Convergence Properties of Gradient Descent Noise Reduction

David Ridout* Kevin Judd

Centre for Applied Dynamics and Optimisation, Department of Mathematics and Statistics, University of Western Australia, Nedlands, WA 6907, Australia

Abstract

Gradient descent noise reduction is a technique that attempts to recover the true signal, or trajectory, from noisy observations of a non-linear dynamical system for which the dynamics are known. This paper provides the first rigorous proof that the algorithm will recover the original trajectory for a broad class of dynamical systems under certain conditions. The proof is obtained using ideas from linearisation theory. Since the first introduction of the algorithm it has been recognised that the algorithm can fail to recover the true trajectory, and it has been suggested that this is a practical or numerical limitation that is a consequence of near tangencies between stable and unstable manifolds. This paper demonstrates through numerical experiments and details of the proof that the situation is worse than expected in that near tangencies impose essential limitations on noise reduction, not just practical or numerical limitations. That is, gradient descent noise reduction will sometimes fail to recover the true trajectory, even with unlimited, perfect computation. On the other hand, the numerical experiments suggest that the gradient descent noise reduction algorithm will always recover a trajectory that is entirely consistent with the evidence provided by the observations, that is, it attains the best that can be achieved given the observations. It is argued that near tangencies will therefore impose the same limitations on any noise reduction algorithm.

1 Introduction

Non-linear noise reduction refers to a collection of techniques for recovering a signal from a time series of measurements of a non-linear dynamical system where the measurements are corrupted by noise. Noise reduction is an important technique, not only

*Present Address: Department of Physics and Mathematical Physics, University of Adelaide, Adelaide, SA 5005, Australia.

for what it does, but also because it is closely related to the important concept of shadowing trajectories ([7, 5, 25]) and techniques of state estimation ([1, 2, 12]).

There have been proposed a number of noise reduction algorithms ([15, 8, 4, 5, 6, 23, 18]), but although these algorithms have been around for quite some time now, there are few rigorous results about their properties. Most of what is known about the algorithms is either the result of numerical experiments or loosely justified. Of particular interest are the conditions that guarantee convergence of an algorithm to the true trajectory. In exploring noise reduction it is useful to make a division into three situations — where one has a perfect model of the system (the system being observed has known dynamics), where one has only an imperfect model, and where there is no model of the dynamics. Surprisingly, it is only the third case where any previous rigorous convergence results have been obtained ([17]). Here we provide rigorous results for the first case. For the second case, which is arguably the usual case and therefore the most important, nothing is known, but it is hoped that the results presented here for perfect models can be extended to imperfect models in the future.

This paper is concerned with the situation where one has a perfect model of the system. We study the properties of a particular algorithm that applies in this situation called the *gradient descent* algorithm ([4]). The major part of the paper will outline in some detail a proof that the gradient descent algorithm converges to the true trajectory under specified conditions. The initial part of the paper describes some numerical experiments that motivate the main restrictions.

The initial numerical experiments are of interest in their own right because they clarify an important phenomenon that we believe has not been fully understood before. From the first introduction of the gradient descent algorithm it has been observed that there are times when the algorithm fails to converge to the true trajectory. This failure has been attributed to *near tangencies* between stable and unstable manifolds, that is, points where the stable and unstable manifold are almost tangent. The supposed role of near tangencies is that they result in a nearly singular derivative in the neighbourhood of these points, which consequently results in slow convergence of the algorithm in these neighbourhoods. The failure of the gradient descent algorithm has therefore been seen as a practical or numerical limitation. Our numerical experiments demonstrate that the situation is worse than supposed and that the mechanism of failure is different from what has been suggested. Our interpretation of the algorithm's failure focuses on the observation that the transverse intersection between stable and unstable manifolds at a near tangency implies the existence of a nearby trajectory that is homoclinic to the true trajectory. We will show that observational noise can make it impossible to distinguish between the true trajectory and its homoclinic cousin, and that the observational noise might be such that the weight of evidence is for the homoclinic cousin being the maximum likelihood trajectory. Consequently, the gradient descent algorithm will sometimes convergence to the incorrect trajectory. This phenomenon was previously

illustrated in earlier work of Judd and Smith ([12]) on indistinguishable states. It is also almost certain that this kind of failure is not a unique feature of gradient descent noise reduction, and should be expected of all noise reduction methods.

The rest of the paper is organised as follows. In section 2, we introduce the gradient descent algorithm and indicate why it should be able to reduce noise. This is followed by the results of some numerical experimentation in section 3 which further investigate the properties of the gradient descent algorithm. By analysing these results and particularly, the reasons why noise reduction sometimes fails, we arrive at the precise mathematical notion of what we mean when we say that the noise reduction has succeeded, and also, which classes of systems can be expected to allow noise reduction to succeed. With these ideas, we can then turn to the theoretical properties of gradient descent in section 4, where noise reduction is rigorously proven to occur for the appropriate class of systems, subject to an additional condition. This is followed by further discussion in section 5.

2 Gradient Descent

In this paper, we will always assume that the dynamics of the system under investigation is known. The system will be assumed to be *discrete* time: $y_{i+1} = f(y_i)$, $i \in \mathbb{Z}$, and the dynamical map, f, will be assumed to be a diffeomorphism from (a subset of) $\mathbb{R}^d$ into itself[1]. Let $\{x_i\}_{i=1}^n$, $x_i \in \mathbb{R}^d$, be the set of experimental measurements of the corresponding states $\{y_i\}_{i=1}^n$. It is convenient to regard this set of observations as a vector in $\mathbb{R}^{nd}$: $x = (x_1, x_2, \ldots, x_n)$. The result of a noise reduction algorithm is therefore another vector in $\mathbb{R}^{nd}$ which will be denoted by $\hat{x}$, which is an estimate of the true trajectory of states $y = (y_1, y_2, \ldots, y_n)$. Because the dynamical map is assumed to be known, there is no loss of generality in assuming that the noise on the measurements is additive, and so

$$x_i = y_i + \delta_i$$

where the δ_i are a realisation of some noise distribution (assumed independently and identically distributed).

We will consider the following form of the gradient descent algorithm for noise reduction. Define the *determinism function* $L : \mathbb{R}^{nd} \to \mathbb{R}$ by

$$L(x) = \frac{1}{2} \sum_{i=1}^{n-1} \|x_{i+1} - f(x_i)\|^2 . \tag{1}$$

The norm used in this definition is completely arbitrary — the standard Euclidean norm is convenient (it has nice analytic properties) and will be used in what follows. Note

[1] In section 4, f will be restricted to act on a compact manifold M for technical reasons. However, noise reduction algorithms are more conveniently discussed in Euclidean space, and it is clear that their action can always be transferred back onto the manifold using the appropriate charts.

that $L(x) = 0$ precisely when the points $x_i \in \mathbb{R}^d$ form a deterministic trajectory of f. Generally, the noisy measurements do not form a deterministic trajectory. The idea behind gradient descent noise reduction is that a trajectory close to the observations can be obtained by minimising L through gradient descent, using the observations x as the starting point. An explicit implementation of this idea, which we will refer to as the gradient descent algorithm, is achieved by solving the set of differential equations

$$\dot{x}(t) = -\nabla L(x(t)), \qquad x(0) = x. \tag{2}$$

The noise reduced trajectory, $\hat{x}$, is then given by $\hat{x} = \lim_{t\to\infty} x(t)$. The "time" variable t used in the gradient descent will be referred to as the *descent time* to distinguish it from the discrete "time" implicit in the iteration $y_i \mapsto y_{i+1} = f(y_i)$.

3 Numerical Experiments

In this section, some numerical results of the gradient descent algorithm are presented and discussed. The basic features of the difference between noise reduced trajectories and the true trajectory is described and the important role of near tangencies is revealed.

The last section indicated that the gradient descent algorithm could be implemented by solving the set of differential equations 2. Our experiments employ the 1-5 stiff integration function `ode15s` of `MATLAB`. The noise reduced trajectories were obtained by letting the descent-time variable increase until convergence appeared to have been established. The Ikeda map ([10]) is used as an example system. When the Ikeda map is expressed as a real function from $\mathbb{R}^2$ into itself, we select the parameters so that

$$I(x,y) = \left(1 + \frac{9}{10}(x\cos\theta - y\sin\theta), \frac{9}{10}(x\sin\theta + y\cos\theta)\right), \tag{3}$$

where $\theta = 2/5 - 6\left(1 + x^2 + y^2\right)^{-1}$.

To display the properties of the gradient descent algorithm we consider a typical example trajectory that has twenty points with initial point $(0.9255, -1.0126)$ and final point $(1.1243, -2.1607)$ (approximately). Gaussian noise with mean zero and standard deviation $1/10$ was added to this trajectory (giving a noisy trajectory) before the gradient descent algorithm was applied (to get a noise-reduced trajectory). Ten different noise realisations were used, giving ten different noisy trajectories and hence ten different noise-reduced trajectories. The magnitudes of the differences between the points of the clean and the ten noise-reduced trajectories are plotted in Figure 1.

There are two features of the distribution of errors revealed in Figure 1 that should be observed. First, there are the obvious "spikes" in the errors around points 3, 8, and 16, which all the noise-reduced trajectories display to varying degrees. Second, the errors are large at the initial and final points of the trajectory but quite small in between, and

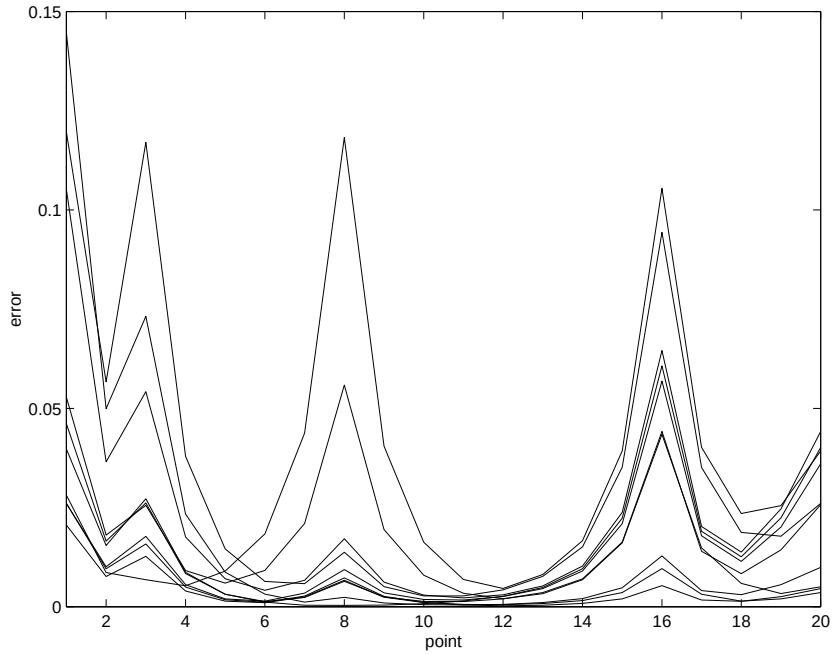

Figure 1: Errors after noise reduction by gradient descent (with ten different noise realisations) for each point of a twenty point trajectory of the Ikeda system, contaminated by Gaussian noise of standard deviation 0.1. This gives a signal to noise ratio of about 25dB.

the errors at the ends are of an order of magnitude comparable to the noise added to the system. These features of the errors are typical of noise reduction algorithms. The large errors at the initial and final points of the trajectory are generally ascribed to the fact that at end points, the algorithm only has forwards *or* backwards iterations (rather than both) to help it locate where the true trajectory should be.

