Ensemble data assimilation for hyperbolic systems

Cecilia González-Tokman*

School of Mathematics and Statistics, University of New South Wales. Sydney, NSW, 2052, Australia.

Brian R. Hunt*

Department of Mathematics and Institute for Physical Sciences and Technology, University of Maryland, College Park, MD 20742, USA.

Abstract

Ensemble based methods are now widely used in applications such as weather prediction, but there are few rigorous results regarding their application. The broad goal of this paper is to provide some theoretical evidence of their applicability in the computational study of dynamical systems in some idealized, yet interesting setting. The specific goal of this paper is to investigate a data assimilation procedure (DAP), an ensemble Kalman filter (EKF), in the context of hyperbolic systems. We show that with appropriate assumptions on observations, for every trajectory on an attractor, the predictions produced by the DAP remain close to the truth for all time provided the ensemble is properly initialized, making the DAP reliable. We deal with the case of one-dimensional unstable direction first, and later extend to higher dimensional unstable spaces. A feature of this approach is that no model linearizations are involved, making it efficient and potentially of interest for applications in high dimensional systems. Lyapunov exponents are also investigated.

1. Introduction

An ensemble of trajectories can yield qualitative and quantitative information about a dynamical system. Computing an ensemble of trajectories with slightly different initial conditions can help judge whether the trajectories are unstable. If so, the growth of the ensemble spread may give a rough quantification of a Lyapunov exponent. For a more precise estimate of an infinite-time Lyapunov exponent, one should periodically rescale the ensemble to remain close to a reference trajectory. By an ensemble method we mean such a method that alternately propagates an ensemble of trajectories for a finite time and adjusts the ensemble in some way that keeps its spread relatively small, in order to maintain local information near a particular trajectory (or pseudotrajectory).

The adjustment step of an ensemble method can be tailored to a particular purpose. Often, and in this article, the adjustment is done within the local linear span of the ensemble; for example, thinking of a 2-member ensemble as a line segment, the adjustment step may choose a subsegment. In this case, over time the ensemble will typically become aligned with the local unstable directions, and can be thought of as approximately sampling a local unstable manifold. This approach has been used, for example, to generate pseudotrajectories that remain close to a chaotic saddle [18, 17].

Here, we regard propagation of the ensemble trajectories as providing information about the linearization of the dynamics in the unstable directions, without relying on analytical lineariza-

*Corresponding author. Tel: +61 29385 7090. Fax: +61 29385 7123.

Email addresses: ceciliagt@unsw.edu.au (Cecilia González-Tokman), bhunt@umd.edu (Brian R. Hunt)
tions, which can be intractable or computationally expensive for high-dimensional systems. Further, the ensemble provides a reduced rank approximation that attempts to describe the most important aspects of the linearization for forecasting, with an ensemble size that may be significantly smaller than the dimension of the state space.

The goal of data assimilation algorithms is to identify a pseudotrajectory of a model system that is consistent with a time series of noisy measurements collected from a real system. These algorithms work iteratively, and combine forecasts with data in real time. See [14] for an overview of data assimilation methods. In ensemble data assimilation, an ensemble of model state vectors is considered. It intends to describe and keep track of position and uncertainty of the system state. See [7] for an overview of ensemble data assimilation. Although ensemble methods perform well in some practical situations [25, 30], not much is known about their stability from a rigorous point of view. Here, we impose an assumption, hyperbolicity, that allows us to rigorously study the performance of a specific data assimilation method. We remark that despite of this assumption being quite restrictive, we hope that this work provides some ground on which the study and assessment of data assimilation methods in more general contexts can rely.

In data assimilation, the propagation and adjustment steps described above are called forecast and analysis, respectively. In the ensemble approach, the forecast step takes as initial conditions the analysis ensemble from the previous cycle and evolves each ensemble member separately according to an appropriate model, generating the background ensemble. Then, the information collected from measurements is used to produce the new analysis ensemble, by adjusting the background ensemble toward the data observed. Since the model state may not be measured directly, in general the analysis step uses an observation function (also called a forward operator) that quantifies what the measurements should be for a given model state. At this step, data is filtered according to the algorithm, and adjustments are made along a space determined by the ensemble members.

An example where data assimilation is used heavily is in forecasting the weather. This is the motivation behind the present work. Weather models are very high dimensional, as the state of the system comprises information about the Earth’s entire atmosphere. At the resolutions currently used, there are millions of variables involved. The main goal is to predict the state of the system in the future; that is, weather forecasting. Data assimilation is required because the current state of the system is not well determined by current observations. In this context, observations are measurements of meteorological variables, such as temperature, at various locations. A successful data assimilation procedure combines information collected from the observations with the forecast generated by the weather model, and produce, at each step, a good approximation to the corresponding state of the system.

