligned}\tag{1.14}$$

Next. define the corresponding *piecewise constant* right continuous interpolations:

$$\begin{aligned}\bar{\mathbf{w}}^0(t) &= \mathbf{w}^0(t_n)\ , \text{ and}\\ \bar{M}^0(t) &= M^0(t_n)\ ,\end{aligned}\tag{1.15}$$

for $t\in[t_n,t_{n+1})$. Equation (1.13) implies that

$$\begin{aligned}\mathbf{w}^0(t) &= \frac{(t_{n+1}-t)}{\gamma(n)}\mathbf{w}(n)+\frac{(t-t_n)}{\gamma(n)}[\mathbf{w}(n)+\gamma(n)\bar{h}(\mathbf{w}(n))+\gamma(n)\xi(n)]\\ &= \mathbf{w}(n)+(t-t_n)\bar{h}(\mathbf{w}(n))+(t-t_n)\xi(n)\ .\end{aligned}\tag{1.16}$$

From (1.10), we have that

$$\begin{aligned}\mathbf{w}(n) &= \mathbf{w}(n-1)+\gamma(n-1)\bar{h}(\mathbf{w}(n-1))+\gamma(n-1)\xi(n-1)\\ &= \mathbf{w}(0)+\sum_{i=0}^{n-1}\gamma(i)\bar{h}(\mathbf{w}(i))+\sum_{i=0}^{n-1}\gamma(i)\xi(i)\ .\end{aligned}\tag{1.17}$$

Substituting (1.17) into (1.16) gives

$$\begin{aligned}\mathbf{w}^0(t) &= \mathbf{w}(0) + \sum_{k=0}^{n-1} \gamma(k)\bar{h}(\mathbf{w}(k)) + \sum_{k=0}^{n-1} \gamma(k)\xi(k) + (t-t_n)\bar{h}(\mathbf{w}(n)) + \\ &\quad (t-t_n)\xi(n) \\ &= \mathbf{w}(0) + \int_0^t \bar{h}(\bar{\mathbf{w}}^0(s))ds + \sum_{k=0}^{n-1} \gamma(k)\xi(k) + (t-t_n)\xi(n)\ , \qquad (1.18)\end{aligned}$$

where

$$\int_0^t \bar{h}(\bar{\mathbf{w}}^0(s))ds = \sum_{k=0}^{n-1} \gamma(k)\bar{h}(\mathbf{w}(k)) + (t-t_n)\bar{h}(\mathbf{w}(n))\ . \qquad (1.19)$$

From (1.14), note that

$$M^0(t) = M^0(t_n) + (t-t_n)\xi(n)\ .$$

Therefore (1.18) may be rewritten as

$$\mathbf{w}^0(t) = \mathbf{w}(0) + \int_0^t \bar{h}(\bar{\mathbf{w}}^0(s))ds + M^0(t)\ . \qquad (1.20)$$

To get our sequence of left shifts, define the functions $\mathbf{w}^n(.)$ and $M^n(.)$ on $(-\infty, \infty)$ by:

$$\mathbf{w}^n(t) = \begin{cases} \mathbf{w}^0(t+t_n) & ,\ t \geq -t_n \\ \mathbf{w}(0) & ,\ t < -t_n \end{cases} \qquad (1.21)$$

and

$$M^n(t) = \begin{cases} M^0(t+t_n) - M^0(t_n) & ,\ t \geq -t_n \\ -M^0(t_n) & ,\ t < -t_n \end{cases} \qquad (1.22)$$

Using (1.20), equation (1.21) becomes (for $t \geq -t_n$)

$$\begin{aligned} \mathbf{w}^n(t) &= \mathbf{w}^0(t+t_n) \\ &= \mathbf{w}(0) + \int_0^{t+t_n} \bar{h}(\bar{\mathbf{w}}^0(s))ds + M^0(t+t_n) \ . \end{aligned} \tag{1.23}$$

From (1.17), we may write

$$\mathbf{w}(0) = \mathbf{w}(n) - \sum_{k=0}^{n-1} \gamma(k)\bar{h}(\mathbf{w}(k)) - \sum_{k=0}^{n-1} \gamma(k)\xi(k) \ ,$$

which is then substituted into (1.23) to give

$$\begin{aligned} \mathbf{w}^n(t) &= \mathbf{w}(n) + \left\{ \int_0^{t+t_n} \bar{h}(\bar{\mathbf{w}}^0(s))ds - \sum_{k=0}^{n-1} \gamma(k)\bar{h}(\mathbf{w}(k)) \right\} + \\ &\quad [M^0(t+t_n) - M^0(t_n)] \\ &= \mathbf{w}(n) + \left\{ \int_0^{t+t_n} \bar{h}(\bar{\mathbf{w}}^0(s))ds - \sum_{k=0}^{n-1} \gamma(k)\bar{h}(\mathbf{w}(k)) \right\} + M^n(t) \ . \end{aligned} \tag{1.24}$$

Noting that

$$\int_0^{t+t_n} \bar{h}(\bar{\mathbf{w}}^0(s))ds = \sum_{k=0}^{n-1} \gamma(k)\bar{h}(\mathbf{w}(k)) + \int_{t_n}^{t+t_n} \bar{h}(\bar{\mathbf{w}}^0(s))ds \ ,$$

equation (1.24) becomes

$$\begin{aligned} \mathbf{w}^n(t) &= \mathbf{w}(n) + \int_{t_n}^{t+t_n} \bar{h}(\bar{\mathbf{w}}^0(s))ds + M^n(t) \\ &= \mathbf{w}(n) + \int_0^t \bar{h}(\bar{\mathbf{w}}^0(t_n+s))ds + M^n(t) \ . \end{aligned} \tag{1.25}$$

Using (1.9), it can be shown that $M^n(t) \to 0$ as $n \to \infty$, for fixed t. Also, it can

be shown that the sequence $\{\mathbf{w}^n(.)\}$ is equicontinuous and uniformly bounded on each finite interval in $(-\infty, \infty)$.

Thus, by the Arzelà-Ascoli theorem, there exists a subsequence, also indexed by n, and a continuous bounded function $\mathbf{w}(.)$ such that

$$\mathbf{w}^n(t) \to \mathbf{w}(t) \text{ as } n \to \infty, \tag{1.26}$$

uniformly on finite time intervals. We may rewrite (1.25) as

$$\begin{aligned} \mathbf{w}^n(t) &= \mathbf{w}^n(0) + \int_0^t \bar{h}(\bar{\mathbf{w}}^0(t_n + s))ds + M^n(t) \\ &= \mathbf{w}^n(0) + \int_0^t \bar{h}(\mathbf{w}^n(s))ds + M^n(t) + e^n(t)\,, \end{aligned} \tag{1.27}$$

where

$$\begin{aligned} e^n(t) &= \int_0^t \{\bar{h}(\bar{\mathbf{w}}^0(t_n + s)) - \bar{h}(\mathbf{w}^n(s))\}ds \\ &= \int_{t_n}^{t+t_n} \{\bar{h}(\bar{\mathbf{w}}^0(u)) - \bar{h}(\mathbf{w}^0(u))\}du\,. \end{aligned}$$

From this, it can be shown that $e^n(t) \to 0$ as $n \to \infty$. From (1.26), it follows that

$$\mathbf{w}^n(0) \to \mathbf{w}(0) \text{ as } n \to \infty.$$

Therefore, (1.27) implies that

$$\mathbf{w}^n(t) \to \mathbf{w}(0) + \int_0^t \bar{h}(\mathbf{w}(s))ds\,, \text{ as } n \to \infty\,. \tag{1.28}$$

Finally, combining (1.26) and (1.28) yields

$$\mathbf{w}(t) = \mathbf{w}(0) + \int_0^t \bar{h}(\mathbf{w}(s))ds \ ,$$

which gives

$$\dot{\mathbf{w}}(t) = \bar{h}(\mathbf{w}(t)) \ , \tag{1.29}$$

the sought-after ODE associated with (1.10). The proof of result (vi) of the main theorem exploits the convergence of $\{\mathbf{w}(n)\}$ to a solution of (1.29), as detailed on pages 42-43 of [1].

1.6 Some simple examples

1.6.1 Example 1

Consider the iteration formula [1]

$$w_{n+1} = w_n + \gamma_n h(w_n) \ , \ w_n \in \mathbb{R} \ , \tag{1.30}$$

where it is assumed that $\{w_n\}$ is bounded, and that $\{\gamma_n\}$ is a sequence of $\mathbb{R}^+$-valued parameters satisfying condition (i) of the main theorem. The function $h(.)$ is a random variable whose probability density function is specified below. Furthermore, let $h_i(.)$, $i = 1, 2$, be continuous bounded $\mathbb{R}$-valued functions on $\mathbb{R}$, and let $\alpha(.)$ be a continuous function on $\mathbb{R}$ with values in [0,1]. Let the random variable $h(w_n)$,

parameterised by w_n, be defined by

$$\begin{aligned} P\{h(w_n) = h_1(w_n) \mid w_0, \cdots, w_n\} &= \alpha(w_n) , \\ P\{h(w_n) = h_2(w_n) \mid w_0, \cdots, w_n\} &= 1 - \alpha(w_n) . \end{aligned} \tag{1.31}$$

(Note that, in the spirit of the main theorem, algorithm (1.30) may be generalised to $\mathbb{R}^r$. That is, $w_n \in \mathbb{R}^r$, $h_i : \mathbb{R}^r \to \mathbb{R}^r$, $i = 1, 2$, and $\alpha : \mathbb{R}^r \to [0, 1]$.)

Our goal is to investigate the long term behaviour of (1.30). Since this difference equation has a random term, it is very difficult, if not outrightly impossible, to obtain an approximation of its long term behaviour by direct techniques. We shall use ODE and compactness methods, outlined in the previous section, to find an associated deterministic *autonomous* ODE. Then, if all the conditions of the main theorem are met, the ODE will have the same asymptotic properties as (1.30).

Define the following quantities:

$$\bar{h}(w) = \alpha(w)h_1(w) + [1 - \alpha(w)]h_2(w) , \tag{1.32}$$

and

$$\xi_n = h(w_n) - \bar{h}(w_n) . \tag{1.33}$$

Then, it is clear that $E[\xi_n \mid w_0, \cdots, w_n] = 0$ wpl. It can further be shown that, since the $h_i(.)$ are bounded, $var\{\xi_n\}$ is bounded uniformly in n.

Define the series

$$M_n = \sum_{i=0}^{n} \gamma_i \xi_i . \tag{1.34}$$

Then we have that

$$\begin{aligned} E[M_n \mid w_0, \cdots, w_n] &= E[\gamma_n \xi_n \mid w_0, \cdots, w_n] + E[\sum_{i=0}^{n-1} \gamma_i \xi_i \mid w_0, \cdots, w_n] \\ &= \gamma_n E[\xi_n \mid w_0, \cdots, w_n] + E[\sum_{i=0}^{n-1} \gamma_i \xi_i \mid w_0, \cdots, w_n] \\ &= E[\sum_{i=0}^{n-1} \gamma_i \xi_i \mid w_0, \cdots, w_n] \\ &= E[M_{n-1} \mid w_0, \cdots, w_n] \ , \end{aligned} \tag{1.35}$$

and

$$E \mid M_n \mid < \infty \ \text{ for each } \ n \ . \tag{1.36}$$

We conclude that $\{M_n\}$ is a martingale sequence. Employing (1.32) and (1.33), we may rewrite (1.30) as

$$w_{n+1} = w_n + \gamma_n \bar{h}(w_n) + \gamma_n \xi_n \ . \tag{1.37}$$

Now assume that

$$\sum_{i=0}^{n} \gamma_i^2 < \infty \ . \tag{1.38}$$

If $\{s_m\}$ is a martingale sequence, Doob's inequality [1] states that

$$P\{sup_{m \geq 0} \mid s_m \mid \geq \epsilon\} \leq \lim_{m \to \infty} \frac{E \mid s_m \mid^2}{\epsilon^2} \ , \ \text{ for each } \ \epsilon > 0 \ . \tag{1.39}$$

Applying this to our problem yields

$$P\{sup_{m \geq n} \mid \sum_{i=n}^{m} \gamma_i \xi_i \mid \geq \epsilon\} \leq \lim_{m} \frac{E \mid \sum_{i=n}^{m} \gamma_i \xi_i \mid^2}{\epsilon^2} \ . \tag{1.40}$$

The RHS of (1.40) may be manipulated to obtain

$$\begin{aligned} \lim_{m\to\infty} \frac{E \mid \sum_{i=n}^{m} \gamma_i \xi_i \mid^2}{\epsilon^2} &\leq \lim_{m\to\infty} \frac{E\left\{\sum_{i=n}^{m} \gamma_i \mid \xi_i \mid\right\}^2}{\epsilon^2} \\ &= \lim_{m\to\infty} \frac{\sum_{j=n}^{m}\sum_{k=n}^{m} \gamma_j\gamma_k E[\mid \xi_j \mid\mid \xi_k \mid]}{\epsilon^2} . \end{aligned} \tag{1.41}$$

Furthermore, let the sequence $\{\xi_l\}$ be such that there exists another sequence $\{R_\xi(j,k)\}$ satisfying

$$E[\mid \xi_j \mid\mid \xi_k \mid] \leq R_\xi(j,k) \text{ for all } \{j,k\} \text{ , and that } \sum_{j,k} R_\xi(j,k) < \infty . \tag{1.42}$$

Employing (1.42) in (1.41) yields

$$\begin{aligned} RHS &\leq \lim_{m\to\infty} \frac{\sum_{j=n}^{m}\sum_{k=n}^{m} \gamma_j\gamma_k R_\xi(j,k)}{\epsilon^2} \\ &\leq \frac{1}{\epsilon^2} \times \lim_{m\to\infty} [\max_{j,k=n..m} \{\gamma_j\gamma_k\}] \times \sum_{j=n}^{m}\sum_{k=n}^{m} R_\xi(j,k) \\ &= \frac{\gamma_{n+N}^2}{\epsilon^2} \times \lim_{m\to\infty} \sum_{j=n}^{m}\sum_{k=n}^{m} R_\xi(j,k) \text{ , where } n \leq N < m \text{ is fixed} \\ &\longrightarrow 0 \text{ as } n \to \infty . \end{aligned} \tag{1.43}$$

From (1.40), this implies that

$$P\{sup_{m\geq n} \mid \sum_{i=n}^{m} \gamma_i \xi_i \mid \geq \epsilon\} \to 0 \text{ as } n \to \infty . \tag{1.44}$$

Equation (1.44) holds for any $\epsilon > 0$ and for any realisation of $\{\xi_i\}$. The interpretation of (1.44) is that, even though ξ_i may not tend to zero as $i \to \infty$, the sequence of tails of the series $\sum_{i=n}^{m} \gamma_i \xi_i$ does converge to zero as $n \to \infty$. This is a relatively

strong condition on the sequence of random variables $\{\xi_i\}$.

Finally, as in the previous section, define the piecewise linear interpolations $w^0(.)$, $M^0(.)$, the piecewise constant interpolations $\bar{w}^0(.)$, $\bar{M}^0(.)$, and the left shifts $w^n(.)$, $M^n(.)$. Then, in exactly the same way as before, it can be shown that the sequence $\{w^n(.)\}$ has a convergent subsequence in $C(-\infty, \infty)$, and that all limits satisfy

$$\dot{z}(t) = \bar{h}(z),$$

the sought-after associated ODE.

1.6.2 Example 2: Numerical Simulations

Consider the two-dimensional Hebbian algorithm [12]

$$\mathbf{w}(n+1) = \mathbf{w}(n) + \gamma(n)y(n)[\mathbf{x}(n) - \mathbf{w}(n)u(n)] \tag{1.45}$$

applied to a linear neuron, where $u(n) = \mathbf{x}^T(n)\mathbf{w}(n)$, $y(n) = u(n)$, and $\mathbf{w}(n) = [w_1(n)\ w_2(n)]^T$. The sequence of learning parameters, $\{\gamma(n)\}$, is defined as in the previous example. Suppose that the neuron receives time-varying input signals which are defined by

$$\mathbf{x}(n) = \begin{cases} \mathbf{k}_1, & \text{with probability } 0.5 \\ \mathbf{k}_2, & \text{with probability } 0.5 \end{cases} \tag{1.46}$$

where $\mathbf{k}_1$ and $\mathbf{k}_2$ are given by

$$\mathbf{k}_1 \stackrel{def}{=} \begin{bmatrix} k_{11} \\ k_{12} \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0.25 \end{bmatrix}, \text{ and } \mathbf{k}_2 \stackrel{def}{=} \begin{bmatrix} k_{21} \\ k_{22} \end{bmatrix} = \begin{bmatrix} -0.25 \\ -0.5 \end{bmatrix}. \tag{1.47}$$

Hence, for $\mathbf{x}(n) = \mathbf{k}_1$, we have that

$$\begin{aligned} w_1(n+1) &= w_1(n) + \gamma(n)[k_{11}w_1(n) + k_{12}w_2(n)] \times \\ &\quad [k_{11} - k_{11}w_1^2(n) - k_{12}w_1(n)w_2(n)] , \\ w_2(n+1) &= w_2(n) + \gamma(n)[k_{11}w_1(n) + k_{12}w_2(n)] \times \\ &\quad [k_{12} - k_{11}w_1(n)w_2(n) - k_{12}w_2^2(n)] . \end{aligned} \tag{1.48}$$

Analogously, for $\mathbf{x}(n) = \mathbf{k}_2$, we obtain

$$\begin{aligned} w_1(n+1) &= w_1(n) + \gamma(n)[k_{21}w_1(n) + k_{22}w_2(n)] \times \\ &\quad [k_{21} - k_{21}w_1^2(n) - k_{22}w_1(n)w_2(n)] , \\ w_2(n+1) &= w_2(n) + \gamma(n)[k_{21}w_1(n) + k_{22}w_2(n)] \times \\ &\quad [k_{22} - k_{21}w_1(n)w_2(n) - k_{22}w_2^2(n)] . \end{aligned} \tag{1.49}$$

Denoting $\mathbf{z}(t) = [z_1(t)\ z_2(t)]^T$, we obtain the associated system of ODEs

$$\frac{d\mathbf{z}(t)}{dt} = \left\{ \mathbf{g}_1(t) - g_2(t) \begin{bmatrix} z_1(t) \\ z_2(t) \end{bmatrix} \right\}, \tag{1.50}$$

where

$$\mathbf{g}_1(t) = 0.5 \begin{bmatrix} \sum_{j=1}^{2} z_j(t)(k_{21}k_{2j} + k_{11}k_{1j}) \\ \sum_{j=1}^{2} z_j(t)(k_{22}k_{2j} + k_{12}k_{1j}) \end{bmatrix},$$

and

$$g_2(t) = 0.5 \left\{ \sum_{m=1}^{2} \sum_{j=1}^{2} z_m(t) z_j(t) [k_{2m}k_{2j} + k_{1m}k_{1j}] \right\}.$$

Following (1.48) and (1.49), it can be shown that the fixed points of the algorithm (1.45) are $(0,0)$, and

$$\begin{cases} \pm(0.894, 0.447) & , \quad \text{if } \mathbf{x}(n) = \mathbf{k}_1 \text{ with probability 1} \\ \pm(0.447, 0.894) & , \quad \text{if } \mathbf{x}(n) = \mathbf{k}_2 \text{ with probability 1}. \end{cases}$$

From this information alone, it is not immediately clear to which fixed point the sequence $\{(w_1(n), w_2(n))\}$ will converge if signals defined by (1.46) are used. To illustrate the ambiguity of this problem, we perform numerical runs of (1.45), with $(w_1(0), w_2(0))$ randomised in [-1, 1]. Fig 1.1 (*left*) shows the results of performing 20 runs of (1.45). The learning parameter used is of the form $\gamma_n = (n+1)^{-0.25}$. Now from (1.50), it can be shown that the associated system of ODEs has two locally asymptotically stable equilibria, viz. $\pm(0.707, 0.707)$, and an unstable equilibrium point at (0,0).

Numerically solving (1.50), using a 4^{th} order Runge-Kutta scheme, and with initial conditions randomised in [0,1], for example, we obtain the results displayed in Fig 1.1 (*right*).

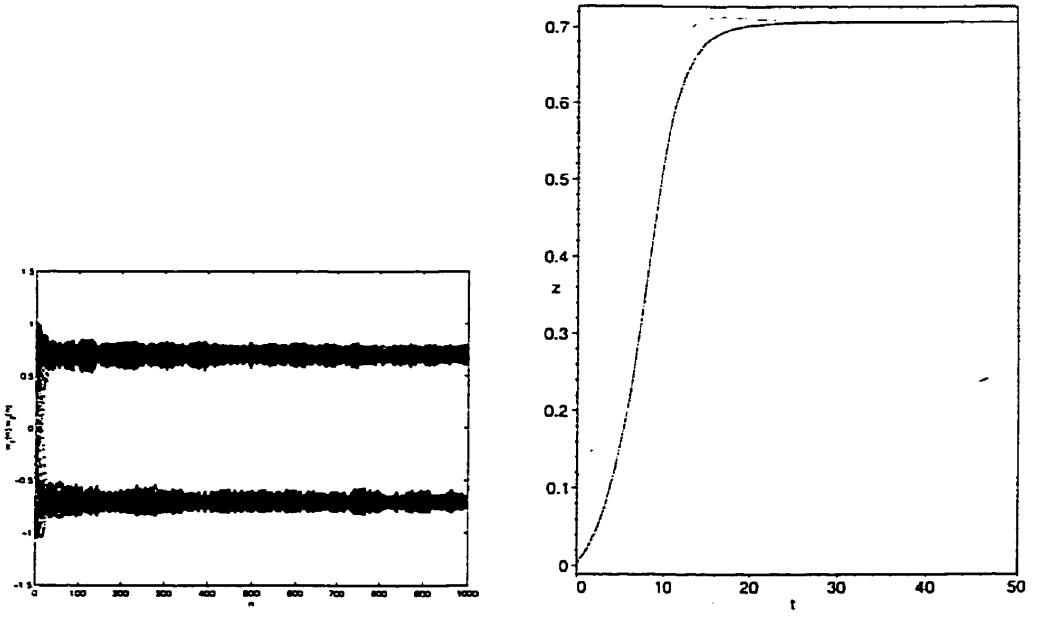

Figure 1.1: Temporal evolution of: *left*: $\{(w_1(n), w_2(n))\}$ *right*: $\{(z_1(t), z_2(t))\}$

1.6.3 Comments on numerics

It is important to note that the Kushner-Clark theorem is *not* applicable to this problem, since the associated system of ODEs has *two* competing locally asymptotically stable equilibria. The numerics suggest that the sequence $\{(w_1(n), w_2(n))\}$ approaches the two locally asymptotically stable equilibria of the associated system of ODEs (see Fig 1.1 (*left*)). Note that Fig 1.1 (*right*) depicts $\{(z_1(t), z_2(t))\}$ evidently approaching the equilibrium point (0.707, 0.707).

The fundamental issue arising from this example is encapsulated in the following set of questions: *How often does* $\{(w_1(n), w_2(n))\}$ *converge to either of the two fixed points? Is this dependent on initial conditions of the algorithm? Given an initial condition, can we say with what probability the algorithm will converge to either of the two fixed points?*

These questions articulate and underpin the core theme of this thesis. As previously stated, this is very much an open problem in the area of stochastic approximation. The thesis seeks to qualitatively and quantitatively investigate this problem.

1.7 The neurobiological connection

The operation of the brain depends on the flow of information through elaborate circuits consisting of networks of neurons or nerve cells. Information is relayed from one cell to the next at specialised points of contact: the *synapses*. Most neurons share certain structural features that make it possible to distinguish three regions of the cell: the *cell body*, the *dendrites*, and the *axon*. The cell body contains the nucleus of the neuron and the biochemical infrastructure for synthesizing enzymes and other molecules essential to the life of the cell. The dendrites provide the main physical surface on which the neuron receives incoming signals. The axon provides the pathway over which signals can travel from the cell body, for long distances, to other parts of the brain and nervous system. Fig 1.2 (*left*) depicts the structure of a pyramidal cell, one of the most common types of cells occurring in the cerebral cortex.

Information is relayed from one neuron to the next via a chemical transmitter generated and stored at the synapses. From the dendrites, through the cell body, the axon, and right up to the synaptic terminals, information is transmitted in the form of electrical signals. The *firing* of a neuron, which is the generation of nerve impulses, reflects the activation of hundreds of synapses by impinging

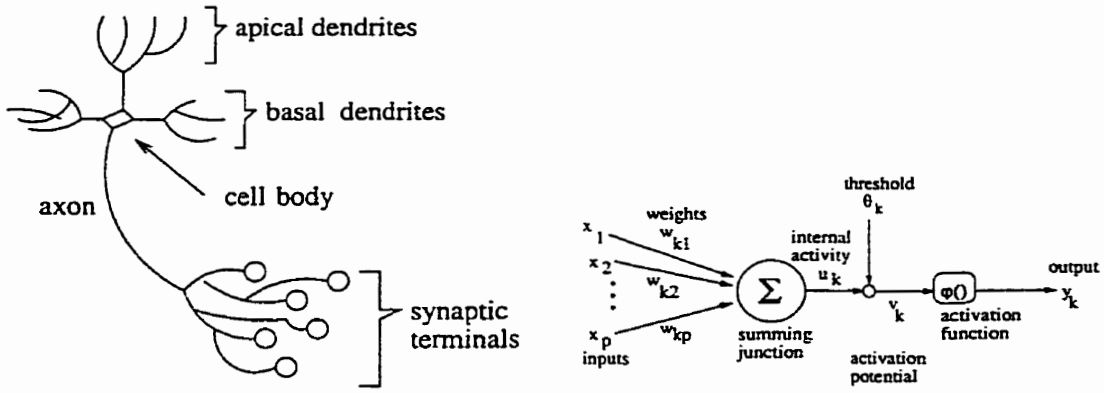

Figure 1.2: *left*: morphology of a pyramidal cell *right*: nonlinear model of a neuron

neurons. Some synapses are *excitatory* in that they tend to promote firing, whereas others are *inhibitory* and so are capable of neutralising signals that otherwise would excite a neuron to fire. At any instant, a neuron has some *threshold*, which an excitation must exceed to initiate an impulse. This, except for the fact and time of its occurrence, is determined by the neuron, not by the excitation. Between the time of arrival of impulses upon a neuron and its own propagated impulse, there is a synaptic *delay*, characteristically more than 5×10^{-4} seconds [10].

1.7.1 Artificial neural networks

An artificial neuron is a mathematical model which attempts to reproduce the functional characteristics of a biological neuron. An interconnection of these neurons is termed an *artificial neural network* (ANN) [12]. It is worth mentioning, at the onset, that artificial neurons and their corresponding networks are *far* from even approximating their biological counterparts. This is not surprising since the brain is indeed a highly complicated structure, the mechanics of which is not yet

fully understood. At the heart of the current surge in artificial neural networks research lies the challenging prospect of mechanising intellectual tasks which previously have been performed only by humans. Already, some significant progress has been made in this direction. Some of the examples that come to mind are [9]: weather forecasting, handwritten-character recognition, automated fingerprint identification, automated disease diagnosis, speech recognition, image processing, security/speaker identification, etc. Many of these tasks involve the ability *to classify or sort data.*

A nonlinear model of a neuron is depicted in Fig 1.2 (*right*). Essentially, the model consists of a finite set of input signals $\mathbf{x} = \{x_1, ..., x_p\}$, a corresponding set of weights $\mathbf{w} = \{w_{k1}, ..., w_{kp}\}$, a linear summing junction, a nonlinear activation function φ, a threshold θ_k, and an output y_k.

Below, we provide clarification of the nomenclature used:

- The linear combiner is an attempt to represent the *integrating property* of the brain. The brain routinely receives a myriad of different signals from the environment, integrates them, and produces a response which, in turn, determines an appropriate course of action to be taken.

- The neuron is assumed to be the k^{th} member of the network under consideration. The notation w_{kj} means that the referenced weight links input j to neuron k.

- u_k is called the *internal activity* of neuron k. It is the linear combiner output

for this particular neuron, viz.

$$u_k = \mathbf{w}_k^T \mathbf{x} \, .$$

- An *activation function*, $\varphi(.)$, is a smooth, nonlinear function that constrains the amplitude of the output signal to a desired range. The smoothness aspect of this function captures the *graded* response characteristic of the brain.

- A *threshold*, θ_k, has the effect of applying an affine transformation to u_k as shown by

$$v_k = u_k - \theta_k \, ,$$

where v_k is the *activation potential* of neuron k. The output of neuron k is given by

$$\begin{aligned} y_k &= \varphi(\sum_{i=1}^{p} w_{ki} x_i - \theta_k) \\ &= \varphi(v_k) \, . \end{aligned}$$

When $\theta_k < 0$, it is called a *bias*. The net input to $\varphi(.)$ may increase or decrease, depending on whether we apply a bias or a threshold term.

- We say that a neuron is linear if φ is the identity function. In that case, we have that $y_k = v_k$.

- There is a wide range of possible nonlinear activation functions. One of the

most commonly used ones is the so-called *logistic function*:

$$\varphi(v) = \frac{1}{1 + \exp(-av)} ,$$

where $\frac{a}{4}$ is the slope of the linear approximation to $\varphi(v)$ at zero, as illustrated in Fig 1.3.

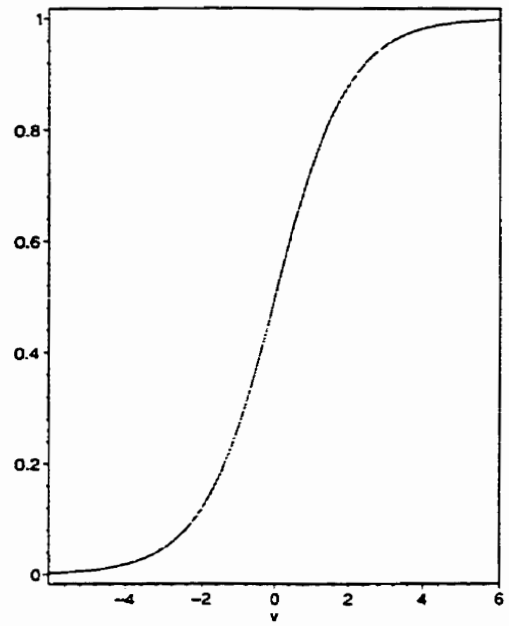

Figure 1.3: The logistic function with $a = 1$

- It is of interest to note that the activation function used in the first neuron model, due to McCulloch and Pitts [10], is a (discontinuous) threshold function:

$$\varphi(v) = \begin{cases} 1 \text{ , if } v \geq 0 \\ 0 \text{ , if } v < 0 \end{cases}$$

This choice of φ fails to represent the graded response aspect of the brain

that was mentioned above. They made the assumption that the brain is characterised by a binary input-output response function, which is contrary to experience.

1.7.2 Learning and generalisation in ANNs

One of the fundamental characteristics of the human brain is its *ability to learn from the environment, and to generalise.* Needless to say, this capability is essential for the survival of the human species. The origin of this adaptive behaviour in biological systems is the subject of Ashby's book [11]. The task of replicating this property in ANNs is currently a subject of intense research.

Learning in ANNs is an iterative procedure (algorithm) by which the free parameters of a network are adapted through a continuing process of stimulation by the environment in which the network is embedded [12].

There is a whole wide range of different learning schemes, each being determined by the manner in which the free parameters are adapted. Normally, the parameters are altered so as to attain a desired design goal such as the minimisation of an error function [13]. A typical learning rule is of the form

$$w_{ji}(n+1) = w_{ji}(n) + \Delta w_{ji}(n),$$

where n denotes the iteration number, and $\Delta w_{ji}(n)$ is the weight update at iteration n. Essentially, the weight update determines the type of learning rule. A realistic learning rule has to converge, in some sense, as $n \to \infty$. This means that the weight vector should tend to a finite limiting value as $n \to \infty$.

Learning (also called *training*) involves repeatedly presenting the network with a finite set of *training examples* and updating the weights until stabilisation is achieved. The weights are updated according to some rule, preferably one that complies with the desired design objectives. Often. three different versions of the same learning rule have been given: the *on-line* version where the modification is calculated after the presentation of each input signal (pattern), the *off-line* version where the previous modifications are averaged over the cycle of all patterns, and the *continuous* version where the discrete changes induced by the off-line algorithm are approximated continuously by a differential equation governing the evolution of the weights in time [13]. In the spirit of research in this field, the updating rule should make neurobiological sense. The concept of generalisation refers to the network producing reasonable outputs for inputs not encountered during training. Of course, these post-training inputs (*validation set*) have to be from the same 'family' as the training examples.

The earliest [1] demonstration that neural networks could actually be trained to perform certain tasks which, hitherto, could only be performed by humans is due to Rosenblatt [14]. He coined the term *perceptron* for the simplest form of a neural network used to classify *linearly separable* patterns. These are patterns that lie on opposite sides of a hyperplane. A more technical definition is given below. Essentially, a perceptron consists of a single McCulloch-Pitts neuron as described in Section 1.6.1. Consider a set of incoming signals, $\{\mathbf{x}\}$, divided into two disjoint classes C_1 and C_2, say. A perceptron will find a separating hyperplane $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$,

[1] The first proof of the so-called Perceptron Convergence Theorem is due to Agmon [16].

such that

$$g(\mathbf{x}) > 0 \text{ if } \mathbf{x} \in C_1 , \text{ and}$$
$$g(\mathbf{x}) \leq 0 \text{ if } \mathbf{x} \in C_2 .$$

The vector $\mathbf{w}$ denotes the free parameters of the network. We say that the two classes of patterns are linearly separable if there exists a $\mathbf{w}$ such that the two inequalities above are satisfied. The algorithm developed by Rosenblatt recursively determines the vector $\mathbf{w}$ for which the inequalities are true. He established the convergence, in a finite number of steps, of an error-correction procedure for an elementary perceptron to a classification or dichotomy of the input signals, providing such a dichotomy exists. Rosenblatt's now famous *Perceptron Convergence Theorem* is stated below [15].

Theorem 1.3 *Consider a set of vectors,* $\mathbf{x}_1, \cdots, \mathbf{x}_N$, *in some fixed finite dimensional Euclidean space, that are assumed to satisfy the single hypothesis that there exists a vector* $\mathbf{w}_*$ *such that*

$$\mathbf{x}_i^T \mathbf{w}_* > \theta > 0 \, , i = 1, \cdots, N.$$

Then consider an infinite sequence

$$\mathbf{x}_{i1}, \ \mathbf{x}_{i2}, \ \mathbf{x}_{i3}, \cdots, \ \ (1 \leq i_k \leq N \ \textit{for every } k) \, ,$$

such that each vector, $\mathbf{x}_1, \cdots, \mathbf{x}_N$, *occurs infinitely often. Now construct a sequence*

of vectors. $\mathbf{w}_0, \mathbf{w}_1, \cdots, \mathbf{w}_N, \cdots$, *recursively as follows:*

$$\mathbf{w}_0 \text{ is arbitrary},$$

and

$$\mathbf{w}_n = \begin{cases} \mathbf{w}_{n-1} & \text{if} \quad \mathbf{w}_{n-1}^T \mathbf{x}_{in} > \theta \\ \mathbf{w}_{n-1} + \mathbf{x}_{in} & \text{if} \quad \mathbf{w}_{n-1}^T \mathbf{x}_{in} \leq \theta \end{cases}$$

Then, the sequence $\{\mathbf{w}_n\}$ *is convergent in the sense that there exists an index* $m < \infty$ *such that*

$$\mathbf{w}_m = \mathbf{w}_{m+1} = \mathbf{w}_{m+2} = \cdots = \tilde{\mathbf{w}} .$$

It is worth mentioning that there are numerous variants of this theorem in the literature (see [12], [17], and references therein). The one stated above is enough to illustrate convergence in one dimension. The Perceptron Convergence Theorem single-handedly revolutionised research in the discipline of artificial neural networks.

Most neural learning algorithms are generally (systems of) difference equations of the form (1.1), and whose long term behaviour may be investigated using stochastic approximation techniques. Needless to say, understanding the asymptotics of these algorithms is essential in designing effective training strategies.

Chapter 2

Dynamics of Ljung's Automatic Classifier

2.1 Introduction

The purpose of this chapter is two-fold. First, we examine the system of ODEs associated with Ljung's automatic classifier [2]. This analysis is performed in the continuous time phase space. In the second half of the chapter, we focus on the discrete phase space. The connection between the dynamics on these two phase spaces is understood, via the celebrated Kushner-Clark theorem [1], only if the associated system of ODEs has a single locally asymptotically stable equilibrium. As previously stated, we are interested in extending the Kushner-Clark theorem to cover the case when the associated system of ODEs has multiple locally asymptotically stable equilibria. In other words, we seek to resolve the question: *to which one of these stable equilibria is the discrete system most likely to converge?* To

tackle this problem, both analytical and numerical techniques will be adopted. We choose to study Ljung's classification algorithm because of its relative simplicity, viz. it classifies a one-dimensional set of data, which is the most basic classification problem.

Ljung [2] considered an automatic classifier which receives scalar valued signals $\{\varphi(t)\}$ that may belong to two a priori unknown classes A and B, with probabilities λ and $1-\lambda$ respectively, and where $\lambda \in [0,1]$ is a free parameter. The signals are distributed according to a specified rule, viz. the doubly-triangular probability density function (*pdf*) shown in Fig. 2.1 and algebraically defined by

$$f_\lambda(\varphi) = \begin{cases} \lambda(\varphi+3) & , \quad -3 \le \varphi < -2 \\ -\lambda(\varphi+1) & , \quad -2 \le \varphi < -1 \\ (1-\lambda)(\varphi-1) & , \quad 1 \le \varphi < 2 \\ (\lambda-1)(\varphi-3) & , \quad 2 \le \varphi < 3 \\ 0 & , \quad \text{otherwise} . \end{cases} \tag{2.1}$$

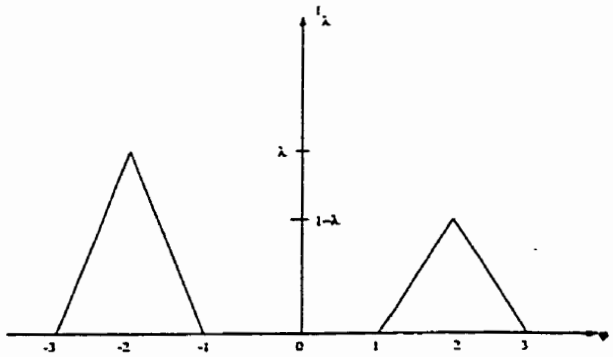

Figure 2.1: pdf of the random signals to be classified

The classifier is tasked with determining a number $c(t)$ such that $\varphi(t)$ is classified

as A if $\varphi(t) \leq c(t)$ and as B otherwise. The number $c(t)$ is given by

$$c(t) = \frac{1}{2}[x_A(t) + x_B(t)] \,, \tag{2.2}$$

where

$$x_A(t) = \begin{cases} x_A(t-1) + \gamma_t[\varphi(t) - x_A(t-1)] & , \quad \text{if } \varphi(t) \leq c(t-1) \\ x_A(t-1) & , \quad \text{otherwise} \end{cases} \tag{2.3}$$

The iterates $\{x_B(t)\}$ are defined in a similar manner. The γ_t are decreasing-to-zero positive real numbers, such that $\sum_{t=0}^{\infty} \gamma_t = \infty$. It is assumed that $\{\varphi(t)\}$ is a sequence of independent identically-distributed (*i.i.d.*) random variables. Furthermore, from Fig. 2.1, it is desirable that $c(t) \to c_0$ as $t \to \infty$, where $c_0 \in (-1, 1)$. This ensures that no samples are misclassified.