It is well known that the large errors near the initial point decay exponentially, at a rate given by the largest non-positive Lyapunov exponent (largest as in closest to zero), and that the growth of errors near the final point is also exponential with the rate given by the smallest non-negative Lyapunov exponent. It is obvious then that to achieve any sort of noise reduction, we must consider *hyperbolic* systems (these have no vanishing Lyapunov exponent). An example of a non-hyperbolic system is that given by $f = \mathrm{id}$, the identity mapping. It is clear that no noise reduction will occur for this system.

The spiking phenomenon observed in Figure 1 around various points of the trajectory is generally attributed to the presence of *tangencies*, where a tangency is a point whose (generalised) stable and unstable manifolds are tangent at that point. That is, at a tangency point, the generalised stable and unstable eigenspaces do *not* span the entire tangent space. The argument usually given to explain the observed spiking phenomenon runs something like the following. Noise reduction exploits the fact that a hyperbolic point has a stable and unstable direction. The stable direction implies a convergence of nearby trajectories moving forward in time, and the unstable direction implies a convergence of nearby trajectories moving backward in time. It is the convergence of nearby trajectories that can be exploited to remove the noise. At a tangency of a two-dimensional system, these directions lose their sharp distinction, and as a consequence, the convergence forward and backward in time is (at least partially) lost, and so some of the noise cannot be removed. (The argument is usually a little more general than this, submitting that near tangencies are enough to cause slow numerical convergence by a similar mechanism as explained below.)

Let us now examine the stable and unstable manifolds in the neighbourhoods of the points that display the error spikes seen in Figure 1, which correspond to the points 3, 8, and 16 of our sample trajectory. These manifolds are shown in Figure 2. Notice that for points 8 and 16, the stable and unstable manifolds appear to be almost tangent to one another at the clean trajectory point, and for point 3, although the manifolds are not tangent there, the angle[2] between them is relatively small. The angles can be numerically approximated easily ([22]) and are (about) 8.2° for point 3, 2.8° for point 8, and 1.2° for point 16. These are not tangency points as the angle is non-zero, so we will refer to them as *near-tangency points*. It should perhaps be mentioned here that the Ikeda map does contain genuine tangency points, but these are expected to be extremely rare by the Multiplicative Ergodic Theorem ([19]). These tangency and near-tangency

[2]This is defined to be the angle between the one-dimensional stable and unstable eigenspaces at the point — see also section 4.

features would be seen in any dynamical system like the Ikeda map where action of the map is to stretch and fold the state space.

Now reconsider the argument given above to explain the spikes in the error distributions. We feel that this argument is unsatisfactory for the following reason. It does not explain how near tangencies can affect the noise reduction procedure, even though trajectories with near tangencies should be expected to be infinitely more common than trajectories with exact tangencies. With noise reduction algorithms that rely on solving algebraic equations (manifold decomposition for instance) rather than differential equations, it can be argued (and usually is) that near tangencies cause *practical* difficulties, in that the matrix which needs inverting becomes badly conditioned. However, these difficulties do not arise with gradient descent. Instead, it has been noted that the gradient descent algorithm "grinds to a halt" around tangencies, meaning that the convergence of the algorithm is very slow, presumably because the cost function (the determinism function L) is locally rather flat. The suggestion here seems to be that the failure of gradient descent noise reduction around tangencies is due to a lack of convergence, another practical difficulty. While these difficulties do occur, and are important, we now argue that there is in fact a theoretical impediment to noise reduction, more fundamentally important than the aforementioned practical problems. It is this, and not a lack of convergence, that causes the spikes in the error distributions seen in Figure 1.

To examine more closely the problem of near tangencies and how they prevent noise reduction, we consider another sample trajectory of the Ikeda map that has only one near tangency with very small angle. The sample trajectory has fifty points, a near tangency of about 3° at point 39 (spatially situated at approximate coordinates $(1.167, 0.485)$), and lesser near tangencies of between 10° and 20° at points 2, 10, and 25. Applying Gaussian noise of standard deviation $1/10$ and then noise-reducing *thirty* different noise realisations, we find error distributions as shown in Figure 3. Notice that the errors are shown on a logarithmic scale so that the distribution shape can be easily examined even when the errors are negligible, although with the logarithmic scale some of the error curves are almost identical over some time periods.

We note the spike at point 39 as well as smaller spikes at other points of the trajectory. It is also apparent that the logarithmic error distribution is roughly piecewise-linear with two different slopes (a negative and a positive one). These slopes represent the (local) *Lyapunov exponents* of the system - this is clear for the pieces of the distribution connecting to the initial and final points of the trajectory, and it will become clear why this is also the case for the points around the near-tangency point shortly. What is of greater interest is that the error distributions around the tangency at point 39 form two quite distinct groups. The jump in the errors around point 39 is sometimes small and sometimes much larger[3].

[3]The errors corresponding to the larger jump are not *resolved* into separate curves in this figure due to the logarithmic scale. In fact, approximately half the distributions show this larger jump.

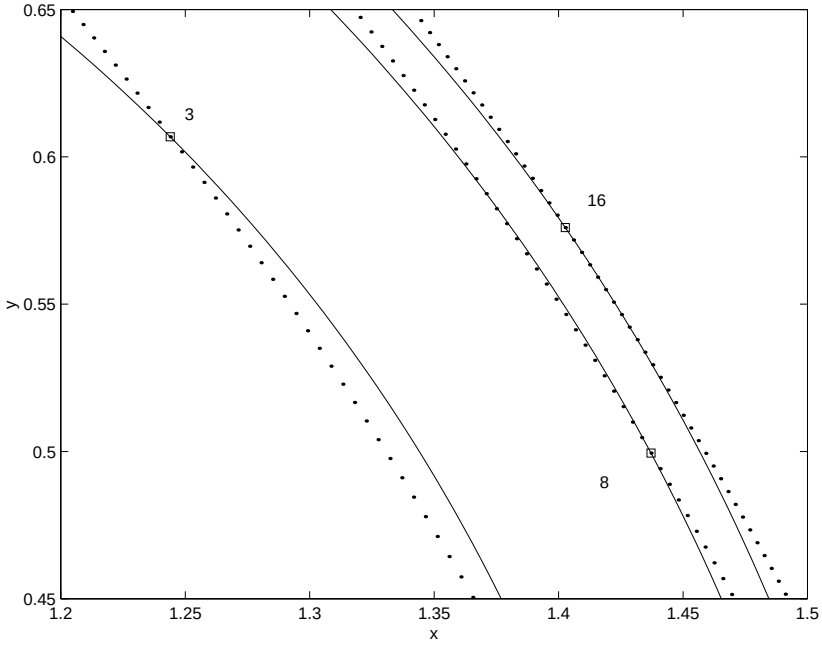

Figure 2: Stable (dotted) and unstable (solid) manifold for points 3, 8 and 16 from the clean Ikeda trajectory considered (see text).

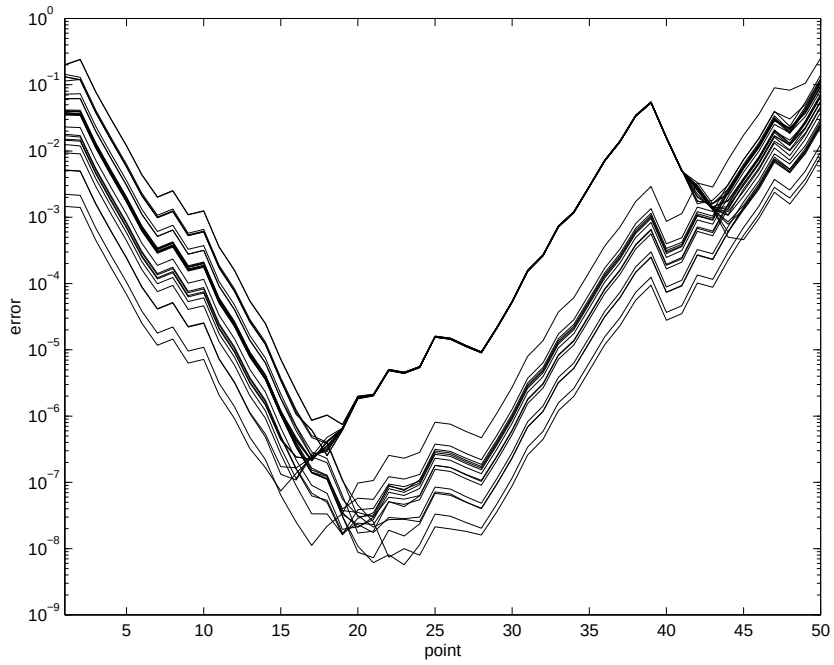

Figure 3: Error distributions after noise reduction by gradient descent (with thirty different noise realisations) for each point of a fifty point trajectory of the Ikeda map, contaminated with Gaussian noise of standard deviation 0.1. Note the logarithmic scale.

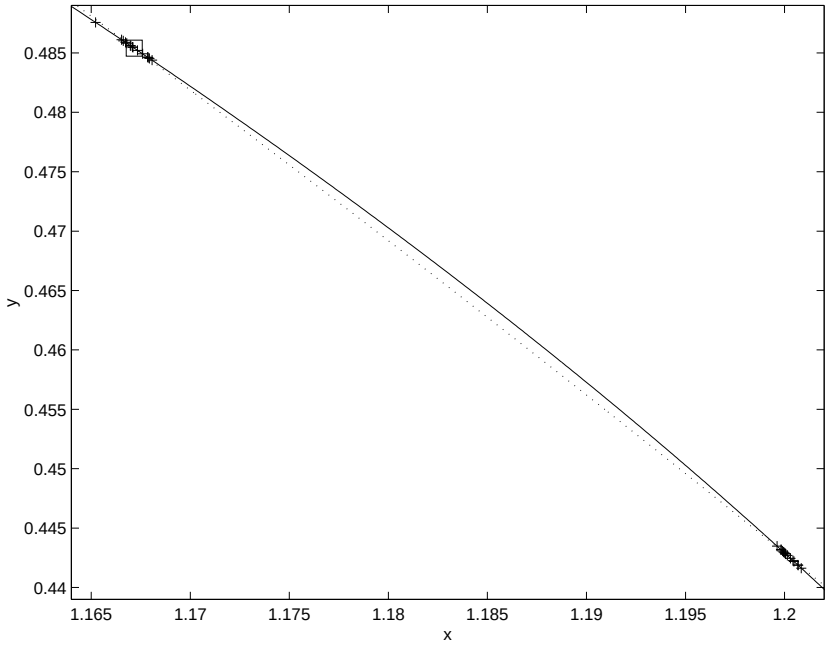

Figure 4: Spatial plot of the 39^{th} points of the noise reduced trajectories of Figure 3 (+) and the 39^{th} point of the correct trajectory (□) with its stable (dotted) and unstable (solid) manifolds.