In this paper, we investigate the stability of a data assimilation algorithm, an ensemble Kalman filter (EKF) studied in [11], assuming the underlying system is uniformly hyperbolic. The Kalman filter was introduced in [13]. It is optimal, in a least square sense, for the case of linear model and observations with white noise. Extensions to the non-linear setting (Extended KF) have been developed, see for example [12]. They involve linearizations and model-dimension matrix inversions, thus making computations costly for high-dimensional systems. Ensemble Kalman filters were introduced in [6], and further developed and tested in [5, 10]. They keep track of a set (ensemble) of trajectories, and are suitable for parallel computations. More recently, methods with deterministic choice of ensemble elements have been developed [1, 4, 20, 27, 28, 29]. The algorithm studied here belongs to this class. We prove that under a genericity assumption on the observation function and for sufficiently small measurement noise, there is an open set of initial conditions for which the ensemble shadows the true trajectory that generated the measurements for all future times. We believe our method of proof can be adapted to other deterministic EKF and the Extended KF. While there are probabilistic results on the stability
of the Extended KF [22], we are not aware of a published result analogous to ours.

Our results are related to the literature, starting with Pecora and Carroll [21] on the synchronization of chaotic systems, which typically uses a simple form of data assimilation called *direct insertion*. For concreteness, we consider the example from [21] of the Lorenz system

\[
\frac{d}{dt} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} -\sigma x + \sigma y \\ \rho x - y - xz \\ -bz + xy \end{bmatrix},
\]

where \( \sigma = 10, b = 8/3, \rho = 28 \). Pecora and Carroll consider two subsystems, the first of which evolves a state vector \([x_1, y_1, z_1]\) according to the equations above, and the second of which evolves two variables, say \([y_2, z_2]\), according to the differential equations with \(x_2\) replaced by \(x_1\) from the first subsystem. In data assimilation terms, we think of the first subsystem as the *truth*, the variable \(x\) as the observation, and the second subsystem as the data assimilation algorithm obtained by direct insertion of the observations into the state vector. Pecora and Carroll say that the systems synchronize if \(y_1 - y_2\) and \(z_1 - z_2\) approach 0 time increases. They find empirically that synchronization does occur for a pair of Lorenz systems if they are coupled by the \(x\) coordinate or the \(y\) coordinate, but not if they are coupled by the \(z\) coordinate. Synchronization in the Lorenz system via the \(x\) coordinate is confirmed rigourously by [9], which relaxes the assumption of continuous-time observations by using direct insertion only at discrete times. However, the results of [21] suggest that stability of data assimilation by direct insertion (or *nudging*, see §2.1.3) depends on the choice of observation function.

In order to track a trajectory of a system \(f : M \to M\), usually referred to as the *truth* and known only through measurements with limited accuracy, it is necessary to be able to adjust forecasts along all unstable directions. The number of these may be much smaller than the total dimension of the system. In this case, we can hope to keep track of the truth by keeping track of the evolution of an ensemble of trajectories surrounding an approximation to the truth, with the number of elements in the ensemble related to the number of unstable directions. Hyperbolic systems possess a well defined number of unstable directions, independent of the trajectory. For this reason, they provide a tractable setting to investigate the properties of EKF. Because the differences between ensemble members determine directions in which EKF can make adjustments, the number of ensemble members must exceed the dimension of the unstable space in order to be able to correct errors in all unstable directions.

A fundamental property of hyperbolic systems is shadowing. This property ensures that for any \(\delta > 0\) there exists some \(\epsilon > 0\) such that every \(\epsilon\)-pseudo-orbit of the system is \(\delta\)-shadowed by a real orbit. An \(\epsilon\)-pseudo-orbit of \(f\) is a sequence \(\{x_n\}_{a < n < b} \subset M\) for which \(\|f(x_n) - x_{n+1}\| < \epsilon\) for all \(a < n < b\), and it is said to be \(\delta\)-shadowed by the orbit of \(x\) if \(\|x_n - f^n(x)\| < \delta\) for all \(a < n < b\). See [15, §18] for a precise statement of the shadowing lemma.