Ljung's contribution to this problem is outlined below. First, he derived a system of ODEs associated with the discrete system, part of which is given in (2.3). Then he demonstrated, by numerically integrating the ODE vector field, that the number of stable equilibria increases from one to two, as λ is increased from 0.5 to 0.99. When $\lambda = 0.5$, the system of ODEs has one globally asymptotically stable equilibrium at $(x_1^*, x_2^*) = (-2, 2)$. Following the Kushner-Clark theorem [1], $(x_A(t), x_B(t)) \to (-2, 2)$ with probability 1 (wp1 for short) as $t \to \infty$, giving the correct classification rule $c^* = 0$. When $\lambda = 0.99$, the system of ODEs has two locally asymptotically stable equilibria at $(x_1^*, x_2^*) = (-2, 2)$ and $(x_1^{**}, x_2^{**}) = (-2.258, -1.372)$. In this case, the Kushner-Clark theorem is not applicable, and the problem of determining the long-term behaviour of $(x_A(t), x_B(t))$, and hence

$c(t)$, is almost intractable. In view of this difficulty, Ljung performed numerical simulations of the discrete system. In 257 of these simulations, the classifier $c(t)$ converged to the undesired value $c^{**} = -1.815$ only *once*, while converging to the desired value $c^{*} = 0$ the rest of the time. When $c(t)$ converges to the desired value, all the samples are correctly classified, and when it converges to the undesired value, 33% of the samples are misclassified. It is interesting to ask *how often* convergence to c^{**} happens. Of course, Ljung had only one 'bad' run in 257 simulations. But, what is the meaning of his results, especially with regard to the convergence of $(x_A(t), x_B(t))$? Can we conclude, from his experiment, that the discrete system converges *with probability one* to $(-2, 2)$? Why, or why not? These questions set the stage for the spirit of this chapter.

We mention that Ljung furnishes neither the γ_t nor the initial condition $(x_A(0), x_B(0))$ that he uses in his simulations leading to his Fig 4. Furthermore, it is important to note that he only performs 1000 iterations. It is quite feasible that his Fig 4 might change if more iterations are carried out. It is also interesting to note that we "failed" to reproduce his Fig 4. Instead, we only managed to get Fig 2.2 (*left*). The learning parameter used is $\gamma_n = (n+10)^{-\alpha}$, with $\alpha = 0.15$ and the initial condition used is $(x_A(0), x_B(0)) = (-0.1, 0.1)$. 240 runs were performed. Similar results were obtained for other choices of α and γ_n. A different sequence $\{\varphi_n\}$ was used for each simulation. Contrary to Ljung's results, we do not obtain the "misbehaving" trajectory, i.e. we do not get the classification rule $c^{**} = -1.815$. For the case $\lambda = 0.5$, we obtained Fig 2.2 (*right*). Note that the classifier is converging to $c^{*} = 0$, as expected.

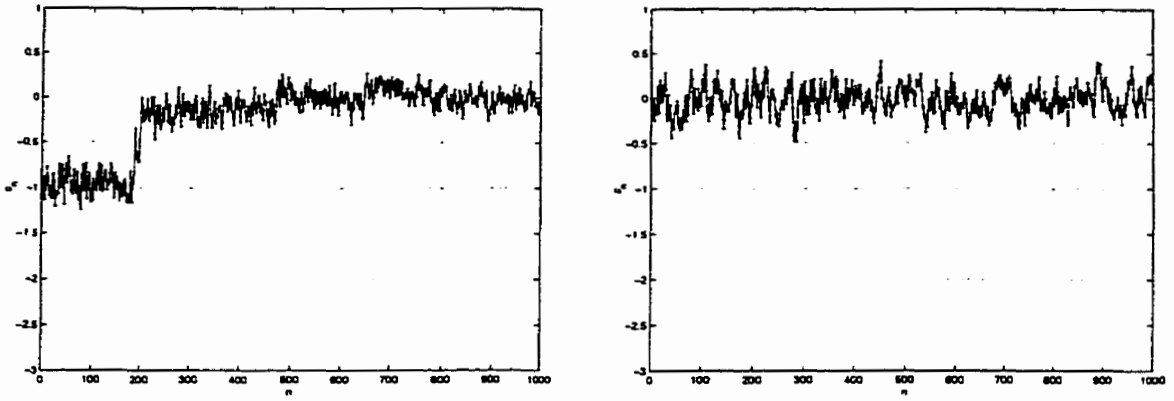

Figure 2.2: time histories of $c(n)$, with: *left*: $\lambda = 0.99$ *right*: $\lambda = 0.5$. In both Figs., note that only one typical run is shown.

Another interesting question is: *What is happening to the associated system of ODEs, as λ is increased from 0.5 to 0.99?* Ljung did not adequately address this problem. Intuitively, the drop in classifier performance *when* $c(t) \to -1.815$ suggests some kind of bifurcation phenomenon. In this chapter, we derive and analyse the associated system of ODEs, which is parameterised by λ. Numerical experiments are conducted to help explain how and why the classifier performance drops as λ is increased from 0.5 to 0.99.

By construction, the iterates are confined to the region $\overline{D} : 3 > x_B \geq x_A > -3$. We will prove this in Section 2.4. The equilibria of the associated system of ODEs are contained in $\overline{D}$. We are interested in tracking the temporal evolution of $(x_A(t), x_B(t))$ inside $\overline{D}$, with a view to delineating probabilistic "domains of attraction" of the two equilibria of the system of ODEs, when $\lambda = 0.99$. In other words, we are trying to establish a prediction rule that will tell us which set of initial points $(x_A(0), x_B(0)) \in \overline{D}$ will lead to $(x_A(t), x_B(t))$ converging to one or the other of the two equilibria. To this end, both analytical and numerical techniques

are employed. Finally, we also study pathological sequences of $\{\varphi(t)\}$ that steer $(x_A(t), x_B(t))$ away from equilibria of the associated system of ODEs, in the region $\overline{D}$.

2.2 Computation of the associated system of ODEs

The sought-after system of ODEs is of the form

$$\begin{aligned} \dot{x}_A &= f_A(x_A, x_B, \lambda) \\ \dot{x}_B &= f_B(x_A, x_B, \lambda) \,, \end{aligned} \tag{2.4}$$

where f_A and f_B are defined by

$$\begin{aligned} f_A(x_A, x_B, \lambda) &\stackrel{def}{=} E_\varphi[\varphi - x_A] \,, \quad \varphi \leq c \\ &= \int_{-3}^{c} (\varphi - x_A) f_\lambda(\varphi) d\varphi \\ &= -x_A \int_{-3}^{c} f_\lambda(\varphi) d\varphi + \int_{-3}^{c} \varphi f_\lambda(\varphi) d\varphi \,, \end{aligned} \tag{2.5}$$

and

$$\begin{aligned} f_B(x_A, x_B, \lambda) &\stackrel{def}{=} E_\varphi[\varphi - x_B] \,, \quad \varphi > c \\ &= \int_{c}^{3} (\varphi - x_B) f_\lambda(\varphi) d\varphi \\ &= -x_B \int_{c}^{3} f_\lambda(\varphi) d\varphi + \int_{c}^{3} \varphi f_\lambda(\varphi) d\varphi \,, \end{aligned} \tag{2.6}$$

and where E_φ denotes the statistical expectation with respect to φ.

Then, using (2.1), (2.5), and (2.6), we obtain the ODEs

$$\dot{x}_A = \begin{cases} 0\ ,\ c < -3 \\ \lambda\{\frac{c^3}{3} + (3 - x_A)\frac{c^2}{2} - 3x_A c - \frac{9}{2}(x_A + 1)\}\ ,\ -3 \le c < -2 \\ -\frac{\lambda}{2}\{\frac{7}{3} + x_A\} + \lambda\{-\frac{c^{-3}}{3} + \frac{c^2}{2}(x_A - 1) - \frac{2}{3} + x_A c\}\ ,\ -2 \le c < -1 \\ -2\lambda - \lambda x_A\ ,\ -1 \le c < 1 \\ -2\lambda - \lambda x_A + (1-\lambda)(\frac{c^3}{3} - \frac{c^2}{2} + \frac{1}{6}) - x_A(1-\lambda)(\frac{c^2}{2} - c + \frac{1}{2})\ ,\ 1 \le c < 2 \\ -2\lambda - \lambda x_A + \frac{5}{6}(1-\lambda) - \frac{1}{2}(1-\lambda)x_A + (\lambda - 1)(\frac{c^3}{3} - \frac{3}{2}c^2 + \frac{10}{3}) \\ -x_A(\lambda - 1)(\frac{c^2}{2} - 3c + 4)\ ,\ 2 \le c < 3 \\ -2\lambda - \lambda x_A + \frac{5}{6}(1-\lambda) - x_A(1-\lambda) - \frac{7}{6}(\lambda - 1)\ ,\ c \ge 3 \end{cases} \tag{2.7}$$

and

$$\dot{x}_B = \begin{cases} -2\lambda - \lambda x_B - 2(\lambda - 1) - x_B(1-\lambda)\ ,\ c < -3 \\ \lambda(\frac{10}{3} - \frac{c^3}{3} - \frac{3}{2}c^2) + \lambda x_B(4 + 3c + \frac{c^2}{2}) - \frac{17}{6}\lambda + 2 + \\ \frac{\lambda}{2}x_B - x_B\ ,\ -3 \le c < -2 \\ \lambda(\frac{c^3}{3} + \frac{c^2}{2} - \frac{1}{6}) - \lambda x_B(\frac{c^2}{2} + c + \frac{1}{6}) + 2(1-\lambda) + (\lambda - 1)x_B\ ,\ -2 \le c < -1 \\ (2 - x_B)(1-\lambda)\ ,\ -1 \le c < 1 \\ (1-\lambda)(\frac{2}{3} - \frac{c^3}{3} + \frac{c^2}{2}) - (1-\lambda)x_B(c - \frac{c^2}{2}) - \frac{7}{6}(\lambda - 1) + \\ \frac{1}{2}(\lambda - 1)x_B\ ,\ 1 \le c < 2 \\ (\lambda - 1)(\frac{3}{2}c^2 - \frac{c^3}{3} - \frac{9}{2}) - (\lambda - 1)x_B(3c - \frac{c^2}{2} - \frac{9}{2})\ ,\ 2 \le c < 3 \\ 0\ ,\ c \ge 3 \end{cases} \tag{2.8}$$

From (2.7) and (2.8), we note that f_A and f_B are piecewise-defined continuous func-

tions. The significance of this fact will be apparent in our discussion on existence and uniqueness of solutions of the associated system of ODEs. in Section 2.3. Now, for $-1 \leq c < 1$, equations (2.7) and (2.8) yield the system of ODEs

$$\begin{aligned} \dot{x}_A &= -\lambda(2 + x_A) \\ \dot{x}_B &= (1 - \lambda)(2 - x_B) \,. \end{aligned} \tag{2.9}$$

Solving (2.9) for equilibria gives

$$(x_1^*, x_2^*) = (-2, 2) \,. \tag{2.10}$$

The Jacobian matrix of (2.9), evaluated at the point (x_1^*, x_2^*), is

$$J(x_1^*, x_2^*) = \begin{pmatrix} -\lambda & 0 \\ 0 & -(1-\lambda) \end{pmatrix} ,$$

which implies that (x_1^*, x_2^*) is asymptotically stable since both eigenvalues of $J(x_1^*, x_2^*)$ are strictly negative.

2.3 Stability and bifurcation analysis of the associated system of ODEs

At the onset, we stress that the analysis in this section is valid for any (piecewise-defined) continuous compactly-supported probability density function, not just f_λ.

To understand (2.5) and (2.6), we consider the family of ODEs given by

$$\begin{aligned} \dot{x_A} &= -x_A \int_{-\infty}^{c} f_\lambda(\varphi)d\varphi + \int_{-\infty}^{c} \varphi f_\lambda(\varphi)d\varphi \stackrel{def}{=} f_A(x_A, x_B, \lambda) \\ \dot{x_B} &= -x_B \int_{c}^{\infty} f_\lambda(\varphi)d\varphi + \int_{c}^{\infty} \varphi f_\lambda(\varphi)d\varphi \stackrel{def}{=} f_B(x_A, x_B, \lambda)\ , \end{aligned} \tag{2.11}$$

where $\int_{-\infty}^{\infty} f_\lambda(\varphi)d\varphi = 1$, and $c \stackrel{def}{=} 0.5(x_A + x_B)$. Recall, from Section 2.2. that f_A and f_B are piecewise-defined continuous functions. Furthermore, via the Fundamental Theorem of Calculus, we have that both $f_A(x_A, x_B, \lambda)$ and $f_B(x_A, x_B, \lambda)$ are differentiable with respect to x_A and x_B, viz.

$$\frac{\partial f_A}{\partial x_A} = 0.5(x_B - x_A)f_\lambda(c) - \int_{-\infty}^{c} f_\lambda(\varphi)d\varphi$$

$$\frac{\partial f_A}{\partial x_B} = 0.5(x_B - x_A)f_\lambda(c)$$

$$\frac{\partial f_B}{\partial x_A} = 0.5(x_B - x_A)f_\lambda(c)$$

$$\frac{\partial f_B}{\partial x_B} = 0.5(x_B - x_A)f_\lambda(c) - \int_{c}^{\infty} f_\lambda(\varphi)d\varphi\ .$$

Note that the above partial derivatives are all piecewise-defined continuous functions. Thus, the family of ODEs given in (2.11), with the initial point $(x_A(0), x_B(0))$ specified, has a unique solution existing on some region $J \subset \mathbb{R}^2$ [25]. In addition, since both $f_A(x_A, x_B, \lambda)$ and $f_B(x_A, x_B, \lambda)$ are bounded, it follows that we have global existence of solutions [25].

The family of ODEs in (2.11) has the equilibrium point(s) implicitly given by

$$(\bar{x}_1, \bar{x}_2) = \left(\frac{P(\bar{c})}{Q(\bar{c})}, \frac{2(1-2\lambda) - P(\bar{c})}{1 - Q(\bar{c})} \right) , \tag{2.12}$$

where $\bar{c} \stackrel{def}{=} 0.5(\bar{x}_1 + \bar{x}_2)$, and

$$\begin{aligned} P(c) &\stackrel{def}{=} \int_{-\infty}^{c} \varphi f_\lambda(\varphi) d\varphi \\ Q(c) &\stackrel{def}{=} \int_{-\infty}^{c} f_\lambda(\varphi) d\varphi \, . \end{aligned} \tag{2.13}$$

Linearising (2.11) about $(\bar{x}_1, \bar{x}_2)$ yields the Jacobian matrix

$$J(\bar{x}_1, \bar{x}_2) = \begin{pmatrix} G - Q & G \\ G & G + Q - 1 \end{pmatrix} , \tag{2.14}$$

where

$$G \stackrel{def}{=} \frac{1}{4}(\bar{x}_2 - \bar{x}_1) f_\lambda(\bar{c}) \geq 0 \, ,$$

and Q is as defined in (2.13) and such that $0 < Q < 1$. Observe that, since (2.14) is symmetric, it follows that both its eigenvalues are real. This implies that $(\bar{x}_1, \bar{x}_2)$ can be a saddle point, a node, but not a spiral or centre. From (2.14), we have that

$$tr \; J(\bar{x}_1, \bar{x}_2) = 2G - 1 \, ,$$

and

$$det \; J(\bar{x}_1, \bar{x}_2) = Q(1 - Q) - G \, .$$

The eigenvalues of (2.14) are given by

$$\begin{aligned}\mu &\stackrel{def}{=} \frac{tr \pm \sqrt{tr^2 - 4 \times det}}{2} \\ &= \frac{1}{2}\left\{2G - 1 \pm \sqrt{4(G^2 + Q^2 - Q) + 1}\right\} .\end{aligned}$$

Thus, $(\bar{x}_1, \bar{x}_2)$ is asymptotically stable if and only if

$$tr\ J(\bar{x}_1, \bar{x}_2) < 0 \ \text{ and } \ det\ J(\bar{x}_1, \bar{x}_2) > 0 ,$$

implying that

$$G < \frac{1}{2} \ \text{ and } \ Q(1-Q) > G .$$

Note that, since $0 \leq Q \leq 1$, we have that $0 \leq Q(1-Q) \leq \frac{1}{4}$ (see Fig 2.3).

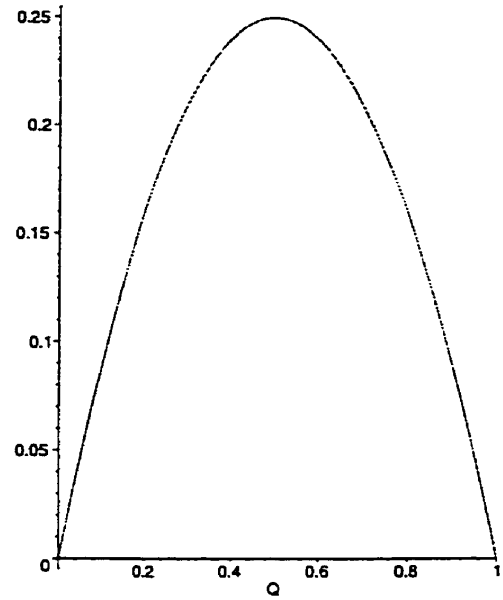

Figure 2.3: $Q(1-Q)$ in the interval $[0, 1]$.

This implies that $(\bar{x}_1, \bar{x}_2)$ is asymptotically stable if and only if $G < Q(1-Q)$, which gives the necessary condition

$$0 < f_\lambda(\bar{c}) \leq (\bar{x}_2 - \bar{x}_1)^{-1} \ . \tag{2.15}$$

If *det* $J(\bar{x}_1, \bar{x}_2) < 0$, then $(\bar{x}_1, \bar{x}_2)$ is a saddle point. Thus, $(\bar{x}_1, \bar{x}_2)$ is a saddle point whenever $Q(1-Q) < G$. If *det* $J(\bar{x}_1, \bar{x}_2) > 0$ and *tr* $J(\bar{x}_1, \bar{x}_2) > 0$, then we have that $Q(1-Q) > G$ and $G > \frac{1}{2}$, which is impossible. Thus, $(\bar{x}_1, \bar{x}_2)$ can not be an unstable node.

In the case *det* $J(\bar{x}_1, \bar{x}_2) < 0$, the stable and unstable eigenspaces are given by

$$E^s \stackrel{def}{=} span\left\{\begin{pmatrix} 1 \\ G^{-1}(\mu_1 - G + Q) \end{pmatrix}\right\}$$

and

$$E^u \stackrel{def}{=} span\left\{\begin{pmatrix} 1 \\ G^{-1}(\mu_2 - G + Q) \end{pmatrix}\right\} \ ,$$

respectively, where

$$\begin{aligned} \mu_1 &\stackrel{def}{=} \frac{1}{2}\left\{2G - 1 - \sqrt{4(G^2 + Q^2 - Q) + 1}\right\} \\ \mu_2 &\stackrel{def}{=} \frac{1}{2}\left\{2G - 1 + \sqrt{4(G^2 + Q^2 - Q) + 1}\right\} \ . \end{aligned}$$

Remark 2.1 *Note that* $Q(1-Q) = 0$ *if* $Q = 0$ *or* $Q = 1$, *which, for* f_λ *given by (2.1), corresponds to* $\bar{c} = -3$ *or* $\bar{c} = 3$, *respectively. The equilibria in these cases are, respectively,* $(\bar{x}_1, \bar{x}_2) = (k_0,\ 2(1-2\lambda))$ *and* $(\bar{x}_1, \bar{x}_2) = (2(1-2\lambda),\ k_1)$, *where the*

values of k_0, $k_1 \in \mathbb{R}$ *are forced by* $0.5(\bar{x}_1 + \bar{x}_2) = \bar{c}$. *The above ODE equilibria are both neutrally stable. Note that these two cases (i.e.* $Q = 0$ *and* $Q = 1$*) correspond to a situation when the classifier fails to distinguish the two clusters* A *and* B.

Recall that the Kushner-Clark theorem [1] shows that, under some conditions, if the associated system of ODEs possesses only *one* locally asymptotically stable equilibrium, then the discrete system will converge wp1 to that equilibrium. However, the theorem does not tell us anything about what happens if the system of ODEs has multiple locally asymptotically stable equilibria. Furthermore, recall that Ljung [2] showed that if λ is increased from 0.5 to 0.99, the number of stable equilibria appears to increase from one to two. Below, we determine how the number and stability of the equilibria vary with λ and briefly discuss their basins of attraction.

To investigate the behaviour of the classifier as λ is varied, we proceed as follows. From (2.12), we have that

$$\bar{c} \stackrel{def}{=} \frac{1}{2}\left\{\frac{P(\bar{c})}{Q(\bar{c})} + \frac{2(1-2\lambda) - P(\bar{c})}{1 - Q(\bar{c})}\right\} .$$

Define

$$\begin{aligned} H_\lambda(c) &\stackrel{def}{=} \left\{\frac{P(c)}{Q(c)} + \frac{2(1-2\lambda) - P(c)}{1 - Q(c)}\right\} - 2c \\ &= \frac{P(c)[1 - Q(c)] + Q(c)[2(1-2\lambda) - P(c)] - 2cQ(c)[1 - Q(c)]}{Q(c)[1 - Q(c)]} , \end{aligned} \tag{2.16}$$

where $P(\cdot)$ and $Q(\cdot)$ are defined in (2.13). The equilibria of (2.11) may be computed

by finding the zeros $\bar{c}$ of the numerator of (2.16)

$$F_\lambda(c) \stackrel{def}{=} P(c)[1-Q(c)] + [2(1-2\lambda) - P(c)]Q(c) - 2cQ(c)[1-Q(c)]$$

and then applying (2.12) to obtain

$$\bar{x}_1 = \frac{\int_{-\infty}^{\bar{c}} \varphi f_\lambda(\varphi) d\varphi}{\int_{-\infty}^{\bar{c}} f_\lambda(\varphi) d\varphi} \quad , \quad \bar{x}_2 = \frac{\int_{\bar{c}}^{\infty} \varphi f_\lambda(\varphi) d\varphi}{\int_{\bar{c}}^{\infty} f_\lambda(\varphi) d\varphi} \, . \tag{2.17}$$

Some computer-generated plots of $F_\lambda(c)$ versus c, for various parameter values are shown in Figs 2.4-2.5 (*left*). Note that the roots of $F_\lambda(c)$ are the $\bar{c}$ values associated with $(\bar{x}_1, \bar{x}_2)$.

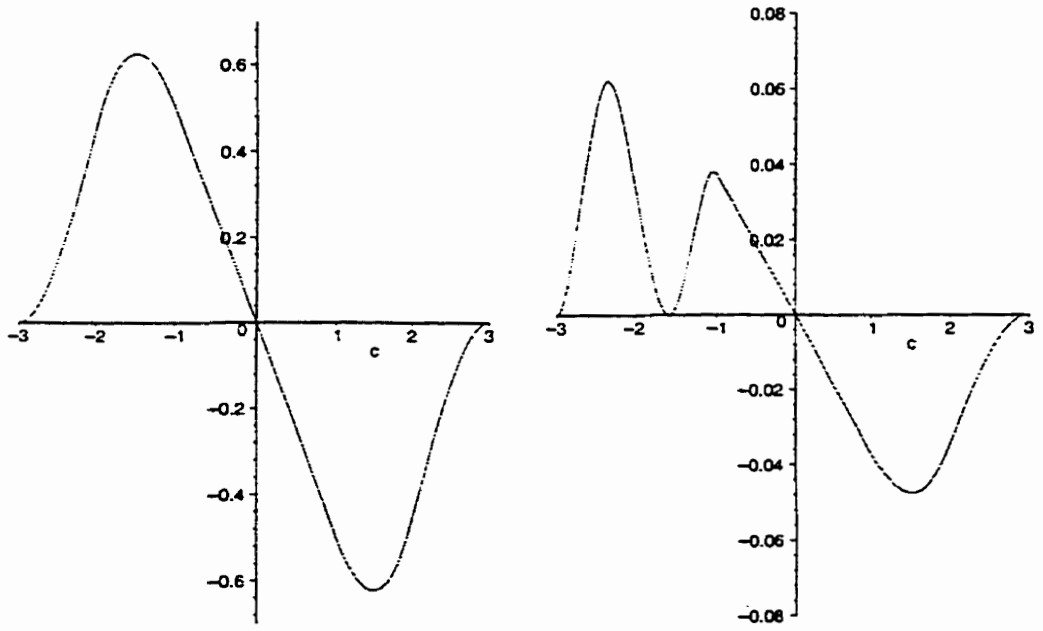

Figure 2.4: $F_\lambda(c)$ with *left*: $\lambda = 0.5$, and *right*: $\lambda = 0.981$

The plots show that the qualitative behaviour of the zeros of $F_\lambda(c)$ is dependent on the parameter λ. When $\lambda = 0.5$, $F_\lambda(c)$ has only one zero, at $\bar{c}_1 = 0$. This

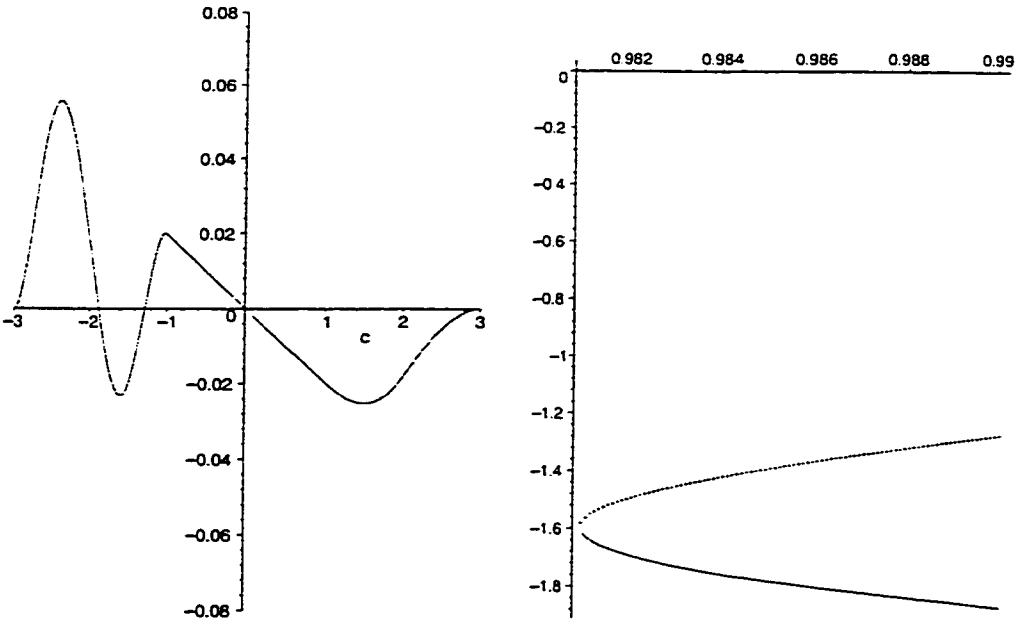

Figure 2.5: *left*: $F_\lambda(c)$ with $\lambda = 0.99$ *right*: Bifurcation of zeros of $F_\lambda(c)$

zero *always* exists, and never loses its stability as λ varies. Numerically, at $\lambda \approx 0.981$, a saddle-node (S-N for short) bifurcation occurs. In other words, a new zero appears, at $\bar{c}_0 \approx -1.600$. As λ increases, this zero branches to yield two new zeros. This bifurcation phenomenon is illustrated in Fig 2.5 (*right*), which is generated as follows: Solve the equation $F_\lambda(c) = 0$ for $c(\lambda)$, where $\lambda \in [0, 1]$. Then plot $c(\lambda)$ versus λ. In Fig 2.5 (*right*), we have only considered a small 'window' of λ values, essentially to accentuate the emergence of the two new zeros.

At $\lambda = 0.99$, we have the two zeros $\{\bar{c}_2, \bar{c}_3\} = \{-1.874, -1.277\}$ as depicted in Fig 2.5 (*left*). We now have three roots of $F_\lambda(c)$, namely: $\{\bar{c}_2,\ \bar{c}_3,\ \bar{c}_1\} = \{-1.874,\ -1.277,\ 0\}$. From (2.17), the corresponding equilibria are $(\bar{x}_{11}, \bar{x}_{12}) = (-2.258,\ -1.372)$, $(\bar{x}_{21}, \bar{x}_{22}) = (-2.033,\ -1.075)$, and $(\bar{x}_{31}, \bar{x}_{32}) = (-2,\ 2)$ respectively. It can be shown, via the computation of respective Jacobian matri-

ces, that $(\bar{x}_{11}, \bar{x}_{12})$ and $(\bar{x}_{31}, \bar{x}_{32})$ are locally asymptotically stable, with eigenvalues $\{-0.5618, -0.0802\}$ and $\{-0.9900, -0.0100\}$ respectively, while $(\bar{x}_{21}, \bar{x}_{22})$ is a saddle point.

We now turn our attention to the problem of delineating the boundary between the basins of attraction of $(\bar{x}_{11}, \bar{x}_{12})$ and $(\bar{x}_{31}, \bar{x}_{32})$, inside the region $\overline{D}$. Note that this boundary is a segment of the stable eigenspace of $(\bar{x}_{21}, \bar{x}_{22})$. From (2.14), we have that

$$J(\bar{x}_{21}, \bar{x}_{22}) = \begin{pmatrix} -0.886 & 0.066 \\ 0.066 & 0.018 \end{pmatrix}, \tag{2.18}$$

with eigenvalues -0.8910 and 0.0224. Corresponding eigenvectors are

$$\begin{aligned} \mathbf{v_1} &= [1 \quad -0.072]^T, \text{ and} \\ \mathbf{v_2} &= [1 \quad 13.834]^T \end{aligned} \tag{2.19}$$

respectively. Therefore, the angles (measured relative to the x_1-axis) associated with $\mathbf{v_1}$ and $\mathbf{v_2}$ are $\alpha_1 = -1.50\ rad$ and $\alpha_2 = 0.07\ rad$, in that order. Thus, the stable eigenspace is given by

$$E^s = span\left\{\begin{pmatrix} 1 \\ -0.072 \end{pmatrix}\right\},$$

while the unstable eigenspace is

$$E^u = span\left\{\begin{pmatrix} 1 \\ 13.834 \end{pmatrix}\right\}.$$

The equation of the stable eigenspace of the saddle point is given by

$$x_2 = -0.0723x_1 - 1.222\ . \tag{2.20}$$

To reconstruct the stable eigenspace, we pick an arbitrary initial point in $\overline{D}$, as close as possible to $(\bar{x}_{21}, \bar{x}_{22})$ and lying on the line given by (2.20). Then, integrate, backwards in time, the corresponding system of ODEs.

Note that, in the invariant region $\overline{D}$, the saddle point $(\bar{x}_{21}, \bar{x}_{22})$ lies in the interval $-2 < c < -1$. Thus, to approximate the stable eigenspace, we need only to consider the following system of ODEs, which is obtained from (2.7) and (2.8),

$$\begin{aligned} \dot{x}_A &= -1.815 - 0.495x_A - 0.33c^3 - 0.495c^2 + x_A(0.495c^2 + 0.99c) \\ \dot{x}_B &= -0.145 + 0.33c^3 + 0.495c^2 - x_B(0.495 + 0.495c^2 + 0.99c) - 0.01x_B\ , \end{aligned} \tag{2.21}$$

where $c \stackrel{def}{=} 0.5(x_A + x_B)$. Now, pick two initial points $(-2.036,\ -1.0748)$ and $(-2.030,\ -1.0750)$, lying to the left and to the right of $(\bar{x}_{21},\ \bar{x}_{22})$ on the line given by (2.20), respectively. Using these initial points, find a particular solution of (2.21) for $t < 0$. This is then plotted in the (x_A, x_B) phase space. The resulting plot gives the boundary between the domains of attraction of $(\bar{x}_{11},\ \bar{x}_{12})$ and $(\bar{x}_{31},\ \bar{x}_{32})$, which is depicted in Fig 2.6. The latter indicates that $(\bar{x}_{31}, \bar{x}_{32})$ has a relatively larger domain of attraction than $(\bar{x}_{11}, \bar{x}_{12})$. We note that the boundary shown in Fig 2.6 is valid only inside the triangular region $\overline{D}$. The part of this boundary lying outside of $\overline{D}$ is a numerical artefact, resulting from using the MAPLE *view* plot function. Essentially, one can only *view* objects inside a box, $[-3, 3]^2$ in our case. MAPLE is

unable to *view* objects inside a triangular region, for example.

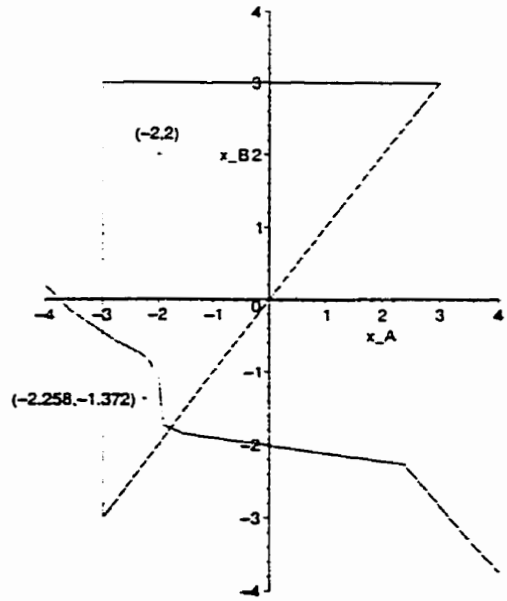

Figure 2.6: The domains of attraction of $(\bar{x}_{11}, \bar{x}_{12}) = (-2.258, -1.372)$ and $(\bar{x}_{31}, \bar{x}_{32}) = (-2, 2)$, inside the invariant region $\overline{D}$ in ODE space

Recall that, by the Kushner-Clark theorem, if the associated system of ODEs has a single locally asymptotically stable equilibrium z_0 with domain of attraction $DA(z_0)$, and for which there exists a compact set $A \subset DA(z_0)$ such that $\mathbf{x}_n \in A$ infinitely often, then wp1 $\mathbf{x}_n \to z_0$ as $n \to \infty$. In the current problem ($\lambda = 0.99$), the associated system of ODEs has *two* locally asymptotically stable equilibria, and a saddle point, as shown previously. The issue at hand is the following: *to which one of these stable equilibria is* $\{(x_A(n), x_B(n))\}$ *likely to converge, as* $n \to \infty$, *and how will this be affected by the choice of initial conditions?* Of course, the same question may be asked if the associated system of ODEs possesses more than two stable equilibria. To rephrase the problem, we seek to develop a feeling for

the probability that the discrete system will converge to the undesired classifier $c^* = -1.815$, given that we choose (according to some rule) our initial conditions inside $\overline{D}$.

It is interesting to note that, in Fig 4 of Ljung's [2] paper, even though his simulations always start in the ODE basin of attraction of $(-2, 2)$ (i.e. $c(0) = 0$), the presence of noise causes the discrete system to go to the other equilibrium, i.e. $(-2.258, -1.372)$. Finally, having delineated the boundary between the equilibria $(\bar{x}_{11}, \bar{x}_{12})$ and $(\bar{x}_{21}, \bar{x}_{22})$ in the ODE phase space, we now turn our attention to investigating how dynamics in this (continuous) space relate to those in the discrete space. To this end, we begin by constructing a probabilistic vector field for the discrete algorithm.

2.4 Construction and analysis of a probabilistic vector field for the discrete algorithm

We begin by recalling that, according to Ljung's classification scheme, $\varphi \in A$ iff $\varphi \leq c$ and $\varphi \in B$ iff $\varphi > c$, where $c \stackrel{def}{=} 0.5(x_A + x_B)$. Furthermore, x_A and x_B are updated depending on whether φ is classified as belonging to A or B. That is

$$\begin{aligned} \varphi \leq c \;\; &\Rightarrow \;\; x_A = x_A + \gamma(t)(\varphi - x_A) \text{ , and} \\ \varphi > c \;\; &\Rightarrow \;\; x_B = x_B + \gamma(t)(\varphi - x_B) \text{ .} \end{aligned} \tag{2.22}$$

Note that only one of either x_A or x_B can be updated at any particular time.

Now, from the definition of c, we have that $x_A < c < x_B$. Therefore

$$\begin{aligned} \varphi \leq c \;\; &\Rightarrow \;\; x_A < \varphi < c \text{ or } \varphi < x_A < c\,, \text{ and} \\ \varphi > c \;\; &\Rightarrow \;\; c < \varphi < x_B \text{ or } \varphi > x_B > c\,. \end{aligned} \tag{2.23}$$

Following (2.22) and (2.23) above, we see that x_A and x_B may be updated *positively* or *negatively*. For example, x_A is updated positively if $\varphi \leq c$ and $\varphi - x_A > 0$, and negatively if $\varphi \leq c$ and $\varphi - x_A < 0$. Based on this observation, we shall construct a probabilistic vector field of the discrete algorithm. We envisage that this vector field will shed some light on the difficult problem of determining the long term behaviour of the algorithm in the case when the associated system of ODEs possesses multiple asymptotically stable equilibria. Consider Ljung's doubly-triangular pdf with $\lambda = 0.99$. In this case, the associated system of ODEs has two locally asymptotically stable equilibria, namely: $(-2, 2)$ and $(-2.258, -1.372)$, and a saddle point at $(-2.033, -1.075)$. To construct the associated vector field, we proceed as follows. Suppose that $-2 < x_A < -1$ and $x_B = 0$. Then we have that $-1 < c < -0.5$. Fig 2.7 (*left*) gives a summary of this scenario.