These two groups are shown spatially (around the tangency point 39) in Figure 4. The large square marks point 39, the "+" signs mark the thirty noise reduced approximations of point 39, and the dotted and solid lines show the stable and unstable manifolds through point 39 (respectively). Note that the groups cluster about the points where the stable and unstable manifolds intersect. These intersection points are called *homoclinic intersection points* because the trajectories of the intersection points converge in forward *and* backward time.

This clustering about homoclinic intersection points occurs because the points are then forced to be close to the stable *and* unstable manifolds of the true point. Iterating forward then means that the error must shrink (because the point is near the stable manifold). The error along the unstable manifold must likewise grow, and to accommodate this shrinking along the stable manifold and growth along the unstable manifold, the unstable manifold "bulges" outwards (and the angle between the stable and unstable manifolds increases). Similarly, upon iterating backwards, the error along the unstable manifold decreases and the error along the stable manifold grows, leading to a bulging of the stable manifold (and a corresponding increase in the angle between the manifolds). This is pictured in Figure 5. Thus the magnitude of the errors *decreases* as we

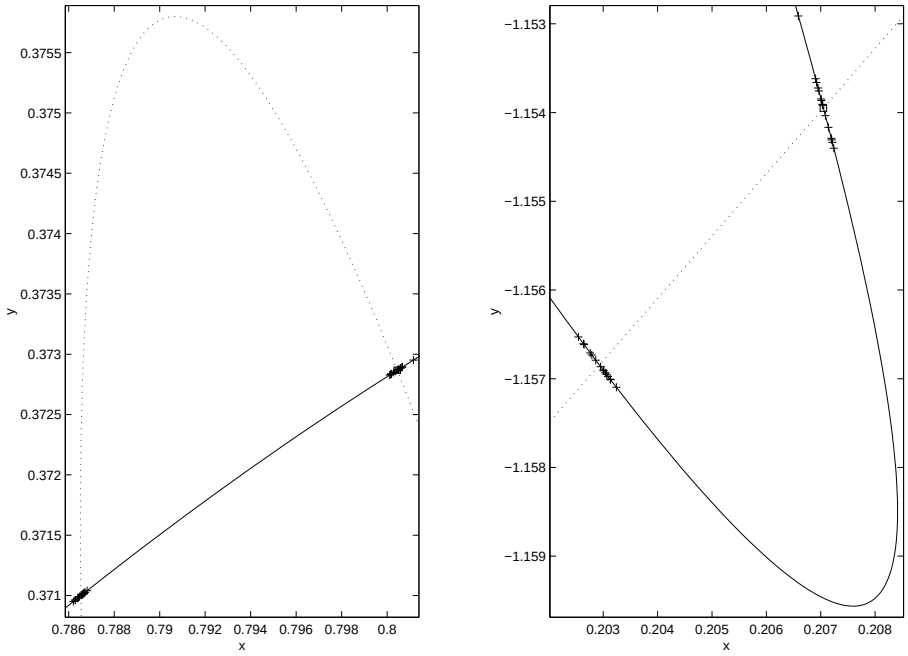

Figure 5: As in Figure 4 but for the 37^{th} point (left) and the 41^{st} point (right). At left, the stable manifold (dotted) "bulges" outward relative to the unstable manifold (solid), whereas at right, the opposite is true.

iterate forwards and backwards (as it must decrease for forward iterations since we are close to the stable manifold, and must likewise decrease for backwards iterations since we are close to the unstable manifold). In fact, this argument also explains why the errors grow and decay *exponentially* around a tangency point at a rate corresponding to the Lyapunov exponents of the system. If the noise reduced points were not near a homoclinic intersection point, then by iterating forwards or backwards, the errors would have to eventually grow. Summarising, it can be said that the trajectories through the two homoclinic intersection points of Figure 4 (one of which is the true point) remain close together and so the noise reduction algorithm chooses one or the other depending on the particular noise realisation.

Another way of saying this is that the trajectories through the homoclinic intersection points are difficult to distinguish on the basis of the given noise realisations. This difficulty can be quantified using the *indistinguishability theory* of Judd and Smith ([12]). The noise distribution used here was Gaussian with standard deviation $1/10$, so the probability that two trajectories y and y' will be indistinguishable given a random

noise realisation is given by (see [12] or [21]):

$$\mathrm{P}\left(y \sim y'\right) = \exp\left\{-25\sum_i \left\|y_i - y_i'\right\|^2\right\}.$$

A plot of (an excellent approximation of) the indistinguishability of the correct trajectory and the nearby trajectories is given in Figure 6 (left). The plot measures the probability of indistinguishability versus the point corresponding to point 39 of the correct trajectory. The two peaks correspond to the homoclinic intersection points (the peak with value 1 is obviously the correct point). The second peak has probability approximately 0.9. Therefore it is very likely that a given noise realisation will be unable to distinguish between the true trajectory and the trajectory through the other homoclinic intersection point. This explains why, in our example, the numbers of noise reduced points clustered around each homoclinic intersection point are approximately equal — the two trajectories are usually indistinguishable so the noise reduction algorithm gives each with approximately equal probabilities.

The large peak around tangencies in the error distribution is therefore due to the algorithm choosing the wrong homoclinic intersection point. This is usually only observed when the angle between the stable and unstable manifold is quite small however. For small angles, the distance between the homoclinic intersection points is expected to be small *compared to the noise level* (and this forces the distances between the forward and backward iterates of the homoclinic intersection points to decay exponentially). Therefore the algorithm is just as likely to converge onto the wrong homoclinic intersection point as the right one. In terms of indistinguishability, this is nicely pictured in Figure 6 (right) where the standard deviation of the noise has been dropped from $1/10$ to $1/50$. The probability that the trajectories through each of the homoclinic intersection points cannot be distinguished drops from 0.9 to about 0.06. At this noise level, the algorithm will only rarely choose the wrong homoclinic intersection point.

In summary then, it seems that noise reduction by gradient descent (and indeed, by any other type of algorithm) is limited by the presence of near-tangency points to noise levels which are smaller than the minimum distance between the points of the clean trajectory and their corresponding homoclinic intersection points (if they exist). In order to prove a result stating that noise reduction is guaranteed to converge onto the clean trajectory (except around the end points of course) as the number of data points is increased to infinity then, it is necessary to restrict our attention to systems without genuine tangency points (that is, the angles between the stable and unstable manifolds must be bounded below) and to sufficiently small noise levels. This is the subject of the rest of this work. Note first however that the requirement that the noise level be small compared to the distances over which the stable and unstable manifolds can intersect (non-trivially) is equivalent to the requirement that we restrict our attention to areas around each point of the trajectory where the non-linear dynamics is qualitatively equivalent to its linearisa-

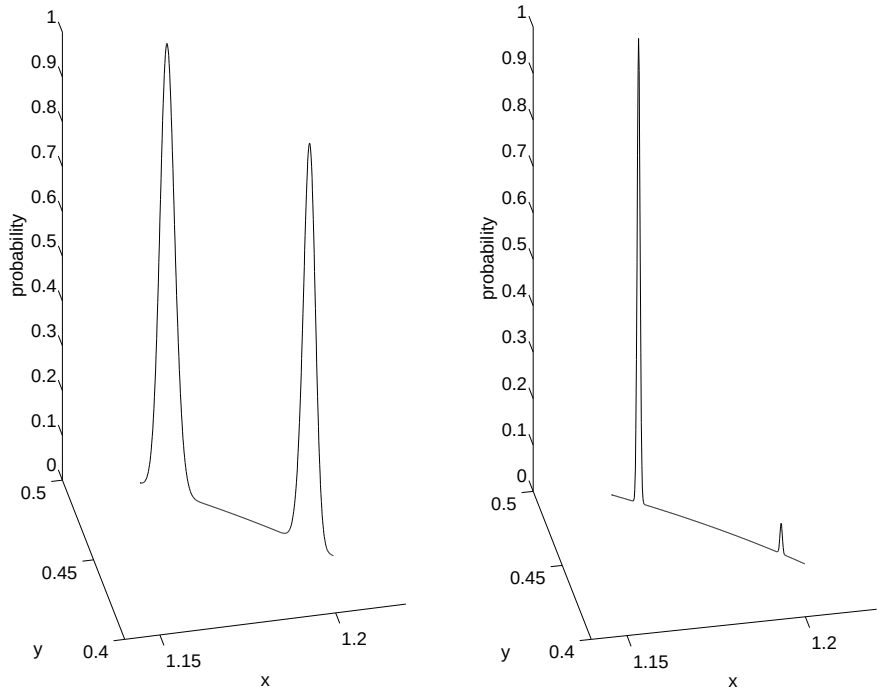

Figure 6: Probability that the fifty point trajectory of the Ikeda map whose 39^{th} point is (x,y) will be indistinguishable from the clean trajectory (whose 39^{th} point is $(1.167, 0.485)$). Here (x,y) varies over a section of the unstable manifold of $(1.167, 0.485)$. At left, the probability is computed assuming Gaussian noise with standard deviation $1/10$. At right, the standard deviation is $1/50$.

tion. That is, it is the essential non-linearity of the system (in the form of the curvature of its invariant manifold families) that stops noise reduction from working. Note also that it follows for *unbounded* noise distributions that we can never guarantee that the noise reduction will work, even for arbitrarily small noise levels, as there will always be a positive probability that the gradient descent will find a homoclinic intersection point. That is, for unbounded noise distributions, the homoclinic intersection points are never completely distinguishable from the true points (this is the geometric reason for Theorem 3 of [17]).