The main result of the paper, presented in Proposition 1 and generalized in Proposition 4, ensures that for hyperbolic attractors with \(k^n < k\) dimensional unstable spaces satisfying a *bunching* condition, and measurement errors bounded by sufficiently small \(\epsilon > 0\), the EKF shadows the true trajectory for all positive time for a non-empty open set of initial \(k\)-member ensembles, under Takens’ genericity conditions [26] for the observation function. This property guarantees that the data assimilation procedure is reliable, in the sense that when appropriately initialized, its trajectory provides an approximation for the true trajectory to within a small error for all future time. In other words, the data assimilation system, driven by the real system, synchronizes with it. Consequences for the approximation of positive Lyapunov exponents of the system are also presented.

The defining property of unstable spaces, and more precisely the existence of invariant unstable cones in some dynamical systems, suggests a convenient way of identifying meaningful ensemble members, by thinking of the differences between ensemble members as vectors in the
the tangent space. Namely, if we intend to track the dynamics near a particular trajectory, we may keep track of the evolution of tangent vectors under the (tangent) dynamics. A consequence of the theorem of Oseledec [19] is that for almost every initial trajectory (with respect to an invariant measure for the system) almost every vector approaches the most unstable direction for that trajectory. If we are careful to orthogonalize and normalize the vectors at each step we can identify the other unstable spaces in the Oseledec filtration as well, through the standard numerical procedure for estimating Lyapunov exponents; see the Appendix of [24]. In [8] an alternative algorithm to identify spaces of the Oseledec splitting is introduced.

The task we just described, including the computation of the tangent map, may be costly, computationally or otherwise. In fact, considerable human time is devoted to linearizing weather models, see p. 215 and Appendix B of [14]. Moreover, dealing with derivatives significantly increases, at the least, storage requirements. As an alternative, we can evolve ensemble vectors according to $f$ instead of $Df$. This is analogous to using the secant method instead of Newton’s method as a root-finding algorithm in calculus or numerical analysis. If the size of the ensemble vectors is small, of order $\epsilon$, the distance between the image of a point and its linearization is of order $\epsilon^2$.\footnote{In general, it would be necessary to use the exponential map to identify tangent vectors with points in the space.} We would like to show that under some circumstances, this procedure indeed produces ensemble vectors that lie inside an unstable cone. When this is the case, if the cones are strictly invariant, once inside the cones, the algorithm would keep successive iterates of the ensemble inside unstable cones. On the one hand, this would allow to adjust errors accumulated in unstable directions. On the other hand, it would permit to iterate the procedure.

We remark that many models of physical systems, while not uniformly hyperbolic, do display features of hyperbolic systems. A deterministic results like ours — that with suitable initial conditions, a data assimilation procedure shadows the truth for all positive time — requires bounded measurement errors and may require the shadowing property as well. Nonetheless, it would be interesting to generalize our results, at least in a statistical sense, to (for example) non-uniformly hyperbolic systems. These share some properties with the systems treated here, such as existence of stable and unstable manifolds for Lyapunov regular points, and some of them also have complete strict (or eventually strict) invariant families of cones as described in [2]. However, in general, stable and unstable spaces do not depend continuously on the base point. Hence our proofs do not extend directly to that setting, and other techniques would be necessary to study the problem in such generality.

The structure of the paper is as follows. In §2, we present the main result, after introducing the setting and discussing the initialization of the ensemble. In §3, properties of the ensemble Kalman filter are established for the case of one-dimensionally unstable hyperbolic systems. These properties are generalized to the case of higher dimensionally unstable hyperbolic systems in §4. The main reason to separate the two cases is to present the control of nonlinear terms in §3, and leave the main complications of the extension to higher dimensionally unstable cases, which lie at the linear level, to §4.

2. Statement of results

2.1. Setting

We start by describing our hypotheses for the forecast model and the observation function.
2.1.1. Model

Throughout the paper, let \( f : M \looparrowright \) be a \( C^3 \) diffeomorphism of a Riemannian manifold\(^2\) of dimension \( N \), with a uniformly hyperbolic attractor of \( A \subset M \). That is, a compact set invariant under \( f \) for which there is an open set \( U \subset M \) such that \( A \subset \text{int}(U) \) and \( \cap_{n \geq 0} f^n(U) = A \). The hyperbolicity condition means that, restricted to \( A \), there is an \( f \)-invariant splitting of the tangent spaces \( T_x M \) into unstable (expanding) and stable (contracting) spaces, \( T_x M = E^u_x \oplus E^s_x \), and constants \( \lambda > 1 > \mu \) such that

\[
\|Df v\| \geq \lambda \|v\| \quad \forall v \in E^u \quad \text{and} \quad \|Df v\| \leq \mu \|v\| \quad \forall v \in E^s.
\]

A reference for hyperbolic systems is [15].