From Fig 2.7 (*left*), if $\varphi < x_A$, the shaded area to the left of x_A represents the probability that x_A is decreased. The unshaded area lying between x_A and -1 gives the probability that x_A is increased. Note that for $\varphi > x_A$, it may happen that $1 < \varphi < 3$. This yields $\varphi > x_B = 0 > c$, implying that x_B would be updated positively. Similarly, we may consider the following cases, for example: $-1 < x_A < 0$ and $x_B = 0$, $-3 < x_A < -2$ and $x_B = 0$, $0 < x_B < 1$ and $x_A = 0$, $1 < x_B < 2$ and $x_A = 0$, $2 < x_B < 3$ and $x_A = 0$, $x_B = 3$ and $x_A = 0$,

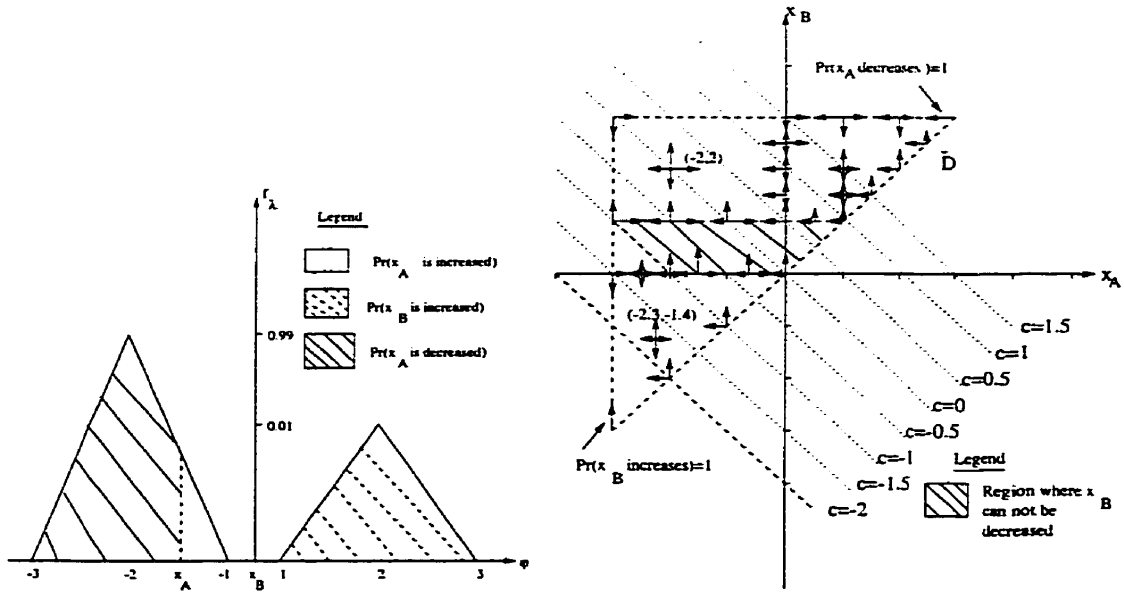

Figure 2.7: *left*: Computing the probabilities that x_A and x_B are decreased or increased, for the case $-2 < x_A < -1$ and $x_B = 0$ *right*: Sketch of the probabilistic discrete vector field with $\lambda = 0.99$. Arrows indicate possible jump directions. The length of each arrow indicates the probability of going in a particular direction.

$x_B = 3$ and $x_A = -3$, $x_A = -3$ and $x_B = 0$, $x_A = -2$ and $x_B = 2$, $x_A = -2.258$ and $x_B = -1.372$. Based on this analysis, we obtain the probabilistic vector field sketch shown in Fig 2.7 (*right*). It is important to note that this is a *time- and place-dependent random vector field with dissipation*, such that, conditioned on the immediately previous position, the current position is independent of all earlier positions. In addition, the magnitude of jumps in this random field diminishes with time. However, this time dependence of jumps is not shown in Fig 2.7 (*right*). Instead, the arrows in the above figure illustrate only the probabilities of going in any one of the four directions at a given point in the discrete space.

Furthermore, Fig 2.7 (*right*) shows that starting on the ODE's stable equilibria

does not guarantee that the algorithm will remain at these equilibria for all time. There is a non-zero probability that $(x_A(t), x_B(t))$ will be knocked off the equilibria. even on the first iteration of the algorithm. The arrows and their magnitudes (not drawn to scale here) represent the probability that the algorithm will go in a particular direction at any point in $\overline{D}$. Also included in the diagram are various c values. Note that, because x_B may be decreased for $-3 < x_A < -2$, any initial point of the algorithm near $(-2, 2)$ may lead to $c(t)$ migrating towards the neighbourhood of $(-2.258, -1.372)$. This observation leads us to the conclusion that the point $(-2, 2)$ is *not* an asymptotically stable equilibrium of the algorithm. In fact, it is not even an equilibrium. Note also that there is no positively invariant region contained *inside* the discrete region $\overline{D}$. The idea of domain of attraction of an equilibrium (of the algorithm) is meaningless in this problem, chiefly as a result of the random nature of the input signals $\{\varphi(t)\}$.

The foregoing discussion leads to the following question: *In discrete space, is $\overline{D}$ positively invariant?* To resolve this question, we need to determine whether the statement $-3 < x_A(t) < c(t) < x_B(t) < 3$ is true for all $t \in \mathbb{Z}^+$. First, recall that $c(t) \stackrel{def}{=} 0.5 \times (x_A(t) + x_B(t))$. Then, we proceed as follows: By construction, we have that $-3 < x_A(1) < c(1) < x_B(1) < 3$. Now suppose that the statement is true for $t = k$, viz. $-3 < x_A(k) < c(k) < x_B(k) < 3$, where $k \in \mathbb{Z}^+$. Consider the case $x_B(t) < \varphi(t) < 3$, and suppose that $\varphi(k) - x_B(k) = \epsilon(k) > 0$, where $\epsilon(k)$ is a small positive jump. We conclude that $x_B(k)$ is increased (i.e. $x_B(k) \leq x_B(k+1)$), while $x_A(k)$ remains intact. In other words,

$$x_B(k+1) \;=\; x_B(k) + \gamma(k)\epsilon(k)$$

$$\begin{aligned} &\leq x_B(k) + \epsilon(k) \\ &= \varphi(k) \\ &< 3 , \end{aligned}$$

and

$$x_A(k+1) = x_A(k) < x_B(k) .$$

This implies that $-3 < x_A(k+1) < c(k+1) < x_B(k+1) < 3$. That is, the statement is true for $t = k+1$.

Next, consider the case $c(t) < \varphi(t) < x_B(t)$. Then, if we suppose that $\varphi(k) - x_B(k) = \epsilon(k) < 0$, where $\epsilon(k)$ is a small negative jump, we have that $x_B(k)$ is decreased while $x_A(k)$ is unaltered. That is,

$$\begin{aligned} x_B(k+1) &= x_B(k) + \gamma(k)\epsilon(k) \\ &\geq x_B(k) + \epsilon(k) \\ &= \varphi(k) \\ &> c(k) , \end{aligned}$$

and

$$x_A(k+1) = x_A(k) .$$

Once again, this yields $-3 < x_A(k+1) = x_A(k) < c(k+1) < x_B(k+1) < 3$, where the rightmost inequality holds because $x_B(k) < 3$ is decreased. The rest of the proof proceeds in exactly the same way as in the previous case.

The remaining cases, namely $x_A(t) < \varphi(t) < c(t)$ and $-3 < \varphi(t) < x_A(t)$, may be treated in a similar way. Therefore, by mathematical induction, the statement $-3 < x_A(t) < c(t) < x_B(t) < 3$ is true for $t = 1$, $t = 2$,...., viz. it is true for all $t \in \mathbb{Z}^+$. We conclude that $\overline{D}$ is positively invariant in discrete space.

As mentioned before, each point in discrete space is characterised by a set of direction vectors, whose lengths give the probability of being 'nudged' in that direction. Also, the sum of all these lengths, at the point under consideration, is always unity. On the other hand, the ODE vector field consists of single arrows at each point on the continuous time phase space. These arrows represent the average paths of the discrete (random) vector field, since the associated system of ODEs (2.11) is constructed by averaging (i.e. taking the expected values with respect to φ) the random terms $\varphi - x_A$ and $\varphi - x_B$. Fig 2.8 shows the expected value vector field and the direction field of the associated system of ODEs. These plots are only valid inside $\overline{D}$. We now give a brief outline of how these two vector fields were constructed.

Recall that, by construction, $-3 < x_A(t) < c(t) < x_B(t) < 3$ for all $t \in \mathbb{Z}^+$. In the expected value vector field, each point (x_A, x_B) in the discrete space $\overline{D}$ is characterised by four arrows pointing up, down, right and left, giving the expected value of a jump in that direction. Let the expected value of a jump to the right of x_A be denoted by $J_r^{x_A}$, and that to its left by $J_l^{x_A}$. Similarly, let the expected value of a jump upwards of x_B be denoted by $J_u^{x_B}$, and that in the downward direction

by $J_d^{x_B}$. Defining c by $c \stackrel{def}{=} \frac{1}{2}(x_A + x_B)$, we have the following

$$J_r^{x_A}(x_A, x_B, \lambda) \stackrel{def}{=} \int_{x_A}^{c} (\varphi - x_A) f_\lambda(\varphi) d\varphi \ , \text{ since } x_A < \varphi < c \ ,$$

$$J_l^{x_A}(x_A, x_B, \lambda) \stackrel{def}{=} \int_{-3}^{x_A} (\varphi - x_A) f_\lambda(\varphi) d\varphi \ , \text{ since } -3 < \varphi < x_A \ ,$$

$$J_u^{x_B}(x_A, x_B, \lambda) \stackrel{def}{=} \int_{x_B}^{3} (\varphi - x_B) f_\lambda(\varphi) d\varphi \ , \text{ since } x_B < \varphi < 3 \ ,$$

and

$$J_d^{x_B}(x_A, x_B, \lambda) \stackrel{def}{=} \int_{c}^{x_B} (\varphi - x_B) f_\lambda(\varphi) d\varphi \ , \text{ since } c < \varphi < x_B \ .$$

Finally, setting $\lambda = 0.99$ in the above formulae and using the MAPLE *fieldplot* routine, we plot and display $J_r^{x_A}$, $J_l^{x_A}$, $J_u^{x_B}$, and $J_d^{x_B}$ in a single diagram, viz. Fig. 2.8 (*right*). The ODE direction field, as displayed in Fig. 2.8 (*left*), was constructed in the following way. Define the vector field of the system of ODEs in (2.11) by

$$\begin{aligned} \dot{x}_A = f_1(x_A, x_B, \lambda) &\stackrel{def}{=} \int_{-\infty}^{c} (\varphi - x_A) f_\lambda(\varphi) d\varphi \ , \\ \dot{x}_B = f_2(x_A, x_B, \lambda) &\stackrel{def}{=} \int_{c}^{\infty} (\varphi - x_B) f_\lambda(\varphi) d\varphi \ . \end{aligned}$$

Then put the above system of equations into the MAPLE *dfieldplot* routine, with $\lambda = 0.99$ and for $-3 < x_A < 3$, $-3 < x_B < 3$. As mentioned previously, the arrows in the direction field of the ODEs represent the *average* paths of the discrete vector field.

The existence of the two locally asymptotically stable equilibria of (2.11), viz. $(-2, 2)$ and $(-2.258, -1.372)$, is evident in both Figs. 2.8 (*left*) and (*right*). In

Fig. 2.8 (*right*). there is some indication that the scale of arrows is smallest around these stable points. In other words, the expected values of the various jumps are minimal (*but still non-zero*) in the regions immediately surrounding $(-2, 2)$ and $(-2.258, -1.372)$. This is true for Fig 2.8 (*left*) as well. Note that both Figs. 2.8 (*left*) and (*right*) will change with t as $\gamma(t)$ does.

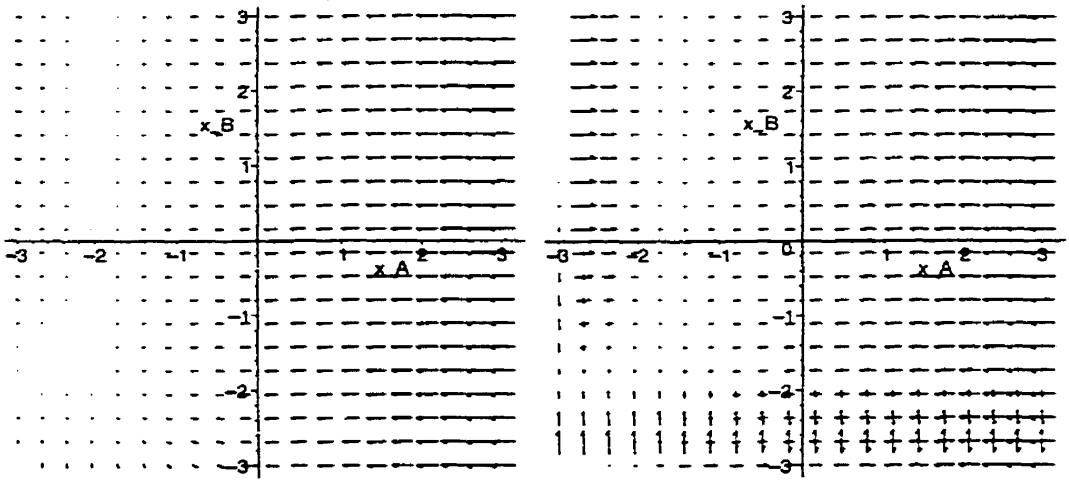

Figure 2.8: *left*: Direction field for the ODE associated with Ljung's algorithm with $\lambda = 0.99$ *right*: Expected value vector field (discrete) for Ljung's algorithm with $\lambda = 0.99$. Both figures reveal the existence of two stable points at $(-2, 2)$ and $(-2.258, -1.372)$.

In the next section, we explore the possibility of characterising and delineating probabilistic "domains of attraction" of (-2,2) and (-2.258,-1.372) inside $\overline{D}$. The term "domain of attraction" is used loosely in this context, for reasons cited previously.

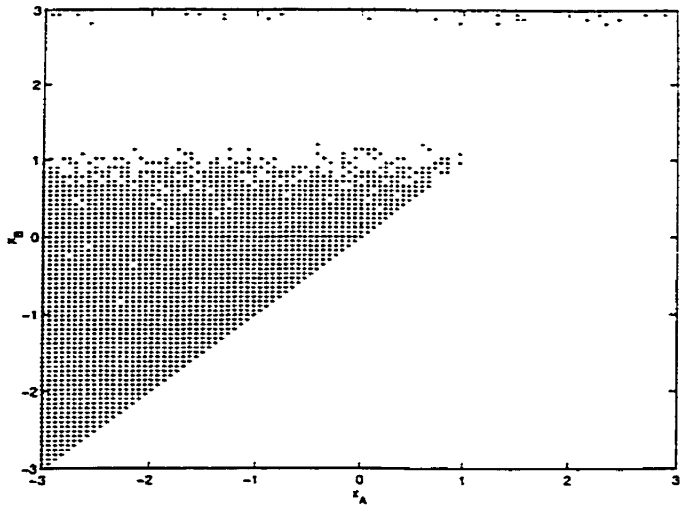

Figure 2.9: "Domain of attraction" of $(-2, 2)$ in discrete space with $\lambda = 0.99$ and $\gamma_n = \frac{0.5}{\sqrt{n+1}}$, as discussed in the text

2.5 A crude experiment to map probabilistic "domains of attraction" for the case $\lambda = 0.99$

The motivation for this numerical exploration arises from the question of whether the choice and location of initial conditions inside $\overline{D}$ has any influence on the subsequent convergence of the discrete algorithm. For $\lambda = 0.99$, we recall that the associated system of ODEs possesses two locally asymptotically stable equilibria. Naturally, it is interesting to ask which one of these two points the algorithm is most likely to converge to, with initial conditions picked inside $\overline{D}$.

In what follows, a brief descriptive outline of the experiment is given. The region $\overline{D}$ is evenly discretised into a 100×100 grid, with each grid point then being used to initialise the algorithm. Subsequently, we run the algorithm N times and compute $(x_A(N), x_B(N))$ each time. Each time, we compute the distance between

$(x_A(N), x_B(N))$ and $(-2, 2)$. If this distance is less than some specified $\delta > 0$. we record the grid point used to initialise the algorithm. Finally, we display all grid points that lead to convergence of the algorithm to within a distance of δ.

A typical result of this experiment is displayed in Fig 2.9. The clear region denotes those grid points that, as initial conditions, yield sequences $\{(x_A(n), x_B(n))\}$ that converge to $(-2, 2)$ to within a distance of $\delta = 0.4$ in $N = 10000$ iterations. The dots denote those points that fail this test. We note that there is a "fuzzy" boundary between the two classes of grid points at roughly $x_B = 1$. It is also worth noting that there are some points, near $x_B = 3$, that do not converge to $(-2, 2)$ to within the stipulated distance. It is suspected that these will disappear as the number of iterations is increased. Fig 2.9 is a rough sketch of the probabilistic "domain of attraction" of the point $(-2, 2)$. Note that we can *not* conclude that those points which did not converge to within the stipulated distance of $(-2, 2)$ converge to $(-2.258, -1.372)$. We are using the term "domain of attraction" with much trepidation, since it does not carry the same conventional meaning as used in the study of deterministic dynamical systems.

It is interesting to note that it is possible that the algorithm may *freeze* at some pseudo-equilibrium point. This all depends on the choice of the learning parameter, γ_n. This problem is explained in the next section.

2.6 Explaining the problem of *freezing* of an algorithm at a pseudo-equilibrium point

Recall that the learning parameter γ_n obeys the constraints: $\gamma_n \to 0$ as $n \to \infty$, and $\sum_{n=0}^{\infty} \gamma_n = \infty$ [1]. Clearly, there is a whole range of γ_n's satisfying these constraints. A central problem in the convergence of neural learning algorithms is the selection of an *appropriate* learning parameter, γ_n. It turns out that if $\gamma_n \to 0$ *too quickly*, then there arises a danger of the algorithm getting frozen at some pseudo-equilibrium point, as $n \to \infty$. The other key problem concerns numerical round-off errors. To avoid these problems, care must be taken choosing a learning parameter that is decaying *neither too quickly nor too slowly.* Note that if the learning parameter converges *too slowly*, then the algorithm converges slowly as $n \to \infty$, resulting in the accumulation of numerical round-off errors over time. In theory, the choice of γ_n, with the above problems in mind, should ensure that the algorithm is not trapped at pseudo-equilibria and that round-off errors are minimal. However, in practice, striking this balance is a remarkably difficult problem.

To motivate and demonstrate the problem of freezing of an algorithm at a pseudo-equilibrium point, we consider the non-autonomous deterministic system

$$x_{n+1} = (1 - \gamma_n)x_n \stackrel{def}{=} f(x_n, n) ,$$

where γ_n is the usual learning parameter. Note that $x^* = 0$ appears to be a stable fixed point of this system, since $\mid f'(0, n) \mid = 1 - \gamma_n < 1$ for all n. The iterates x_n

may be expressed in closed form as

$$\begin{aligned} x_n &= x_0 \prod_{k=0}^{n-1} (1-\gamma_k) \\ &= x_0 e^{\sum_{k=0}^{n-1} \ln(1-\gamma_k)} \\ &\sim x_0 e^{-\sum_{k=0}^{n-1} \gamma_k} . \end{aligned}$$

Depending on how quickly γ_n goes to zero, two cases arise. These are discussed below.

Case 1

If $\gamma_k \to 0$ slowly enough (e.g. $\gamma_k = \frac{1}{k+1}$), then $\sum_{k=0}^{n-1} \gamma_k \to \infty$, as $n \to \infty$. Thus, we have that $x_n \to 0$ as $n \to \infty$. This behaviour may be observed for any $\gamma_k = \frac{1}{(k+1)^p}$, where $p \in [0, 1]$. To illustrate this case, assume that $\gamma_k = \frac{1}{k+1}$. Then from above, we have that

$$\begin{aligned} x_n &\sim x_0 e^{-\sum_{k=0}^{n-1} \frac{1}{k+1}} \\ &= x_0 e^{-\sum_{k=0}^{n-1} \frac{1}{(\frac{k+1}{n})} \times \frac{1}{n}} \\ &\sim x_0 e^{-\int_{\frac{1}{n}}^{1} \frac{1}{t} dt} \\ &= \frac{x_0}{n} \to 0 \text{ as } n \to \infty . \end{aligned}$$

which shows that x_n approaches 0 relatively slowly (not exponentially).

Case 2

If $\gamma_k \to 0$ too quickly (e.g. $\gamma_k = \frac{1}{(k+1)^2}$), then $\sum_{k=0}^{n-1} \gamma_k \to \Gamma \neq \infty$ (where Γ is a constant). Then, we have that $x_n \to x_0 e^{-\Gamma} \neq 0$. In this case, the sequence $\{x_n\}$ has

frozen at a pseudo-equilibrium point. This occurs for any $\gamma_k = \frac{1}{(k+1)^p}$ with $p > 1$.

To reinforce the meaning of the phenomenon of *freezing*, consider the continuous time ODE

$$\frac{dx}{dt} = a(t)x(t) \,, \tag{2.24}$$

where $a(t)$ is a real, time-varying, and decreasing-to-zero parameter that is analogous to γ_n in the discrete domain. The point $x^* = 0$ is an equilibrium of (2.24). Then, the issue is the following: *Is it feasible that if* $a(t) \to 0$ *fast enough, then* $x(t)$ *does not approach* 0 *as* $t \to \infty$*?* The answer to this is *YES*, as the examples below illustrate.

Example 1

Suppose that $a(t) = -k$, where k is a positive real constant. Then

$$x(t) = x_0 e^{-kt} \to 0 \text{ as } t \to \infty \,. \tag{2.25}$$

Example 2

Suppose that $a(t) = -\frac{1}{t}$. Then we have that

$$x(t) = x_0 e^{\left(\int_{t_0}^{t} a(s)ds\right)} = \frac{x_0 t_0}{t} \to 0 \text{ as } t \to \infty \,. \tag{2.26}$$

Example 3

Suppose $a(t) = -\frac{1}{t^2}$. Then we get

$$x(t) = x_0 e^{\left(-\frac{1}{t_0}+\frac{1}{t}\right)} \to x_0 e^{-\frac{1}{t_0}} \neq 0 \,. \tag{2.27}$$

Remark 2.2 *In example 3, we have exhibited an $a(t)$ such that $x(t)$ does not approach 0 as $t \to \infty$. Instead, in this example, $x(t)$ 'converges' to a pseudo-equilibrium point of the ODE. A similar situation would occur for any $a(t) = -\frac{1}{t^p}$ with $p > 1$. Finally, we remark that example 3 highlights the significance of judiciously choosing $a(t)$, and hence the learning parameter in the discrete domain, so that the solution of the ODE (algorithm) converges to a true equilibrium point.*

2.7 Evolution of Dirac probability distributions under Ljung's algorithm, with a doubly triangular pdf for the input signals

In this section, we perform a numerical study to determine how an initial condition with unit probability temporally evolves, using Ljung's automatic classifier. Recall that our phase space is the upper triangular region $\overline{D} : -3 \leq x_A \leq x_B \leq 3$. The algorithm employed to perform this investigation is outlined below.

- Divide $[-3,3]^2$ into a grid of $npts \times npts$ cells, where typically, $npts = 100$. Each cell is of size $del = \frac{6}{npts}$ and has a midpoint (x_A, x_B) which is considered to hold the mass of the cell. Phase space is the upper triangular region $\overline{D}$.

- Initially, assign unit mass to a point $(x_A(0), x_B(0))$, and a mass of zero to the rest of the points. Store this "information" in a matrix P.

- Determine where the point (x_A, x_B) could go after a single iteration as follows. Note that one must know the current value of γ_n, the training parameter.

Where (x_A, x_B) goes depends upon the φ value that is picked (according to the doubly-triangular pdf). Since either x_A or x_B can be changed, one must examine all grid cells that lie in the same horizontal and vertical lines as (x_A, x_B).

- Suppose that we consider x_A as getting updated. Pick a grid cell and determine its left and right endpoints, say x_1, x_2. The following issue arises: *For what values of φ does x_A get mapped to x_1, x_2?* Solve for these φ values, and denote them by φ_1, φ_2 respectively. Note that one must check the feasibility of these φ values, as a result of the 'geometry' of the doubly-triangular pdf. The φ values are computed as follows:

$$x_1 = x_A(n) + \gamma_n\{\varphi_1 - x_A(n)\} ,$$

which yields

$$\varphi_1 = \frac{\{x_1 - x_A(n)\}}{\gamma_n} + x_A(n) . \tag{2.28}$$

Similarly, we get

$$\varphi_2 = \frac{\{x_2 - x_A(n)\}}{\gamma_n} + x_A(n) . \tag{2.29}$$

Then, the probability of choosing this range of φ values is given by

$$prob = \int_{\varphi_1}^{\varphi_2} f_\lambda(\varphi) d\varphi , \tag{2.30}$$

where f_λ is Ljung's doubly-triangular pdf. As mentioned above, it is crucial to note that the values φ_1 and φ_2 given by (2.28) and (2.29) may not be

feasible. In other words, they may not lie in the range $[-3, 3]$ nor may they lie in the range $[-3, c]$ responsible for updating x_A. This must be checked in each case, and appropriately remedied. Also, one must check that the total sum of the probabilities of transferring to the various horizontal and vertical cells is unity.

- Store the updated mass in a matrix PP, which is initialised with zero entries. Iterate PP according to

$$PP[i, j] = PP[i, j] + prob \times P[k, l] ,$$

where i, j are the indices of the point being mapped to and k, l are the indices of the point being mapped from.

- Set $P[i, j] = PP[i, j]$, and repeat the computations, recursively.

Some plots resulting from simulations of the above algorithm are shown in Figs. 2.10-2.13. Fig 2.10 (*left*) is a 3-dimensional "impulse" plot showing all those grid points $(x_A, x_B) \in \overline{D}$ at which the mass is greater than some arbitrary threshold, $\bar{t} = 0.001$ in this case, after 1000 iterations of the algorithm, with the unit mass initialised at $(0, 0)$, and $\lambda = 0.5$. We note that the separation of the mass "humps" is a numerical artefact, dependent on the value of the threshold $\bar{t} > 0$. The smaller the threshold, the bigger the separation of the "humps", and vice-versa. The height of each "impulse" at each grid point gives the mass at that point. In theory, after N iterations of the algorithm, the total sum of masses at all the grid points in phase space $\overline{D}$ should be unity. In practice, however, because of the thresholding

mentioned above, this sum will only be approximately unity.

In Fig 2.10 (*left*), it is clear that the entire mass migrates towards $(-2, 2)$. This behaviour is observed no matter where the algorithm is initialised inside $\overline{D}$. Recall that, for $\lambda = 0.5$, the associated system of ODEs has exactly one globally asymptotically stable equilibrium at $(-2, 2)$. Fig 2.10 (*right*) depicts all those grid points (x_A, x_B) at which the mass is greater than the threshold $\bar{t} = 0.001$ after 1000 iterations of the algorithm, with the unit mass initialised at (0.0), and for $\lambda = 0.99$ (essentially, this captures the results of Ljung's simulations in Fig 4 of his paper [2]). In this case, the associated system of ODEs possesses two locally asymptotically stable equilibria, located at $(-2, 2)$ and $(-2.258, -1.372)$. It is clear that there is migration of a major portion of mass towards $(-2, 2)$ and a minor one towards $(-2.258, -1.372)$. Fig 2.11 (*left*) shows the results of running the algorithm, starting with a unit mass positioned at $(-2, 2)$, with $\lambda = 0.99$. In this case, the mass remains centred around $(-2, 2)$. Fig 2.11 (*right*) depicts the results when a unit mass is initialised at $(-2.258, -1.372)$, with $\lambda = 0.99$. That is, we start the algorithm on the equilibrium point (of the associated system of ODEs) $(-2.258, -1.372)$. Surprisingly, it is evident that a major portion of the mass migrates towards $(-2, 2)$, while only a relatively minor portion remains centred around $(-2.258, -1.372)$.

Figs. 2.12 - 2.13 show projections, on the (x_A, x_B) plane, of Figs. 2.10 - 2.11. Here, each dot on the (x_A, x_B) plane denotes mass at a particular grid point in $\overline{D}$. In all the plots, we used $\gamma_n = n^{-0.25}$, which has a relatively slow rate of convergence. This is desirable in the sense that it minimises the chances of the algorithm getting

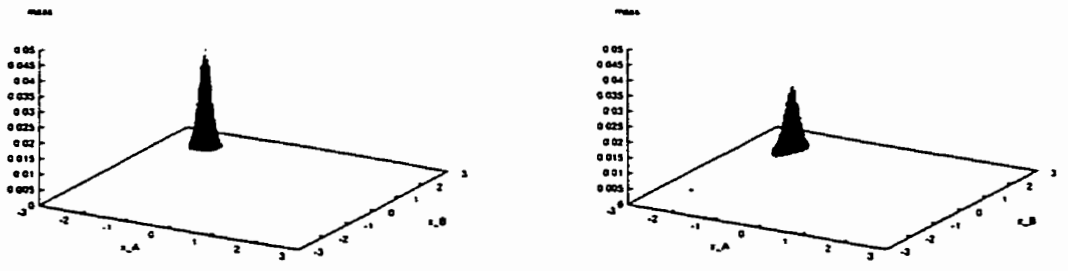

Figure 2.10: 3-dimensional "impulse" plot of all grid points whose associated masses are greater than $\bar{t} = 0.001$ after 1000 iterations of the algorithm, as described in the text, with *left*: $\lambda = 0.5$, and *right*: $\lambda = 0.99$. In both cases, the unit mass is initialised at the grid point $(0, 0)$. As discussed in the text, the height of each "impulse" gives the mass at the grid point under consideration.

trapped at some pseudo-equilibrium point, as previously discussed. Furthermore, we have only shown plots for fixed λ values. If we change the value of λ, the variation in the plots is minimal - but a description of how this happens is beyond the scope of the thesis.

Comments on simulation results

- The simulation results suggest that the density of mass around the two equilibria (inside $\overline{D}$) of the associated ODE is almost independent of the initial condition of the algorithm. This mass density is indicative of the probability that the algorithm will converge to each respective equilibrium point, given some arbitrary initial point $(x_A(0), x_B(0)) \in \overline{D}$. Thus, the achievement of the algorithm is that, given some arbitrary initial point $(x_A(0), x_B(0)) \in \overline{D}$, we are able to say, in relative terms, what the probability of landing on (-2,2) or (-2.258,-1.372) is. The problem of numerically quantifying this probability is

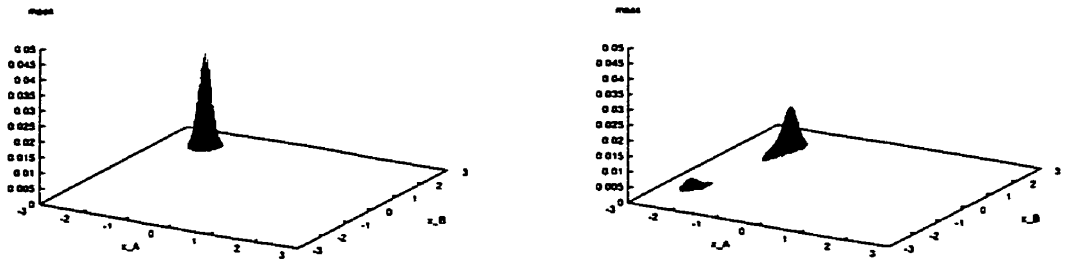

Figure 2.11: 3-dimensional "impulse" plot of all grid points whose associated masses are greater than $\bar{t} = 0.001$ after 1000 iterations of the algorithm, as described in the text, with *left*: $(x_A(0), x_B(0)) = (-2, 2)$, and *right*: $(x_A(0), x_B(0)) = (-2.258, -1.372)$. In both cases, $\lambda = 0.99$.

straightforward, since each dot or "impulse" represents a fraction of the unit mass.

- Even though it is clear that the algorithm can converge to either of the equilibria, it is possible to "perversely" pick individual sequences of φ's that steer the algorithm elsewhere. These sequences have probability zero. We discuss this phenomenon below.

Consider the annotated picture of Ljung's doubly-triangular pdf shown in Fig 2.14 (*left*). Suppose that $x_A = -2$, and $-1 < x_B < 0$, giving $-2 < c < -1$. Now, let $d = \min(-1, x_B)$. Then, artificially restrict $\varphi \in [c, d]$ and construct a new *pdf*, g_λ, which mimics f_λ. That is,

$$g_\lambda(\varphi) = \begin{cases} -\lambda(\varphi - d) + \frac{1}{d-c} + \frac{\lambda}{2}(c - d) & , \quad \varphi \in [c, d] \\ 0 & , \quad \text{elsewhere} \end{cases}$$

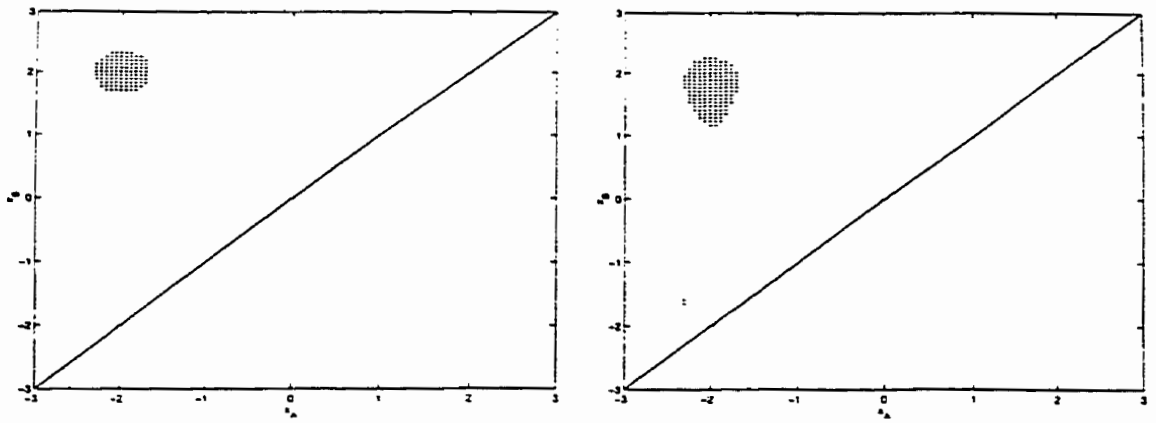

Figure 2.12: Grid points whose associated masses are greater than $\bar{t} = 0.001$ after 1000 iterations of the algorithm, as described in the text, with *left*: $\lambda = 0.5$, and *right*: $\lambda = 0.99$. Note that both these Figs. are projections of Figs. 2.10 - 2.11 on the (x_A, x_B) phase plane. In both cases, the unit mass is initialised at the grid point (0.0).

Note that restricting φ's to the closed interval $[c, d]$ ensures that x_B is decreased for all n. The probability that x_B is decreased is given by

$$Pr(x_B \text{ is decreased}) = \int_c^d g_\lambda(\varphi) d\varphi \ .$$

Select individual infinite sequences of φ's from the closed interval $[c, d]$ according to g_λ and feed these into Ljung's algorithm. Set $\lambda = 0.5$. This ensures that the associated ODE has only one globally asymptotically stable equilibrium, viz. $(-2, 2)$. Then, *depending on the location of* $(x_A(0), x_B(0)) \in \overline{D}$, the iterates (x_A, x_B) will not converge to $(-2, 2)$ since x_B is always decreasing. Furthermore, if $\gamma_n \to 0$ 'fast enough', then (x_A, x_B) will freeze at some pseudo-equilibrium point (x_A^*, x_B^*), say. To numerically illustrate the effect of selecting φ's as described above, we proceed as follows. Initialise the algorithm at $(x_A(0), x_B(0)) = (-2, -0.5)$. Then choose a

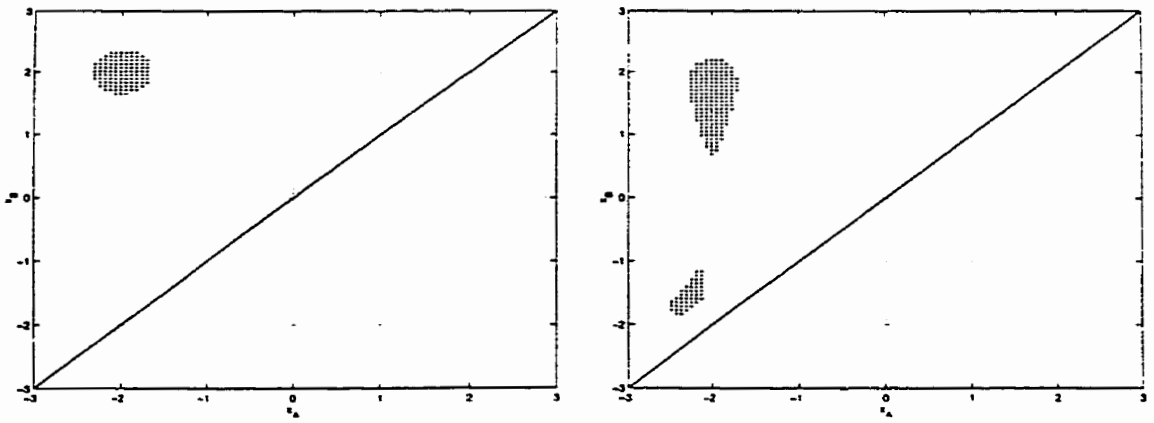

Figure 2.13: Grid points whose associated masses are greater than $\bar{t} = 0.001$ after 1000 iterations of the algorithm, as described in the text. The unit mass is initialised at *left*: $(-2, 2)$ *right*: $(-2.258, -1.372)$. In both cases, we set $\lambda = 0.99$

sequence of φ's according to g_λ. Feed this sequence into Ljung's algorithm and iterate. The results of this simulation corroborate our theoretic predictions. That is. $x_A = -2$ is unchanged, while x_B decreases progressively as $n \to \infty$, finally freezing at $x_B \approx -2$, which lies on the line $x_A = x_B$ (a part of the boundary of $\overline{D}$). Fig 2.14 (*right*) displays the results of this experiment. Fig 2.15 is a schematic diagram of the observed dynamics of the iterates (x_A, x_B) as $n \to \infty$. It is important to stress, at this point, that whether $x_B \to -2$ or not is dependent on the $\{\varphi_n\}$ sequence. For example, if we set $\varphi_n = \varphi_0$, a constant for all n, then it is clear that $x_B \to \varphi_0$ as $n \to \infty$. By the same token, if $\varphi_n \to -2$, then it follows that $x_B \to -2$ as $n \to \infty$.