4 Analytic Results

We consider now the gradient descent algorithm from a theoretical point of view. The aim is to show that for systems where the angle between the stable and unstable manifolds is bounded below (*uniformly hyperbolic systems*), we can guarantee that for sufficiently small noise levels, the noise-reduced trajectories are excellent approximations of the original clean trajectory, and that as the length of the trajectories tends to infinity, the noise-reduced trajectories converge onto the clean trajectory everywhere except near the end points. As mentioned above, this result cannot be generalised to arbitrary noise levels, and the amount of noise that can be accommodated corresponds to the neighbourhoods of each point in which the dynamics and their linearisation are in qualitative agreement. Hence we shall begin by studying linear(ised) dynamical systems. First however, some general properties of gradient descent need to be addressed. In particular, we need to show that the gradient descent algorithm outlined above actually converges (Proposition 2 below). While this seems to be taken as obvious in the literature, the arguments usually given there are not complete, as they ignore the fact that the fixed points of the gradient descent are not isolated.

An outline of the proof that gradient descent does indeed give a satisfactory approximation of the true trajectory is as follows. We study the properties of the gradient descent algorithm for a linear dynamical system, for the reason mentioned above. The linearity of the system translates into a linear gradient descent algorithm, and this fact allows us to derive analytic bounds for the errors between the noise-reduced trajectory and (any suitable) candidate for the true trajectory (see Proposition 4). The proof of these bounds generalises immediately to a suitable linearisation of a general (uniformly hyperbolic) dynamical system, and this linearisation corresponds to the linearisation of the general gradient descent algorithm, about some fixed point (which we may take to be the true trajectory).

Noting that this gradient descent algorithm is in fact topologically conjugate to its linearisation, we construct a commutative diagram relating the gradient descent flow with its linearisation. The idea now is to use this commutative diagram to translate the analytic bounds we have derived for the linearised case, to the full non-linear case,

thus proving that the gradient descent algorithm achieves noise reduction. To do this, we need some quantitative information about the topological conjugacy between the gradient descent flow and its linearisation. This seems to be quite difficult. Instead, we introduce a semi-conjugacy whose properties are more amenable to analysis. A generalisation of the Hartman-Grobman Theorem ([20, 16, 21]) and Condition 7 below, then provide this information.

This proof is fairly long, so a few details have been omitted. These omissions are explicitly noted in what follows however. In particular, we have omitted a result concerning the existence and Hölder continuity of some conjugacies needed in section 4.3. The existence follows from the extension of the Hartman-Grobman Theorem mentioned above ([16]), and it should be plausible at least that the conjugacies implied by this theorem are Hölder continuous (this is certainly true for the standard Hartman-Grobman Theorem). All the relevant details can be found in [21].

4.1 General Properties of Gradient Descent

We now suppose that $f : M \to M$ is a C^2-diffeomorphism defining a discrete dynamical system on a d-dimensional manifold M which will be assumed smooth and *compact*. However, as we are only concerned with small noise levels, we can (and will) always locally identify M with $\mathbb{R}^d$. As in section 2, trajectories of length n are given by vectors in $\mathbb{R}^{nd}$: $x = (x_1, \ldots, x_n)$, $x_i \in \mathbb{R}^d$ and the gradient descent algorithm consists of solving equation 2:

$$\dot{x}(t) = -\nabla L(x(t)), \qquad x(0) = x,$$

and letting the descent-time t tend to infinity. Here, x represents the noisy trajectory, and L is the determinism function defined by equation 1. Now, $L(x) = 0$ if and only if the x_i form a deterministic trajectory for f, and clearly the deterministic trajectories are critical points of L. Conversely, by differentiating L:

$$\frac{\partial L}{\partial x_i} = \begin{cases} -df(x_1)^*(x_2 - f(x_1)) & \text{if } i = 1 \\ (x_i - f(x_{i-1})) - df(x_i)^*(x_{i+1} - f(x_i)) & \text{if } i = 2, \ldots, n-1 \\ (x_n - f(x_{n-1})) & \text{if } i = n \end{cases} \qquad (4)$$

,

it is easily checked that these are the only critical points (here * denotes matrix transposition). If these critical points were *isolated*, then the gradient descent would have to converge to one of them, regardless of the initial point ([9]). However, the deterministic trajectories are not isolated — they form a smooth manifold parameterised continuously by the first coordinate (for instance). Therefore, more consideration is required before convergence to a deterministic trajectory can be claimed.

Choose a deterministic trajectory y. This is a fixed point of the gradient descent flow. With $q = (q_1, \ldots, q_n) \equiv \nabla L : \mathbb{R}^{nd} \to \mathbb{R}^{nd}$ defined by equation 4, the *linearisation* of the

gradient descent flow about the fixed point y is given by

$$\dot{w}(t) = -dq(y)\,w(t), \qquad w(0) = x - y. \tag{5}$$

Note that we need to assume that f is C^2 so that $dq(y)$ exists. A quick computation shows that $dq(y)$ has the block-tridiagonal structure:

$$dq(y) = \begin{pmatrix} A_1^*A_1 & -A_1^* & & & & & \\ -A_1 & I + A_2^*A_2 & -A_2^* & & & & \\ & -A_2 & I + A_3^*A_3 & -A_3^* & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -A_{n-2} & I + A_{n-1}^*A_{n-1} & -A_{n-1}^* & \\ & & & & -A_{n-1} & I & \end{pmatrix}, \tag{6}$$

where $A_i = df(y_i)$. The deterministic trajectories for the linearised gradient descent are defined to be those satisfying

$$w_{i+1} = A_i w_i = df(y_i)\,w_i \quad \Rightarrow \quad w_{i+1} = df^i(y_1)\,w_1$$

for all $i = 1, \ldots, n-1$. The effect of the linearised gradient descent is now easy to determine.

Proposition 1 *The linearised gradient descent algorithm given by solving equation 5 and letting $t \to \infty$, is equivalent to projecting orthogonally onto the subspace of deterministic trajectories.*

Proof: Clearly the set of deterministic trajectories form a linear subspace. Now $dq(y)$ is symmetric (by inspection, but also because it is a *Hessian* matrix for L), and a quick computation shows that it is in fact positive semi-definite:

$$\langle w, dq(y)\,w \rangle = \sum_{i=1}^{n-1} \|A_i w_i - w_{i+1}\|^2 \geqslant 0. \tag{7}$$

The solution of the linearised gradient descent equations is of course

$$w(t) = e^{-dq(y)t} w(0),$$

and since $dq(y)$ is positive semi-definite, it follows that

$$e^{-dq(y)t} \longrightarrow \mathscr{P}$$

as $t \to \infty$, where $\mathscr{P}$ is the orthogonal projection onto $\ker dq(y)$. That is, $w(t) \to \hat{w} = \mathscr{P} w(0)$. It remains to show then, that $\ker dq(y)$ is the deterministic subspace. But,

$\langle w, dq(y)\,w\rangle = 0 \iff dq(y)\,w = 0$ (because $dq(y)$ possesses a symmetric square root). Hence this is an easy consequence of equation 7. ■

We are now in a position to reconsider the convergence of the gradient descent algorithm. Essentially, the fact that the fixed points are not isolated is compensated for by the fact that the gradient descent algorithm approaches the fixed point set *orthogonally*. This does not seem to be, however, a direct consequence of the fact that gradient descent algorithms always pass through surfaces of constant "height" (for us, constant L) orthogonally - a little sketching will imply that generally this orthogonality need not be continued to the fixed point set.

Proposition 2 *The gradient descent algorithm is guaranteed to converge onto a deterministic trajectory.*

Proof: It follows from Proposition 1 that for each fixed point, the linearised dynamics has a centre eigenspace ($\ker dq(y)$) and a stable eigenspace which is the orthogonal complement of the centre eigenspace. By the Centre Manifold Theorem ([24]), the non-linear gradient descent flow then possesses centre and stable manifolds, tangent to these respective eigenspaces. The stable manifold is clearly the set of all initial conditions which give y after gradient descent. Now, the set of deterministic trajectories of the non-linear system may be represented as the graph of a smooth function $\left(f, f^2, \ldots, f^{n-1}\right) : M \to M^{n-1}$, and so forms a smooth submanifold of M^n (the n-fold Cartesian product of M). As q is constant (zero) on this submanifold, its tangent space at y is contained in $\ker dq(y)$, the centre eigenspace. But, both these linear spaces have dimension d, so they are equal. Hence the submanifold of deterministic trajectories is tangent at y to the centre eigenspace, and as this submanifold is clearly invariant under q, the deterministic trajectories form a centre manifold for the non-linear gradient descent. But y was an arbitrary deterministic trajectory, so it follows that the submanifold of deterministic trajectories is a centre manifold for *every* fixed point of the non-linear gradient descent. We denote this submanifold by $\mathscr{W}_c$.

Note that $\mathscr{W}_c$ is closed (since L is continuous) hence compact. It is also smooth, so it follows that the centre eigenspace at each point of $\mathscr{W}_c$ varies continuously with the point. Each stable eigenspace is the orthogonal complement of the corresponding centre eigenspace so these also vary continuously with the point. Therefore, there is a continuous splitting along the compact invariant set $\mathscr{W}_c$ into stable and centre eigenspaces. By the Generalised Centre Manifold Theorem ([24]), the (local) generalised stable manifolds corresponding to each point in $\mathscr{W}_c$ vary continuously. These are of course just the stable manifolds for each fixed point. It follows now that there is an open neighbourhood of $\mathscr{W}_c$ which is *laminated* by stable manifolds, meaning that the (disjoint) union of these stable manifolds contains the entire neighbourhood. Any point in this neighbourhood will therefore end up on $\mathscr{W}_c$ after the gradient descent algorithm has been completed.

By extending this to global manifolds and making use of the compactness of M once more, it can be concluded that the global stable manifolds laminate all of M^n. Hence every point of M^n (corresponding to every noisy trajectory) belongs to a (unique) stable manifold, so the gradient descent algorithm must always converge to a point on the centre manifold. But, we have already proved that the centre manifold consists of points fixed under the non-linear gradient descent. These points correspond to deterministic trajectories, completing the proof. ■

4.2 Linear Dynamical Systems

We start by investigating the case where the dynamical map f is linear (and acts on $\mathbb{R}^d$). For clarity, this linear map will be denoted by A (indicating that we think of it as a matrix) rather than f. As hyperbolicity is necessary for noise reduction, we assume that A is a hyperbolic matrix with stable and unstable eigenspaces denoted by E_s and E_u respectively. The corresponding eigenprojections are denoted by P_s and P_u. These are complementary but not generally orthogonal. The following simple result is needed (the proof is very easy and may be found in [21]).