**Remark 1.** This is not the usual definition of hyperbolicity, in the sense that we are already working with a metric adapted to \( f \). This assumption simplifies some calculations, but does not restrict the scope of the paper since adapted metrics always exist. Moreover, even if a metric (e.g. Euclidean) is not adapted to a hyperbolic system \( f \), it will be adapted to a suitable power.

The dependence of stable and unstable spaces with respect to the base point is in general H"older continuous. It is convenient, and a frequent additional hypothesis in smooth hyperbolic theory, to impose a so-called bunching condition to ensure this dependence is in fact \( C^1 \). For each \( x \in M \), let \( \lambda_x > 1 > \mu_x \geq \nu_x \) be, respectively the weakest expansion along \( E^u_x \), the weakest contraction along \( E^s_x \), and the strongest contraction along \( E^s_x \). We make the following assumption. For every \( x \in M \),

\[
\mu_x \lambda_x^{-1} \nu_x^{-1} < 1.
\]

For the remainder of the paper, we also assume that for each \( f \) periodic point \( x \) of period \( k \leq 2N + 1 \) the eigenvalues of \( Df^k_x \) are distinct. This open and dense property in \( \text{Diff}(M) \) is needed for Takens' embedding theorem to apply; see Theorem 2.1.2 for the statement.

2.1.2. Observation function

For \( f \) fixed, we consider generic \( C^2 \) real-valued observation functions \( h : M \to \mathbb{R} \) in the sense of the following theorem, which will be repeatedly used in this paper.

**Theorem** (Takens embedding theorem, [26]). Let \( f : M \looparrowright \) as in §2.1.1. Then, for smooth proper\(^3\) functions \( h : M \to \mathbb{R} \), it is a generic property that the map

\[
x \mapsto (h(x), h(f(x)), h(f^2(x)), \ldots, h(f^{2N}(x)))
\]

is an embedding, i.e. one-to-one proper immersion.

We note that this result has been (or may be) refined in a couple of ways that may be relevant for concrete applications. On the one hand, generalizations of Theorem 2.1.2, such as those in [23] may be useful. In short, they allow to reduce the number of measurements from \( 2N + 1 \) to \( 2 \dim A + 1 \), where \( A \) is the attractor of \( f \) under consideration. This would improve the estimates significantly, as errors grow exponentially with the number of steps considered.

On the other hand, there may be multiple observations available at each step, say \( l \) scalar measurements. To extend our results to this setting, a multidimensional version of Takens’

---

\(^2\)To avoid technical difficulties, we assume \( M \) is a Euclidean space, a cylinder or a torus so there is no need to make use of the exponential map to identify tangent vectors with points in the space.

\(^3\)A function is proper if the inverse image of every compact set is compact. This is always the case for smooth functions if the domain is a compact manifold.
theorem is needed. Such an extension may be established following Takens’ original proof.
Therefore our proof could be adapted to, in some appropriate sense, generic \( h : M \to \mathbb{R}^l \), and reduce the number of forecast steps considered by a factor of \( l \). It is also possible to reduce to the one-dimensional case by assimilating observations sequentially, as discussed in [16, §7.4] or [29, §3]. We do not present all the details here.

2.1.3. Data assimilation

Given a trajectory \( \{x_n\} \) of \( f \), which we call the truth, and a corresponding sequence of observations \( y_n \approx h(x_n) \), the goal of a data assimilation procedure (DAP) is to produce a sequence of approximations \( x_n^a \) to the truth, using only knowledge of \( f, h \), and the observations. Once \( x_{n-1}^a \) is determined, using only the observations up to and including \( y_{n-1} \), the model \( f \) is applied to produce a background estimate \( x_n^b \), and then the analysis \( x_n^a \) is determined from the background and the observation \( y_n \). Often the analysis equation takes the form

\[
x_n^a = x_n^b + g_n (y_n - h(x_n^b)),
\]

where \( g_n \) is called the gain; different DAPs determine the gain in different ways. If \( x \) is a Euclidean vector, \( h \) is the linear function that assigns a vector to one of its coordinates, and \( g_n \) is the gradient of \( h \), then (1) is the discrete-time direct insertion method described in the introduction. More generally, if \( h \) is arbitrary and \( g_n \) is a constant (which is a parameter of the method) times the gradient of \( h \) evaluated at \( x_n^b \), then (1) is often called nudging.