Note that, since x_B is decreasing, it may be shown that there exists an $N > 0$ such that $x_B < -1$ for all $n \geq N$. This implies that the interval $[c, d]$, from which the φ's are sampled, shrinks as $n \to \infty$. Furthermore, this implies that

$$Pr(x_B \text{ is decreased}) \to 0 \text{ as } n \to \infty .$$

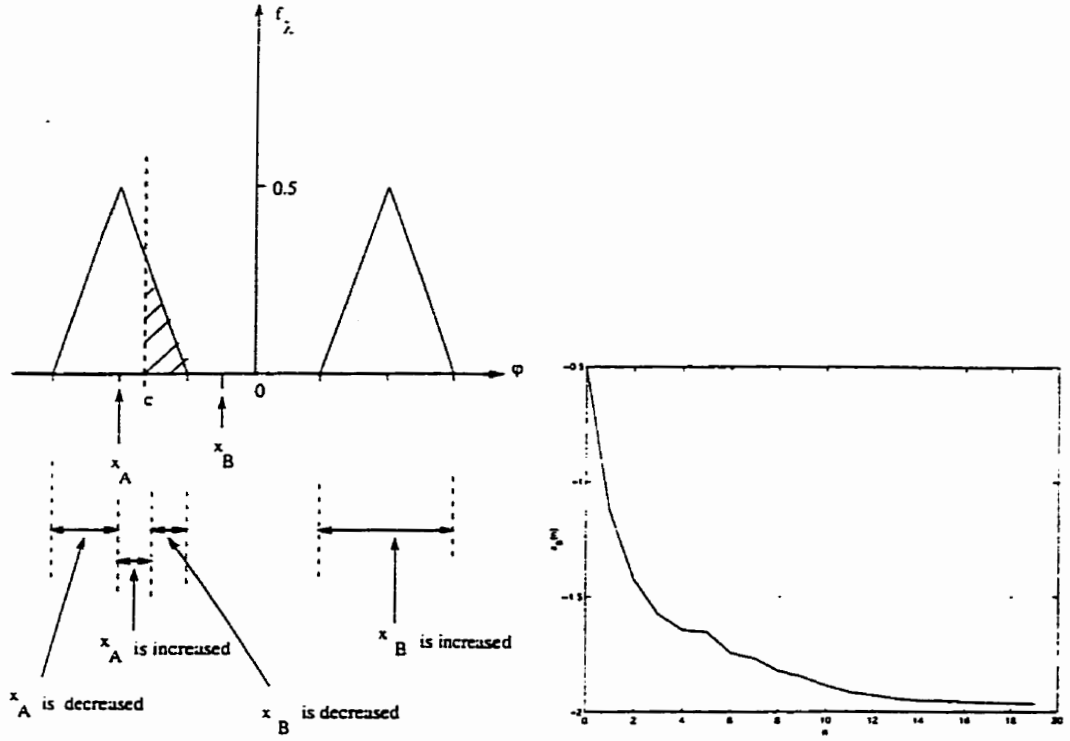

Figure 2.14: *left*: Relevant regions that are used in the construction of a pathological sequence of inputs to steer the algorithm away from the globally asymptotically stable equilibrium $(-2, 2)$, when $\lambda = 0.5$ *right*: Evolution of the sequence $\{x_B(n)\}$ resulting from the use of a pathological sequence of inputs $\{\varphi_n\}$, as described in the text.

Note that the numerics suggest that *both* c and x_B are moving towards -2 as $n \to \infty$, and that $c < x_B$ for all n. That is, there is *always* a non-zero probability, no matter how small, that x_B is decreased.

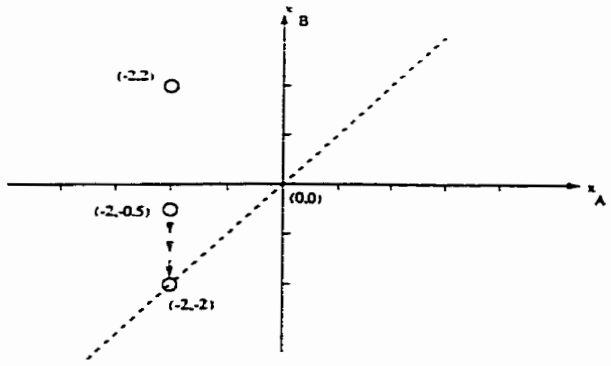

Figure 2.15: Dynamics of (x_A, x_B) using a pathological $\{\varphi_n\}$ sequence as input to the algorithm

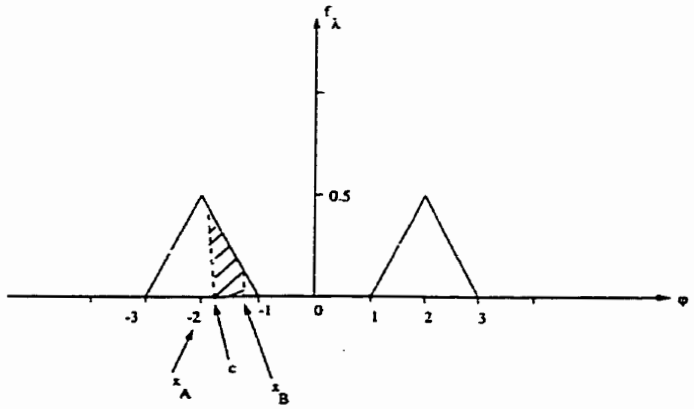

Figure 2.16: Illustration of the fact that for $x_B < -1$, the interval $[c, d]$ diminishes

2.8 Investigation of *pathological* sequences of inputs

This section seeks to elaborate on the idea of selectively choosing φ's so that the iterates (x_A, x_B) are steered away from the equilibrium point $(-2, 2)$ of the associated system of ODEs. As before, set $x_A(n) = -2$, $-1 < x_B(n) < 0$. Then we get that $-1.5 < c_n < -1$. We will select φ's such that $x_B(n)$ is decreased for all n.

Recall that $x_B(n)$ is decreased if $c < \varphi < x_B$. The sought-after sequence of φ's is constructed by randomly choosing $\{\varphi_n\}$ according to g_λ. As discussed before, we have that $d_n \sim x_B(n)$.

Consider the sequence

$$\varphi = (\varphi_1, \varphi_2, ..., \varphi_n) , \tag{2.31}$$

where $\varphi_n \in [c_n, d_n]$. Now, let P_n denote the probability of choosing φ_n from $[c_n, d_n]$. This is given by

$$P_n = \int_{c_n}^{d_n} g_\lambda(\varphi) d\varphi , \tag{2.32}$$

where g_λ is a *pdf* that mimics Ljung's doubly-triangular pdf. The probability of realising (2.31) is given by the product

$$\beta_n = P_1 \times P_2 \times ... \times P_n . \tag{2.33}$$

As mentioned above, $x_B(n)$ decreases if and only if

$$-1 + \frac{1}{2} x_B(n) < \varphi < x_B(n) . \tag{2.34}$$

Recall that $x_B(n)$ is updated according to the rule

$$x_B(n+1) = x_B(n) + \gamma_n[\varphi_n - x_B(n)] . \tag{2.35}$$

From (2.34) and (2.35), we obtain the following inequalities

$$x_B(n+1) < x_B(n) , \tag{2.36}$$

and

$$x_B(n+1) > (1-0.5\gamma_n)x_B(n) - \gamma_n \ . \tag{2.37}$$

From (2.37), we obtain

$$x_B(2) > (1-0.5\gamma_1)x_B(1) - \gamma_1 \ , \tag{2.38}$$

and

$$x_B(3) > (1-0.5\gamma_2)[(1-0.5\gamma_1)x_B(1) - \gamma_1] - \gamma_2 \ . \tag{2.39}$$

The right hand side of (2.37) may be expressed in closed form as

$$\begin{aligned} x_B(n+1) &> x_B(1)\prod_{j=1}^{n}(1-0.5\gamma_j) + \sum_{k=1}^{n-1}\{-\gamma_k \prod_{j=k+1}^{n}(1-0.5\gamma_j)\} - \gamma_n \\ &\sim x_B(1)e^{-\frac{n^{1-\alpha}}{1-\alpha}} - \sum_{k=1}^{n-1} k^{-\alpha} e^{-\frac{1}{1-\alpha}\{n^{1-\alpha}-(k+1)^{1-\alpha}\}} - n^{-\alpha} \ , \end{aligned} \tag{2.40}$$

where the last line is obtained by setting $\gamma_n = n^{-\alpha}$, $\alpha \in (0,1)$ and by asymptotically expanding $\prod_{j=1}^{n}(1-\gamma_j)$ as follows. Denote

$$F_n = \prod_{j=1}^{n}(1-\gamma_j) \ .$$

Then we have that

$$\begin{aligned} \ln F_n &= \sum_{j=1}^{n} \ln(1-j^{-\alpha}) \\ &\approx -\sum_{j=1}^{n} j^{-\alpha} \ , \text{ for } n \text{ large enough} \end{aligned}$$

$$
\begin{aligned}
&= -n^{1-\alpha} \sum_{j=1}^{n} \frac{1}{(j/n)^{\alpha}} \cdot \frac{1}{n} \\
&\sim -n^{1-\alpha} \int_{\frac{1}{n}}^{1} t^{-\alpha} dt \\
&= \frac{n^{1-\alpha}}{1-\alpha} [n^{\alpha-1} - 1] \\
&\approx -\frac{n^{1-\alpha}}{1-\alpha} \text{ , for large } n \text{ ,}
\end{aligned}
$$

leading to

$$F_n \sim e^{-\frac{n^{1-\alpha}}{1-\alpha}} .$$

Returning to (2.36) and (2.37), we note that (2.36) implies that $x_B(n)$ is monotonically decreasing. Now, from (2.37), if we can show that $x_B(n)$ is bounded below by "something", then we will be able to conclude, using the *monotone convergence theorem*, that $x_B(n)$ converges. In fact, showing that $x_B(n)$ is bounded below is quite difficult, since the bound depends on the particular choice of the sequence $\{\varphi_n\}$, as we show below.

From (2.32), we have that

$$P_n = \int_{c_n}^{d_n} g_\lambda(\varphi) d\varphi \sim \int_{0.5(x_B-2)}^{x_B} g_\lambda(\varphi) d\varphi \, . \tag{2.41}$$

From this, it is clear that if $x_B \to -2$, then $P_n \to \int_{-2}^{-2} g_\lambda(\varphi) d\varphi = 0$. That is, the probability of choosing any φ_n from $[c_n, d_n]$ approaches zero as $n \to \infty$. Furthermore, this yields $\beta_n \to 0$ as $n \to \infty$. The issue is whether or not $x_B(n)$ converges to -2. Clearly, as discussed previously, this depends on the $\{\varphi_n\}$ sequence. Suppose

that $\varphi_n \to \varphi^* > -2 \in [c, d]$. Then

$$P_n \sim \int_{0.5(\varphi^*-2)}^{\varphi^*} g_\lambda(\varphi) d\varphi \ .$$

Clearly, P_n does *not* approach 0 as $n \to \infty$ in this case. Note that P_n is the probability of picking the n^{th} element of the sequence shown in (2.31) from the closed interval $[c_n, d_n]$. Now, because $\lim_{n\to\infty} \sum_{i=1}^{n} P_i = 1$, we conclude that $\beta_n \to 0$, as $n \to \infty$. In other words, the probability of realising any particular pathological sequence $\varphi = (\varphi_1, \varphi_2, \ldots, \varphi_n)$ approaches zero as $n \to \infty$.

Remark 2.3 *Even though it is feasible to select a pathological sequence of inputs that steers the iterates away from the equilibrium point of the associated ODE, the probability of realising such a sequence approaches zero as $n \to \infty$. This implies that the fraction of sequences $\{\varphi_n\}$ that steer iterates away from the equilibrium approaches zero as $n \to \infty$.*

2.9 A look at the special case $\lambda = 1$

When $\lambda = 1$, the doubly-triangular *pdf* of the previous section degenerates to

$$f_\lambda(\varphi) = \begin{cases} \varphi + 3 & , \ -3 < \varphi \leq -2 \\ -(\varphi + 1) & , \ -2 < \varphi \leq -1 \\ 0 & , \ \text{otherwise} \ . \end{cases} \tag{2.42}$$

We seek to determine what the classifier $c(t)$ (previously defined in Section 2.1) converges to in this case. Intuitively, we expect that $c(t) \to -2$ as $t \to \infty$. However,

we have to investigate whether Ljung's model reproduces this. As before, we obtain the following system of ODEs:

$$\dot{x}_A = \begin{cases} 0 & , \quad c < -3 \\ \frac{c^3}{3} + (3 - x_A)\frac{c^2}{2} - 3x_A c - \frac{9}{2}(x_A + 1) & , \quad -3 \leq c < -2 \\ -\frac{1}{2}(\frac{7}{3} + x_A) - \frac{c^3}{3} + \frac{1}{2}(x_A - 1)c^2 - \frac{2}{3} + x_A c & , \quad -2 \leq c < -1 \\ -(x_A + 2) & , \quad c \geq -1 \end{cases} \tag{2.43}$$

and

$$\dot{x}_B = \begin{cases} -(x_B + 2) & , \quad c < -3 \\ \frac{5}{2} - \frac{c^3}{3} - \frac{3}{2}c^2 + (\frac{7}{2} + 3c + \frac{c^2}{2})x_B & , \quad -3 \leq c < -2 \\ \frac{c^3}{3} + \frac{c^2}{2} - \frac{1}{6} - (\frac{c^2}{2} + c + \frac{1}{2})x_B & , \quad -2 \leq c < -1 \\ 0 & , \quad c \geq -1 \end{cases} \tag{2.44}$$

Now for $c \geq -1$, we have the system

$$\begin{aligned} \dot{x}_A &= -(x_A + 2) \\ \dot{x}_B &= 0 \end{aligned} \tag{2.45}$$

which has a line of equilibrium points $(x_1^*, x_2^*) = (-2, k)$, where $k > 0$. Hence, the classifier converges to

$$c^* = \frac{k}{2} - 1 \,, \tag{2.46}$$

where the value of k depends on the initial condition. Note that $k \geq 0$, since $c^* \geq -1$. It is clear that an unusual bifurcation has occurred, since we now have infinitely many equilibria - all lying on the line $x_A = -2$. This is shown in Fig 2.17.

From (2.45), we have that $\dot{x}_A > 0$ if $x_A < -2$ and $\dot{x}_A < 0$ if $x_A > -2$. This

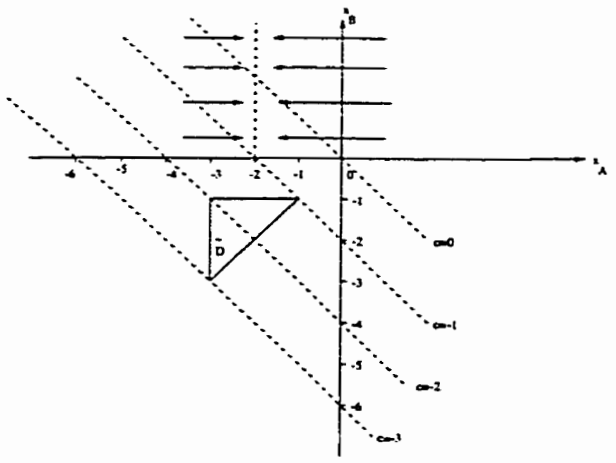

Figure 2.17: Plot of phase space of the associated system of ODEs (with $\lambda = 1$) showing a continuum of equilibria, as discussed in the text

implies that the equilibrium $(-2, k)$ is neutrally stable. Also, we observe that the invariant region is now given by $\overline{D} : -1 > x_B \geq x_A > -3$. Thus, $(x_1^*, x_2^*) \notin \overline{D}$ for all $k \geq 0$.

In a similar manner, it can be shown that the associated system of ODEs (2.43)-(2.44) yields one locally asymptotically stable equilibrium and a saddle point, viz. $(-2.333, -1.667) \in \overline{D}$ and $(-2.000, -0.667) \notin \overline{D}$, respectively. Hence, the classifier $c(t)$ converges to -2 as $t \to \infty$.

Remark 2.4 *From the foregoing considerations, it is evident that the special case* $\lambda = 1$ *represents a breakdown of Ljung's model. The case* $\lambda = 0$ *displays similar behaviour. In this case, the classifier* $c(t)$ *converges to* 2, *the mean value of the right hand component triangle of the pdf.*

2.10 Closing remarks

The numerical simulations performed in this chapter do help in developing a feeling for the long term behaviour of the $\{(x_A(n), x_B(n))\}$ sequence, *provided we are given some starting value* $(x_A(0), x_B(0)) \in \overline{D}$. However, there is an inherent limitation in this approach, viz. tracking of *individual* trajectories of the algorithm will *not* give any illuminating picture, since $\{(x_A(n), x_B(n))\}$ is not ergodic in most applications. To be able to make any general comments about the convergence of iterates, there is a need to encompass all possible starting values $(x_A(0), x_B(0))$. One way to do this is to define a probability density function over all possible initial states of the algorithm, and then track the temporal evolution of this density. This idea is the subject of the next chapter.

Chapter 3

Evolution of Densities of Algorithms

3.1 Introduction

Consider the usual [1] noise-driven neural learning algorithm

$$x_{n+1} = S_n(x_n) + \gamma_n \xi_n \ , \tag{3.1}$$

where $S_n(x_n) \stackrel{def}{=} x_n + \gamma_n \bar{h}(x_n)$, and $\{\xi_n\}$ is a sequence of *i.i.d.* random variables with common density g. The rest of the symbols retain their usual connotations. Note that the map S_n explicitly depends on n. Furthermore, for $\bar{h}(\cdot)$ bounded, (3.1) gives $|x_{n+1} - x_n| = \gamma_n |\bar{h}(x_n) + \xi_n| \to 0$ as $n \to \infty$, since $\gamma_n \to 0$ as $n \to \infty$. However, this observation does not imply that $\{x_n\}$ is a Cauchy sequence. To see this, consider, for example, the sequence $\{x_n\}$ generated by the terms $x_n = \sum_{k=1}^{n} \frac{1}{k}$,

i.e. the individual terms of $\{x_n\}$ are partial sums of the harmonic series. Then we have that $|x_{n+1} - x_n| = \frac{1}{n+1} \to 0$ as $n \to \infty$. On the other hand, clearly, $x_n \to \infty$ as $n \to \infty$. Alternatively, we may write (3.1) as $x_{n+1} - x_n = \gamma_n[\bar{h}(x_n) + \xi_n] \to 0$ as $n \to \infty$. This implies that (3.1) approaches the identity map I : $x_{n+1} = x_n$ as $n \to \infty$. The latter is known to have infinitely many fixed points.

In this Chapter, and essentially throughout the thesis, we are interested in the long term behaviour of the $\{x_n\}$ sequence. For example, *do the iterates converge to a unique fixed point, for all initial points* x_0*?* It is well-known [1] that if the associated ODE

$$\dot{x} = \bar{h}(x) \tag{3.2}$$

has exactly one globally asymptotically stable equilibrium $\hat{x}$, then, wp1, (3.1) converges to $\hat{x}$. However, if (3.2) possesses more than one stable equilibrium, not much is known about the convergence of (3.1) to these stable equilibria. It is the primary goal of this chapter to get a glimpse of what goes on in this latter situation.

As with any random process (see [26],[27]), it is not possible to obtain meaningful convergence results of (3.1) by simply tracking individual trajectories as is the case for ergodic transformations. Instead, more meaningful results may be obtained by an examination of the evolution of densities of $\{x_n\}$ [5]. The utility of this approach lies in the fact that it takes into account *all* possible initial states. An initial probability density function f_0 is defined over all the possible initial states of the algorithm. The idea is then to find how this prescribed density of initial states evolves over time. Ultimately, one hopes to find a (limiting) *stationary density* f_* for the sequence of densities $\{f_n\}$. It is this stationary density that

may hold the key to understanding the intricacies of the convergence properties of an algorithm, especially if the associated ODE has multiple locally asymptotically stable equilibria.

In this chapter, we will study the evolution of densities, under the action of the so-called *Frobenius-Perron operator* corresponding to the map S_n, of *both* the full nonlinear algorithm (3.1) as well as its linearised counterpart

$$x_{n+1} = \alpha_n x_n + \gamma_n \xi_n \ , \tag{3.3}$$

where $\alpha_n = 1 + \gamma_n \bar{h}'(\hat{x})$ and $\hat{x}$ is an equilibrium point of (3.2). A brief description of the above operator is given in Section 3.3.

We now digress to show how (3.3) is derived from (3.1). Let $\hat{x}$ be a stable equilibrium point of the associated ODE (3.2). Furthermore, assume that

$$x_n \stackrel{def}{=} \hat{x} + \epsilon_n \ ,$$

where $\epsilon_n \to 0$ as $n \to \infty$. Taking a Taylor expansion of $\bar{h}(\cdot)$ about $\hat{x}$, we have that

$$\bar{h}(x_n) = \epsilon_n \bar{h}'(\hat{x}) + O(\epsilon_n^2) \ , \text{ for } n \text{ large enough} \ .$$

Plugging the last expression back into (3.1), and ignoring higher order terms in ϵ_n, gives

$$\epsilon_{n+1} = [1 + \gamma_n \bar{h}'(\hat{x})]\epsilon_n + \gamma_n \tilde{\xi}_n \ ,$$

where $\tilde{\xi}_n \stackrel{def}{=} h(\hat{x} + \epsilon_n, \varphi_n) - \bar{h}(\hat{x} + \epsilon_n)$. Setting $\alpha_n \stackrel{def}{=} 1 + \gamma_n \bar{h}'(\hat{x})$ yields the desired

result, where ϵ_n is a dummy variable. Note that $\alpha_n \in (0,1)$, and that $\alpha_n \to 1^-$ as $n \to \infty$.

To give a flavour of the spirit of this chapter, consider the unperturbed linearised algorithm given by

$$x_{n+1} = (1 - \lambda n^{-\alpha}) x_n \ , \qquad (*)$$

where $\lambda > 0$, $\alpha \in (0,1)$. The iterates x_n may be expressed in closed form as follows:

$$x_n = x_1 \prod_{k=1}^{n-1} (1 - \lambda k^{-\alpha}) \ .$$

We wish to characterise the long term bahaviour of x_n. To this end, we proceed in the following way. Let

$$B_n \stackrel{def}{=} \prod_{k=1}^{n-1} (1 - \lambda k^{-\alpha}) \ ,$$

which gives

$$\begin{aligned} \ln B_n &= \sum_{k=1}^{n-1} \ln(1 - \lambda k^{-\alpha}) \\ &\sim -\lambda \sum_{k=1}^{n-1} k^{-\alpha} \ , \text{ for } k \text{ sufficiently large.} \end{aligned}$$

In view of this, we get

$$B_n \sim e^{-\lambda \sum_{k=1}^{n-1} k^{-\alpha}} \ .$$

Now, it may be shown that

$$\sum_{k=1}^{n-1} k^{-\alpha} \sim \frac{n^{1-\alpha}}{1-\alpha} \ ,$$

which implies that

$$x_n \sim x_1 e^{-\frac{\lambda}{1-\alpha}n^{1-\alpha}} \to 0 \text{ as } n \to \infty .$$

Next, let us perturb (*) by adding a forcing term to it, viz.

$$x_{n+1} = (1 - \lambda n^{-\alpha})x_n + n^{-\alpha} r_n , \qquad (**)$$

where r_n is a deterministic forcing term. For this perturbed algorithm, Chung [19] proved that if $r_n \to 0$ faster than $n^{-\epsilon}$, where $\epsilon > 0$, then $x_n \to 0$ as $n \to \infty$. The following theorem [19] summarises his results:

Theorem 3.1 *Suppose that* $\{b_n\}$, $n \geq 1$, *is a sequence of real numbers such that for* $n \geq n_0$,

$$b_{n+1} \leq \left(1 - \frac{c_n}{n^s}\right) b_n + \frac{c'}{n^t} ,$$

where $s \in (0,1)$, $s < t$, $c_n \geq c > 0$, $c' > 0$. *Then*

$$\overline{\lim}_{n\to\infty} n^{t-s} b_n \leq \frac{c'}{c} .$$

A proof of this result may be found in [19].

Fabian [18] considered the special case $\alpha = 1$ of (**), and proved that $x_n \to 0$ if and only if $n^{-1}\sum_{j=1}^{n} r_j \to 0$ as $n \to \infty$. His result is consistent with intuition, since $n^{-1}\sum_{j=1}^{n} r_j$ may be interpreted as the mean of the sequence $\{r_k\}_{k=1}^{n}$. In other words, he showed that if the mean of $\{r_k\}_{k=1}^{n}$ converges to zero, then the iterates x_n will converge to zero as well, as $n \to \infty$. During the course of this work, we have been interested in generalising Fabian's result to the case when $\alpha \neq 1$, viz. proving

the conjecture that $x_n \to 0$ if and only if $n^{-\alpha} \sum_{j=1}^{n} r_j \to 0$ as $n \to \infty$. However, we suspect that the latter may not be true, since the term $n^{-\alpha} \sum_{j=1}^{n} r_j$ is not the usual mean of the sequence $\{r_k\}_{k=1}^{n}$.

The ultimate goal is to examine the long term behaviour of the sequence $\{x_n\}_{n=1}^{\infty}$ in the case when $\{r_n\}$ is a sequence of random variables, with some prescribed probability density function. In this case, it is not guaranteed that $r_n \to 0$ as $n \to \infty$. However, we point out that if the sequence $\{r_n\}$ of random variables is compactly supported, and

$$x_{n+1} = (1 - \lambda n^{-\alpha})x_n + n^{-\gamma} r_n \ , \text{ where } \gamma > \alpha \ ,$$

then Theorem 3.1 guarantees that $x_n \to 0$ as $n \to \infty$.

In light of the above examples, we shall now proceed to formulate a framework for the analysis of sequences of densities generated by iteration of the Frobenius-Perron operator.

3.2 The Space $L_D^1(X, \Lambda, \mu)$

Conventionally, most authors (see [5], for example) formulate the analysis of the sequence of densities $\{f_n\}_{n=1}^{\infty}$ in the space L_D^1 of probability density functions. However, as will be shown in Section 3.4, this space is not always appropriate. For completeness, we give a brief description of L_D^1. Let (X, Λ, μ) be a measure space, where Λ is a σ-algebra in X, and where μ is a measure on Λ. Consider the set Γ of

all real-valued functions f whose absolute values are integrable over X, i.e.

$$\int_X | f | d\mu < \infty ,$$

where, in most practical applications, μ is Lebesgue measure. Note that Γ is a linear space, since every finite linear combination of integrable functions is integrable. Now define the subset $\Gamma_0 \subset \Gamma$ by $\Gamma_0 = \{f \in \Gamma \mid f = 0 \ a.e.\}$. Then the real valued functional p defined below is a norm on the factor space Γ/Γ_0, where Γ/Γ_0 is denoted by $L^1(X, \Lambda, \mu)$:

$$p : L^1 \to \mathbb{R} \ , \ p(f) = \int_X | f | d\mu \ , \text{ for all } f \in L^1 \ . \tag{3.4}$$

It is usual to denote $p(f)$ by $\|f\|_1$. $L^1(X, \Lambda, \mu)$ is a metric space with the metric given by

$$\rho(f, g) = \|f - g\|_1 \ , \text{ for all } f, g \in L^1(X, \Lambda, \mu) \ . \tag{3.5}$$

The "conventional" space on which densities are defined is denoted by $L^1_D(X, \Lambda, \mu) \subset L^1(X, \Lambda, \mu)$, and defined by

$$L^1_D(X, \Lambda, \mu) = \left\{ f \mid f \geq 0 \text{ and } \int_X f d\mu = 1 \right\} \ . \tag{3.6}$$

It is well-known that the space L^1_D is complete with respect to the L^1 norm. The utility of this fact will become apparent in later sections of this Chapter.

Definition 3.1 *Any function $f \in L^1_D(X, \Lambda, \mu)$ is called a density.*

Remark 3.1 *It turns out that, under some conditions, the sequence of densities*

$\{f_n\}_{n=1}^{\infty}$ converges to elements outside of the space $L_D^1(X, \Lambda, \mu)$. This necessitates the definition of a broader space. the space of all objects to which one can associate a distribution, *that includes $L_D^1(X, \Lambda, \mu)$. We will return to this issue in Sections 3.4 and 3.5.*

3.3 The Frobenius-Perron Operator

Suppose that we have a non-singular, measurable transformation $S : X \to X$ on a measure space (X, Λ, μ). For our purposes, S shall be defined as the deterministic part of the algorithm under consideration. For example, in equation (3.3), it is given by $S(x_n) = \alpha_n x_n$. Let $f \in L_D^1(X, \Lambda, \mu)$ be an arbitrary density. The Frobenius-Perron operator, $P : L_D^1 \to L_D^1$, describes the evolution of f induced by S. In other words, if f defines the distribution of initial conditions, i.e. points $x_0 \in X$, then Pf gives the resulting distribution of points $x_1 = Sx_0$. Define the action of P on f as follows

$$\int_A Pf(x)\mu(dx) = \int_{S^{-1}(A)} f(x)\mu(dx) \text{ , for } A \in \Lambda \text{ .} \tag{3.7}$$

This relationship uniquely defines P (see [5] for details). From (3.7), it may be shown that P has the following properties.

1. P is a linear operator. That is,

$$P(\lambda_1 f_1 + \lambda_2 f_2) = \lambda_1 P f_1 + \lambda_2 P f_2 \text{ ,} \tag{3.8}$$

for all $f_1, f_2 \in L^1$, $\lambda_1, \lambda_2 \in \mathbb{R}$.

2. For $f \in L^1$,

$$Pf \geq 0 \text{ if } f \geq 0 \text{ on } X. \tag{3.9}$$

3.

$$\int_X Pf(x)\mu(dx) = \int_X f(x)\mu(dx) \; . \tag{3.10}$$

4. If $S_n = S \circ \overset{\text{n times}}{\cdots} \circ S$ and P_n is the Frobenius-Perron operator corresponding to S_n, then $P_n = P^n$, where P is the Frobenius-Perron operator corresponding to S.

5. In the special case when $X = \mathbb{R}$, choosing $A = [0, x]$, we have that

$$Pf(x) = \frac{d}{dx} \int_{S^{-1}([0,x])} f(u)du \; .$$

Note that (3.9) and (3.10) imply that Pf is also a density.

Definition 3.2 *Any function $f \in L^1_D(X, \Lambda, \mu)$ that satisfies $Pf = f$ is called a stationary density of P.*

To illustrate the computation of the Frobenius-Perron operator, consider the map

$$x_{n+1} = \alpha x_n \; , \;\; \alpha \in (0, 1) \; ,$$

where $S(x_n) \overset{def}{=} \alpha x_n$, and $X = \mathbb{R}$. For any $f \in L^1_D$, the associated Frobenius-Perron operator is

$$Pf(x) \;\; = \;\; \frac{d}{dx} \int_{S^{-1}([a,x])} f(u)du$$

$$\begin{aligned} &= \frac{d}{dx}\int_{\frac{a}{\alpha}}^{\frac{x}{\alpha}} f(u)du \\ &= \frac{1}{\alpha}f(\frac{x}{\alpha}) , \end{aligned}$$

where $[a, x] \subset \mathbb{R}$. In light of this, we get

$$\begin{aligned} f_n(x) &\stackrel{def}{=} P^n f(x) \\ &= \frac{1}{\alpha^n}f(\frac{x}{\alpha^n}) . \end{aligned}$$

3.4 An example showing the inadequacy of L_D^1

To motivate the need for a broader space of densities that encompasses L_D^1, consider the iteration of a uniform pdf

$$f(x) = \begin{cases} 1 \quad , \quad x \in (-\frac{1}{2}, \frac{1}{2}) \\ 0 \quad , \quad \text{elsewhere} \end{cases} \tag{3.11}$$

under the Frobenius-Perron operator defined above. The action of P on f generates a sequence of densities, $\{f_n\}$, defined by

$$f_n(x) \stackrel{def}{=} P^n f(x) = \begin{cases} \frac{1}{\alpha^n} \quad , \quad x \in (-\frac{\alpha^n}{2}, \frac{\alpha^n}{2}) \\ 0 \quad , \quad \text{elsewhere} . \end{cases} \tag{3.12}$$

Suppose that $m < n$, and consider the densities f_m and f_n. As illustrated in Fig 3.1, we wish to compute the L^1 distance between f_m and f_n. We obtain

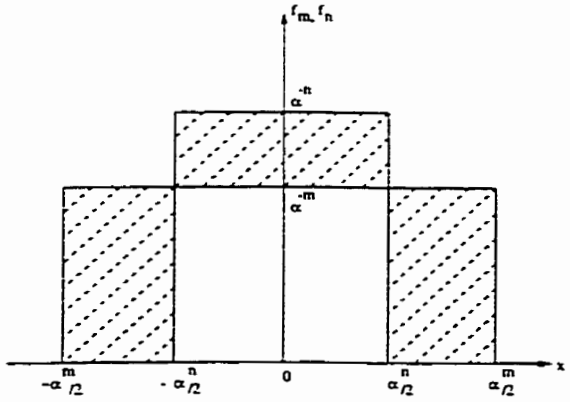

Figure 3.1: Iteration of a uniform pdf f under P associated with the map $S(x) = \alpha x$, $\alpha \in (0,1)$.

$$\begin{aligned} \|f_n - f_m\|_1 &= \|P^n f(x) - P^m f(x)\|_1 \\ &= 2(1-\alpha^{n-m}) \\ &< 2(1-\alpha^n) \to 2 \text{ as } n \to \infty . \end{aligned} \tag{3.13}$$

Thus, $\{f_n\}$ is *not* Cauchy in L^1_D. This implies that L^1_D is not an appropriate space for the analysis of evolution of densities under the action of this Frobenius-Perron operator. What is needed is to view these densities as special cases of distributions. We examine this idea in detail in the next section, where we show that the limiting distribution of the sequence $\{f_n\}$ is the so-called "Dirac distribution".

In view of the above example, we now formally introduce the space of normalised positive distributions, denoted by $\overline{D}'_+(X)$. The motivation for the work in the next section comes from a paper of Forte and Vrscay [20]. Essentially, they look at a similar space, but with more general distributions. In this thesis, we focus on a subset of their space, the space of normalised positive distributions mentioned above.

3.5 A Suitable Metric Space of Distributions

Let X be a compact and connected subset of $\mathbb{R}$. Distributions are defined as continuous linear functionals over a suitable space of test functions, denoted as $D(X)$. The space of distributions on X, denoted by $D'(X)$, is the set of all continuous linear functionals on $D(X)$. In other words, the set of $F : D(X) \to \mathbb{R}$ such that

1. For any sequence of test functions $\{\psi_\nu(t)\}_{\nu=1}^{\infty}$ that converges in $D(X)$ to $\psi(t)$, the sequence of numbers $\{F(\psi_\nu)\}_{\nu=1}^{\infty}$ converges to the real number $F(\psi)$ in the usual sense [22].

2. $F(c_1\psi_1 + c_2\psi_2) = c_1F(\psi_1) + c_2F(\psi_2)$, $c_1, c_2 \in \mathbb{R}$, $\psi_1, \psi_2 \in D(X)$.

The space $D'(X)$ includes the following as special cases:

a. Functions $f \in L^1$, for which the corresponding distributions are given by

$$F(\psi) = \int_X f(x)\psi(x)dx \text{ , for all } \psi \in D(X) .$$

b. The Dirac distribution, $\delta(x-a)$, which may be defined in the distributional sense as follows: For a point $a \in X$, $F(\psi) = \psi(a)$, for all $\psi \in D(X)$. This is often symbolically expressed as

$$F(\psi) = \int_X \psi(x)\delta(x-a)dx .$$

In this thesis, we are interested in the space of *normalised positive distributions*, denoted as $\overline{D}'_+(X)$, and defined by

$$\overline{D}'_+(X) = \{F \in D'(X) |\ F(1) = 1\ ,\ \text{and}\ F(\psi) \geq 0\ \forall \psi \in C^\infty_+(X)\}\ , \qquad (3.14)$$

where $C^\infty_+ = \{\psi \in C^\infty |\ \psi(x) \geq 0\ ,\ x \in X\}$. The physical motivation for the assumptions $F(\psi) \geq 0$ and $F(1) = 1$ is that we wish to interpret the values of the distributions as probability measures, in a manner analogous to integrals of density functions in L^1_D. In fact, note that $L^1_D(X) \subset \overline{D}'_+(X)$.

In the following analysis involving the space $\overline{D}'_+(X)$, we shall restrict our test functions to a subset of C^∞_+, namely, positive C^∞ functions that are Lip_1 on X, viz. $D(X) \stackrel{def}{=} Lip^+_1(X)$, where

$$Lip^+_1(X) = \{\psi \in Lip_1(X) |\ \psi(x) \geq 0\ ,\ \text{for all}\ x \in X\}\ , \qquad (3.15)$$

and where

$$Lip_1(X) = \{\psi : X \to \mathbb{R} |\ |\psi(x_1) - \psi(x_2)| \leq d(x_1, x_2)\ ,\ \text{for all}\ x_1, x_2 \in X\}\ .$$

The following property is very important in formulating a representation theory for distributions in $\overline{D}'_+(X)$.

Theorem 3.2 *For any distribution $F \in \overline{D}'_+(X)$, there exists a sequence of test*

functions $f_n \in Lip_1^+(X)$, $n = 1, 2, \cdots$, *such that for all* $\psi \in Lip_1^+(X)$,

$$\begin{aligned} \lim_{n\to\infty} F_n(\psi) &= \lim_{n\to\infty} \int_X f_n(x)\psi(x)dx \\ &\stackrel{def}{=} F(\psi) . \end{aligned}$$

This result is a rather simple specialisation of a theorem for the case $F \in D'(X)$, which was stated in [20]. By recourse to the above result, it will be convenient to express the distribution $F \in D'_+(X)$ symbolically as

$$F(\psi) = \int_X f(x)\psi(x)dx ,$$

even though there may not exist a pointwise function $f(x)$ which defines F (e.g. the Dirac distribution). *For notational convenience, given* $f \in L^1_D$, *we will write "*$f \in \overline{D}'_+$*" meaning that one can associate a distribution* $F \in D'_+(X)$ *to* f. (In the same way, we can write "$\delta \in D'_+(X)$", where δ is the Dirac distribution function.)