Lemma 3 *Suppose that $a_1,\ldots,a_n$ is a set of non-negative numbers satisfying $a_j \leqslant C\kappa^{j-i}a_i$ for all $j \geqslant i$ where $a_1 > 0$, and $0 \leqslant \kappa < 1$ and $C > 0$ are constants. Then,*

$$\frac{\left(\sum_{j=1}^n a_j\right)^2}{\sum_{j=1}^n a_j^2} \leqslant \frac{1+(2C-1)\,\kappa}{1-\kappa}\,.$$

The gradient descent algorithm for a linear dynamical system is equivalent to projecting orthogonally onto the subspace of deterministic trajectories (Proposition 1). We shall investigate the theoretical properties of the gradient descent algorithm by deriving quantitative information about this orthogonal projection, $\mathscr{P}$. The quantitative information that we have however, is in the form of the following well-known inequalities:

$$\|A^n v_s\| \leqslant C_s \mu^n \|v_s\| \qquad \text{and} \qquad \|A^n v_u\| \geqslant C_u \nu^n \|v_u\|\,, \tag{8}$$

which hold for all $n \geqslant 0$, $v_s \in E_s$, $v_u \in E_u$, and $\mu < 1 < \nu$ such that μ (ν) is larger (smaller) than any of the moduli of the eigenvalues of A inside (outside) the unit circle, and for some constants $C_s \geqslant 1$ and $0 < C_u \leqslant 1$ depending only on μ and ν respectively ([21]). (We will refer to such μ and ν as *hyperbolicity bounds*.) The analysis of this information and how it pertains to the projection $\mathscr{P}$ is complicated by the fact that the stable and unstable eigenspaces of A need not be orthogonal. It will be convenient to

consider the minimal angle between these subspaces. For two subspaces E and E' of a Euclidean space, the minimal angle θ is defined to be the acute angle satisfying

$$\cos\theta = \sup\left\{ \frac{\langle x, x'\rangle}{\|x\|\,\|x'\|} : x \in E \setminus \{0\} \text{ and } x' \in E' \setminus \{0\} \right\}.$$

The norm $\|\cdot\|$ denotes the Euclidean norm (on $\mathbb{R}^d$ or $\mathbb{R}^{nd}$). We also define the norm $\|\cdot\|_\infty$ on $\mathbb{R}^{nd}$ by $\|x\|_\infty = \max_i \|x_i\|$, and a norm $\|\cdot\|_*$ on the linear maps from $\mathbb{R}^{nd}$ to $\mathbb{R}^d$ by $\|T\|_* = \sup_{\|x\|_\infty=1} \|Tx\|$.

Proposition 4 *Suppose that A is a hyperbolic linear operator from $\mathbb{R}^d$ into itself, with stable and unstable eigenprojections P_s and P_u respectively, and $\mathscr{P}$ is the orthogonal projection in $\mathbb{R}^{nd}$ onto $\mathscr{E}_c$, the subspace of deterministic trajectories for A. Then, if $\mu < 1 < \nu$ are hyperbolicity bounds for A, and C_s and C_u are the associated constants, then the following bounds hold:*

$$\|P_s \pi_i \mathscr{P}\|_* \leqslant \dim E_s \frac{C_s \mu^{i-1}}{\sin\phi} \left(\frac{\sqrt{1+(2C_s-1)\,\mu}}{\sin\phi\sqrt{1-\mu}} + \frac{\sqrt{1+\left(2C_u^{-1}-1\right)\nu^{-1}}}{\tan\phi\sqrt{1-\nu^{-1}}} \right)$$

$$\|P_u \pi_i \mathscr{P}\|_* \leqslant \dim E_u \frac{C_u^{-1} \nu^{-(n-i)}}{\sin\phi} \left(\frac{\sqrt{1+\left(2C_u^{-1}-1\right)\nu^{-1}}}{\sin\phi\sqrt{1-\nu^{-1}}} + \frac{\sqrt{1+(2C_s-1)\,\mu}}{\tan\phi\sqrt{1-\mu}} \right)$$

where π_i projects out the i^{th} point of a trajectory ($\pi_i x = x_i$), and ϕ is the minimal angle between E_s and E_u.

Proof: Let $\mathfrak{E}_s$ and $\mathfrak{E}_u$ be the deterministic trajectories whose points are in E_s and E_u respectively. That is, let

$$\mathfrak{E}_s = \left\{ \begin{pmatrix} v \\ Av \\ A^2 v \\ \cdots \\ A^{n-1} v \end{pmatrix} : v \in E_s \right\} \qquad \text{and} \qquad \mathfrak{E}_u = \left\{ \begin{pmatrix} v \\ Av \\ A^2 v \\ \cdots \\ A^{n-1} v \end{pmatrix} : v \in E_u \right\}.$$

Since A is hyperbolic, $\mathbb{R}^d = E_s \oplus E_u$, and this induces the decomposition $\mathscr{E}_c = \mathfrak{E}_s \oplus \mathfrak{E}_u$, since $\mathscr{E}_c$ is the subspace of deterministic trajectories of A in $\mathbb{R}^{nd}$. If $\pi(E, E')$ denotes the projection onto the subspace E parallel to the subspace E'[4], then $\mathscr{P}$ may be decomposed as

$$\mathscr{P} = \pi\left(\mathfrak{E}_s, \mathfrak{E}_s^\perp\right) + \pi\left(\mathfrak{E}_s^\perp, \mathfrak{E}_s\right) \tag{9}$$

[4]That is, $\pi(E, E')$ is the unique projection with image E and kernel E'.

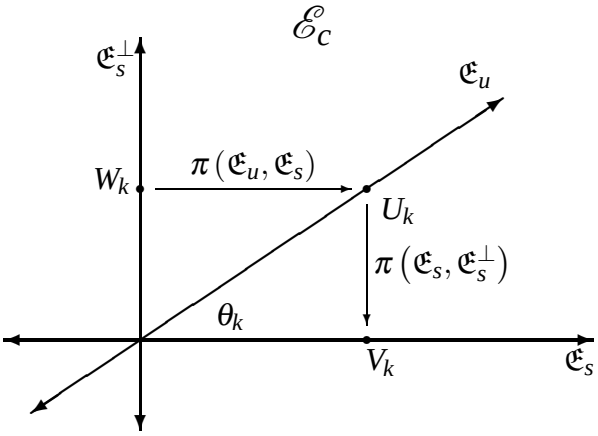

Figure 7: Construction of "unstable trajectories", $\{U_k\}$, and "stable trajectories", $\{V_k\}$, from the basis $\{W_k\}$ of $\mathfrak{E}_s^\perp$.

where $^\perp$ denotes orthogonal complementation. This obviously corresponds to a decomposition of $\mathscr{E}_c$ into $\mathfrak{E}_s \oplus \mathfrak{E}_s^\perp$, so the idea is to rewrite $\pi\left(\mathfrak{E}_s^\perp, \mathfrak{E}_s\right)$ so that it involves $\mathfrak{E}_u$. The constructions which achieve this are indicated schematically in Figure 7 for convenience.

We take an orthogonal basis, $\{W_k\}$ for $\mathfrak{E}_s^\perp$. Each W_k may be uniquely decomposed as $U_k + V_k$ where $U_k \in \mathfrak{E}_u$ and $V_k \in \mathfrak{E}_s$, and if θ_k is the acute angle between U_k and V_k (or $\pi/2$ if $V_k = 0$), then we have $\|W_k\| = \|U_k\| \sin\theta_k = \|V_k\| \tan\theta_k$. The reason why we chose the W_k to be an orthogonal basis of $\mathfrak{E}_s^\perp$ is that the (orthogonal) projection onto this subspace may be expanded as:

$$\pi\left(\mathfrak{E}_s^\perp, \mathfrak{E}_s\right) = \sum_{k=1}^{d_u} \frac{W_k W_k^*}{\|W_k\|^2} = \sum_{k=1}^{d_u} \|W_k\|^{-2} \left(U_k U_k^* - U_k V_k^* - V_k U_k^* + V_k V_k^*\right),$$

where $d_u = \dim \mathfrak{E}_s^\perp = \dim \mathfrak{E}_u = \dim E_u$ (recall also that * denotes transposition). But, $P_u \pi_i(\mathfrak{E}_s) = \{0\}$, so the first term in the decomposition of $\mathscr{P}$ (equation 9) is annihilated by $P_u \pi_i$, as are two of the terms in the above expansion. A quick calculation gives

$$P_u \pi_i \mathscr{P} x = \sum_{k=1}^{d_u} \|W_k\|^{-2} A^{i-1} u_k \sum_{j=1}^{n} \left(\left\langle A^{j-1} u_k, x_j \right\rangle - \left\langle A^{j-1} v_k, x_j \right\rangle\right),$$

where $u_k = \pi_1 U_k \in E_u$ and $v_k = \pi_1 V_k \in E_s$. This expresses $\mathscr{P}$ in terms of vectors from the stable and unstable eigenspaces of A. The Cauchy-Schwarz inequality now gives us the bounds:

$$\|P_u \pi_i \mathscr{P} x\|_* \leqslant \sum_{k=1}^{d_u} \frac{\left\|A^{i-1} u_k\right\|}{\|U_k\| \sin\theta_k} \sum_{j=1}^{n} \left(\frac{\left\|A^{j-1} u_k\right\|}{\|U_k\| \sin\theta_k} + \frac{\left\|A^{j-1} v_k\right\|}{\|V_k\| \tan\theta_k} \right) \|x\|_\infty . \qquad (10)$$

(If any of the V_k are zero, the corresponding v_k are zero, and so the second term in the parentheses above is zero.)

Consider now the term $\sum_{j=1}^{n} \left\|A^{j-1}v_k\right\| / \|V_k\| = \left(\sum_{j=1}^{n} \left\|A^{j-1}v_k\right\|\right) / \left[\sum_{j=1}^{n} \left\|A^{j-1}v_k\right\|^2\right]^{1/2}$. If $a_j = \left\|A^{j-1}v_k\right\|$, then we have $a_j = \left\|A^{j-1}v_k\right\| \leqslant C_s\mu^{j-i}\left\|A^{i-1}v_k\right\| = C_s\mu^{j-i}a_i$ where $\mu < 1$ is a (stable) hyperbolicity bound for A, and $C_s \geqslant 1$ is the associated constant. By Lemma 3 then,

$$\sum_{j=1}^{n} \frac{\left\|A^{j-1}v_k\right\|}{\|V_k\|} \leqslant \left[\frac{1+(2C_s-1)\,\mu}{1-\mu}\right]. \tag{11}$$

Similarly, if $a_j = \left\|A^{n-j}u_k\right\|$, then $a_j \leqslant C_u^{-1}\nu^{-(j-i)}a_i$ where $\nu > 1$ is an (unstable) hyperbolicity bound for A and $C_u \leqslant 1$ is the associated constant. Therefore,

$$\sum_{j=1}^{n} \frac{\left\|A^{j-1}u_k\right\|}{\|U_k\|} \leqslant \left[\frac{1+\left(2C_u^{-1}-1\right)\nu^{-1}}{1-\nu^{-1}}\right]. \tag{12}$$

Noting that we also have $\left\|A^{i-1}u_k\right\| / \|U_k\| \leqslant \left\|A^{i-1}u_k\right\| / \left\|A^{n-1}u_k\right\| \leqslant C_u^{-1}\nu^{-(n-i)}$, we derive from equations 10, 11, and 12, the estimate

$$\|P_u\pi_i\mathscr{P}\|_* \leqslant \sum_{k=1}^{d_u} \frac{C_u^{-1}\nu^{-(n-i)}}{\sin\theta_k}\left(\frac{\sqrt{1+\left(2C_u^{-1}-1\right)\nu^{-1}}}{\sin\theta_k\sqrt{1-\nu^{-1}}} + \frac{\sqrt{1+(2C_s-1)\,\mu}}{\tan\theta_k\sqrt{1-\mu}}\right).$$

This bound expresses the norm of $P_u\pi_i\mathscr{P}$ in terms of the constants μ, ν, C_s, C_u and d_u — which depend on the hyperbolic linear operator A and not on the length of the trajectory n — and the angles θ_k. As the θ_k are angles between the *trajectories* U_k and V_k, they will generally vary with n. It remains then to show that they are bounded away from zero, so that $\sin\theta_k$ and $\tan\theta_k$ do not vanish as n tends to infinity. If ϕ is the minimal angle between the eigenspaces E_s and E_u (which only depends on A), then a simple computation using the Cauchy-Schwarz inequality for sums shows that $|\cos\theta_k| \leqslant \cos\phi$. Therefore, $\sin\theta_k \geqslant \sin\phi$ and $\tan\theta_k \geqslant \tan\phi$, so substitution gives the required unstable bound. The stable bound (for $P_s\pi_i\mathscr{P}$) is derived using the same technique, with s and u interchanged. ■

The relevance of this result is seen by noting that if x denotes the noisy trajectory, $\hat{x}$ the noise-reduced trajectory, and y the clean trajectory, then the error in comparing the noise-reduced and clean trajectories at the i^{th} point is

$$\|\pi_i(\hat{x}-y)\| = \|\pi_i\mathscr{P}(x-y)\| \leqslant \|P_s\pi_i\mathscr{P}(x-y)\| + \|P_u\pi_i\mathscr{P}(x-y)\|.$$

If the noise distribution is bounded (by ε say), then Proposition 4 states that the error in comparing the noise-reduced and clean trajectories at the i^{th} point satisfies

$$\|\pi_i(\hat{x}-y)\| \leqslant \left(K_s\mu^{i-1} + K_u\nu^{-(n-i)}\right)\varepsilon \tag{13}$$

where K_s and K_u are constants independent of i or n, the length of the trajectory. It follows now that these errors can be made small everywhere (except near the end points) by taking the trajectory length sufficiently long. So we have proved the following result.

Theorem 5 *Let A be a hyperbolic linear operator defining a discrete dynamical system on $\mathbb{R}^d$, $x \in \mathbb{R}^{nd}$ be a noisy trajectory, and $\hat{x}$ be the noise reduced trajectory given by the gradient descent algorithm. If the noise distribution is* bounded, *then the points of* any *deterministic trajectory that could be the true trajectory, differ from the points of $\hat{x}$ by an amount which tends to zero as n, the length of the trajectories, tends to infinity, except for points near the initial and final points. The errors at these points remain* bounded *as $n \to \infty$.*

We have already remarked that a corresponding result for unbounded noise distributions is untenable — the errors cannot be absolutely bounded. However, the proof of Proposition 4 can be trivially adapted to show that for these distributions, the *root-mean-square* errors at each point of the trajectory are bounded by the same expressions as before, but with ε denoting the standard deviation of the noise distribution. This also extends to confidence levels. For unbounded noise distributions, the errors can be bounded "on average".

4.3 Non-linear Dynamical Systems

We now turn to the problem of generalising Theorem 5 to non-linear dynamical systems. Of course, the systems under consideration must be hyperbolic, and the results of section 3 show that we must restrict further to systems where the angle between stable and unstable manifolds is bounded below. An important class of systems which satisfy this requirement is the class of *uniformly* hyperbolic dynamical systems ([13]). These are systems which possess a hyperbolic set (each point of the set has complementary generalised stable and unstable eigenspaces) which is invariant and compact.

First, we consider the linearisation of such a system. If $f : M \to M$ is uniformly hyperbolic, and y is a deterministic trajectory for f, then Proposition 1 asserts that the effect of the linearised gradient descent flow is to project orthogonally onto the subspace

$$\mathscr{E}_c^{\mathscr{O}} = \left\{ \begin{pmatrix} v \\ df(y_1)v \\ df^2(y_1)v \\ \vdots \\ df^{n-1}(y_1)v \end{pmatrix} : v \in \mathbb{R}^d \right\}.$$

We can think of these trajectories as deterministic trajectories for a linear system where the linear operator changes with each iteration. It is easy now to generalise Proposition

4 to this case. We still have hyperbolicity bounds $\mu < 1 < \nu$ (which are independent of the point of the uniformly hyperbolic set), the minimal angle, ϕ, between (generalised) eigenspaces is still non-zero, and the estimates of equation 8 are replaced by

$$\|df^n(p)\,x_s\| \leqslant C_s\mu^n\,\|x_s\| \qquad \text{and} \qquad \|df^n(p)\,x_u\| \geqslant C_u\nu^n\,\|x_u\|\,, \tag{14}$$

where p is an arbitrary point of the uniformly hyperbolic set, x_s and x_u are elements of the generalised stable and unstable eigenspaces (respectively) at p, and C_s and C_u are constants depending only on μ and ν respectively ([21]). These estimates are consequences of the Multiplicative Ergodic Theorem ([19]), and the fact that the constants C_s and C_u may be chosen independent of the point p is due to working over a uniformly hyperbolic set, which is compact by definition.

Proposition 6 *Suppose that f is a C^2-diffeomorphism of a smooth compact d-dimensional manifold M possessing an invariant uniformly hyperbolic set Λ with splitting into stable and unstable eigenspaces $E_s(p)$ and $E_u(p)$, $p \in \Lambda$, and that y is a deterministic trajectory of length n for f. If $\mathscr{P}$ is the orthogonal projection (in $\mathbb{R}^{nd}$) onto $\mathscr{E}_c(y)$, the subspace of deterministic trajectories for the system linearised about y, and $P_s^{(i)}$ and $P_u^{(i)}$ are the stable and unstable projections onto $E_s(y_i)$ and $E_u(y_i)$ for $i = 1,\ldots,n$ (respectively), then the following bounds hold:*

$$\left\|P_s^{(i)}\pi_i\mathscr{P}\right\|_* \leqslant d_s\frac{C_s\mu^{i-1}}{\sin\phi}\left(\frac{\sqrt{1+(2C_s-1)\,\mu}}{\sin\phi\sqrt{1-\mu}}+\frac{\sqrt{1+\left(2C_u^{-1}-1\right)\nu^{-1}}}{\tan\phi\sqrt{1-\nu^{-1}}}\right)$$

$$\left\|P_u^{(i)}\pi_i\mathscr{P}\right\|_* \leqslant d_u\frac{C_u^{-1}\nu^{-(n-i)}}{\sin\phi}\left(\frac{\sqrt{1+\left(2C_u^{-1}-1\right)\nu^{-1}}}{\sin\phi\sqrt{1-\nu^{-1}}}+\frac{\sqrt{1+(2C_s-1)\,\mu}}{\tan\phi\sqrt{1-\mu}}\right)$$

where $\mu < 1 < \nu$ are hyperbolicity bounds for $f|_\Lambda$, C_s and C_u are the associated constants, d_s and d_u are the common dimensions of the $E_s(p)$ and $E_u(p)$ (respectively), and ϕ is the minimal angle between $E_s(p)$ and $E_u(p)$, $p \in \Lambda$.

Proof: This proof is the same as that of Proposition 4 with a few modifications. In particular, A^m is replaced by $df^m(y_1)$ throughout. The subspaces $\mathfrak{E}_s$ and $\mathfrak{E}_u$ are then the trajectories in $\mathscr{E}_c(y)$ whose first point belongs to $E_s(y_1)$ and $E_u(y_1)$ respectively. The invariance of the $E_s(p)$ and the $E_u(p)$ given by the Multiplicative Ergodic Theorem and the fact that y was chosen to be a deterministic trajectory for f, show that $\mathfrak{E}_s$ and $\mathfrak{E}_u$ consist of trajectories whose points stay in stable and unstable eigenspaces (respectively). Hence, $\mathscr{E}_c(y) = \mathfrak{E}_s \oplus \mathfrak{E}_u$. Given an orthogonal basis of $\mathfrak{E}_s^{\perp}$ say, the construction of stable and unstable trajectories can proceed as in the proof of Proposition 4, and these can be used to derive the analogue of equation 10. Equation 14 and Lemma 3 are then used to

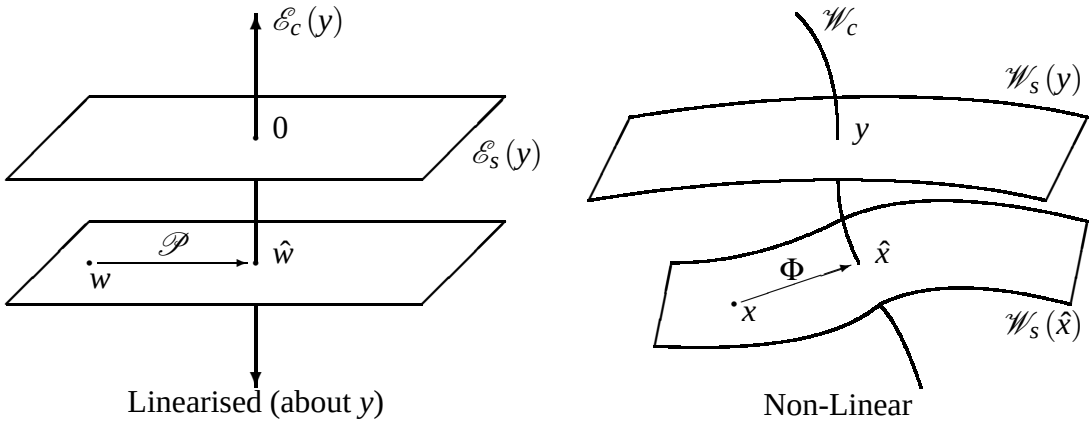

Figure 8: Stable manifold and eigenspace laminations in trajectory space $\mathbb{R}^{nd}$

simplify this expression, noting that because Λ is compact, the constants C_s and C_u may be chosen independently of the points of the trajectory y, and hence independent of n. The resulting expression still contains angles between stable and unstable trajectories — these are dealt with in exactly the same manner as in the proof of Proposition 4, noting that the angles between the $E_s(p)$ and the $E_u(p)$ are uniformly bounded away from zero. ■

Recall the proof of Proposition 2. There it was shown that for a non-linear system, the set of deterministic trajectories, $\mathscr{W}_c$, forms a centre manifold for *every* fixed point of the gradient descent flow, and there is a lamination of stable manifolds, $\{\mathscr{W}_s(y) : y \in \mathscr{W}_c\}$, orthogonal to this common centre manifold. The situation is exactly the same in the linearised case — here there is a *subspace* of deterministic trajectories which forms a centre eigenspace, and a lamination of stable eigenspaces given by the family of $(n-1)d$-dimensional hyperplanes parallel to $\ker \mathscr{P} = \mathscr{E}_c(y)$. These laminations are indicated in Figure 8. It would seem plausible then, that the non-linear gradient descent flow and its linearisation about some fixed point are qualitatively similar, that is, topologically conjugate, despite the presence of a centre manifold. This is in fact true ([14, 21])[5].