In [20], a metric was introduced over the space $D'(X)$. Following this treatment, we introduce a metric over the space $\overline{D}'_+(X)$:

$$d_{\overline{D}'_+}(f,g) = sup_{\psi\in Lip_1^+(X)} \left\{ \left| \int_X (f-g)(x)\psi(x)dx \right| \right\} , \text{ for all } f,g \in \overline{D}'_+(X) . \quad (3.16)$$

A major difference is the use of Lip_1^+ test functions in this metric, as opposed to test functions inside the unit C^∞ ball used in [20]. Our restriction to normalised positive distributions permits the use of Lip_1^+ functions, as we now show.

Given two test functions $\psi_1(x)$, $\psi_2(x)$ such that $\psi_1(x) = \psi_2(x) + c$, where $c \in \mathbb{R}$

and $f, g \in \overline{D}'_+(X)$, then

$$\int_X f(x)\psi_2(x)dx - \int_X g(x)\psi_2(x)dx = \int_X f(x)\psi_1(x)dx - \int_X g(x)\psi_1(x)dx \ .$$

In other words, the metric will not be affected by translations in the test functions. This allows us to use Lip_1 functions as is done for probability measures [21].

Theorem 3.3 *The metric space* $(\overline{D}'_+(X), d_{\overline{D}'_+})$ *is complete.*

Proof

Let $\{f_n\}_{n=1}^{\infty}$ be a Cauchy sequence in $(\overline{D}'_+(X), d_{\overline{D}'_+})$. In other words, for any $\epsilon > 0$, there exists an $N(\epsilon)$ such that $d_{\overline{D}'_+}(f_n, f_m) < \epsilon$, for all $n, m > N(\epsilon)$. From the definition of $d_{\overline{D}'_+}$ in (3.16), it follows that for any fixed $\psi \in Lip_1^+(X)$, the sequence of real numbers $\{t_n(\psi)\}_{n=1}^{\infty}$, where

$$t_n(\psi) = \int_X f_n(x)\psi(x)dx \ ,$$

is a Cauchy sequence on $\mathbb{R}$. The latter is true since, for any $\tilde{\psi} \in Lip_1^+(X)$, we have that $|t_n(\tilde{\psi}) - t_m(\tilde{\psi})| \leq sup_{\psi \in Lip_1^+(X)} |t_n(\psi) - t_m(\psi)| = d_{\overline{D}'_+}(f_n, f_m) < \epsilon$, for all $n, m > N(\epsilon)$. Let $\bar{t}(\psi)$ denote the limit of this sequence. Note that $\bar{t}(\psi) \geq 0$ for each $\psi \in Lip_1^+(X)$, since $\{t_n(\psi)\}_{n=1}^{\infty}$ is non-negative. By setting $F(\psi) = \bar{t}(\psi)$, we define a continuous linear functional F on $Lip_1^+(X)$. Furthermore, since $t_n(1) = 1$, it follows that $\bar{t}(1) = 1$. Therefore $F(1) = 1$, implying that $F \in \overline{D}'_+(X)$. This procedure can be easily extended to all C_+^{∞} test functions on X by noting (via the Mean Value Theorem) that $M^{-1} \times \psi \in Lip_1^+(X)$, where $M = \|\psi'\|_{\infty}$. Therefore, the metric space $(\overline{D}'_+(X), d_{\overline{D}'_+})$ is complete. □

We now illustrate the use of the metric space $(\overline{D}'_+(X), d_{\overline{D}'_+})$ in the investigation of the Frobenius-Perron operator P corresponding to the linear map $S(x) = \alpha x$, $\alpha \in (0,1)$. Note that P will now have to be a mapping from $\overline{D}'_+(X)$ to itself. We proceed in a manner analogous to that described in Section 3.3, with particular reference to the example considered in that section. Now, for any $f \in \overline{D}'_+(X)$, the distribution $q = Pf$ is defined by the linear functional

$$\begin{aligned} Q(\psi) &= \int_X q(x)\psi(x)dx \\ &= \int_X (Pf)(x)\psi(x)dx \\ &= \alpha^{-1}\int_{X_\alpha} f(\frac{x}{\alpha})\psi(x)dx \\ &= \int_X f(y)\psi(\alpha y)dy\,, \end{aligned} \tag{3.17}$$

where $\psi \in D(X)$. In the penultimate line, $X_\alpha \stackrel{def}{=} \{y \in X \mid y = S(x),\ x \in X\}$.

Theorem 3.4 *P is contractive in $(\overline{D}'_+(X), d_{\overline{D}'_+})$.*

Proof

Suppose that $\psi \in Lip_1^+(X)$, and define $\tilde{\psi}(y) = \alpha^{-1}\psi(\alpha y)$. Then

$$\begin{aligned} |\tilde{\psi}(x) - \tilde{\psi}(y)| &= \alpha^{-1}|\psi(\alpha x) - \psi(\alpha y)| \\ &\leq |x - y|\,, \end{aligned}$$

which implies that $\tilde{\psi} \in Lip_1^+(X)$. In addition, define

$$\tilde{L} = \{\tilde{\psi} \in Lip_1^+(X) |\ \tilde{\psi}(x) = \alpha^{-1}\psi(\alpha x)\,,\ \text{for some}\ \psi \in Lip_1^+(X)\}\,.$$

Then, (3.16) yields

$$\begin{aligned} d_{\overline{D}'_+}(Pf, Pg) &= sup_{\psi \in Lip_1^+(X)}\{|\int_X [f-g](y)\psi(\alpha y)dy|\} \\ &= sup_{\tilde{\psi}\in\tilde{L}}\{\alpha|\int_X [f-g](y)\tilde{\psi}(y)dy|\} \\ &\leq \alpha \sup_{\psi \in Lip_1^+(X)}\{|\int_X [f-g](y)\psi(y)dy|\}\ , \ \text{since}\ \ \tilde{L} \subset Lip_1^+(X) \\ &= \alpha d_{\overline{D}'_+}(f,g)\ , \end{aligned} \tag{3.18}$$

which gives the desired result. By Banach's Fixed Point Theorem, there exists a unique fixed point of the operator P in the metric space $(\overline{D}'_+(X), d_{\overline{D}'_+})$. □

This resolves the difficulty encountered in Section 3.4. Our limiting density is a distribution. Together, theorems 3.3 and 3.4 imply that there exists a *unique* $f_* \in \overline{D}'_+(X)$ such that

1. $Pf_* = f_*$, and

2. $d_{\overline{D}'_+}(P^n f, f_*) \to 0$, as $n \to \infty$ for any $f \in \overline{D}'_+(X)$.

The last statement follows from the observation that

$$d_{\overline{D}'_+}(P^n f, f_*) = d_{\overline{D}'_+}(P(P^{n-1}f), P(P^{n-1}f_*)) \leq \alpha^n d_{\overline{D}'_+}(f, f_*) \to 0\ , \ \text{as}\ \ n \to \infty\ .$$

We now show that $f_* = \delta(x)$, the Dirac delta "function". Let F be the Dirac distribution, i.e.

$$\begin{aligned} F(\psi) &= \int_X \delta(x)\psi(x)dx \\ &= \psi(0)\ . \end{aligned}$$

Then, from (3.17), Q, the distribution associated with the Frobenius-Perron operator, is

$$\begin{aligned} Q(\psi) &= \int_X \delta(y)\psi(\alpha y)dy \\ &= \psi(0) , \end{aligned}$$

which gives the desired result.

3.6 Densities of linearised algorithms

The previous section has given us a complete description of the linear map

$$x_{n+1} = \alpha x_n \ , \ \alpha \in (0,1) \ , \tag{3.19}$$

where $S(x_n) \stackrel{def}{=} \alpha x_n$. The main thrust of this section is to investigate the long term behaviour of sequence of densities $\{f_n\}$ for generalisations and noise-driven versions of (3.19). This section illustrates that the set $\{f_n\}$ does converge to elements outside of L^1_D, and hence the need for a broader space $\overline{D}'_+(X)$ that includes L^1_D. However, in the rest of this chapter, we shall formulate the analysis of $\{f_n\}$ in L^1_D, following the treatment of Mackey and Lasota [5]. Analysis in the space of distributions $\overline{D}'_+(X)$ is difficult and beyond the scope of this thesis. Besides, it appears that the space L^1_D is sufficient for the study of random processes, viz. limiting densities do not converge to Dirac-like distributions.

The most general linearised algorithm that we will consider is of the form

$$x_{n+1} = \alpha_n x_n + \gamma_n \xi_n \ , \tag{3.20}$$

where $\{\xi_n\}$ is a sequence of *i.i.d.* random variables, and $\alpha_n = 1 + \gamma_n \bar{h}'(\hat{x}) < 1$, so that $\hat{x}$ is a stable equilibrium point of the ODE

$$\dot{x} = \bar{h}(x) \ .$$

The parameter γ_n is called the gain of the algorithm (also referred to as the *learning parameter* in neural networks literature) and has the usual properties, viz.

$$\gamma_n \to 0 \ \text{ as } \ n \to \infty \ , \ \text{ and } \ \sum_n \gamma_n = \infty \ .$$

For latter purposes, we now take a brief detour to show that $\lim_{n\to\infty}(\prod_{j=1}^{n} \alpha_j) = 0$. Write

$$\alpha_n ... \alpha_1 = \prod_{j=1}^{n} [1 + b\gamma_j] \ , \tag{3.21}$$

where $b \stackrel{def}{=} \bar{h}'(\hat{x}) < 0$. Now denote the right hand side of (3.21) by R_n. Thus we have that

$$R_n = e^{\left\{\sum_{j=1}^{n} \ln[1+b\gamma_j]\right\}} \ , \tag{3.22}$$

which yields

$$R_n \sim e^{\left\{b\sum_{j=1}^{n} \gamma_j\right\}} \ . \tag{3.23}$$

If $\{\gamma_k\}_{k=1}^{\infty}$ satisfies the usual conditions (see Chapter 1, Section 1.4), we have that

$\sum_{j=1}^{n} \gamma_j \to \infty$ as $n \to \infty$. This yields

$$R_n \to 0 \text{ as } n \to \infty \,, \tag{3.24}$$

implying that $\lim_{n\to\infty}(\prod_{j=1}^{n} \alpha_j) = 0$. Below, we study the following linearised (see Section 3.1 for details) variations of algorithm (3.1):

i. $x_{n+1} = \alpha_n x_n$,

ii. $x_{n+1} = \alpha x_n + \gamma_n \xi_n$, where $\alpha \in (0,1)$,

iii. $x_{n+1} = \alpha_n x_n + \gamma_n \xi_n$, and

iv. $x_{n+1} = \alpha_n x_n + \xi_n$ (unattenuated noise amplitude) .

3.6.1 Algorithm (i)

Applying result 5 on page 91, it can be shown that the n^{th} iterate of the density $f(x)$ is given by

$$P_n f(x) = \alpha_n^{-1} f(\alpha_n^{-1} x) \,, \tag{3.25}$$

where P_n is the n^{th} Frobenius-Perron operator. Note that P_n acting on f contracts it in the x direction and expands it in the y direction. In other words, the RHS of (3.25) gets more "spiked" around the origin, as $n \to \infty$. Each α_n gives a different operator, leading to a sequence of operators $\{P_1, P_2, ..., P_n\}$. Now, iterating these operators, (3.25) becomes

$$P_n P_{n-1} ... P_1 f(x) = \frac{1}{\alpha_n ... \alpha_1} f\left(\frac{x}{\alpha_n ... \alpha_1}\right) \,. \tag{3.26}$$

Consider an arbitrary bounded interval $[-A, A] \subset \mathbb{R}$, where $A > 0$. Then

$$\int_{-A}^{A} P_n P_{n-1} \ldots P_1 f(x) dx = \int_{-\frac{A}{\alpha_n \ldots \alpha_1}}^{\frac{A}{\alpha_n \ldots \alpha_1}} f(x) dx \,, \tag{3.27}$$

which approaches $\int_{-\infty}^{\infty} f(x) dx = 1$ as $n \to \infty$. In other words, for any $A > 0$, there exists an n such that the right hand side of (3.27) is approximately unity. This suggests that the sequence of densities $\{f_n\}_{n=1}^{\infty}$ converges to $\delta(x)$, where the convergence is understood in the sense of distributions. We illustrate this in the following way: Suppose $f(x) = f_0 > 0$ for $x \in [-\frac{1}{2f_0}, \frac{1}{2f_0}]$, and zero elsewhere. Then

$$P_n \ldots P_1 f(x) = \begin{cases} 0 \,, \text{ for } x \notin [-\frac{\alpha_1 \ldots \alpha_n}{2f_0}, \frac{\alpha_1 \ldots \alpha_n}{2f_0}] \,, \\ \frac{f_0}{\alpha_n \ldots \alpha_1} \,, \text{ for } x \in [-\frac{\alpha_1 \ldots \alpha_n}{2f_0}, \frac{\alpha_1 \ldots \alpha_n}{2f_0}] \,. \end{cases}$$

Since $\frac{f_0}{\alpha_n \ldots \alpha_1} \to \infty$ as $n \to \infty$, this suggests that $\{f_n\}_{n=1}^{\infty}$ converges to $\delta(x)$. To close, note that (i) may be re-expressed as

$$x_{n+1} = x_n + \gamma_n \bar{h}'(\hat{x}) x_n \,,$$

giving the associated ODE

$$\dot{z} = z \bar{h}'(\hat{x}) \,, \tag{3.28}$$

where $\bar{h}'(\hat{x}) < 0$. We will return to this ODE in Section 3.6.5, where, in addition, we shall give an interpretation of the above results.

3.6.2 Algorithm (iii)

This represents the most general linear model that we will study. It will be straightforward to look at (ii) and (iv) once we have analysed this algorithm. Let $F_n(p)$ be the Laplace transform of the pdf of x_n, and $G(p)$ be that of the pdf of ξ_n. Then, $F_n(\alpha_n p)$ is the transform of the pdf of $\alpha_n x_n$, and $G(\gamma_n p)$ is that of the pdf of $\gamma_n \xi_n$. Following [3] and [4], we then have that

$$F_{n+1}(p) = F_n(\alpha_n p) G(\gamma_n p) \ . \tag{3.29}$$

Now, (3.29) may be expressed in closed form as follows

$$F_n(p) = F_0\left(p \prod_{j=0}^{n-1} \alpha_j\right) \prod_{k=0}^{n-1} G\left(\gamma_k p \prod_{i=k+1}^{n-1} \alpha_i\right) \ . \tag{3.30}$$

Assume that the pdf of x_0 is given by

$$f_{x_0}(z) = \delta(z - x_0) \ .$$

Then the pdf of $\prod_{j=0}^{n-1} \alpha_j x_0$ is

$$\left(\prod_{j=0}^{n-1} \alpha_j\right)^{-1} \delta\left(\frac{z - x_0 \prod_{j=0}^{n-1} \alpha_j}{\prod_{j=0}^{n-1} \alpha_j}\right) \ . \tag{3.31}$$

The Laplace transform of (3.31) is then

$$\int_0^\infty e^{-sx} \left(\prod_{j=0}^{n-1} \alpha_j\right)^{-1} \delta\left(\frac{x}{\prod_{j=0}^{n-1} \alpha_j} - x_0\right) dx = e^{-sx_0 \prod_{j=0}^{n-1} \alpha_j} \ . \tag{3.32}$$

Equations (3.30) and (3.32) give

$$F_0\left(p\prod_{j=0}^{n-1}\alpha_j\right) = e^{-px_0\prod_{j=0}^{n-1}\alpha_j}\ . \tag{3.33}$$

Whence, (3.30) reduces to

$$F_n(p) = e^{-px_0\prod_{j=0}^{n-1}\alpha_j}\prod_{k=0}^{n-1} G\left(\gamma_k p\prod_{i=k+1}^{n-1}\alpha_i\right)\ . \tag{3.34}$$

We shall choose a pdf of $\gamma_n\xi_n$ such that the product in (3.34) is easily computed. Following [3], one possible form, *which allows Gaussian and other more general distributions*, is

$$G(p) = e^{\beta p^\eta}\ ,\ \beta\in\mathbb{R}\ . \tag{3.35}$$

which yields

$$G\left(\gamma_k p\prod_{i=k+1}^{n-1}\alpha_i\right) = e^{\{\beta\gamma_k^\eta p^\eta(\prod_{i=k+1}^{n-1}\alpha_i)^\eta\}}\ .$$

Therefore, assuming that (3.35) holds, we get

$$\prod_{k=0}^{n-1} G\left(\gamma_k p\prod_{i=k+1}^{n-1}\alpha_i\right) = \exp\{\beta p^\eta\sum_{k=0}^{n-1}\gamma_k^\eta\left(\prod_{i=k+1}^{n-1}\alpha_i\right)^\eta\}\ . \tag{3.36}$$

Substituting (3.36) back into (3.34) yields

$$F_n(p) = \exp\{-px_0\prod_{j=0}^{n-1}\alpha_j + \beta p^\eta\sum_{k=0}^{n-1}\gamma_k^\eta\left(\prod_{i=k+1}^{n-1}\alpha_i\right)^\eta\}\ . \tag{3.37}$$

The density of x_n is thus given by

$$f_n(x) = \mathcal{L}^{-1}\{\exp\{-px_0 \prod_{j=0}^{n-1} \alpha_j + \beta p^\eta \sum_{k=0}^{n-1} \gamma_k^\eta \left(\prod_{i=k+1}^{n-1} \alpha_i \right)^\eta \}\} , \tag{3.38}$$

where $\mathcal{L}^{-1}$ denotes the inverse Laplace transform. For $\eta = 2$, (3.38) becomes

$$\begin{aligned} f_n(x) &= 0.5 \left\{ \frac{\pi}{\beta \sum_{k=0}^{n-1} \gamma_k^2 \left(\prod_{i=k+1}^{n-1} \alpha_i \right)^2} \right\}^{\frac{1}{2}} U\left(x - x_0 \prod_{j=0}^{n-1} \alpha_j \right) \times \\ &\quad \exp\left\{ - \frac{\left(x - x_0 \prod_{j=0}^{n-1} \alpha_j \right)^2}{\beta \sum_{k=0}^{n-1} \gamma_k^2 \left(\prod_{i=k+1}^{n-1} \alpha_i \right)^2} \right\} , \end{aligned} \tag{3.39}$$

where U is the Heaviside step function. From the analysis in equations (3.21)-(3.24), we recall that

$$\prod_{j=0}^{n-1} \alpha_j \to 0 \text{ as } n \to \infty . \tag{3.40}$$

Let us now examine the quantities in the denominator of (3.39), viz.

$$T_n \stackrel{def}{=} \beta \sum_{k=0}^{n-1} \gamma_k^2 \left(\prod_{i=k+1}^{n-1} \alpha_i \right)^2 . \tag{3.41}$$

From (3.21)-(3.24), it follows that

$$\left(\prod_{i=k+1}^{n-1} \alpha_i \right)^2 \sim \exp\left\{ \frac{2b}{1-\alpha} [(n-1)^{1-\alpha} - (k+1)^{1-\alpha}] \right\} ,$$

where we have set $\gamma_i \stackrel{def}{=} i^{-\alpha}$, $\alpha \in (0,1)$, and $b \stackrel{def}{=} \bar{h}'(\hat{x}) < 0$. Consequently, we have that

$$\begin{aligned} T_n &\sim \beta \sum_{k=0}^{n-1} \gamma_k^2 \exp\left\{\frac{2b}{1-\alpha}[(n-1)^{1-\alpha} - (k+1)^{1-\alpha}]\right\} \\ &= \beta \exp\left\{\frac{2b}{1-\alpha}(n-1)^{1-\alpha}\right\} \sum_{k=0}^{n-1} \gamma_k^2 \exp\left\{-\frac{2b}{1-\alpha}(k+1)^{1-\alpha}\right\} . \end{aligned}$$

The leading behaviour of the partial sum on the RHS of the preceding equation is found by examining the corresponding integral

$$I_n \stackrel{def}{=} \int_0^{n-1} (x+1)^{-2\alpha} \exp\left\{-\frac{2b}{1-\alpha}(x+1)^{1-\alpha}\right\} dx ,$$

where, without loss of generality, we have set $\gamma_k \stackrel{def}{=} (k+1)^{-\alpha}$. Integrating I_n by parts yields

$$I_n \sim -\frac{n^{-\alpha}}{2b} \exp\left\{-\frac{2b}{1-\alpha}n^{1-\alpha}\right\} ,$$

which implies that

$$\begin{aligned} T_n &\sim \beta \exp\left\{\frac{2b}{1-\alpha}(n-1)^{1-\alpha}\right\} \times \left\{-\frac{n^{-\alpha}}{2b} \exp\left(-\frac{2b}{1-\alpha}n^{1-\alpha}\right)\right\} \\ &= -\frac{\beta}{2b} n^{-\alpha} \longrightarrow 0 \text{ as } n \to \infty . \end{aligned} \tag{3.42}$$

We may express (3.39) as

$$f_n(x) = \begin{cases} 0.5 \times \left\{\frac{\pi}{T_n}\right\}^{\frac{1}{2}} \times \exp\left\{-\frac{\left(x - x_0 \prod_{j=0}^{n-1} \alpha_j\right)^2}{T_n}\right\} & , \quad x > x_0 \prod_{j=0}^{n-1} \alpha_j \\ 0.5 \times \{\frac{\pi}{T_n}\}^{0.5} & , \quad x = x_0 \prod_{j=0}^{n-1} \alpha_j \\ 0 & , \quad x < x_0 \prod_{j=0}^{n-1} \alpha_j \end{cases}$$

Furthermore, for $x > x_0 \prod_{j=0}^{n-1} \alpha_j$, (3.42) implies that

$$\lim_{n \to \infty} f_n(x) = 0 \ ,$$

since $\{\frac{\pi}{T_n}\}^{0.5}$ approaches infinity at a relatively much slower rate than $\exp\left\{-\frac{(x - x_0 \prod_{j=0}^{n-1} \alpha_j)^2}{T_n}\right\}$ approaches zero. On the other hand, when $x = x_0 \prod_{j=0}^{n-1} \alpha_j$, we have that

$$\lim_{n \to \infty} f_n(x) = \infty \ .$$

The above analysis suggests that $\{f_n\}_{n=1}^{\infty}$ converges to $\delta(x)$.

3.6.3 Algorithm (ii)

Two cases arise, viz.

(a) $\alpha > 1$ (unstable equilibrium), and

(b) $\alpha \in (0, 1)$ (stable equilibrium).

Before looking at these separate cases, we derive $f_n(x)$ for this algorithm. We make the same assumptions as in the determination of $f_n(x)$ for algorithm (iii) above.

We obtain

$$F_{n+1}(p) = F_n(\alpha p)G(\gamma_n p) .$$

which may be expressed in closed form as

$$F_n(p) = F_0(\alpha^n p) \prod_{k=0}^{n-1} G(\gamma_k \alpha^{n-k-1} p) .$$

The sought-after density is thus

$$f_n(x) = 0.5 \left\{ \frac{\pi}{T_n} \right\}^{\frac{1}{2}} U(x - \alpha^n x_0) \exp\left\{ -\frac{(x - \alpha^n x_0)^2}{T_n} \right\} , \tag{3.43}$$

where

$$T_n \stackrel{def}{=} \beta \alpha^{2n} \sum_{k=0}^{n-1} \gamma_k^2 \alpha^{-2(k+1)} . \tag{3.44}$$

Case (a): Recall that

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \ldots , \mid x \mid < 1 ,$$

which gives

$$-\ln(1-x) = x + \frac{x^2}{2} + \frac{x^3}{3} + \ldots , \mid x \mid < 1 .$$

In the special case $\gamma_k \stackrel{def}{=} (k+1)^{-\frac{1}{2}}$, we have that

$$\sum_{k=0}^{n-1} \gamma_k^2 x^{k+1} = -\ln(1-x) .$$

Now setting $x \stackrel{def}{=} \frac{1}{\alpha^2}$, and using the previous notation, we get

$$T_n = -\beta\alpha^{2n}\ln(1-\frac{1}{\alpha^2}) \to \infty \text{ . as } n \to \infty \text{ ,} \tag{3.45}$$

since $\ln(1-\frac{1}{\alpha^2}) < 0$. Thus, from (3.43) and using the same analysis as before, we have that

$$f_n(x) = 0.5\left\{\frac{\pi}{T_n}\right\}^{\frac{1}{2}} U(x-\alpha^n x_0)\exp\left\{-\frac{(x-\alpha^n x_0)^2}{T_n}\right\} \to 0 \text{ as } n \to \infty \text{ .} \tag{3.46}$$

Following our previous analysis, we conclude that $\{f_n\}_{n=1}^{\infty}$ converges pointwise to zero.

Case (b): Note that we can no longer employ the geometric sum used in Case (a), since $\frac{1}{\alpha^2} > 1$ in the present case. Equation (3.45) is not valid for this case. Let

$$T_n \stackrel{def}{=} \beta\alpha^{2n}\sum_{k=0}^{n-1}\gamma_k^2\left(\frac{1}{\alpha^2}\right)^{k+1} \text{ .}$$

In the special case $\gamma_k \stackrel{def}{=} (k+1)^{-\frac{1}{2}}$, and $x \stackrel{def}{=} \frac{1}{\alpha^2}$, we have that

$$T_n = \beta\alpha^{2n}\sum_{k=1}^{n}\frac{x^k}{k} \text{ .} \tag{3.47}$$

Now suppose that $f(k) \stackrel{def}{=} \frac{x^k}{k}$. Then, employing the Maclaurin-Euler sum formula gives

$$T_n \sim \beta\alpha^{2n}\int_1^n \frac{1}{t}e^{t\ln x}dt \text{ , for } x > 0 \text{ ,} \tag{3.48}$$

which may be integrated by parts to show that $T_n \to 0$ as $n \to \infty$. From (3.43), this suggests that $\{f_n\}_{n=1}^{\infty}$ converges to $\delta(x)$.

3.6.4 Algorithm (iv)

With the usual assumptions on the densities of ξ_n and x_0, the density of algorithm (iv) is given by

$$f_n(x) = \mathcal{L}^{-1}\left\{\exp\left(-px_0\prod_{j=0}^{n-1}\alpha_j + \beta p^{\eta}\sum_{k=0}^{n-1}\left(\prod_{i=k+1}^{n-1}\alpha_i\right)^{\eta}\right)\right\} ,$$

which, for $\eta = 2$, yields

$$f_n(x) = 0.5\left\{\frac{\pi}{T_n}\right\}^{\frac{1}{2}} U(x - x_0\prod_{j=0}^{n-1}\alpha_j)\exp\left\{-\frac{(x - x_0\prod_{j=0}^{n-1}\alpha_j)^2}{T_n}\right\} ,$$

where U is the Heaviside step function and

$$T_n \stackrel{def}{=} \beta\sum_{k=0}^{n-1}\left(\prod_{i=k+1}^{n-1}\alpha_i\right)^2 \sim -\frac{\beta n^{\alpha}}{2b} > 0 , \tag{3.49}$$

where α and b are as previously defined and the asymptotic expansion is arrived at in a way similar to that performed for algorithm (iii). Consequently, we obtain that

$$T_n \longrightarrow \infty , \text{ as } n \to \infty ,$$

and thus

$$f_n(x) \longrightarrow 0 , \text{ as } n \to \infty .$$

To close our analysis of linearised algorithms, we now consider the usual noise-driven algorithm

$$x_{n+1} = a_n x_n + \delta_n \xi_n \ , \tag{3.50}$$

where

$$a_n = 1 + \gamma_n \bar{h}'(\hat{x}) \ , \ \ \gamma_n = \frac{1}{(n+1)^r} \ , \text{ and } \ \delta_n = \frac{1}{(n+1)^s} \ , r \neq s \ .$$

In other words, the noise amplitude δ_n and the parameter a_n decay at different rates. The problem is to determine the stationary density of (3.50), assuming that we know the relative sizes of r and s. We have already derived the following results:

(a) $0 < r < 1$, $r = s \Rightarrow \{f_n\}_{n=1}^{\infty}$ converges to $\delta(x)$, in the distributional sense,

(b) $s = 0$ (i.e. unattenuated noise amplitude): $f_n(x) \overset{pointwise}{\longrightarrow} 0$, as $n \to \infty$.

Now, we seek to examine the long term behaviour of the density sequence $\{f_n\}_{n=1}^{\infty}$ of (3.50) for cases (c) and (d) below, inclusive of the special case.

(c) $r > s$, special case: $r = 2s$,

(d) $r < s$.

From previous results and with the usual assumptions on the densities of ξ_n and x_0, we have that the density of (3.50) is given by

$$f_n(x) = \mathcal{L}^{-1} \left\{ \exp \left(-p x_0 \prod_{j=0}^{n-1} a_j + \beta p^{\eta} \sum_{k=0}^{n-1} \delta_k^{\eta} \left(\prod_{i=k+1}^{n-1} a_i \right)^{\eta} \right) \right\} \ . \tag{3.51}$$

For $\eta = 2$, this yields

$$f_n(x) = 0.5 \left\{ \frac{\pi}{T_n} \right\}^{\frac{1}{2}} U(x - x_0 \prod_{j=0}^{n-1} \alpha_j) \exp \left\{ - \frac{(x - x_0 \prod_{j=0}^{n-1} \alpha_j)^2}{T_n} \right\} , \tag{3.52}$$

where U is the Heaviside function and

$$T_n \stackrel{def}{=} \beta \sum_{k=0}^{n-1} \delta_k^2 \left(\prod_{i=k+1}^{n-1} \alpha_i \right)^2 = \beta \sum_{k=0}^{n-1} (k+1)^{-2s} \left(\prod_{i=k+1}^{n-1} \alpha_i \right)^2 . \tag{3.53}$$

It can be shown that

$$T_n \sim - \frac{\beta}{2b} n^{r-2s} > 0 , \tag{3.54}$$

which implies that

$$T_n \to \begin{cases} -\frac{\beta}{2b} & , \ r = 2s \\ +\infty & , \ r > 2s \\ 0 & , \ r < 2s \end{cases} \tag{3.55}$$

as $n \to \infty$. This gives

$$f_n(x) \to \begin{cases} 0.5 \left\{ -\frac{2b\pi}{\beta} \right\}^{\frac{1}{2}} U(x) \exp \left\{ \frac{2bx^2}{\beta} \right\} & , \ r = 2s \\ 0 & , \ r > 2s \end{cases} \tag{3.56}$$

For $r < 2s$, the above results suggest that $\{f_n\}_{n=1}^{\infty}$ converges, in the distributional sense, to $\delta(x)$. This is an improvement on Chung's result (see Theorem 3.1 in Section 3.1) in the sense that we now have a tighter bound on the parameter r to guarantee that $x_n \to 0$ as $n \to \infty$.

3.6.5 Closing remarks

Algorithm (iii) may be expressed in the form

$$x_{n+1} = x_n + \gamma_n \bar{h}'(\hat{x})x_n + \gamma_n \xi_n \ ,$$

giving the associated ODE

$$\dot{z} = z\bar{h}'(\hat{x}) \ ,$$

where $\bar{h}'(\hat{x}) < 0$. Note that $\hat{z} = 0$ is the stable equilibrium of this ODE. It is straightforward to show that the ODE solution is given by

$$z(t) = z(0)e^{t\bar{h}'(\hat{x})} \to 0 \ , \ \text{as} \ t \to \infty \ .$$

Therefore from the results in Sections 3.6.1-3.6.4, we conclude that the behaviours of the ODE solutions and the sequence of densities are in accord only for algorithms (i) and (iii). In these cases, the stationary density of $\{x_n\}_{n=1}^{\infty}$ is concentrated about $\hat{z} = 0$, the stable fixed point of the associated ODE.

We have derived stationary densities of linearised algorithms of the general form given by algorithm (iii). Note that the stationary densities so derived are *localised*, about $\hat{x}$. To obtain a full stationary density of $\{x_n\}_{n=1}^{\infty}$, one needs to piece together the various local stationary densities which correspond to different stable equilibria of the associated ODE. If we assume that ξ_n is characterised by a pdf of the form

$$G(p) = e^{\beta p^2} \ , \ \beta \in \mathbb{R} \ ,$$

then, depending on the amplitude of ξ_n, we obtain two different types of local stationary densities, viz. $\delta(x)$, and an exact replica of the pdf of ξ_n. In the case of the $\delta(x)$ stationary density, it is possible that different masses may be 'sitting' on each of the stable equilibria. The almost intractable problem is to quantify masses that are associated with each localised $\delta(x)$ distribution. If achieved, this would give an answer to the question: *to which one of the stable equilibria is the algorithm most likely to converge?* However, mention must be made that this is a very difficult problem, beyond the scope of this thesis.

3.7 Densities of nonlinear algorithms

Consider an algorithm in the form

$$x_{n+1} = S_n(x_n) + \gamma_n \xi_n \ , \tag{3.57}$$

where S explicitly depends on n (refer to equation (3.1) to see an example of such an S). In order to study this problem, let us first look at a simpler version, viz.

$$x_{n+1} = S(x_n) + \gamma_n \xi_n \ , \tag{3.58}$$

where S does *not* depend on n. From (3.58) above, let

$$z_n = \gamma_n \xi_n \ . \tag{3.59}$$

The density of z_n is given by

$$G_n(x) = \frac{1}{\gamma_n} g(\frac{x}{\gamma_n}) , \tag{3.60}$$

where g is the pdf of the random variable ξ_n. Suppose that $f_n \in L_D^1(X)$ is the density of x_n. By (3.58), x_{n+1} is the sum of two independent random variables. Note that $S(x_n)$ and z_n are independent since in calculating $x_1, ..., x_n$, we only need $\xi_0, ..., \xi_{n-1}$. Let $w : \mathbb{R} \to \mathbb{R}$ be an arbitrary, bounded, measurable function. Then, the mathematical expectation of $w(x_{n+1})$ is

$$E[w(x_{n+1})] = \int_{\mathbb{R}} w(x) f_{n+1}(x) dx . \tag{3.61}$$

Furthermore, using (3.58) and the fact that the joint density of (x_n, z_n) is $\frac{1}{\gamma_n} f_n(y) g(\frac{z}{\gamma_n})$, we have that

$$E[w(x_{n+1})] = \frac{1}{\gamma_n} \int_{\mathbb{R}} \int_{\mathbb{R}} w(S(y) + z) f_n(y) g(\frac{z}{\gamma_n}) dy dz . \tag{3.62}$$

Now using the change of variables $x = S(y) + z$, $y = y$, we obtain the Jacobian

$$\frac{\partial(z, y)}{\partial(x, y)} = 1 .$$

Thus, (3.62) is transformed to

$$E[w(x_{n+1})] = \frac{1}{\gamma_n} \int_{\mathbb{R}} \int_{\mathbb{R}} w(x) f_n(y) g(\frac{x - S(y)}{\gamma_n}) dx dy . \tag{3.63}$$

Equating (3.61) and (3.63) yields

$$\int_{\mathbb{R}} w(x)f_{n+1}(x) = \int_{\mathbb{R}} w(x)\left\{\frac{1}{\gamma_n}\int_{\mathbb{R}} f_n(y)g(\frac{x-S(y)}{\gamma_n})dy\right\}dx \ ,$$

which gives

$$f_{n+1}(x) = \frac{1}{\gamma_n}\int_{\mathbb{R}} f_n(y)g(\frac{x-S(y)}{\gamma_n})dy \ . \tag{3.64}$$

From (3.64), define the operator $\overline{P}_n : L^1_D \to L^1_D$ by

$$\overline{P}_n f(x) = \frac{1}{\gamma_n}\int_{\mathbb{R}} f(y)g(\frac{x-S(y)}{\gamma_n})dy \ , \tag{3.65}$$

for $f \in L^1_D$.

We now digress for a moment to introduce the *Koopman operator* (see [5] for details), which is adjoint to the Frobenius-Perron operator.

Definition 3.3 *Let (X, Λ, μ) be a measure space, $S : X \to X$ a non-singular transformation, and $f \in L^\infty$. The operator $U : L^\infty \to L^\infty$ defined by*

$$Uf(x) = f(S(x))$$

is called the Koopman operator with respect to S.

As a result of the non-singularity of S, U is well-defined since $f_1(x) = f_2(x)$ *a.e.* implies that $f_1(S(x)) = f_2(S(x))$ *a.e.*. Listed below are some essential properties of U:

1. $U(\lambda_1 f_1 + \lambda_2 f_2) = \lambda_1 U f_1 + \lambda_2 U f_2$ for all $f_1, f_2 \in L^\infty$, $\lambda_1, \lambda_2 \in \mathbb{R}$;

2. For every $f \in L^\infty$,

$$\|Uf\|_{L^\infty} \leq \|f\|_{L^\infty} .$$

We note that [5] refers to any operator satisfying this property (e.g. P) as *contractive*, which is not the usual definition of contractive operators.

3. For every $f_1 \in L^1_D$, $f_2 \in L^\infty$.

$$< Pf_1, f_2 > = < f_1, Uf_2 >$$

so that U is adjoint to the Frobenius-Perron operator.

Suppose that S is non-singular. Therefore the Frobenius-Perron and Koopman operators, P and U respectively, corresponding to S exist. Furthermore, let

$$h_{n,x}(y) \stackrel{def}{=} g(\frac{x-y}{\gamma_n}) . \tag{3.66}$$

Then (3.65) and (3.66) yield

$$\begin{aligned} \overline{P}_n f(x) &= \frac{1}{\gamma_n} \int_{\mathbb{R}} f(y) h_{n,x}(S(y)) dy \\ &= \gamma_n^{-1} < f, U h_{n,x} > , \text{ since } h_{n,x}(S(y)) = U h_{n,x}(y) \\ &= \gamma_n^{-1} < Pf, h_{n,x} > , \text{ since } P \text{ and } U \text{ are adjoint operators} \\ &= \frac{1}{\gamma_n} \int_{\mathbb{R}} g(\frac{x-y}{\gamma_n}) Pf(y) dy . \end{aligned} \tag{3.67}$$

Using the change of variable $\frac{x-y}{\gamma_n} = t$, (3.67) becomes

$$\overline{P}_n f(x) = \int_{\mathbb{R}} g(t) P f(x - \gamma_n t) dt \ . \tag{3.68}$$

Consequently, since $\gamma_n \to 0$ as $n \to \infty$, we should expect that, in some sense,

$$lim_{n\to\infty} \overline{P}_n f(x) = \int_{\mathbb{R}} g(t) P f(x) dt = P f(x) \ . \tag{3.69}$$

The theorem below ascertains the validity of (3.69).