We exploit this qualitative equivalence by constructing a commutative diagram relating the non-linear and linearised gradient descent flows. The non-linear gradient descent equations define a flow φ^t which converges (given any initial condition) as $t \to \infty$ by Proposition 2. The pointwise limit of φ^t as $t \to \infty$ therefore defines a function Φ. Clearly Φ represents the effect of the non-linear gradient descent algorithm, just as the orthogonal projection $\mathscr{P}$ represents the effect of the linearised algorithm. We consider

[5]We will not make direct use of this fact, however, but mention it as motivation for the construction that follows. The proof appearing in [14] is quite involved and we were unable to extend it to get any quantitative information about the conjugacy. It has therefore been omitted.

functions $\mathscr{H}$, H, and h_i ($i = 1, \ldots, n$) which make the following diagram commute:

$$\begin{array}{ccccc} \mathbb{R}^{nd} & \xrightarrow{\Phi} & \mathscr{W}_c & \xrightarrow{\pi_i} & \mathbb{R}^d \\ \mathscr{H}\downarrow & & \uparrow H & & \uparrow h_i \\ \mathbb{R}^{nd} & \xrightarrow{\mathscr{P}} & \mathscr{E}_c(y) & \xrightarrow{\pi_i} & \mathbb{R}^d \end{array} \tag{15}$$

(this involves choosing a deterministic trajectory, y, about which to linearise). Of course, it is not just enough to know that these functions exist. To generalise Proposition 6 to non-linear systems, it is necessary to demand that the h_i take points near the stable and unstable eigenspaces of the linearised system to points near the stable and unstable manifolds of the non-linear system, and that the additional errors induced by using $\mathscr{H}$ and the h_i to switch between the non-linear and linearised spaces can be *bounded* as the length of the trajectories tend to infinity. In this way, the behaviour of the noise reduction procedure will be maintained and the required convergence result will follow easily.

Consider $H : \mathscr{E}_c(y) \to \mathscr{W}_c$. To make the right square of diagram 15 commute, it follows that the function H must decompose as $H = (h_1, \ldots, h_n)$. As H maps deterministic trajectories for the linearised system onto deterministic trajectories for the non-linear system, its action is entirely determined by what it does to the first point of the trajectory. That is, each h_i is determined by h_1:

$$h_i = f^{i-1} \circ h_1 \circ \left[df^{i-1}(y_1) \right]^{-1} = f^{i-1} \circ h_1 \circ df^{-(i-1)}(y_i).$$

Note that the action of h_i on a neighbourhood of y_i will be to map the unstable eigenspace for y_i back onto the unstable eigenspace for y_1, distort it somewhat (the action of h_1), and then map them forward to a neighbourhood of y_i again. For i large enough then (and provided that h_1 is chosen to be continuous and close to the identity say), the resulting set should be an excellent approximation (at least locally) of the generalised unstable manifold of y_i. In fact, there is a choice for h_1 which makes each h_i map each unstable eigenspace locally onto the corresponding local generalised unstable eigenspace *exactly*. This is a direct consequence of a generalisation of the Hartman-Grobman Theorem due to Kurata ([16]), which essentially states that around each point of a uniformly hyperbolic set, there are open neighbourhoods in which the dynamics is topologically conjugate to the linearised dynamics. Because the points need not be fixed (or periodic) as in the standard Hartman-Grobman Theorem, the conjugacies do not map each neighbourhood into itself, but rather into the neighbourhood corresponding to the next iterate. A detailed proof of Kurata's theorem may be found in [21]. With this choice, h_1 and hence each h_i is a local homeomorphism, and hence so is H. Furthermore, the domain of H can be naturally extended to the product of the domains of the h_i, so H maps a neighbourhood of y homeomorphically onto another neighbourhood of y.

Consider now the left square of diagram 15. As Φ is the identity on $\mathscr{W}_c$ and $\mathscr{P}$ is the identity on $\mathscr{E}_c(y)$, it follows that $\mathscr{H}$ and H^{-1} must coincide on $\mathscr{W}_c$. It would be very convenient if defining $\mathscr{H}$ to be H^{-1} (defined on a neighbourhood of y) made diagram 15 commute, at least around y. However, this does not seem to be the case. Instead, note that

$$\mathscr{P}\circ\mathscr{H} = H^{-1}\circ\Phi \qquad \Rightarrow \qquad \mathscr{H} = (I-\mathscr{P})\circ\mathscr{H} + H^{-1}\circ\Phi,$$

and that $(I-\mathscr{P})\circ\mathscr{H}$ takes values in $\mathscr{E}_s(y)$ whereas $H^{-1}\circ\Phi$ takes values in $\mathscr{E}_c(y)$. In fact, it is clear that the commutativity requirement will still be fulfilled if the $\mathscr{H}$ appearing on the *right* of this last equality is replaced by any function mapping $\mathscr{W}_c$ onto $\mathscr{E}_c(y)$. A convenient choice is the homeomorphism H^{-1}, as it is the only function satisfying this requirement whose properties we know. That is, we define

$$\mathscr{H} = (I-\mathscr{P})\circ H^{-1} + H^{-1}\circ\Phi. \tag{16}$$

$\mathscr{H}$ therefore maps a neighbourhood of y into another neighbourhood of y, and satisfies $\mathscr{P}\circ\mathscr{H} = \mathscr{H}\circ\Phi = H^{-1}\circ\Phi$ (whenever this makes sense). Geometrically, $\mathscr{H}$ takes the centre manifold $\mathscr{W}_c$ onto the centre eigenspace $\mathscr{E}_c(y)$, *and* maps each stable manifold of the non-linear lamination onto some stable eigenspace of the linearised lamination (see Figure 8). The term $H^{-1}\circ\Phi$ specifies *which* stable eigenspace corresponds to a particular stable manifold, and the term $(I-\mathscr{P})\circ H^{-1}$ specifies *where* on the stable eigenspace each point of the stable manifold is mapped.

It remains to consider the distortions induced by $\mathscr{H}$ and the h_i. That is, any stretching or contracting of distances caused by switching between the non-linear and linearised gradient descents. These will introduce extra factors and/or terms into our estimates for the errors we accrue when we noise-reduce (and so we need to control them). All these functions are *continuous* (Φ is continuous because the stable manifolds in the lamination vary continuously) on their respective domains, so this distortion can be made arbitrarily *small* by restricting their domains to be sufficiently small (this in turn corresponds to restricting the noise level to be sufficiently small). However, the generalisation of Theorem 5 to non-linear systems must address the behaviour as the length of the relevant trajectories, n, tend to infinity. Therefore it is necessary to know how the distortion varies with n.

Quantitative information for the various h_i comes in the form of their Hölder continuity. In [21], it is shown that there exist $\alpha,\beta>0$ such that

$$\left\|h_i(u) - h_i\left(u'\right)\right\| \leqslant \beta\left\|u-u'\right\|^{\alpha}$$

for each i, every u,u' belonging to the domain of h_i. Furthermore, α and β may be chosen independently of i (and thus n). That is, the distortion induced by using the h_i does not depend on the length of trajectory used. We omit the proof of these statements

because they follow from (a modification of) the proof of the Generalised Hartman-Grobman Theorem, which is too long to include here. Given these facts however, it easily follows that H is Hölder continuous with respect to the norm $\|\cdot\|_\infty$ and with the same constants α and β, and so is H^{-1} (because the inverses of the h_i are also Hölder continuous). What we would like to do now, is show that $\mathscr{H}$ is also Hölder continuous (with respect to $\|\cdot\|_\infty$) with constants independent of n. But this seems to be quite difficult, largely because we have no concrete quantitative information about the non-linear gradient descent flow Φ. Instead, we introduce a condition on Φ which suffices for our needs.

Condition 7 *Let $y_1 \in \mathbb{R}^d$ define deterministic trajectories $y^{(n)} \in \mathscr{W}_c \subset \mathbb{R}^{nd}$ (for each n) by $y_{i+1}^{(n)} = f\left(y_i^{(n)}\right)$, $i = 1,\ldots,n-1$, and let*

$$\mathscr{B}_\varepsilon\left(y^{(n)}\right) = \left\{x \in \mathbb{R}^{nd} : \left\|x - y^{(n)}\right\|_\infty \leqslant \varepsilon\right\}.$$

Then, for $\varepsilon > 0$ (denoting the noise level) sufficiently small but fixed, the function

$$\Omega_\varepsilon(n) = \sup_{x \in \mathscr{B}_\varepsilon\left(y^{(n)}\right)} \left\|\Phi(x) - y^{(n)}\right\|_\infty$$

is bounded above.

This condition does not guarantee that $\mathscr{H}$ is Hölder continuous. For that we would need to specify how $\Omega_\varepsilon(n)$ varies with ε. But it does put a bound on the size of the terms that $\mathscr{H}$ introduces into our estimates for the errors after noise reduction. Assuming the gradient descent satisfies Condition 7 then, we have for a clean trajectory $y \in \mathscr{W}_c$, a noisy trajectory x, the noise-reduced trajectory $\hat{x}$, and noise-bound ε, that (using the commutative diagram 15, the Hölder continuity of the h_i, and Proposition 6):

$$\begin{aligned}
\|\hat{x}_i - y_i\| &= \|\pi_i\Phi(x) - \pi_i\Phi(y)\| \\
&= \|h_i(\pi_i\mathscr{P}\mathscr{H}(x)) - h_i(\pi_i\mathscr{P}\mathscr{H}(y))\| \\
&\leqslant \beta\|\pi_i\mathscr{P}(\mathscr{H}(x) - \mathscr{H}(y))\|^\alpha \\
&\leqslant \beta\left(K_s\mu^{i-1} + K_u\nu^{-(n-i)}\right)^\alpha \|\mathscr{H}(x) - \mathscr{H}(y)\|_\infty^\alpha,
\end{aligned}$$

where K_s and K_u are constants bounded above in n, and (using the Hölder continuity of H^{-1} and Condition 7):

$$\begin{aligned}
\|\mathscr{H}(x) - \mathscr{H}(y)\|_\infty &\leqslant \left\|(I - \mathscr{P})\left(H^{-1}(x) - H^{-1}(y)\right)\right\|_\infty + \left\|H^{-1}\circ\Phi(x) - H^{-1}\circ\Phi(y)\right\|_\infty \\
&\leqslant \|I - \mathscr{P}\|_\infty\beta\|x - y\|_\infty^\alpha + \beta\|\Phi(x) - y\|_\infty^\alpha \\
&\leqslant \|I - \mathscr{P}\|_\infty\beta\varepsilon^\alpha + \beta\Omega_\varepsilon(n)^\alpha,
\end{aligned}$$

where $\|I-\mathscr{P}\|_\infty \leqslant 1+\sup_i \|\pi_i \mathscr{P}\|_* \leqslant 1+\sup_i \left(\|P_s \pi_i \mathscr{P}\|_* + \|P_u \pi_i \mathscr{P}\|_*\right)$ which is bounded above in n, and so finally,

$$\begin{aligned}\|\hat{x}_i - y_i\| &\leqslant \beta\left(K_s\mu^{i-1} + K_u\nu^{-(n-i)}\right)^\alpha \|\mathscr{H}(x)-\mathscr{H}(y)\|_\infty^\alpha \\ &\leqslant \beta^{1+\alpha}\left[\left(K_s\mu^{i-1}+K_u\nu^{-(n-i)}\right)\left(\|I-\mathscr{P}\|_\infty \varepsilon^\alpha + \Omega_\varepsilon(n)^\alpha\right)\right]^\alpha,\end{aligned}$$

which is bounded above in n. Thus, as n increases, the exponential decay of the terms μ^{i-1} and $\nu^{-(n-i)}$ for $i \sim n/2$, mean that the errors converge to zero away from the end points. That is, we have proved the following theorem:

Theorem 8 *Let f be a C^2-diffeomorphism of a smooth compact d-dimensional manifold M possessing an invariant uniformly hyperbolic set Λ and satisfying Condition 7, $x \in \mathbb{R}^{nd}$ be a noisy trajectory of the (non-linear) system, and $\hat{x}$ be the noise reduced trajectory given by the gradient descent algorithm. If the noise distribution is* bounded *by $\varepsilon > 0$ sufficiently small, then the points of* any *deterministic trajectory that could be the true trajectory, differ from the points of $\hat{x}$ by an amount which tends to zero as n, the length of the trajectories, tends to infinity, except for points near the initial and final points. The errors at these points are bounded.*

This result essentially states that the gradient descent algorithm is a good noise reduction algorithm for non-linear dynamical systems with an invariant uniformly hyperbolic set (that is, one without genuine tangencies), provided the noise level is sufficiently small. We do, however, rely on Condition 7 being satisfied. When does this condition hold? Perhaps a better question to ask would be: How could this condition possibly fail to hold? For a consequence of failure would be that the errors at the initial and final points could grow without bound as the length of the trajectory increases. This is certainly at odds with the numerical experiments of section 3, although these experiments are of course, not even remotely exhaustive. However, we know from Bowen's Shadowing Theorem ([3]), that for sufficiently small noise, there is a *unique* deterministic trajectory that could produce any given noisy trajectory *of infinite length*. One would hope that a respectable noise reduction algorithm would converge (pointwise, not uniformly) onto this unique trajectory as the length of trajectory tends to infinity. It seems reasonable therefore to *conjecture* that for any uniformly hyperbolic dynamical system, Condition 7 is satisfied. Of course, the compactness of our manifold M means that M has a finite diameter, so we can always claim that the conjecture holds in this limited sense. However, this is clearly not as satisfactory as we would like.

We would also like to mention that these results also clarify the role of the noise level in noise reduction processes. Theorem 5 essentially states that gradient descent noise reduction will work asymptotically for any hyperbolic linear dynamical system, *regardless of the noise level*. Clearly one should not expect the same result to be true

for non-linear systems — it would, for instance, be rather amazing if we could recover a signal (asymptotically) when the noise level far exceeds the size of the attractor that the signal comes from. What Theorem 8 states (and the results of section 3 demonstrate) is that recovery may be achieved if the noise level is smaller than the size of the neighbourhoods in which the non-linear dynamics is qualitatively equivalent to its linearisation. That is, the noise level must be smaller than the smallest distance between a trajectory point and its closest homoclinic intersection point.

5 Discussion and Conclusions

This paper has demonstrated two important results: one by numerical means and another by analytical means.

The first result, shown using numerical experiments, is that the failure of the gradient descent algorithm is a little worse than supposed in earlier studies, in that the failure is a theoretical consequence of the combination of near tangencies and sufficiently large noise levels, and does not require the presence of an exact tangency anywhere in the system. Instead, it is the presence of a nearby homoclinic intersection point which can cause the failure. As indistinguishability theory states that the two trajectories passing through the actual point and its nearby homoclinic intersection point are both consistent with the noisy data (for sufficiently large noise), this implies that nearby homoclinic intersection points (that is, near tangencies) will have a similar effect on any other conceivable noise reduction algorithm. Therefore, the presence of near tangencies is a fundamental theoretical limitation which can cause any noise reduction algorithm to fail.

The second result is the proof of the convergence of the gradient descent algorithm under specified conditions. The proof relies on two facts. First, that a (semi-)conjugacy ($\mathscr{H}$) can be constructed between the gradient descent flow of a uniformly hyperbolic system and its linearisation about some fixed point, and second, that analytic bounds for the errors between the noise-reduced and true trajectories can be derived for the linearised gradient descent flow. This confirms (among other things), a loosely justified expectation of state estimation theory that appears in Judd and Smith ([12], Dictum 1).

These results are significant not only for what they say about noise reduction by gradient descent, but what they also imply about shadowing trajectories and state estimation. Recently Judd ([11]) has shown using numerical experiments that the gradient descent algorithm (extended to the imperfect model case) is superior to the Extended Kalman Filter for estimating the state of nonlinear systems. Furthermore, finding shadowing trajectories has recently been recognised as an important technique for assessing the quality of imperfect models. Admittedly, the results presented here only deal with the perfect model scenario, but it is hoped that, and seems likely that, these results will generalise to parametrised models, and to imperfect models to some extent.

6 Acknowledgements

The authors would like to thank G Froyland for significant contributions, and M Shub for answering some of our technical questions regarding his theory of Hölder foliations. DR was supported by an University Postgraduate Award and a Jean Rogerson Scholarship.

References

[1] B D O Anderson and J B Moore. *Linear Optimal Control*. Prentice-Hall, New Jersey, 1971.

[2] B D O Anderson and J B Moore. *Optimal Filtering*. Prentice-Hall, New Jersey, 1979.

[3] R Bowen. *On Axiom A Diffeomorphisms*, volume 35 of *Regional Conference Series in Mathematics*. American Mathematical Society, Providence, 1978.

[4] M Davies. Noise Reduction Schemes for Chaotic Time Series. *Physica D*, 79:174–192, 1994.

[5] J D Farmer and J J Sidorowich. Optimal Shadowing and Noise Reduction. *Physica D*, 47(3):373–392, 1991.

[6] P Grassberger, R Hegger, H Kantz, C Schaffrath, and T Schreiber. On Noise Reduction Methods for Chaotic Data. *CHAOS*, 3(2):127–141, 1993.

[7] C Grebogi, S Hammel, J A Yorke, and T Sauer. Shadowing of Physical Trajectories in Chaotic Dynamics: Containment and Refinement. *Physical Review Letters*, 65:1527–1530, 1990.

[8] S M Hammel. A Noise Reduction Method for Chaotic Systems. *Physics Letters A*, 148(8,9):421–428, September 1990.

[9] M W Hirsch and S Smale. *Differential Equations, Dynamical Systems and Linear Algebra*, volume 60 of *Pure and Applied Mathematics*. Academic Press, New York, 1974.

[10] K Ikeda. Multiple-valued Stationary State and its Instability of the Transmitted Light by a Ring Cavity System. *Optics Communications*, 30(2):257–261, August 1979.

[11] K Judd. Nonlinear State Estimation, Indistinguishable States and the Extended Kalman Filter. Submitted, 2001.

[12] K Judd and L Smith. Indistinguishable States I: Perfect Model Scenario. *Physica D: Nonlinear Phenomena*, 151(2-4):125–141, May 2001.

[13] A Katok and B Hasselblatt. *Introduction to the Modern Theory of Dynamical Systems*, volume 54 of *Encyclopaedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1998.

[14] U Kirchgraber and K J Palmer. *Geometry in the Neighbourhood of Invariant Manifolds of Maps and Flows and Linearisation*, volume 233 of *Pitman Research Notes in Mathematics*. Longman Scientific and Technical, Essex, 1990.

[15] E J Kostelich and J A Yorke. Noise Reduction in Dynamical Systems. *Physical Review A*, 38(3):1649–1652, August 1988.

[16] M Kurata. Hartman's Theorem for Hyperbolic Sets. *Nagoya Mathematical Journal*, 67:41–52, 1977.

[17] S P Lalley. Beneath the Noise, Chaos. *Annals of Statistics*, 27(2):461–479, 1999.

[18] S P Lalley. Removing the Noise from Chaos plus Noise. In A I Mees, editor, *Nonlinear Dynamics and Statistics: Proceedings of an Isaac Newton Institute Workshop*. Birkhauser, Boston, 2000.

[19] V I Oseledec. A Multiplicative Ergodic Theorem. Lyapunov Characteristic Numbers for Dynamical Systems. *Transactions of the Moscow Mathematical Society*, 19:197–221, 1968.

[20] C C Pugh. On a Theorem of P Hartman. *American Journal of Mathematics*, 91:363–367, 1969.

[21] D Ridout. Convergence Properties of Noise Reduction by Gradient Descent. Master's thesis, The University of Western Australia, 2001. Available from `www.cado.uwa.edu.au/Reports.php3`.

[22] D Ruelle. Ergodic Theory of Differentiable Dynamical Systems. *Publications Mathematiques. Institut des Hautes Etudes scientifiques*, 50:27–58, 1979.

[23] T Schreiber. An Extremely Simple Nonlinear Noise Reduction Method. *Physical Review E*, 47(4):2401–2404, 1993.

[24] M Shub. *Global Stability of Dynamical Systems*. Springer-Verlag, New York, 1987.

[25] L A Smith. Accountability in Ensemble Prediction. In *Predictability*, volume 1 of *ECMWF Workshop Proceedings*, pages 351–368, Shinfield Park, Reading, UK, 1996. ECMWF.