Theorem 3.5 *For the system defined by (3.58),*

$$lim_{n\to\infty} \| \overline{P}_n f - P f \|_1 = 0 \ , \ for \ all \ f \in L^1_D(X) \ ,$$

where P is the Frobenius-Perron operator corresponding to S and $\overline{P}_n$ is given by (3.68).

The proof of this theorem is similar to that of Theorem 10.6.1 in [5], and we choose to omit it (see [5] for details). A corollary of the above theorem is given below.

Corollary 3.1 *Suppose that S and g are given and that we have a sequence $\{f_n\}$ generated by $f_{n+1} = \overline{P}_n f_n$, and such that*

$$\| f_n - f_* \|_1 \to 0 \ as \ n \to \infty \ . \tag{3.70}$$

Then f_* is a stationary density for the Frobenius-Perron operator corresponding to S. That is, $P f_* = f_*$.

Proof of Corollary

Write

$$\overline{P}_n f_* = f_{n+1} + \overline{P}_n(f_* - f_n)\,. \tag{3.71}$$

Then

$$\begin{aligned}
\|\overline{P}_n f_* - f_*\|_1 &= \|f_{n+1} + \overline{P}_n(f_* - f_n) - f_*\|_1 \\
&\leq \|f_{n+1} - f_*\|_1 + \|\overline{P}_n(f_* - f_n)\|_1 \\
&\leq \|f_{n+1} - f_*\|_1 + \|f_* - f_n\|_1 \text{ , using the contractivity [5] of } \overline{P}_n \\
&\leq \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon \text{ , for all } n > N \text{ ,}
\end{aligned}$$

which implies that

$$\|\overline{P}_n f_* - f_*\|_1 \to 0 \text{ as } n \to \infty. \tag{3.72}$$

However, Theorem 3.5 also gives

$$\|\overline{P}_n f_* - P f_*\|_1 \to 0 \text{ as } n \to \infty. \tag{3.73}$$

Combining (3.72) and (3.73) yields

$$P f_* = f_*\,. \tag{3.74}$$

□

We return to our original algorithm given in (3.57), viz.

$$x_{n+1} = S_n(x_n) + \gamma_n \xi_n \ , \tag{3.75}$$

where now S explicitly depends on n. As before, the noise amplitude is damped and $\{\xi_n\}$ is a sequence of *i.i.d.* random variables with common density g. Suppose that the Frobenius-Perron operator associated with S_n is P_n, and is such that

$$P_n \overset{\|\cdot\|_1}{\to} P \text{ as } n \to \infty \ , \tag{3.76}$$

where P is a limiting operator. In a similar manner to the derivation at the beginning of this section, it may be shown that the Frobenius-Perron operator corresponding to (3.75) is given by

$$\overline{P}_n f(x) = \int_{\mathbb{R}} g(t) P_n f(x - \gamma_n t) dt \ , \text{ for all } f \in L^1_D \ .$$

The following issue arises: *Assuming (3.76), does it follow that* $\overline{P}_n \overset{\|\cdot\|_1}{\to} P$ *as* $n \to \infty$*?* In what follows, we state and prove an equivalent of Theorem 3.5. The following result [5] will be useful:

Lemma 3.1 *For every* $f \in L^1$, $I \subseteq \mathbb{R}$ *bounded or not,*

$$\lim_{h \to 0} \int_I |f(x+h) - f(x)| dx = 0 \ .$$

The proof of this result may be found in [5].

Theorem 3.6 *For the system defined by (3.75),*

$$lim_{n\to\infty}\|\overline{P}_n f - P_n f\|_1 = 0 \text{ , for all } f \in L^1_D(X) \text{ ,}$$

where P_n is the Frobenius-Perron operator corresponding to S_n and $\overline{P}_n$ is the operator corresponding to (3.75).

Proof

Write

$$\overline{P}_n f(x) - P_n f(x) = \int_{\mathbb{R}} g(t)\{P_n f(x - \gamma_n t) - P_n f(x)\}dt \text{ .}$$

Pick an arbitrarily small $\delta > 0$. Since g is an integrable function on $\mathbb{R}$, there must exist an $r > 0$ such that

$$\int_{|y|\geq r} g(y)dy \leq \frac{\delta}{4} \text{ .} \qquad (*)$$

To compute an upper bound for $\|\overline{P}_n f - P_n f\|_1$, we proceed as follows.

$$\begin{aligned} \|\overline{P}_n f - P_n f\|_1 &\leq \int_{\mathbb{R}}\int_{\mathbb{R}} g(y)|P_n f(x - \gamma_n y) - P_n f(x)|dxdy \\ &\stackrel{def}{=} I_1 + I_2 \text{ ,} \end{aligned}$$

where

$$I_1 \stackrel{def}{=} \int_{\mathbb{R}}\int_{|y|\leq r} g(y)|P_n f(x - \gamma_n y) - P_n f(x)|dydx \text{ ,}$$

and

$$I_2 \stackrel{def}{=} \int_{\mathbb{R}}\int_{|y|\geq r} g(y)|P_n f(x - \gamma_n y) - P_n f(x)|dydx \text{ .}$$

We consider each of these integrals in turn. First, since $P_n f$ is integrable, Lemma 3.1 implies that there exists an $N > 0$ such that

$$\int_{\mathbb{R}} |P_n f(x - \gamma_n y) - P_n f(x)| dx < \frac{\delta}{2} , \text{ for all } n \geq N \text{ and for } |y| \leq r .$$

Hence

$$I_1 \leq \frac{\delta}{2} \int_{|y| \leq r} g(y) dy \leq \frac{\delta}{2} \int_{\mathbb{R}} g(y) dy = \frac{\delta}{2} . \tag{3.77}$$

For I_2, we use the triangle inequality to write

$$I_2 \leq \int_{\mathbb{R}} \int_{|y| \geq r} g(y) P_n f(x) dy dx + \int_{\mathbb{R}} \int_{|y| \geq r} g(y) P_n f(x - \gamma_n y) dy dx . \tag{3.78}$$

Using the change of variables $v = y$ and $z = x - \gamma_n y$, we get the Jacobian

$$\frac{\partial(x, y)}{\partial(z, v)} = 1 ,$$

which yields

$$\begin{aligned} \int_{\mathbb{R}} \int_{|y| \geq r} g(y) P_n f(x - \gamma_n y) dy dx &= \int_{\mathbb{R}} \int_{|v| \geq r} g(v) P_n f(z) dv dz \\ &= \int_{|v| \geq r} g(v) dv \times \int_{\mathbb{R}} P_n f(z) dz \\ &= \int_{|v| \geq r} g(v) dv \text{ (since } \int_{\mathbb{R}} P_n f(z) dz = 1) \\ &\leq \frac{\delta}{4} \text{ (by assumption (*) above) ,} \end{aligned} \tag{3.79}$$

and

$$\begin{aligned}\int_{\mathbb{R}} \int_{|y|\geq r} g(y) P_n f(x) dy dx &= \int_{|y|\geq r} g(y) dy \times \int_{\mathbb{R}} P_n f(x) dx \\ &\leq \frac{\delta}{4} . \end{aligned} \tag{3.80}$$

Hence, combining (3.78), (3.79) and (3.80) gives

$$I_2 \leq \frac{\delta}{2} . \tag{3.81}$$

Therefore, (3.77) and (3.81) imply that

$$\lim_{n\to\infty} \|\overline{P}_n f - P_n f\|_1 = 0 .$$

□

An equivalent of Corollary 3.1 is stated and proved below.

Corollary 3.2 *Suppose that S_n and g are given and that we have a sequence $\{f_n\}$ generated by $f_{n+1} = \overline{P}_n f_n$ and such that $f_n \to f_*$ as $n \to \infty$. In other words, given an arbitrary $\epsilon > 0$, there exists an N such that*

$$\|f_n - f_*\|_1 < \frac{\epsilon}{2} , \text{ for all } n > N . \tag{3.82}$$

Furthermore, suppose that P_n, the Frobenius-Perron operator associated with S_n, is such that $P_n \to P$ as $n \to \infty$, where P is a limiting operator. Then f_* is a stationary density for P, viz. $Pf_* = f_*$.

Proof of Corollary

Write

$$\overline{P}_n f_* = f_{n+1} + \overline{P}_n(f_* - f_n) .$$

This gives

$$\begin{aligned} \|\overline{P}_n f_* - f_*\|_1 &\leq \|f_{n+1} - f_*\|_1 + \|\overline{P}_n(f_* - f_n)\|_1 \\ &\leq \|f_{n+1} - f_*\|_1 + \|f_* - f_n\|_1 \\ &\leq \epsilon \text{ , for all } n > N \text{ ,} \end{aligned}$$

which implies that

$$\|\overline{P}_n f_* - f_*\|_1 \to 0 \text{ as } n \to \infty. \tag{3.83}$$

Now, from (3.76) and Theorem 3.6, we get that

$$\|P_n f_* - P f_*\|_1 \to 0 \text{ and } \|\overline{P}_n f_* - P_n f_*\|_1 \to 0 \text{ as } n \to \infty .$$

This leads to

$$\|\overline{P}_n f_* - P f_*\|_1 \to 0 \text{ as } n \to \infty ,$$

which, in combination with (3.83), gives $P f_* = f_*$. □

This result tells us nothing about a specific f_*. In fact, in the particular case given by (3.75), f_* is not unique since P is the Frobenius-Perron operator associated with the identity map I introduced in Section 3.1.

3.7.1 An example where the associated ODE has two stable and one unstable equilibria

Consider the deterministic algorithm

$$x_{n+1} = x_n + \gamma_n \bar{h}(x_n) \stackrel{def}{=} S_n(x_n) , \tag{3.84}$$

where we assume that

$$\bar{h}(x_n) = -x_n(x_n^2 + 2x_n + 0.5) .$$

The associated ODE is thus given by

$$\dot{x} = -x(x^2 + 2x + 0.5) \stackrel{def}{=} g(x) .$$

This ODE has two stable equilibria, $x_1^* = 0$ and $x_3^* = -1.707$, and a single unstable equilibrium, viz. $x_2^* = -0.293$. The multipliers of the stable equilibria are given by $g'(0) = -0.5$ and $g'(-1.707) = -2.414$, implying that -1.707 is the relatively more attractive of the two. The graph depicting (3.84) is shown in Fig 3.2.

The following question arises: *to which one of x_1^* or x_3^* is $\{x_n\}$ likely to converge as $n \to \infty$, given an arbitrary starting value x_0?*

Note that with the given $\bar{h}$, we have that $S_n : [0, x_3^*] \to [0, x_3^*]$. Thus, we focus on the dynamics inside the box outlined in Fig 3.2. The Frobenius-Perron operator associated with the transformation S_n is formulated below. Pick an interval $[x, 0] \subset$

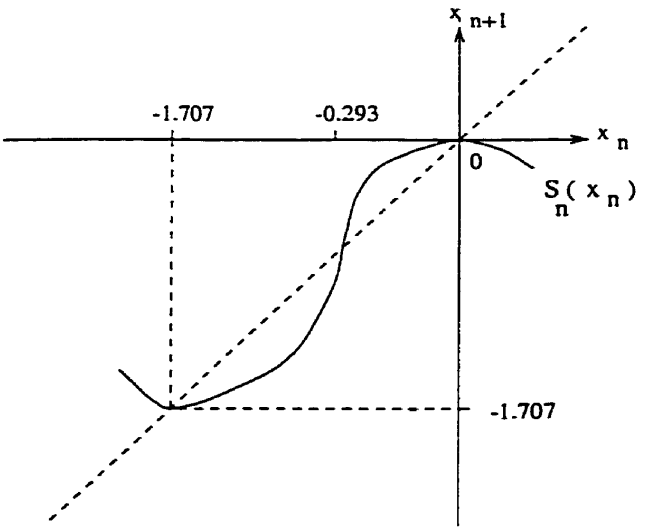

Figure 3.2: x_{n+1} with a cubic $\bar{h}$

$[-1.707, 0]$ so that the counter-image of $[x, 0]$ under S_n is given by

$$\begin{aligned} S_n^{-1}([x,0]) &= [g_{1,n}(x), 0] \cup [g_{2,n}(x), -1 + (2\gamma_n)^{-0.5}\sqrt{2+\gamma_n}] \cup \\ &\quad [g_{3,n}(x), -1 - (2\gamma_n)^{-0.5}\sqrt{2+\gamma_n}] , \end{aligned} \tag{3.85}$$

where $g_{i,n}(x)$, $i = 1...3$ are the three solutions of the equation $x = S_n(y)$. Following [5], the Frobenius-Perron operator corresponding to S_n is defined by

$$\begin{aligned} P_n f(x) &= \frac{d}{dx} \int_{S_n^{-1}([x,0])} f(u)du \\ &= \sum_{j=1}^{3} f(g_{j,n}(x)) g'_{j,n}(x) . \end{aligned} \tag{3.86}$$

This tells us how S_n transforms a given density f into a new one $P_n f$. To see how (3.86) works, pick an initial density $f(x) = 0.586$ for $x \in [-1.707, 0]$, and zero

elsewhere. Then (3.86) becomes

$$P_n f(x) = 0.586 \sum_{j=1}^{3} g'_{j,n}(x) \text{ , for } x \in [-1.707, 0] \text{ .} \tag{3.87}$$

Substituting this expression for $P_n f$ in place of f on the right hand side of (3.86) yields

$$P_2 P_1 f(x) = 0.586 \sum_{k=1}^{3} \left\{ \sum_{j=1}^{3} g'_{j,1}(g_{k,2}(x)) \right\} g'_{k,2}(x) \text{ .} \tag{3.88}$$

Similarly, we obtain that

$$P_3 P_2 P_1 f(x) = 0.586 \sum_{l=1}^{3} \sum_{k=1}^{3} \sum_{j=1}^{3} g'_{j,1}(g_{k,2}(g_{l,3}(x))) g'_{k,2}(g_{l,3}(x)) g'_{l,3}(x) \text{ .}$$

Given an initial density $f \in L^1_D$, we seek to determine $\lim_{n\to\infty} P_n ... P_1 f$. This limit, if it exists, gives the sought-after stationary density. With this in mind, we return to the cubic equation $x = S_n(y)$, which gives

$$y^3 + 2y^2 + \frac{1}{2\gamma_n}(\gamma_n - 2)y + \frac{x}{\gamma_n} = 0 \text{ .} \tag{3.89}$$

We wish to find the three roots of (3.89), viz. $g_{1,n}(x)$, $g_{2,n}(x)$, and $g_{3,n}(x)$. Now, we may write (3.89) as

$$y^3 + a_1 y^2 + a_2 y + a_3 = 0 \text{ ,}$$

where $a_1 = 2$, $a_2 = \frac{1}{2\gamma_n}(\gamma_n - 2)$,and $a_3 = \frac{x}{\gamma_n}$. Let $Q = \frac{1}{9}(3a_2 - a_1^2)$, $R = \frac{1}{54}(9a_1 a_2 - 27a_3 - 2a_1^3)$, $S = (R + \sqrt{(Q^3 + R^2)})^{\frac{1}{3}}$, and $T = (R - \sqrt{(Q^3 + R^2)})^{\frac{1}{3}}$. Furthermore, let $D = Q^3 + R^2$ be the discriminant. Now, it may be shown that,

for $x \in [-1.707, 0]$, we have that $D < 0$ if and only if

$$(1458x + 972 + 378\gamma_n)^2 < (62.5 + \frac{225}{\gamma_n} + \frac{270}{\gamma_n^2} + \frac{108}{\gamma_n^3}) .$$

Clearly, for any fixed $x \in [-1.707, 0]$, the LHS of this inequality approaches a fixed positive real number as $n \to \infty$. On the other hand, the RHS of the inequality approaches $+\infty$ as $n \to \infty$. Therefore, there will be a "cross-over" value, γ_N say, for which the above inequality holds for all $n \geq N$. From the definitions of Q and R above, we have that

$$\frac{R}{\sqrt{-Q^3}} = \beta_n(7\gamma_n + 18 + 27x) ,$$

where

$$\beta_n \stackrel{def}{=} -0.5\sqrt{\frac{\gamma_n}{(2.5\gamma_n + 3)^3}} \to 0^- \text{ as } n \to \infty. \tag{3.90}$$

Now, since there exists an N such that $D < 0$ (for $x \in [-1.707, 0]$) for all $n \geq N$, it is well-known [6] that all the roots of (3.89) are real and distinct *for all* $n \geq N$. These roots are given by

$$g_{1,n}(x) \stackrel{def}{=} \frac{2}{3}\sqrt{\frac{6 + 5\gamma_n}{2\gamma_n}} \cos(\frac{\theta}{3}) - \frac{2}{3} , \tag{3.91}$$

$$g_{2,n}(x) \stackrel{def}{=} \frac{2}{3}\sqrt{\frac{6 + 5\gamma_n}{2\gamma_n}} \cos(\frac{\theta}{3} + \frac{2\pi}{3}) - \frac{2}{3} , \tag{3.92}$$

and

$$g_{3,n}(x) \stackrel{def}{=} \frac{2}{3}\sqrt{\frac{6 + 5\gamma_n}{2\gamma_n}} \cos(\frac{\theta}{3} + \frac{4\pi}{3}) - \frac{2}{3} , \tag{3.93}$$

where

$$\theta \stackrel{def}{=} \arccos\{\beta_n(7\gamma_n + 18 + 27x)\} . \tag{3.94}$$

It may be shown that only $g_{3,n}(x)$ maps the closed interval $[-1.707, 0]$ to itself, for all n. Hence, $g_{3,n}(x)$ is the only root we need to use in evaluating the Frobenius-Perron operator corresponding to S_n. From (3.86), the sought-after operator is given by

$$P_n f(x) = f(g_{3,n}(x)) g'_{3,n}(x) , \tag{3.95}$$

for an arbitrary initial density $f \in L^1_D$. We choose an initial pdf given by: $f(x) = 0.586$ for $x \in [-1.707, 0]$, and zero elsewhere. Then (3.95) becomes

$$P_n f(x) = 0.586 g'_{3,n}(x) . \tag{3.96}$$

Hence we get the following iterates

$$P_2 P_1 f(x) = 0.586 g'_{3,1}(g_{3,2}(x)) g'_{3,2}(x) ,$$

and

$$P_3 P_2 P_1 f(x) = 0.586 g'_{3,1}(g_{3,2}(g_{3,3}(x))) g'_{3,2}(g_{3,3}(x)) g'_{3,3}(x) .$$

In a similar way, higher iterates of $P_n ... P_1 f(x)$ may be computed. Now, analytically determining $\lim_{n\to\infty} P_n ... P_1 f(x)$ is intractable. However, a numerical approach is feasible. Fig 3.3 shows $P_1 f(x)$, $P_2 P_1 f(x)$, and $P_4 P_3 P_2 P_1 f(x)$ respectively, for $x \in [-1.707, 0]$. The learning parameter used is $\gamma_n \stackrel{def}{=} (n+3)^{-0.99}$. The plots suggest that $P_n ... P_1 f(x)$ converges, in the distributional sense, to two Dirac distri-

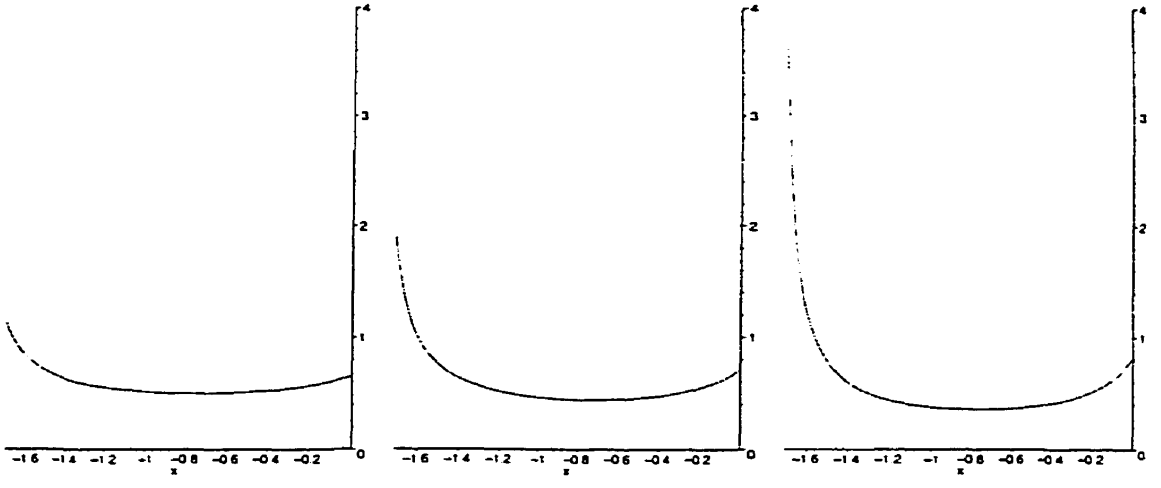

Figure 3.3: *left:* $P_1 f(x)$ *centre:* $P_2 P_1 f(x)$ *right:* $P_4 P_3 P_2 P_1 f(x)$

butions centred at the two stable equilibria as $n \to \infty$. It should be noted that these two Dirac distributions emerge at different rates. The one located at -1.707 has a relatively bigger mass compared to the one located at the origin. This distribution of mass is clearly related to the magnitudes of the multipliers of the two stable equilibria. However, it is not clear how one might go about actually quantifying these masses. Finally, note that these numerical results validate the idea of linearisation of $\bar{h}(x)$ about each stable point before deriving the local densities, as previously done.

3.7.2 A stationary density for the perturbed operator $\overline{P}_n$?

This section investigates the behaviour of the perturbed system

$$x_{n+1} = S_n(x_n) + \gamma_n \xi_n \,, \tag{3.97}$$

where $S_n(x_n) \stackrel{def}{=} x_n + \gamma_n \bar{h}(x_n)$ and $\bar{h}(x_n)$ is as defined in the previous example. As usual, assume that $\{\xi_n\}$ is a sequence of *i.i.d.* random variables, each with density g. The density of the random variable $\gamma_n \xi_n$ is $\gamma_n^{-1} g(\gamma_n^{-1} x)$. The Frobenius-Perron operator associated with (3.97) is given by [5]

$$\overline{P}_n f(x) = \gamma_n^{-1} \int_{\mathbb{R}} f(y) g(\frac{x - S_n(y)}{\gamma_n}) dy \ , \tag{3.98}$$

where $f \in L_D^1$ is an arbitrary initial density of all the possible initial states of (3.97). From (3.98), we may find $\overline{P}_1 f$, $\overline{P}_2 \overline{P}_1 f$,...,$\overline{P}_n \overline{P}_{n-1} ... \overline{P}_1 f$. The issue is to characterise

$$\lim_{n \to \infty} \overline{P}_n \overline{P}_{n-1} ... \overline{P}_1 f \ , \tag{3.99}$$

provided that such a limit exists. Now assume that

$$g(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \ , \tag{3.100}$$

a mean zero and unit variance Gaussian distribution. Furthermore, assume that the initial density f is given by

$$f(x) = \begin{cases} \frac{1}{3} & , \ x \in [-2, 1] \\ 0 & , \ \text{elsewhere} \ . \end{cases} \tag{3.101}$$

Computation of $\overline{P}_n \overline{P}_{n-1} ... \overline{P}_1 f$

The analytic computation of the iterates $\overline{P}_n ... \overline{P}_1 f$, for $n > 1$, is not usually possible. However, a numerical approach to this computation is available. Nonetheless, the

latter approach is not that straightforward and easy, as it involves the evaluation of iterated integrals of increasing complexity. We now outline one possible way of performing the numerical computations of iterates of the operator $\overline{P}_n$. First, compute

$$\overline{P}_1 f(x) = \gamma_1^{-1} \int_{A_1} f(y) g(\frac{x - S_1(y)}{\gamma_1}) dy \ , \tag{3.102}$$

where $A_1 = [-15, 15] \subset \mathbb{R}$. Then approximate (3.102) by a polynomial, using the MAPLE *interp* function. This polynomial has a finite support. The choice of the size of this support is arbitrary. Denote this polynomial fit by $w_1(x)$. Next, compute

$$\overline{P}_2 f(x) = \gamma_2^{-1} \int_{A_1} f(y) g(\frac{x - S_2(y)}{\gamma_2}) dy \ ,$$

which gives

$$\overline{P}_2 \overline{P}_1 f(x) \approx \gamma_2^{-1} \int_{A_1} w_1(x) g(\frac{x - S_2(y)}{\gamma_2}) dy \ . \tag{3.103}$$

In a similar way, compute $\overline{P}_3 \overline{P}_2 \overline{P}_1 f$, $\overline{P}_4 \overline{P}_3 \overline{P}_2 \overline{P}_1 f$, etc. Even after the polynomial fits have been determined, the (symbolic) numerical integration of iterates of the perturbed operator is still relatively intensive and consumes a substantial amount of CPU time. Figs 3.4-3.5 show the first four of these iterates. The learning parameter used is $\gamma_n = (n+4)^{-0.999}$.

3.7.3 Comments on numerics

The plots in Figs 3.4-3.5 suggest that the stationary density of the sequence of perturbed operators consists of two Dirac distributions, μ_1 and μ_2 say, centred at $\bar{x}_1 = -1.707$ and $\bar{x}_2 = 0$ respectively, the two locally asymptotically stable equilibria

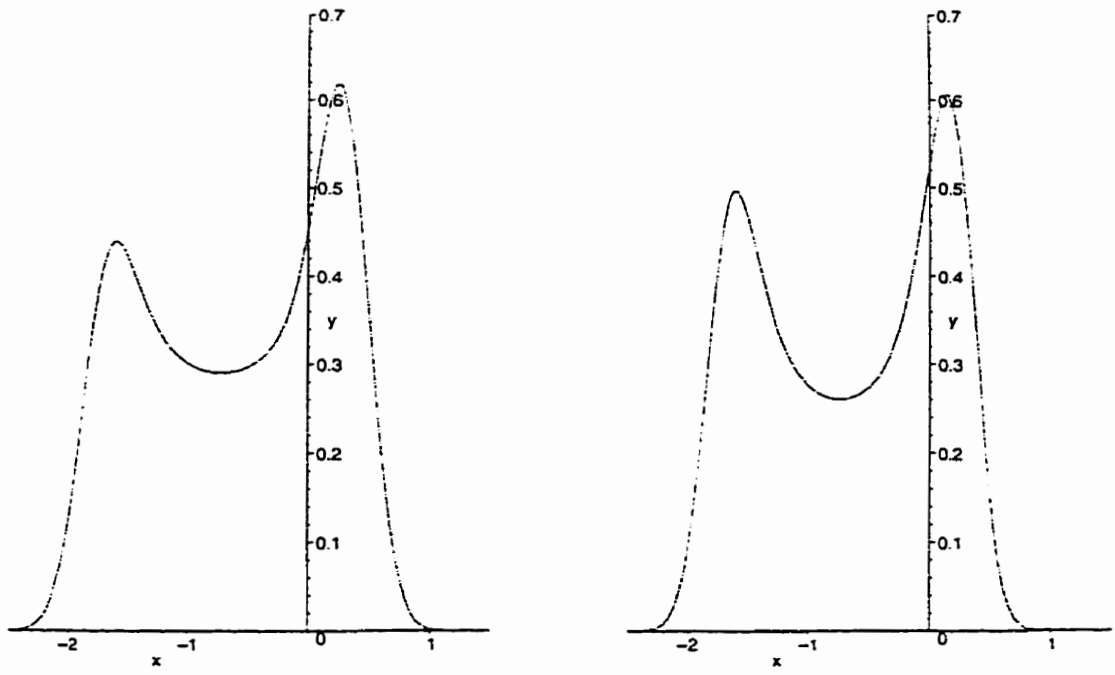

Figure 3.4: *left*: $\overline{P}_1 f(x)$ *right*: $\overline{P}_2 \overline{P}_1 f(x)$

of the associated ODE. We may express the stationary density as $\mu \stackrel{def}{=} \mu_1 + \mu_2$, where the component Dirac masses are

$$m_1 \stackrel{def}{=} \sup_{\psi \in D(X)} \left\{ \int_{-\infty}^{\infty} \psi(t) \mu_1(t - \bar{x}_1) dt \right\} = \sup_{\psi \in D(X)} \{\psi(\bar{x}_1)\} ,$$

and

$$m_2 \stackrel{def}{=} \sup_{\psi \in D(X)} \left\{ \int_{-\infty}^{\infty} \psi(t) \mu_2(t - \bar{x}_2) dt \right\} = \sup_{\psi \in D(X)} \{\psi(\bar{x}_2)\} ,$$

such that $m_1 + m_2 = 1$, and where $D(X)$ is a suitable space of test functions, for example $D(X) \stackrel{def}{=} Lip_1^+(X)$. In this example, it is not immediately obvious what the values of m_1 and m_2 will be, primarily as a result of the difficulty of constructing the set of functions $\{\psi |\ \psi \in Lip_1^+(X)\}$. Also, as previously discussed

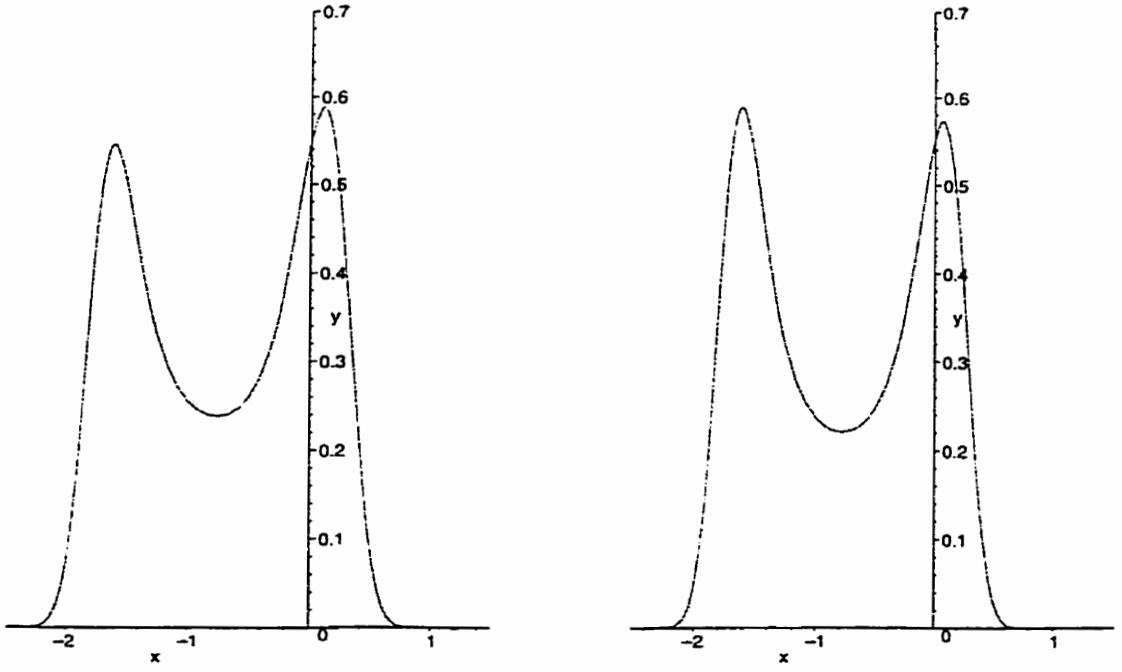

Figure 3.5: *left*: $\overline{P}_3\overline{P}_2\overline{P}_1 f(x)$ *right*: $\overline{P}_4\overline{P}_3\overline{P}_2\overline{P}_1 f(x)$

in Section 3.5, analysis in the space $\overline{D}'_+(X)$ is daunting. Finally, the result in Section 3.7.2 is in accord with the one found in Section 3.7.1, thereby justifying the idea of linearising $\bar{h}$ about a stable equilibrium prior to performing a local analysis of the stationary density. The local stationary densities obtained before, via the linearisation procedure, are Dirac distributions centred at the stable equilibria of the associated system of ODEs.

3.8 Closing remarks

Chapter 3 has primarily been a systematic study of densities of both nonlinear and linearised algorithms (see Sections 3.6 and 3.7). Our results suggest that densities

of linearised algorithms studied here do reproduce the same local characteristics as the nonlinear equations from which they are derived.

We recall that Mackey and Lasota [5] studied algorithms of the form

$$x_{n+1} = S(x_n) + \epsilon\xi_n \ , \ \ 0 < \epsilon \ll 1 \ ,$$

where $S(x_n)$ is *not* explicitly dependent on n, and where $\{\xi_n\}$ is a sequence of *i.i.d.* random variables. Our work in this direction is new, because we look at more general algorithms, i.e.

$$x_{n+1} = S_n(x_n) + \gamma_n\xi_n \ ,$$

where now $S_n(x_n)$ is explicitly dependent on n, and $\{\gamma_n\}$ is a decreasing-to-zero sequence of positive real numbers such that $\sum_n \gamma_n = \infty$. Thus, our proof of Theorem 3.6 is new, and more general than that in [5].

Using the notion of density, we have been able to improve Chung's result (see Theorem 3.1) by computing a tighter upper bound on the parameter r. Finally, it is important to emphasize that the numerical experiments performed in this Chapter are indispensable - they provide tremendous insights into a difficult problem.

Chapter 4

Some Generalisations of Ljung's Problem

4.1 Introduction

In Chapter 2, a doubly-triangular *pdf* was assumed for the input signals $\{\varphi(t)\}$. All the analysis performed in that chapter was based on the characteristics of this single *pdf*. Needless to say, this particular *pdf* is very basic and does simplify a lot of the computations carried out in Chapter 2. A particular feature of the *pdf* which makes the computations straightforward is the fact that the two component triangles do *not* overlap. The present chapter generalises the problem to include overlapping components of the *pdf*. Essentially, we look at two examples, viz.: (a) a *pdf* consisting of two overlapping triangles, and (b) one consisting of two Gaussian distributions. The primary purpose of this chapter is to investigate the effect of changing the *pdf* of $\{\varphi(t)\}$ on the dynamics of the discrete algorithm, inside the

invariant region $\overline{D}$. As before, both analytical and numerical techniques will be adopted.

4.2 The *overlapping* doubly-triangular *pdf*

Suppose that the input signals, to be classified by Ljung's algorithm, are distributed according to the *pdf*

$$f_{\lambda,\sigma}(\varphi) = \begin{cases} \frac{1-\lambda}{\sigma_1^2}(\varphi + 1 + \sigma_1) & , \ -1-\sigma_1 < \varphi < -1 \\ \frac{1-\lambda}{\sigma_1^2}(\sigma_1 - 1 - \varphi) & , \ -1 < \varphi < 1-\sigma_2 \\ \frac{1-\lambda}{\sigma_1^2}(\sigma_1 - 1 - \varphi) + \frac{\lambda}{\sigma_2^2}(\varphi - 1 + \sigma_2) & , \ 1-\sigma_2 < \varphi < -1+\sigma_1 \\ \frac{\lambda}{\sigma_2^2}(\varphi - 1 + \sigma_2) & , \ -1+\sigma_1 < \varphi < 1 \\ \frac{\lambda}{\sigma_2^2}(1 + \sigma_2 - \varphi) & , \ 1 < \varphi < 1+\sigma_2 \\ 0 & , \ \text{elsewhere} \end{cases} \tag{4.1}$$

where $\lambda \in [0,1]$ and $\sigma_1,\ \sigma_2 \in \mathbb{R}^+$. Furthermore, assume that $\sigma_1 > 2 - \sigma_2$. The latter ensures that the two component triangles of (4.1) are overlapping in the interval $[1-\sigma_2, -1+\sigma_1]$, as depicted in Fig 4.1. Note that the left triangle has area $1-\lambda$, while the one on the right hand side has area λ. The *pdf* under consideration is a superposition of these two component triangles. As usual, it is assumed that $\{\varphi(t)\}$ is a sequence of *i.i.d.* random variables.

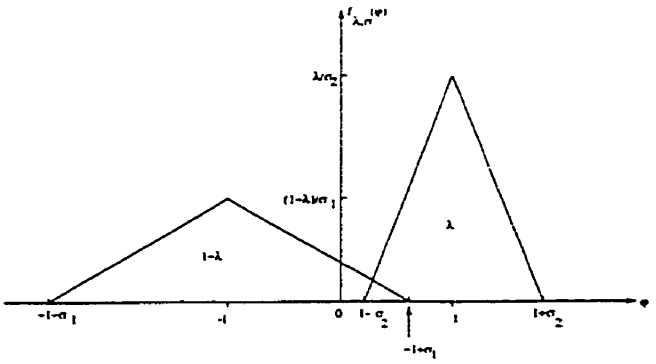

Figure 4.1: pdf of the random variable to be classified by classifier.

4.2.1 Computation of the associated system of ODEs

As before, we obtain the following associated system of ODEs

$$\begin{aligned} \dot{x}_A &= -x_A \int_{-1-\sigma_1}^{c} f_{\lambda,\sigma}(\varphi)d\varphi + \int_{-1-\sigma_1}^{c} \varphi f_{\lambda,\sigma}(\varphi)d\varphi \,, \\ \dot{x}_B &= -x_B \int_{c}^{1+\sigma_2} f_{\lambda,\sigma}(\varphi)d\varphi + \int_{c}^{1+\sigma_2} \varphi f_{\lambda,\sigma}(\varphi)d\varphi \,, \end{aligned} \tag{4.2}$$

where $c \stackrel{def}{=} 0.5(x_A + x_B)$. Note that it is desirable to have $c \in [1-\sigma_2, -1+\sigma_1]$.

4.2.2 Stability and bifurcation analysis of the associated system of ODEs

As in Chapter 2, we now investigate the stability and bifurcations of equilibria of (4.2) as λ varies, for fixed σ_1 and σ_2. The system of ODEs specified in (4.2) has the equilibrium point $(\bar{x}_A, \bar{x}_B)$, implicitly given by

$$(\bar{x}_A,\ \bar{x}_B) = \left(\frac{P(\bar{c})}{Q(\bar{c})}\ ,\ \frac{E_\varphi - P(\bar{c})}{1-Q(\bar{c})} \right)\ ;$$

where $\bar{c} \stackrel{def}{=} 0.5(\bar{x}_A + \bar{x}_B)$, and

$$\begin{aligned} P(c) &\stackrel{def}{=} \int_{-1-\sigma_1}^{c} \varphi f_{\lambda,\sigma}(\varphi) d\varphi \ , \\ Q(c) &\stackrel{def}{=} \int_{-1-\sigma_1}^{c} f_{\lambda,\sigma}(\varphi) d\varphi \ , \text{ and} \\ E_\varphi &\stackrel{def}{=} \int_{-1-\sigma_1}^{1+\sigma_2} \varphi f_{\lambda,\sigma}(\varphi) d\varphi \ . \end{aligned} \tag{4.3}$$

Define

$$\begin{aligned} H_{\lambda,\sigma}(c) &\stackrel{def}{=} \frac{P(c)}{Q(c)} + \frac{E_\varphi - P(c)}{1 - Q(c)} - 2c \\ &= \frac{P(c)[1 - Q(c)] + Q(c)[E_\varphi - P(c)] - 2cQ(c)[1 - Q(c)]}{Q(c)[1 - Q(c)]} \ , \end{aligned} \tag{4.4}$$

where $P(c)$, $Q(c)$, and E_φ are as defined in (4.3). To analyse the parameter-dependent dynamics of the zeros of (4.4), we need only look at the numerator

$$F_{\lambda,\sigma}(c) \stackrel{def}{=} P(c)[1 - Q(c)] + Q(c)[E_\varphi - P(c)] - 2cQ(c)[1 - Q(c)] \ .$$

Figs 4.2-4.3 (*left*) show plots of $F_{\lambda,\sigma}(c)$ versus c, for some selected parameter values. It is evident that a saddle-node (S-N for short) bifurcation occurs as $\lambda \to 0^+$, leading to the birth of two new equilibria. We also note that, in contrast to Ljung's non-overlapping doubly-triangular *pdf* example of Chapter 2, no bifurcation of equilibria occurs as $\lambda \to 1^-$. This will be exemplified later in this section. Instead, the only zero of $F_{\lambda,\sigma}(c)$ shifts towards $\bar{c} = 1$, as $\lambda \to 1^-$.

For $\lambda > 0.056$ and for $\{\sigma_1, \sigma_2\} = \{1.0, 1.5\}$, the function $F_{\lambda,\sigma}(c)$ has only one root, namely $\bar{c}_3$. This root, which is parameterised by λ, *always* exists, as $\lambda \to 0^+$.

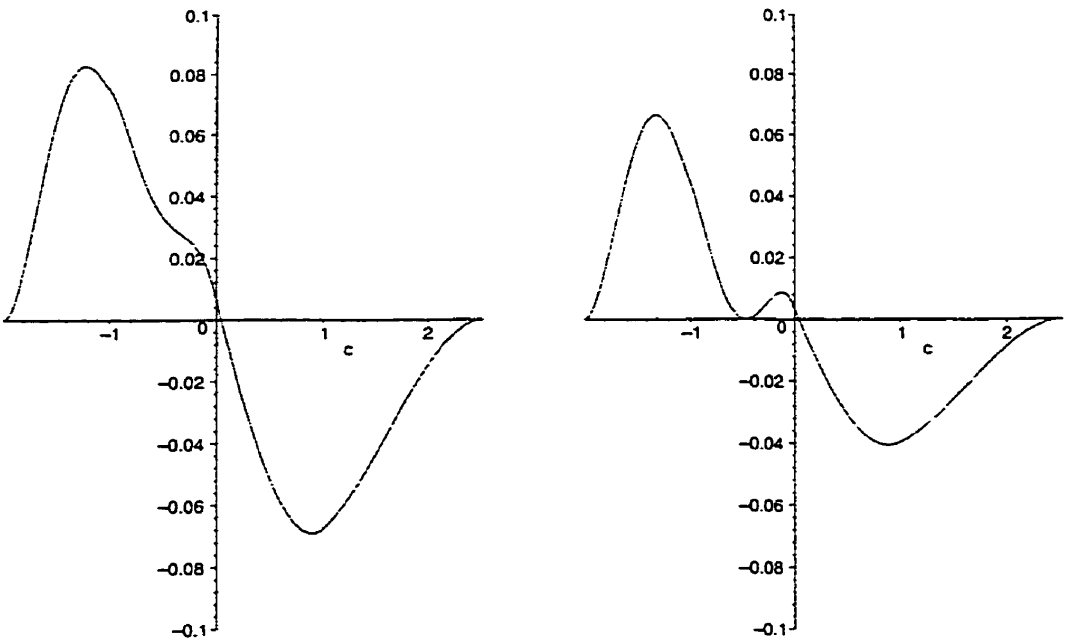

Figure 4.2: $F_{\lambda,\sigma}(c)$ with *left*: $\lambda = 0.1$, and *right*: $\lambda = 0.056$. In both cases, $\sigma_1 = 1.0$ and $\sigma_2 = 1.5$.

Numerically, at $\lambda \approx 0.056$, a S-N bifurcation occurs. In other words, a new zero appears at $\bar{c}_0 \approx -0.483$. As $\lambda \to 0^+$, $\bar{c}_0$ branches into two new zeros, viz. $\{\bar{c}_1, \bar{c}_2\}$. This is depicted in the bifurcation diagram Fig 4.3 (*right*), which is generated as follows: Fix $\{\sigma_1, \sigma_2\} = \{1.0, 1.5\}$, and solve $F_{\lambda,\sigma}(c) = 0$ for $c(\lambda)$, where $\lambda \in [0, 1]$. Then plot $c(\lambda)$ versus λ. Note that in Fig 4.3 (*right*) we have only considered a small 'window' of λ values, i.e. $\lambda \in (0, 0.056)$, essentially to accentuate the emergence of the two new zeros $\bar{c}_1(\lambda)$ and $\bar{c}_2(\lambda)$. The root $\bar{c}_3(\lambda) \approx 0.040$ for $\lambda \in (0, 0.056)$, and is not shown in Fig 4.3 (*right*).

It is important to point out that $\bar{c}_3$ does not lose its stability with the emergence of $\bar{c}_1$ and $\bar{c}_2$. At $\lambda = 0.03$, we have that $\{\bar{c}_1, \bar{c}_2\} = \{-0.830, -0.162\}$. Consequently, we now have a total of three *different* roots of $F_{\lambda,\sigma}(c)$, viz. $\{\bar{c}_1, \bar{c}_2, \bar{c}_3\} =$

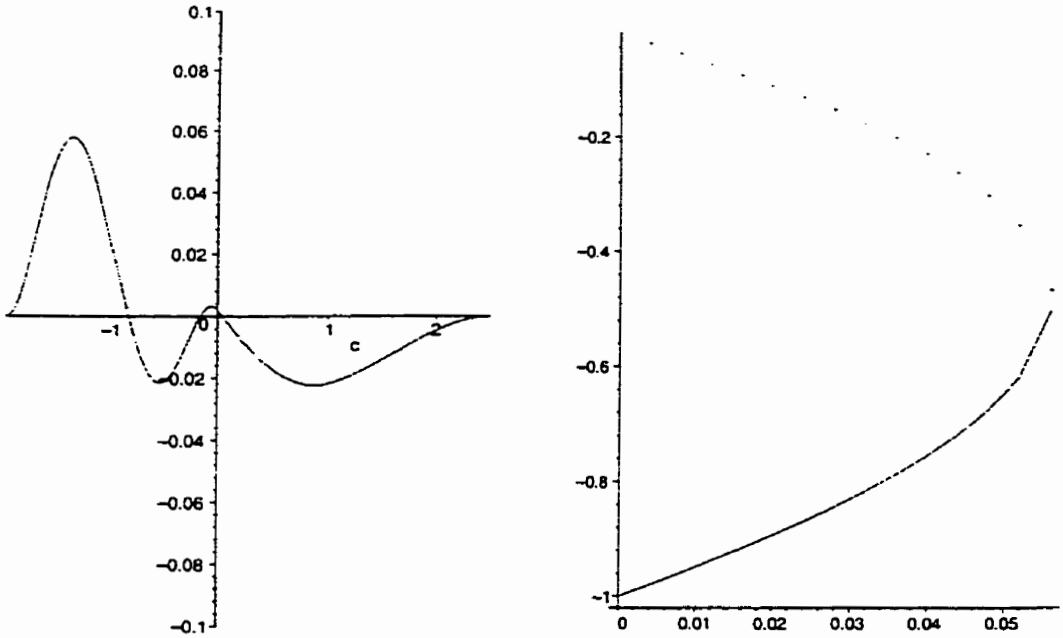

Figure 4.3: *left*: $F_{\lambda,\sigma}(c)$ with $\lambda = 0.03$, $\sigma_1 = 1.0$, and $\sigma_2 = 1.5$ *right*: Plot of zeros, $\bar{c}_1(\lambda)$ and $\bar{c}_2(\lambda)$, of $F_{\lambda,\sigma}(c)$ with $\sigma_1 = 1.0$, $\sigma_2 = 1.5$, for $\lambda \in (0, 0.056)$. A line drawn upward from $\lambda = 0.03$ will cut the curve at the two zeros of $F_{\lambda,\sigma}(c)$ shown on the left figure.

$\{-0.830, -0.162, 0.040\}$. The corresponding equilibria are $(x_{11}, x_{12}) = (-1.2347, -0.4254)$, $(x_{21}, x_{22}) = (-1.0113, 0.6875)$, and $(x_{31}, x_{32}) = (-0.9983, 1.0791)$, respectively. It may be shown, via the computation of Jacobian matrices for (4.2), that (x_{11}, x_{12}) and (x_{31}, x_{32}) are locally asymptotically stable, with eigenvalues $\{-0.5492, -0.1250\}$ and $\{-0.9682, -0.0243\}$, respectively. Similarly, (x_{21}, x_{22}) is a saddle point, with eigenvalues $\{-0.8946, 0.0318\}$. Now, assuming that $c \in [-1 + \sigma_1, 1]$, (4.2) simplifies to

$$\dot{x}_A = -\frac{(1-\lambda)}{2\sigma_1^2} x_A[1 - (1+\sigma_1)^2] - \frac{(1-\lambda)}{\sigma_1} x_A[1 + \sigma_1]$$

$$
\begin{aligned}
&+\frac{(1-\lambda)}{2\sigma_1^2}x_A[(1-\sigma_2)^2-1]-\frac{(1-\lambda)}{\sigma_1^2}x_A(-1+\sigma_1)(2-\sigma_2)\\
&-\frac{x_A}{2}[-\frac{(1-\lambda)}{\sigma_1^2}+\frac{\lambda}{\sigma_2^2}][(-1+\sigma_1)^2-(1-\sigma_2)^2]\\
&-\frac{(1-\lambda)}{\sigma_1^2}x_A(-1+\sigma_1)(-2+\sigma_1+\sigma_2)-\frac{\lambda}{\sigma_2^2}x_A(-1+\sigma_2)(-2+\sigma_1+\sigma_2)\\
&-\frac{\lambda}{2\sigma_2^2}x_A[c^2-(-1+\sigma_1)^2]-\frac{\lambda}{\sigma_2^2}x_A(-1+\sigma_2)(c+1-\sigma_1)\\
&+\frac{(1-\lambda)}{3\sigma_1^2}[-1+(1+\sigma_1)^3]+\frac{(1-\lambda)}{2\sigma_1^2}(1+\sigma_1)[1-(1+\sigma_1)^2]\\
&+\frac{1}{3}[-\frac{(1-\lambda)}{\sigma_1^2}+\frac{\lambda}{\sigma_2^2}][(-1+\sigma_1)^3-(1-\sigma_2)^3]\\
&+\frac{1}{2}[\frac{(1-\lambda)}{\sigma_1^2}(-1+\sigma_1)+\frac{\lambda}{\sigma_2^2}(-1+\sigma_2)][(-1+\sigma_1)^2-(1-\sigma_2)^2]\\
&+\frac{\lambda}{3\sigma_2^2}[c^3-(-1+\sigma_1)^3]+\frac{\lambda}{2\sigma_2^2}(-1+\sigma_2)[c^2-(-1+\sigma_1)^2]\,,
\end{aligned}
$$

and

$$
\begin{aligned}
\dot{x}_B \;=\; &-\frac{\lambda}{2\sigma_2^2}x_B(1-c^2)-\frac{\lambda}{\sigma_2^2}x_B(-1+\sigma_2)(1-c)\\
&+\frac{\lambda}{2\sigma_2^2}x_B[(1+\sigma_2)^2-1]-\frac{\lambda}{\sigma_2}x_B(1+\sigma_2)\\
&+\frac{\lambda}{3\sigma_2^2}(1-c^3)+\frac{\lambda}{2\sigma_2^2}(-1+\sigma_2)(1-c^2)\\
&-\frac{\lambda}{3\sigma_2^2}[(1+\sigma_2)^3-1]+\frac{\lambda}{2\sigma_2^2}(1+\sigma_2)[(1+\sigma_2)^2-1]\,. \qquad (4.5)
\end{aligned}
$$

Returning to an analysis of the limiting case $\lambda \to 1^-$, we consider the following example. Assume that $\sigma_1 = 1.00$, $\sigma_2 = 1.50$, and $\lambda = 0.99$. Thus, (4.5) yields the

Jacobian matrix

$$J(x_A, x_B) = \begin{pmatrix} F_{11}(x_A, x_B) & F_{21}(x_A, x_B) \\ F_{12}(x_A, x_B) & F_{22}(x_A, x_B) \end{pmatrix} , \tag{4.6}$$

where

$$\begin{aligned} F_{11}(x_A, x_B) &\stackrel{def}{=} -0.065 + 2.5 \times 10^{-11}(x_A + x_B)^2 - 0.055(x_A + x_B) \\ &\quad -0.110x_A(x_A + x_B + 1) , \\ F_{21}(x_A, x_B) &\stackrel{def}{=} -0.110x_A(x_A + x_B + 1) + 0.055(x_A + x_B)^2 + 0.055(x_A + x_B) , \\ F_{12}(x_A, x_B) &\stackrel{def}{=} 0.110x_B(x_A + x_B + 1) - 0.055(x_A + x_B)^2 - 0.055(x_A + x_B) , \end{aligned}$$

and

$$F_{22}(x_A, x_B) \stackrel{def}{=} -0.935 - 2.5 \times 10^{-11}(x_A + x_B)^2 + 0.055(x_A + x_B) + 0.110x_B(x_A + x_B + 1) .$$

Also, using the above parameter values, we get the following equilibria for (4.5): $(x_{11}, x_{12}) = (-2.794, -5.109)$, $(x_{21}, x_{22}) = (0.439, 1.471)$, and $(x_{31}, x_{32}) = (0.787, 2.138)$. However, since (4.5) holds only for $c \in [-1+\sigma_1, 1] = [0, 1]$, the only valid equilibrium point is (x_{21}, x_{22}), with eigenvalues $\{-0.1680, -0.5018\}$ and corresponding eigenvectors $\left\{ \begin{pmatrix} 0.7568 \\ 0.6536 \end{pmatrix} , \begin{pmatrix} -0.6536 \\ 0.7568 \end{pmatrix} \right\}$. As discussed earlier in this Section, recall that $F_{\lambda,\sigma}(c)$ has only one zero as $\lambda \to 1^-$. Thus, since we are taking $\lambda = 0.99$, there is only one equilibrium in the invariant region $\overline{D} : -1 - \sigma_1 < x_A \leq x_B < 1 + \sigma_2 = \overline{D} : -2 < x_A \leq x_B < 2.5$. In light of this, we deduce that (x_{21}, x_{22}) is globally

asymptotically stable. Thus, following the Kushner-Clark theorem [1], we conclude that the algorithm converges *wp1* to (x_{21}, x_{22}).

4.2.3 Numerical Simulations

In a manner similar to Chapter 2, it may be shown that the invariant region for (4.2) is given by $\overline{D} : -1 - \sigma_1 < x_A \leq x_B < 1 + \sigma_2$. Figs 4.4-4.5 show simulations of $c(t)$ for specified parameter values and initial conditions. In all the plots, we set $\sigma_1 = 1.0$, $\sigma_2 = 1.5$, and $\gamma_n = (n + 10)^{-\alpha}$, where $\alpha = 0.25$. The free variables are λ and the initial condition $(x_A(0), x_B(0))$. Note that 10^5 iterations of the algorithm are performed. and that only one "run" is shown. A different sequence $\{\varphi_n\}$ is used for each simulation of the algorithm. In Fig 4.4, when $\lambda = 0.1$, the associated ODE has only one stable equilibrium, and thus, in accordance with the Kushner-Clark theorem [1], we expect the algorithm to converge *wp1* to (x_{31}, x_{32}). On the other hand, when $\lambda = 0.03$, a S-N bifurcation has occurred, leading to the emergence of a second stable equilibrium, viz. (x_{11}, x_{12}). In this case, as has been previously pointed out in Chapter 2, the Kushner-Clark theorem is not applicable.

The numerics suggest that, for $\lambda \leq 0.03$, Ljung's algorithm converges, with a relatively high frequency, to (x_{31}, x_{32}). Note that, because (x_{11}, x_{12}) and (x_{31}, x_{32}) are in relative proximity, the algorithm is likely to oscillate between the two equilibria, until γ_n is sufficiently small. This is why we need a comparatively high number of iterations, i.e. 10^5, in this example.

Finally, we track the evolution (inside $\overline{D}$) of a unit mass, as outlined in the algorithm introduced in Chapter 2. We use the same parameters and initial conditions

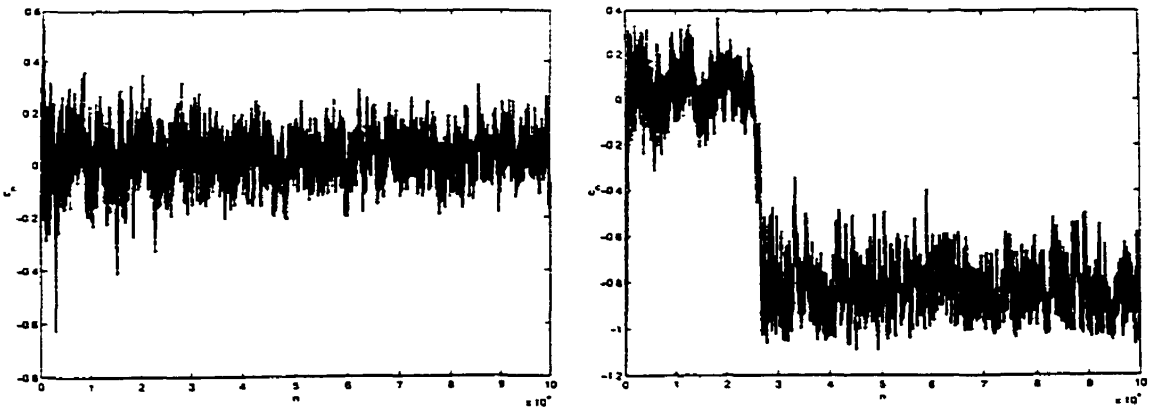

Figure 4.4: typical time histories of $c(n)$, with *left*: $\lambda = 0.1$ *right*: $\lambda = 0.03$. In both cases, $(x_A(0), x_B(0)) = (-0.1, 0.1)$.

as in Figs 4.4-4.5. The invariant region $\overline{D}$ is discretised into $npts \times npts$ cells, where $npts = 100$. Then, 1000 iterations of the algorithm are performed. The results of these simulations are shown in Figs 4.6-4.7.

For the case $\lambda = 0.03$, the numerics suggest that, given any $(x_A(0), x_B(0)) \in \overline{D}$, Ljung's algorithm converges, with a relatively high frequency, to the stable equilibrium point $(-1.2347, -0.4254)$. These simulation results also suggest that the density of mass surrounding each of the two stable equilibria is almost *independent* of the grid point used to initialise the algorithm.

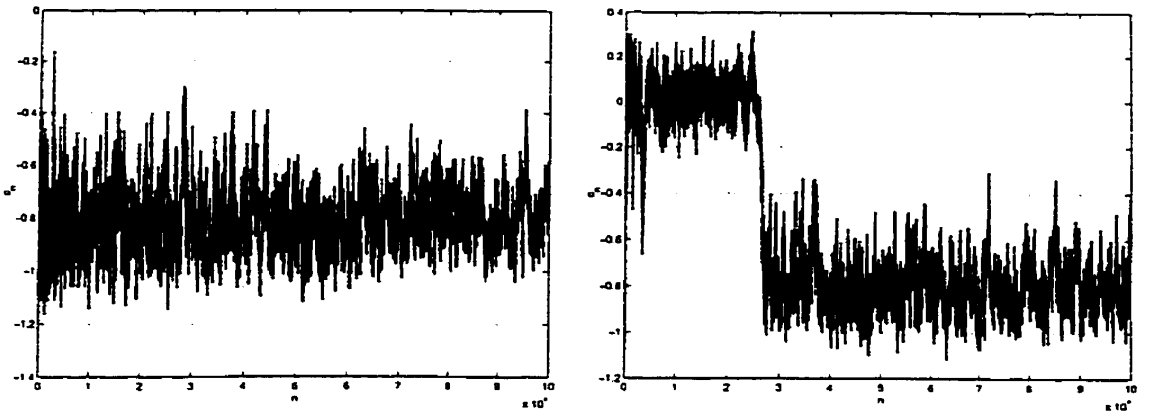

Figure 4.5: typical time histories of $c(n)$, with *left*: $(x_A(0), x_B(0)) = (-1.2347, -0.4254)$ *right*: $(x_A(0), x_B(0)) = (-0.9983, 1.0791)$. In both cases, $\lambda = 0.03$.

4.2.4 Degeneracy at $\lambda = 1$

When $\lambda = 1$, the *pdf* given in (4.1) degenerates to

$$f_{\lambda,\sigma}(\varphi) = \begin{cases} \frac{1}{\sigma_2^2}(\varphi - 1 + \sigma_2) & , \quad 1 - \sigma_2 \leq \varphi < 1 \\ \frac{1}{\sigma_2^2}(1 - \varphi + \sigma_2) & , \quad 1 \leq \varphi < 1 + \sigma_2 \\ 0 & , \quad \text{otherwise} \end{cases} \tag{4.7}$$

This is illustrated in Fig 4.8 (*left*). As usual, we seek to determine the long-term behaviour of Ljung's algorithm, and hence $c(t)$. First, we obtain the associated system of ODEs, viz.:

$$\begin{aligned} \dot{x}_A &= -x_A \int_{-1-\sigma_1}^{c} f_{\lambda,\sigma}(\varphi) d\varphi + \int_{-1-\sigma_1}^{c} \varphi f_{\lambda,\sigma}(\varphi) d\varphi \,, \\ \dot{x}_B &= -x_B \int_{c}^{1+\sigma_2} f_{\lambda,\sigma}(\varphi) d\varphi + \int_{c}^{1+\sigma_2} \varphi f_{\lambda,\sigma}(\varphi) d\varphi \,. \end{aligned} \tag{4.8}$$

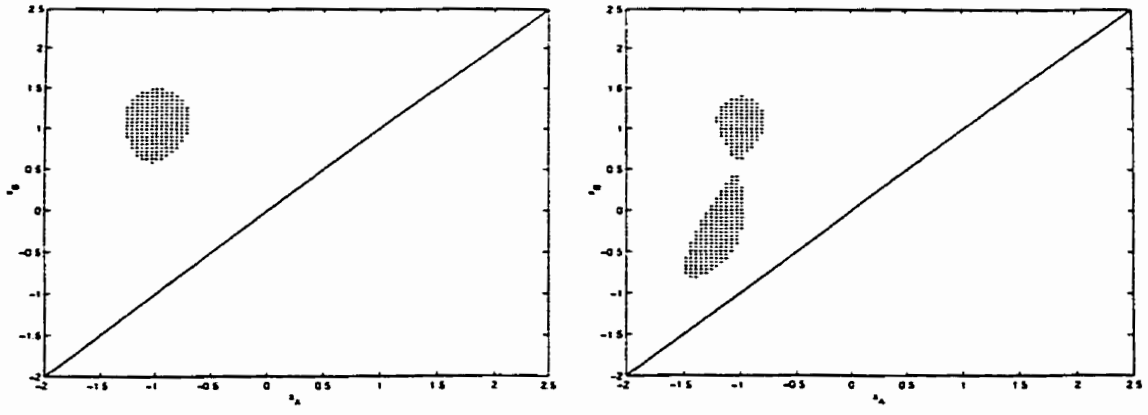

Figure 4.6: Grid points in $\overline{D}$ whose associated masses are greater than $\bar{t} = 0.001$ after 1000 iterations of the algorithm described in Chapter 2, with *left*: $\lambda = 0.1$, and *right*: $\lambda = 0.03$. In both cases, the unit mass is initialised at the grid point $(-0.1, 0.1)$.

Simplifying this system of ODEs yields

$$\dot{x}_A = \begin{cases} 0\ , \quad c < 1 - \sigma_2 \\ \frac{1}{3\sigma_2^2}[c^3 - (1-\sigma_2)^3] + \frac{1}{2\sigma_2^2}(\sigma_2 - 1 - x_A)[c^2 - (1-\sigma_2)^2] \\ -\frac{1}{\sigma_2^2}x_A(\sigma_2 - 1)(c - 1 + \sigma_2)\ , \quad 1 - \sigma_2 < c \leq 1 \\ \frac{1}{3\sigma_2^2}[1 - (1-\sigma_2)^3] + \frac{1}{2\sigma_2^2}(\sigma_2 - 1 - x_A)[1 - (1-\sigma_2)^2] - \frac{1}{\sigma_2}(\sigma_2 - 1)x_A \\ -\frac{1}{3\sigma_2^2}(c^3 - 1) + \frac{1}{2\sigma_2^2}(x_A + 1 + \sigma_2)(c^2 - 1) - \\ \frac{1}{\sigma_2^2}(1 + \sigma_2)(c - 1)x_A\ , \quad 1 < c \leq 1 + \sigma_2 \\ \frac{1}{3\sigma_2^2}[1 - (1-\sigma_2)^3] + \frac{1}{2\sigma_2^2}(\sigma_2 - 1 - x_A)[1 - (1-\sigma_2)^2] - \frac{1}{\sigma_2}(\sigma_2 - 1)x_A \\ -\frac{1}{3\sigma_2^2}[(1+\sigma_2)^3 - 1] + \frac{1}{2\sigma_2^2}(\sigma_2 + 1 + x_A)[(1+\sigma_2)^2 - 1] - \\ \frac{1}{\sigma_2}(1 + \sigma_2)x_A\ , \quad c \geq 1 + \sigma_2 \end{cases} \tag{4.9}$$

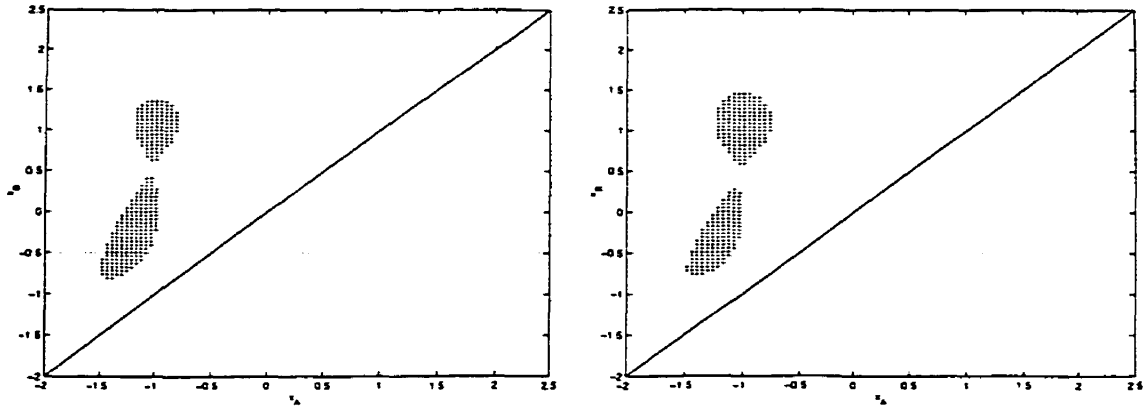

Figure 4.7: Grid points in $\overline{D}$ whose associated masses are greater than $\bar{t} = 0.001$ after 1000 iterations of the algorithm, with $\lambda = 0.03$. The unit mass is initialised at the point *left*: $(x_A, x_B) = (-1.2347, -0.4254)$, and *right*: $(x_A, x_B) = (-0.9983, 1.0791)$.

and

$$\dot{x}_B = \begin{cases} \frac{1}{3\sigma_2^2}[1-(1-\sigma_2)^3] + \frac{1}{2\sigma_2^2}(\sigma_2 - 1 - x_B)[1-(1-\sigma_2)^2] - \frac{1}{\sigma_2}(\sigma_2 - 1)x_B \\ -\frac{1}{3\sigma_2^2}[(1+\sigma_2)^3 - 1] + \frac{1}{2\sigma_2^2}(\sigma_2 + 1 + x_B)[(1+\sigma_2)^2 - 1] \\ -\frac{1}{\sigma_2}(\sigma_2 + 1)x_B \ , \quad c < 1 - \sigma_2 \\ \frac{1}{3\sigma_2^2}(1-c^3) + \frac{1}{2\sigma_2^2}(\sigma_2 - 1 - x_B)(1-c^2) - \frac{1}{\sigma_2^2}(\sigma_2 - 1)(1-c)x_B \\ -\frac{1}{3\sigma_2^2}[(1+\sigma_2)^3 - 1] + \frac{1}{2\sigma_2^2}(\sigma_2 + 1 + x_B)[(1+\sigma_2)^2 - 1] \\ -\frac{1}{\sigma_2}(1+\sigma_2)x_B \ , \quad 1 - \sigma_2 \leq c < 1 \\ -\frac{1}{3\sigma_2^2}[(1+\sigma_2)^3 - c^3] + \frac{1}{2\sigma_2^2}(\sigma_2 + 1 + x_B)[(1+\sigma_2)^2 - c^2] \\ -\frac{1}{\sigma_2^2}(1+\sigma_2)(1+\sigma_2 - c)x_B \ , \quad 1 \leq c < 1 + \sigma_2 \\ 0 \ , \ \text{otherwise} \ . \end{cases} \tag{4.10}$$

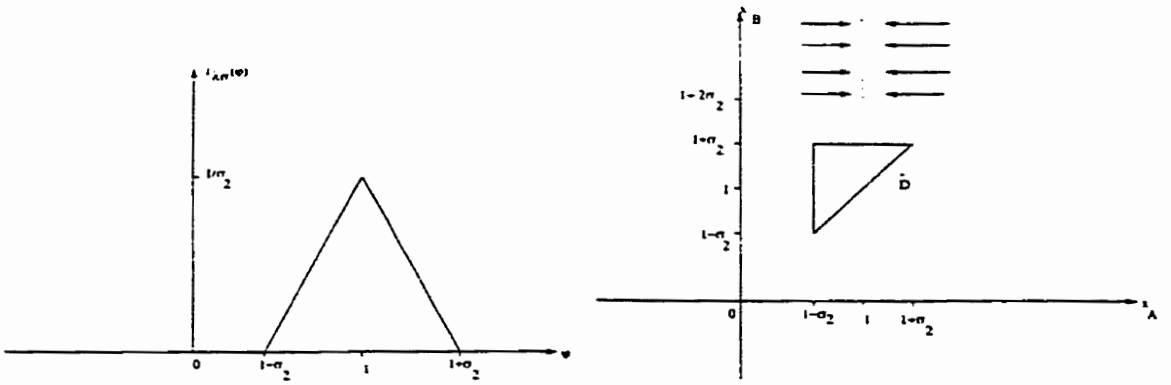

Figure 4.8: *left*: pdf with $\lambda = 1$ *right*: Plot of phase space of the associated system of ODEs, with $\lambda = 1$, showing the continuum of equilibrium points, as described in the text.

For $c \geq 1 + \sigma_2$, we obtain the system

$$\begin{aligned} \dot{x}_A &= \frac{1}{3\sigma_2^2}[1-(1-\sigma_2)^3] + \frac{1}{2\sigma_2^2}(\sigma_2 - 1 - x_A)[1-(1-\sigma_2)^2] - \frac{1}{\sigma_2}(\sigma_2 - 1)x_A \\ &\quad -\frac{1}{3\sigma_2^2}[(1+\sigma_2)^3 - 1] + \frac{1}{2\sigma_2^2}(\sigma_2 + 1 + x_A)[(1+\sigma_2)^2 - 1] - \frac{1}{\sigma_2}(1+\sigma_2)x_A \,, \\ \dot{x}_B &= 0 \,, \end{aligned} \tag{4.11}$$

which consists of a line of equilibria given by $(\bar{x}_A, \bar{x}_B) = (1, k)$, where $k > 1 + 2\sigma_2$. Thus, the classifier converges to the classification rule $\bar{c} = \frac{k}{2} + \frac{1}{2}$. Recalling that the invariant region is given by $\overline{D} : 1 - \sigma_2 < x_A \leq x_B < 1 + \sigma_2$, we deduce that $(1, k) \in \overline{D}$ only if $1 < k < 1 + \sigma_2$. It is clear that an unusual bifurcation has occurred, resulting in the emergence of a continuum of equilibria, all lying on the line $x_A = 1$. This degeneracy of the classifier is shown in Fig 4.8 (*right*). Once again, this signals a breakdown of Ljung's classification model. Similar behaviour may also be observed for the parameter value $\lambda = 0$.

4.3 The "*double Gaussian*" *pdf*

Suppose that the input signals, to be presented to Ljung's algorithm, are distributed according to the *pdf*

$$f_{\lambda,\sigma}(x) = \lambda f_1(x) + (1-\lambda) f_2(x) \ , \ \ \lambda \in [0,1] \ , \tag{4.12}$$

where f_1 and f_2 are normalised Gaussians centred about -1 and 1, with variances σ_1 and σ_2 respectively. That is,

$$\begin{aligned} f_1(x) &= \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2\sigma_1^2}(x+1)^2} \ , \ \text{and} \\ f_2(x) &= \frac{1}{\sigma_2\sqrt{2\pi}} e^{-\frac{1}{2\sigma_2^2}(x-1)^2} \ , \end{aligned} \tag{4.13}$$

where $\sigma_1, \ \sigma_2 \in \mathbb{R}^+$. For the parameter values $\sigma_1 = \sigma_2 = 0.1$ and $\lambda = 0.3$, the pdf defined by (4.12) is depicted in Fig 4.9. As usual, we obtain the following system of ODEs:

$$\begin{aligned} \dot{x}_A &= -x_A \int_{-\infty}^{c} f_{\lambda,\sigma}(x)dx + \int_{-\infty}^{c} x f_{\lambda,\sigma}(x)dx \ , \\ \dot{x}_B &= -x_B \int_{c}^{\infty} f_{\lambda,\sigma}(x)dx + \int_{c}^{\infty} x f_{\lambda,\sigma}(x)dx \ . \end{aligned} \tag{4.14}$$

This system has the equilibrium point implicitly given by

$$(\bar{x}_A, \bar{x}_B) = \left(\frac{P(\bar{c})}{Q(\bar{c})}, \frac{E_x - P(\bar{c})}{1 - Q(\bar{c})} \right) \ , \tag{4.15}$$

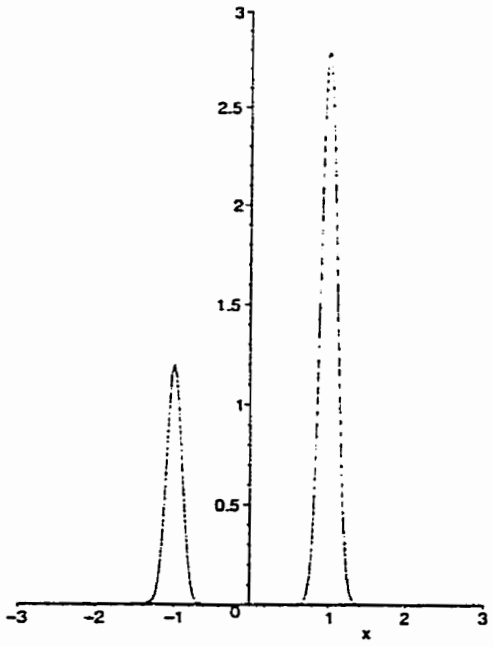

Figure 4.9: Double Gaussian pdf with $\sigma_1 = \sigma_2 = 0.1$, and $\lambda = 0.3$

where $\bar{c} \stackrel{def}{=} 0.5(\bar{x}_A + \bar{x}_B)$, and

$$\begin{aligned} P(c) &\stackrel{def}{=} \int_{-\infty}^{c} x f_{\lambda,\sigma}(x)dx \ , \\ Q(c) &\stackrel{def}{=} \int_{-\infty}^{c} f_{\lambda,\sigma}(x)dx \ , \text{ and} \\ E_x &\stackrel{def}{=} \int_{\mathbb{R}} x f(x)dx = 1 - 2\lambda \ . \end{aligned} \tag{4.16}$$

Furthermore, linearisation of (4.14) yields the Jacobian matrix

$$J(x_A, x_B) = \begin{pmatrix} -\int_{-\infty}^{c} f_{\lambda,\sigma}(x)dx & 0.25(x_B - x_A)f_{\lambda,\sigma}(c) \\ +0.25(x_B - x_A)f_{\lambda,\sigma}(c) & \\ 0.25(x_B - x_A)f_{\lambda,\sigma}(c) & -\int_{c}^{\infty} f_{\lambda,\sigma}(x)dx \\ & +0.25(x_B - x_A)f_{\lambda,\sigma}(c) \end{pmatrix} \tag{4.17}$$

To simplify (4.17), we need to approximate the integrals $\int_{-\infty}^{c} f_{\lambda,\sigma}(x)dx$ and $\int_{c}^{\infty} f_{\lambda,\sigma}(x)dx$. To this end, we shall begin by focussing on the limiting case $\sigma_1, \sigma_2 \to 0^+$. Note that, in this case, the two component Gaussians are tightly packed about their respective means, and we shall see they behave like the two separated component triangles in Chapter 2. Note also that when σ_1 and σ_2 are large, it is not possible to perform analytic approximations of the above integrals. Now, for σ_1, σ_2 arbitrarily small, there are three cases to consider, viz.:

i. $-1 \leq c \leq 1$,

ii. $c > 1$, and

iii. $c < -1$.

These cases describe three different regions of the (x_A, x_B) phase space, as illustrated in Fig 4.10.

Case i

Employing (4.12), we have that

$$\begin{aligned} \int_{-\infty}^{c} f_{\lambda,\sigma}(x)dx &= \lambda \int_{-\infty}^{c} f_1(x)dx + (1-\lambda)\int_{-\infty}^{c} f_2(x)dx \\ &= \frac{\lambda}{\sigma_1\sqrt{2\pi}} \int_{-\infty}^{c} e^{-\left(\frac{x+1}{\sigma_1\sqrt{2}}\right)^2} dx + \frac{(1-\lambda)}{\sigma_2\sqrt{2\pi}} \int_{-\infty}^{c} e^{-\left(\frac{x-1}{\sigma_2\sqrt{2}}\right)^2} dx \,. \end{aligned} \tag{4.18}$$

Let

$$I_1(c) \stackrel{def}{=} \int_{-\infty}^{c} e^{-\left(\frac{x+1}{\sigma_1\sqrt{2}}\right)^2} dx \,.$$

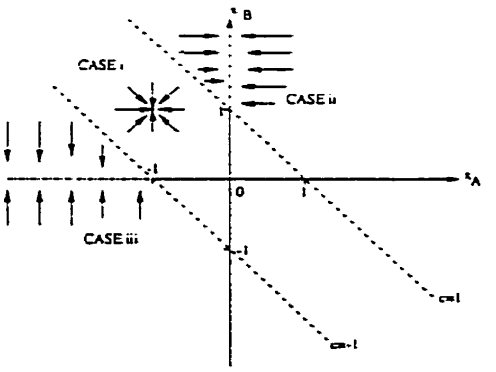

Figure 4.10: The three regions of phase space corresponding to the three cases, as discussed in the text.

Using the change of variables $u = \frac{x+1}{\sigma_1\sqrt{2}}$, and keeping in mind that $-1 \leq c \leq 1$, we get that

$$\begin{aligned} I_1(c) &= \sigma_1 \int_{-\infty}^{\frac{c+1}{\sigma_1\sqrt{2}}} e^{-u^2} du \\ &= \sigma_1\sqrt{2} \int_{-\infty}^{0} e^{-u^2} du + \sigma_1\sqrt{2} \int_{0}^{\frac{c+1}{\sigma_1\sqrt{2}}} e^{-u^2} du \\ &= \sigma_1\sqrt{\frac{\pi}{2}} + \sigma_1\sqrt{2} \int_{0}^{\frac{c+1}{\sigma_1\sqrt{2}}} e^{-u^2} du \, . \end{aligned}$$

Furthermore, using the change of variables $s = \frac{2\sigma_1^2 u^2}{(c+1)^2}$, we obtain

$$\begin{aligned} \sigma_1\sqrt{2} \int_{0}^{\frac{c+1}{\sigma_1\sqrt{2}}} e^{-u^2} du &= \frac{\sigma_1\sqrt{2}}{4} \left(\frac{c+1}{\sigma_1}\right)^2 \int_0^1 u^{-1} e^{-\frac{1}{2}\left(\frac{c+1}{\sigma_1}\right)^2 s} ds \\ &= \frac{1}{2}(c+1) \int_0^1 s^{-\frac{1}{2}} e^{-\frac{1}{2}\left(\frac{c+1}{\sigma_1}\right)^2 s} ds \, . \end{aligned} \tag{4.19}$$

Finally, letting $y = \frac{1}{2}\left(\frac{c+1}{\sigma_1}\right)^2 s$ in (4.19) gives

$$\begin{aligned}\frac{1}{2}(c+1)\int_0^1 s^{-\frac{1}{2}} e^{-\frac{1}{2}\left(\frac{c+1}{\sigma_1}\right)^2 s} ds &= \frac{\sigma_1}{\sqrt{2}}\int_0^{-\frac{1}{2}\left(\frac{c+1}{\sigma_1}\right)^2} y^{-\frac{1}{2}} e^{-y} dy \\ &\sim \frac{\sigma_1}{\sqrt{2}}\Gamma(\frac{1}{2}) \text{ , as } \sigma_1 \to 0^+ \\ &= \sigma_1\sqrt{\frac{\pi}{2}} \text{ .} \end{aligned} \tag{4.20}$$

In view of this, we conclude that

$$I_1(c) \sim \sigma_1\sqrt{2\pi} \text{ .} \tag{4.21}$$

Similarly, it may be shown that

$$\begin{aligned} I_2(c) &\stackrel{def}{=} \int_{-\infty}^{c} e^{-\left(\frac{x-1}{\sigma_2\sqrt{2}}\right)^2} dx \\ &= \sigma_2\sqrt{2}\int_{-\infty}^{\frac{c-1}{\sigma_2\sqrt{2}}} e^{-u^2} du \\ &\to 0 \text{ , as } \sigma_2 \to 0^+ \text{ ,} \end{aligned}$$

since $c - 1 < 0$. Hence, we have that

$$\begin{aligned}\int_{-\infty}^{c} f_{\lambda,\sigma}(x)dx &\sim \frac{\lambda}{\sigma_1\sqrt{2\pi}} \times \sigma_1\sqrt{2\pi} \text{ , as } \sigma_1,\ \sigma_2 \to 0^+ \\ &= \lambda \text{ .} \end{aligned} \tag{4.22}$$

By the same token, it may be shown that

$$\int_c^\infty f_{\lambda,\sigma}(x)dx \sim 1-\lambda \text{ , as } \sigma_1,\ \sigma_2 \to 0^+ \text{ .} \tag{4.23}$$

Therefore, for $-1 < c < 1$ and $\sigma_1,\ \sigma_2 \to 0^+$, (4.17) simplifies to

$$\begin{aligned} J(x_A, x_B) &\sim \begin{pmatrix} -\lambda + 0.25(x_B - x_A)f_{\lambda,\sigma}(c) & 0.25(x_B - x_A)f_{\lambda,\sigma}(c) \\ 0.25(x_B - x_A)f_{\lambda,\sigma}(c) & -(1-\lambda) + 0.25(x_B - x_A)f_{\lambda,\sigma}(c) \end{pmatrix} \\ &\to \begin{pmatrix} -\lambda & 0 \\ 0 & -(1-\lambda) \end{pmatrix} \text{ , as } \sigma_1,\ \sigma_2 \to 0^+ \text{ .} \end{aligned} \tag{4.24}$$

Now, it may be shown that, for $\sigma_1,\ \sigma_2 \to 0^+$ and $-1 < c < 1$, (4.14) yields

$$\begin{aligned} \dot{x}_A &\sim -\lambda(x_A + 1) \\ \dot{x}_B &\sim (1-\lambda)(1 - x_B) \text{ ,} \end{aligned} \tag{4.25}$$

giving the equilibrium point $(-1, 1)$ whose components are the means of the two Gaussians constituting the *pdf* $f_{\lambda,\sigma}$. Furthermore, the Jacobian matrix associated with (4.14), and evaluated at $(-1, 1)$, is given by

$$J(-1,1) = \begin{pmatrix} -\lambda & 0 \\ 0 & -(1-\lambda) \end{pmatrix} \text{ ,}$$

which is similar to (4.24). Thus, (-1,1) is asymptotically stable and yields the classification rule $\bar{c} = 0$. We conclude that, for $\sigma_1,\ \sigma_2 \to 0^+$ and $-1 < c < 1$, the

system of ODEs in (4.14) always possesses a locally asymptotically stable equilibrium point, viz. (-1,1). This is similar to the result obtained in Chapter 2. for the doubly-triangular *pdf*. There, we "recovered" the means (of the component triangular distributions) -2 and 2. Note that, as $\sigma_1,\ \sigma_2 \to 0^+$, the two Gaussians making up the *pdf* degenerate to two Dirac distributions located at -1 and 1.

Case ii

After performing the usual sequence of "operations" (see Case *i*) on (4.18), let $I_1(c) \stackrel{def}{=} \sigma_1\sqrt{2}\int_{-\infty}^{\frac{c+1}{\sigma_1\sqrt{2}}} e^{-u^2}du$, $I_2(c) \stackrel{def}{=} \sigma_2\sqrt{2}\int_{-\infty}^{\frac{c-1}{\sigma_2\sqrt{2}}} e^{-u^2}du$, $I_3(c) \stackrel{def}{=} \sigma_1\sqrt{2}\int_{\frac{c+1}{\sigma_2\sqrt{2}}}^{\infty} e^{-u^2}du$, and $I_4(c) \stackrel{def}{=} \sigma_2\sqrt{2}\int_{\frac{c-1}{\sigma_2\sqrt{2}}}^{\infty} e^{-u^2}du$. Similar to Case *i*, it may be shown that, as $\sigma_1, \sigma_2 \to 0^+$, $I_1(c) \sim \sigma_1\sqrt{2\pi}$, $I_2(c) \sim \sigma_2\sqrt{2\pi}$, $I_3(c) \to 0$, and $I_4(c) \to 0$. In view of these asymptotic expansions, we have that, for $\sigma_1, \sigma_2 \to 0^+$,

$$\int_{-\infty}^{c} f_{\lambda,\sigma}(x)dx \sim \frac{\lambda}{\sigma_1\sqrt{2\pi}} \times \sigma_1\sqrt{2\pi} + \frac{1-\lambda}{\sigma_2\sqrt{2\pi}} \times \sigma_2\sqrt{2\pi} = 1\ ,$$

and

$$\int_{c}^{\infty} f_{\lambda,\sigma}(x)dx \to 0\ .$$

Whence, (4.17) yields

$$J(x_A, x_B) \sim \begin{pmatrix} -1 + 0.25(x_B - x_A)f_{\lambda,\sigma}(c) & 0.25(x_B - x_A)f_{\lambda,\sigma}(c) \\ 0.25(x_B - x_A)f_{\lambda,\sigma}(c) & 0.25(x_B - x_A)f_{\lambda,\sigma}(c) \end{pmatrix}\ ,$$

which implies that

$$J(x_A, x_B) \sim \begin{pmatrix} -1 & 0 \\ 0 & 0 \end{pmatrix}\ .$$

Thus, the linearisation of (4.14) becomes

$$\begin{aligned} \dot{x}_A &= -x_A \\ \dot{x}_B &= 0 , \end{aligned}$$

which has infinitely many neutrally stable equilibria of the form $(0, k)$, where k is limited by the fact that $\bar{c} = \frac{k}{2} > 1$.

Case iii

It can be shown, in a manner similar to the two preceding cases, that (as $\sigma_1, \sigma_2 \to 0^+$) $\int_{-\infty}^{c} f_{\lambda,\sigma}(x)dx \to 0$ and $\int_{c}^{\infty} f_{\lambda,\sigma}(x)dx \sim 1$. Thus, (4.17) gives

$$J(x_A, x_B) \sim \begin{pmatrix} 0.25(x_B - x_A)f_{\lambda,\sigma}(c) & 0.25(x_B - x_A)f_{\lambda,\sigma}(c) \\ 0.25(x_B - x_A)f_{\lambda,\sigma}(c) & -1 + 0.25(x_B - x_A)f_{\lambda,\sigma}(c) \end{pmatrix} .$$

which implies that

$$J(x_A, x_B) \sim \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} .$$

This gives the linearisation of (4.14)

$$\begin{aligned} \dot{x}_A &= 0 \\ \dot{x}_B &= -x_B , \end{aligned}$$

which has infinitely many neutrally stable equilibria of the form $(k, 0)$, where $k < -2$.

Summary of the three cases

We recall that the foregoing analysis, of the system of ODEs given in (4.14), is valid only for the limiting case $\sigma_1, \sigma_2 \to 0^+$.

Case i

The linearisation of (4.14) always possesses a locally asymptotically stable equilibrium, viz. (-1,1).

Case ii

The linearisation of (4.14) has infinitely many neutrally stable equilibria of the form $(0, k)$, where $k > 2$.

Case iii

The linearisation of (4.14) has infinitely many neutrally stable equilibria of the form $(k, 0)$, where $k < -2$.

4.3.1 Stability and bifurcation analysis of the associated system of ODEs

From (4.15), define the function

$$\begin{aligned} H_{\lambda,\sigma}(c) &\stackrel{def}{=} \frac{P(c)}{Q(c)} + \frac{(1-2\lambda) - P(c)}{1 - Q(c)} - 2c \\ &= \frac{P(c)[1-Q(c)] + Q(c)[(1-2\lambda) - P(c)] - 2cQ(c)[1-Q(c)]}{Q(c)[1-Q(c)]}, \end{aligned} \tag{4.26}$$

where $P(c)$ and $Q(c)$ are as defined in (4.16). The equilibria of (4.14) may be computed by finding the zeros $\bar{c}$ of the numerator

$$F_{\lambda,\sigma}(c) \stackrel{def}{=} P(c)[1-Q(c)] + [(1-2\lambda) - P(c)]Q(c) - 2cQ(c)[1-Q(c)]$$

and then applying (4.15). Some computer-generated plots of $F_{\lambda,\sigma}(c)$ versus c, for various parameter values, are shown in Figs 4.11-4.12 (*left*).

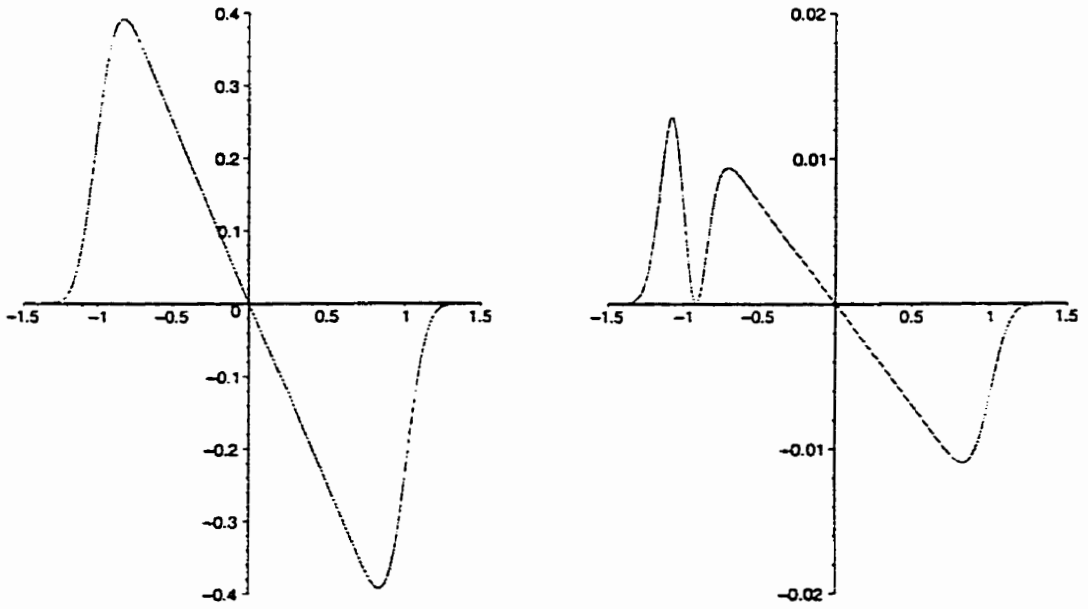

Figure 4.11: $F_{\lambda,\sigma}(c)$ with *left*: $\lambda = 0.5$, and *right*: $\lambda = 0.993$. In both cases, $\sigma_1 = \sigma_2 = 0.1$.

As before, the numerics suggest that the qualitative behaviour of the zeros of $F_{\lambda,\sigma}(c)$ is dependent on the parameters σ_1, σ_2, and λ. When $\lambda = 0.5$ and $\sigma_1 = \sigma_2 = 0.1$, $F_{\lambda,\sigma}(c)$ has only one zero at $\bar{c}_1 = 0$. From (4.15), the corresponding equilibrium point, which is globally asymptotically stable, is (-1,1). Numerically, at $\lambda \approx 0.993$, a S-N bifurcation occurs. In other words, a new zero appears, at

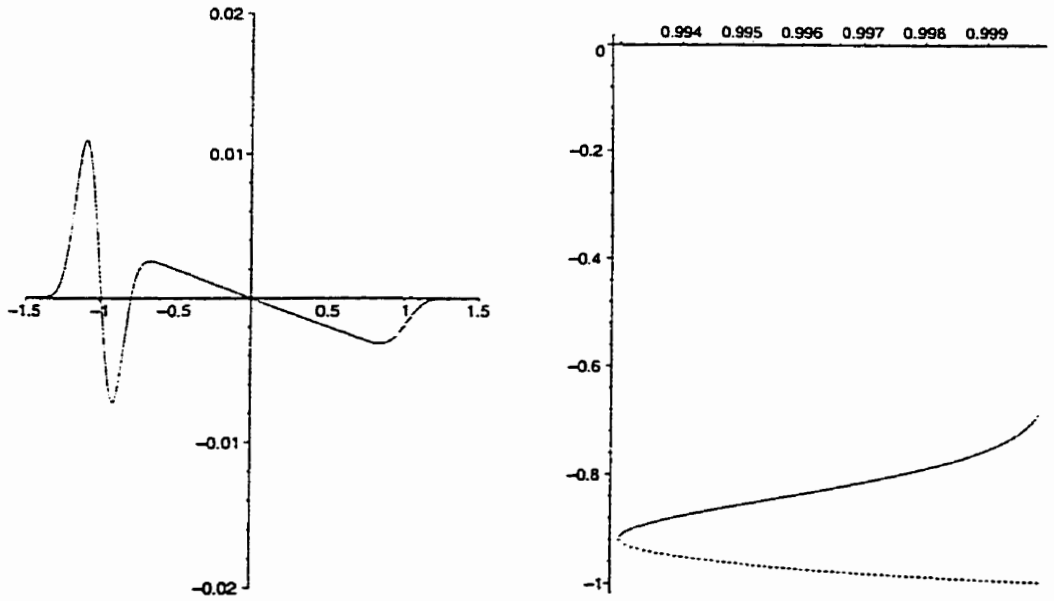

Figure 4.12: *left*: $F_{\lambda,\sigma}(c)$ with $\lambda = 0.998$, $\sigma_1 = \sigma_2 = 0.1$ *right*: Bifurcation of zeros of $F_{\lambda,\sigma}(c)$ with $\sigma_1 = \sigma_2 = 0.1$

$\bar{c}_0 \approx -0.911$. As λ increases, this zero branches to yield two zeros, viz. $\{\bar{c}_2, \bar{c}_3\}$. This is depicted in the bifurcation diagram Fig 4.12 (*right*), which is generated in exactly the same way as before.

At $\lambda = 0.998$, we have that $\{\bar{c}_2, \bar{c}_3\} = \{-0.790, -0.988\}$. The corresponding equilibria, one stable and the other a saddle point, may be computed from (4.15). In view of the Kushner-Clark theorem [1], for fixed $\sigma_1 = \sigma_2 = 0.1$ and for $\lambda < 0.99$, it follows that the discrete system converges *wp1* to $(-1, 1)$. We now focus on the case $\lambda > 0.99$.

Consider the case $\lambda = 0.998$, which is illustrated in Fig 4.12 (*left*). The equilibria are $(x_{11}, x_{12}) = (-1, 1)$, $(x_{21}, x_{22}) = (-1.004, -0.575)$, and $(x_{31}, x_{32}) = (-1.073, -0.904)$, with eigenvalues $\{-0.9980, -0.0020\}$, $\{-0.9360, 0.0294\}$, and

$\{-0.5060, -0.1612\}$, respectively. The corresponding eigenvectors are $\left\{\begin{pmatrix} 1.0000 \\ 0.0000 \end{pmatrix}, \begin{pmatrix} 0.0000 \\ 1.0000 \end{pmatrix}\right\}$, $\left\{\begin{pmatrix} 0.9988 \\ -0.0484 \end{pmatrix}, \begin{pmatrix} 0.0484 \\ 0.9988 \end{pmatrix}\right\}$, and $\left\{\begin{pmatrix} 0.7939 \\ -0.6081 \end{pmatrix}, \begin{pmatrix} 0.6081 \\ 0.7939 \end{pmatrix}\right\}$. From this, we conclude that (x_{11}, x_{12}) and (x_{31}, x_{32}) are locally asymptotically stable, while (x_{21}, x_{22}) is a saddle point. As usual, the issue is what the long-term behaviour of the algorithm is going to be, given the existence of two "competing" stable equilibria. To address this problem, the following numerical experiments are performed:

a. $c(n)$ versus n, given an arbitrary initial grid point $(x_A(0), x_B(0)) \in \overline{D} : -5 < x_A \leq x_B < 5$, the invariant region.

b. The usual simulation of a unit Dirac mass initialised at some arbitrary grid point in $\overline{D} : -1.209 < x_A \leq x_B < 1.201$.

4.3.2 Numerical Simulations

Figs 4.13-4.14 below show various plots of $c(n)$ versus n, for some prescribed parameter values. In all the plots, the learning parameter used is of the form $\gamma_n = (n + 10)^{-\alpha}$, with $\alpha = 0.15$. In each case, five different runs of the algorithm are performed.

The simulation results suggest that, with a relatively high frequency, the discrete algorithm converges to the equilibrium point $(-1, 1)$. It is interesting to note that, from Fig 4.13 (*right*), initialising the algorithm at the stable equilibrium $(-1.073, -0.904)$ still leads to convergence to $(-1, 1)$. Note that for the parameters

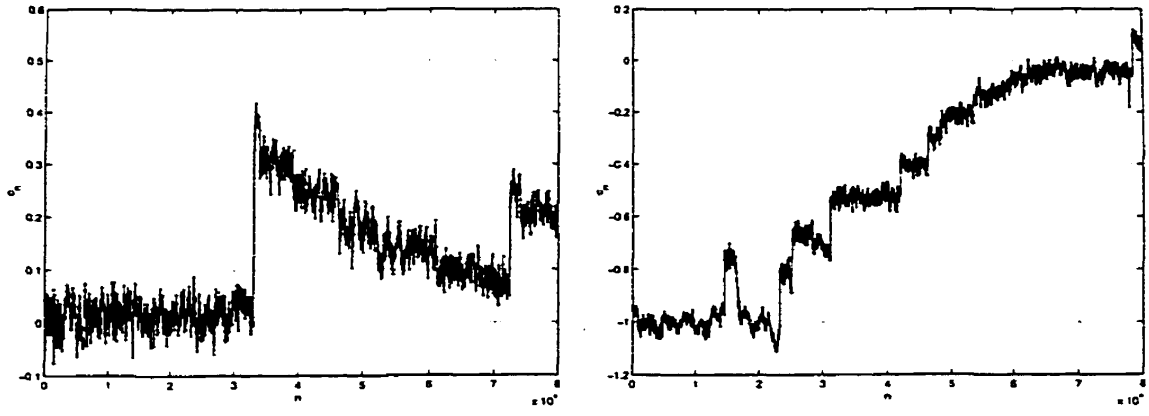

Figure 4.13: time histories of c_n, with *left*: $(x_A(0), x_B(0)) = (-0.9999999998, 0.9999999)$ *right*: $(x_A(0), x_B(0)) = (-1.072578826, -0.9043046980)$. In both cases, $\lambda = 0.998$ and $\sigma_1 = \sigma_2 = 0.1$.

used in Fig 4.14 (*right*), the system of ODEs given in (4.14) has only one globally asymptotically stable equilibrium, at $(-1, 1)$.

Finally, we simulate a unit mass inside $\overline{D}$: $-1.209 < x_A \leq x_B < 1.201$. The simulation results for this experiment are shown in Figs 4.15-4.16. Note that when $\lambda = 0.5$ ($\sigma_1 = \sigma_2 = 0.1$), the system of ODEs (4.14) has exactly one globally asymptotically stable equilibrium. This fact is reflected in Figs 4.14 (*right*) and 4.15 (*left*). Furthermore, it is interesting to observe that initialising the algorithm at the stable equilibrium $(-1.0726, -0.9043)$ does not guarantee that it will remain there for all time. In fact, the simulations suggest that there is a reasonably high probability that, with this stable point used to initialise the algorithm, the algorithm converges to the equilibrium $(-1, 1)$. This is borne out by Figs 4.16 (*left*) and 4.13 (*right*). We also mention that, for the double Gaussian pdf, the algorithm was implemented by finding a polynomial fit to the pdf, which is then used to compute the associated cumulative distribution function. This approach helps optimise the runtime of the

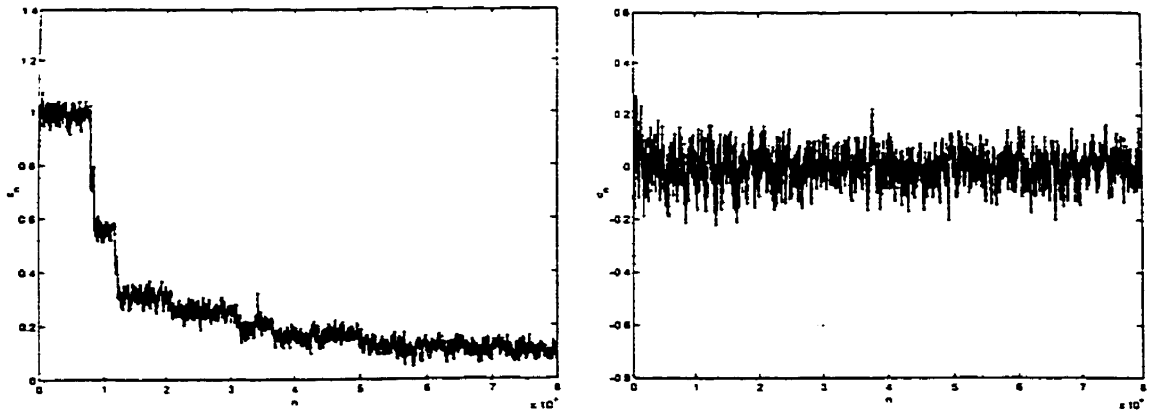

Figure 4.14: time histories of c_n, with *left*: $\lambda = 0.998$ and $\sigma_1 = \sigma_2 = 0.1$ *right*: $\lambda = 0.5$ and $\sigma_1 = \sigma_2 = 0.25$. In both cases, $(x_A(0), x_B(0)) = (-3.00, 2.99)$.

algorithm.

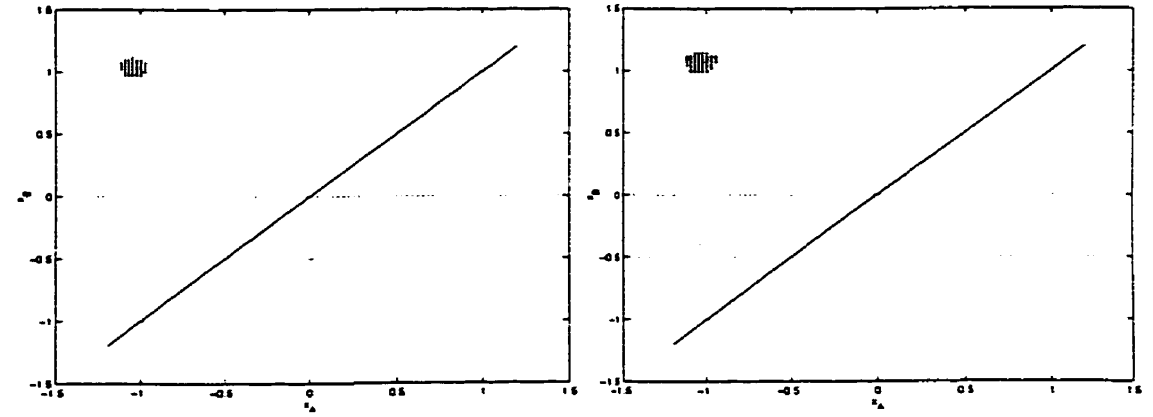

Figure 4.15: Grid points in $\overline{D}$ whose associated masses are greater than $\bar{t} = 0.001$ after 1000 iterations of the algorithm, with *left*: $\lambda = 0.5$, and *right*: $\lambda = 0.998$. In both cases, the unit mass is initialised at the point $(x_A, x_B) = (-1.0, 1.0)$.

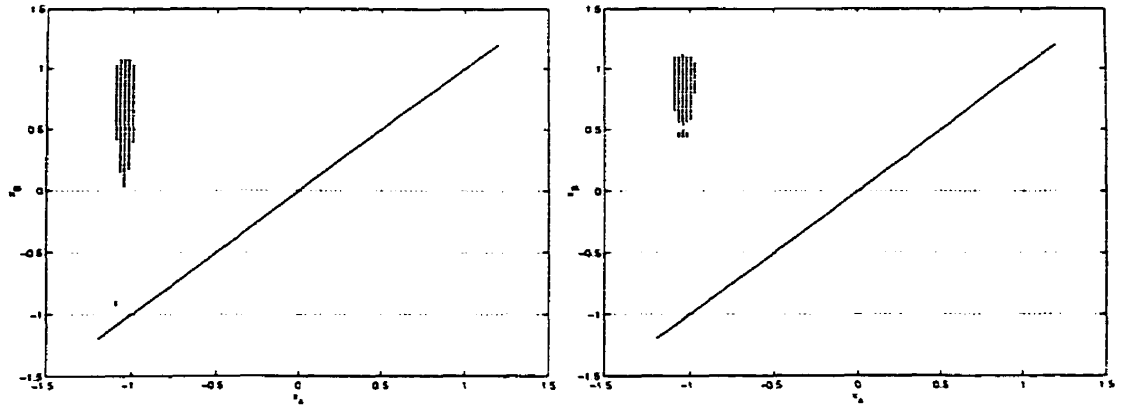

Figure 4.16: Grid points in $\overline{D}$ whose associated masses are greater than $\bar{t} = 0.001$ after 1000 iterations of the algorithm, with the unit mass initialised at the point *left*: $(x_A, x_B) = (-1.073, -0.904)$, and *right*: $(x_A, x_B) = (-0.1, 0.1)$. In both cases, $\lambda = 0.998$.

4.4 Closing remarks

It is interesting to note that, in situations where the *pdf* consists of overlapping components, the extra stable fixed point becomes quite significant, viz. the 'likelihood' of the algorithm converging to the 'wrong' classifier seems to significantly increase. This observation is partly accounted for by the fact that the 'basins of attraction' (in discrete space) of the fixed points of the associated ODE overlap and the fixed points themselves are in relative proximity. As a result, the algorithm tends to oscillate between the fixed points, and takes longer to settle down to just one of them.

Chapter 5

Conclusion and Future Research Directions

Throughout this thesis, we have strongly emphasized the importance of both analytical and numerical approaches in addressing the problem under consideration. This interplay between theory and numerics is not only crucial, but is also the only hope for adequately tackling this problem. Needless to say, there are certain areas of this problem where analytic work alone is almost impossible. In the same vein, there are a lot of instances where numerical work by itself is insufficient.

We have made a couple of significant contributions to the understanding of this problem. Apart from defining an appropriate space $\overline{D}'_+(X)$ for densities (see Chapter 3), we have also been able to sharpen and articulate some existing ideas about the possible use of the notion of stationary density in the analysis of random algorithms of the type considered in the thesis. The main problem encountered in this endeavour stems from the computational aspects of the exercise. Essentially,

we found that computing the iterates $\overline{P}_n\overline{P}_{n-1}...\overline{P}_1 f$ of the perturbed Frobenius-Perron operator (see Chapter 3, section 3.7) presents some serious numerical difficulties, mainly because it involves the computation of iterated integrals of increasing complexity. Compounded with this, the way the learning parameter γ_n is defined renders the convergence of $\overline{P}_n\overline{P}_{n-1}...\overline{P}_1 f$, as $n \to \infty$, painfully slow. Clearly, a potential problem for further investigation is the *efficient* computation of $\lim_{n\to\infty} \overline{P}_n\overline{P}_{n-1}...\overline{P}_1 f$.

However, in spite of the aforementioned difficulties, we did obtain some indications, depending on the amplitude of $\{\xi_n\}$, that the stationary density of $\{x_n\}_{n=1}^{\infty}$ consists of Dirac distributions located at the stable equilibria of the associated ODE. Furthermore, we obtained indications that working with linearised algorithms gives the same *local* stationary densities (in the neighbourhoods of the stable equilibria of the associated ODE) as when one looks at the full nonlinear algorithms. Encouraging as this might be, it must be pointed out that formulating the analysis of the sequence of densities $\{f_n\}_{n=1}^{\infty}$ in the metric space $(\overline{D}'_+(X), d_{\overline{D}'_+})$ still remains a tremendously daunting problem. It is interesting to note that, even after $\lim_{n\to\infty} \overline{P}_n\overline{P}_{n-1}...\overline{P}_1 f$ has been computed, it is still a challenging problem determining what the probability of convergence (of the discrete system) to any particular one of the myriad of locally stable equilibria of the associated ODE is, especially if the stationary density consists of Dirac distributions positioned at these stable equilibria.

We have also developed a computer algorithm (see Chapter 2) that crudely quantifies the probability of the discrete algorithm converging to a stable equilib-

rium of the associated ODE, in the case when the latter has multiple locally stable equilibria. It must be emphasized that this algorithm is limited by the fact that only those values of iterates (of the Dirac mass) that are greater than some arbitrary (threshold) value are considered. This clearly implies that the "probability" so calculated can only be a very crude estimate of the true value.

In Chapter 3, we proved the completeness of the metric space $(\overline{D}'_+(X), d_{\overline{D}'_+})$. Furthermore, for the linear map $S(x) = \alpha x$, $\alpha \in (0,1)$, we showed that the associated Frobenius-Perron operator P is contractive in $(\overline{D}'_+(X), d_{\overline{D}'_+})$. The open challenge is to show that the nonlinear operator $\overline{P}_n$, of Section 3.7, is contractive in this metric space as well. In Chapter 3, we studied densities of algorithms of the form

$$x_{n+1} = S_n(x_n) + \gamma_n \xi_n \ ,$$

where $S_n(x_n) \stackrel{def}{=} x_n + \gamma_n \bar{h}(x_n)$. These algorithms are more general than those considered by Mackey and Lasota [5], viz.

$$x_{n+1} = S(x_n) + \epsilon \xi_n \ , \ \ 0 < \epsilon \ll 1 \ ,$$

where $S(\cdot)$ is *not* explicitly dependent on n. As a result, our proof of Theorem 3.6 in Section 3.7 is new - it is more general than that of Mackey and Lasota [5]. In addition, in Chapter 3, we improved Chung's result (see Theorem 3.1) by establishing a tighter bound on the parameter r.

Finally, it must be noted that, throughout this thesis, we have assumed $\{\xi_n\}$ to be a sequence of *i.i.d.* random variables. This assumption does simplify some

potentially 'murky' conceptual and computational territory (see Chapter 3). Nevertheless, it is interesting to ask how realistic this assumption is, and to also explore situations where it fails outrightly. Certainly, it is true that, in the majority of applications, this assumption does *not* hold.

Bibliography

[1] Kushner H.J. and Clark D.S. *Stochastic Approximation Methods for Constrained and Unconstrained Systems.* Springer-Verlag, New York (1978).

[2] Ljung L. *Analysis of Recursive Stochastic Algorithms* IEEE Transactions on Automatic Control (1977), **AC-22**, 551-575.

[3] Halley J.M. and Iwasa Y. *Extinction Rate of a Population under both Demographic and Environmental Stochasticity* Theoretical Population Biology (1998), **53**, No. 1, 1-15.

[4] Karlin S. and Taylor H.P. *A First Course in Stochastic Processes*, 2nd ed., Academic Press, London (1975).

[5] Mackey C.M. and Lasota A. *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*, 2nd ed., Springer-Verlag (1994).

[6] Abramowitz M. and Stegun I.A. (eds.) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover Publications Inc., New York (1965).

[7] Robbins H. and Monro S., *A stochastic approximation method* Ann. Math. Statist. (1951), **22**, 400-407.

[8] Fort J.C. and Pagès G., *Convergence of stochastic algorithms: From the Kushner-Clark theorem to the Lyapunov functional method* Adv. Appl. Prob. (1996), **28**, 1072-1094.

[9] Nilsson N.J. *Learning Machines: Foundations of trainable pattern-classifying systems.* McGraw-Hill (1965).

[10] McCulloch W.S., and Pitts W. *A logical calculus of the ideas immanent in nervous activity* Bulletin of Mathematical Biophysics (1943), **5**, 115-133.

[11] Ashby W.R. *Design for a Brain.* Wiley, New York (1952).

[12] Haykin S. *Neural Networks: A Comprehensive Foundation.* Macmillan (1994).

[13] Baldi P.F., and Hornik K. *Learning in Linear Neural Networks: A Survey* IEEE Transactions on neural networks (1995), **6**, no.4, 837-858.

[14] Rosenblatt F. *The Perceptron: A probabilistic model for information storage and organization in the brain* Psychological Review (1958), **65**, 386-408.

[15] Novikoff A.B. *On convergence proofs for perceptrons* Proc. Symp. on Math. Theory of Automata (1962: Polytechnic Institute of Brooklyn, New York), 615-622.

[16] Agmon S. *The relaxation method for linear inequalities* Canadian Journal of Mathematics (1954), **6**, No. 3, 382-392.

[17] Minsky M.L. and Papert S.A. *Perceptrons.* Expanded Edition. MIT Press, Cambridge, MA. (1988).

[18] Fabian V. *On asymptotic normality in stochastic approximation* The Annals of Mathematical Statistics (1968), Vol. 39, No. 4, 1327-1332.

[19] Chung K.L. *On a stochastic approximation method* Ann. Math. Statist. (1954), **25**, 463-483.

[20] Vrscay E. and Forte B. *Theory of generalised fractal transforms* Fractal Image Encoding and Analysis (edited by Yuval Fisher), Springer-Verlag, New York (1996), 145-168.

[21] Hutchinson J. *Fractals and Self-similarity* Indiana University Journal of Mathematics (1981), **30**, 713-747.

[22] Zemanian A. H. *Distribution Theory and Transform Analysis: An Introduction to Generalised Functions, with Applications.* McGraw-Hill, New York (1965).

[23] Kushner H. J., and Yin G. *Stochastic Approximation Algorithms and Applications.* Springer-Verlag, New York (1997).

[24] Benveniste A., Métivier M., and Priouret P. *Adaptive Algorithms and Stochastic Approximations.* Springer-Verlag (1987)

[25] Miller R. K. and Michel A. N. *Ordinary Differential Equations.* Academic Press (1982)

[26] Yaglom A. M. *An Introduction to the Theory of Stationary Random Functions.* Prentice-Hall (1962).

[27] Van Kampen N. G. *Stochastic Processes in Physics and Chemistry.* North-Holland (1992).