An important feature of the DAP we consider is that the gain depends on the model dynamics. Ensemble Kalman Filters are nonlinear extensions of the Kalman Filter, which is formulated for a linear model \( F \) and linear observation function \( H \). In the context of a scalar observation, the Kalman Filter can be written

\[
x_n^b = F x_{n-1}^a \tag{2}
\]
\[
P_n^b = FP_{n-1}^a F^\top \tag{3}
\]
\[
K_n = P_n^b H^\top (H P_n^b H^\top + \epsilon^2)^{-1} \tag{4}
\]
\[
x_n^a = x_n^b + K_n (y_n - H x_n^b) \tag{5}
\]
\[
P_n^a = (I - K_n H) P_n^b \tag{6}
\]

Here \( P \) represents an auxiliary matrix that is interpreted as a covariance for the uncertainty in the corresponding estimate \( x \). The parameter \( \epsilon \) is generally taken as an estimate of the typical size of the measurement error \( y_n - H x_n \); indeed, if these errors are normally distributed with mean 0 and variance \( \epsilon^2 \), then the Kalman Filter is statistically optimal. Below, we will treat \( \epsilon \) as a parameter that can be chosen separately from the measurement errors, which we assume to be bounded.

In an Ensemble Kalman Filter, at each time step a background ensemble and analysis ensemble, each consisting of \( k > 1 \) model states, are computed. The estimates \( x_n^b \) and \( x_n^a \) are replaced by ensemble means \( \bar{x}_n^b \) and \( \bar{x}_n^a \), and the matrices \( P_n^b \) and \( P_n^a \) are regarded as sample covariances of the corresponding ensembles. For example if we subtract the mean from each of the background ensemble members and form the resulting column vector displacements into a matrix \( X_n \), then \( P_n^b = X_n X_n^\top / (k-1) \). Notice that (4) can then be rewritten

\[
K_n = X_n (H X_n)^\top (H X_n (H X_n)^\top + (k-1)\epsilon^2)^{-1}.
\]

For a nonlinear observation function \( h \), we can replace \( H X_n \) with the \( k \)-dimensional row vector obtained by applying \( h \) to each background ensemble member and subtracting the mean. The linear time evolution equations (2) and (3) of the Kalman Filter are replaced by simply applying \( f \) to each analysis ensemble member, and (5) is applied to the ensemble mean.
remains to complete the algorithm is an equation that determines a set of analysis ensemble displacements from the mean whose sample covariance is consistent with (6). There are many possibilities, and the one we use is the ETKF of \([4]\) as formulated in \([11]\): the matrix of analysis ensemble displacements is given by

\[
X_n(I + (HX_n)^\topHX_n/[(k - 1)\epsilon^2])^{-1/2},
\]

where \(I\) is the \(k \times k\) identity matrix and by the \(-1/2\) power of a matrix we mean the symmetric square root of its inverse.

While the equations above are written in a form similar to the standard equations that allow vector observation functions, some further simplifications can be made in the scalar case. With some additional notation, we present these equations more formally in the following section.

2.2. Main result

**Definition 1.** A DAP with initial ensemble \(E\) is called \(c\)-reliable if the predictions it produces eventually shadow the true trajectory within error \(c\). We say that a family of DAPs depending on a parameter \(\epsilon\), and having initial ensembles \(\{E_\epsilon\}_{\epsilon > 0}\), is \(O(\epsilon) - \text{reliable}\) if there is some constant \(c\) independent of \(\epsilon\) such that for all \(\epsilon > 0\) sufficiently small, the DAP associated to parameter \(\epsilon\) with initial ensemble \(E_\epsilon\) is \(cc\)-reliable.

Consider \(f : M \ni \epsilon\) as in \(\S 2.1.1\) and \(h : M \to \mathbb{R}\) as in \(\S 2.1.2\). The main result of this paper concerns the reliability of a family of DAPs associated to \(f\) and \(h\), provided they are properly initialized. This family is called the \(k\)-member ensemble Kalman filter (kMEKF). Here we include the definition for the reader’s convenience. (See \(\S\S 4.1\) and 4.2 for further discussion, and relation with the formulation of \(\S 2.1.3\).

Let \(\epsilon > 0\). The kMEKF corresponding to \(\epsilon\) is constructed using the following iterative procedure. After step \(n - 1\), we start with an ensemble of \(k\) vectors with mean \(\mu_{n-1}\) and displacement vectors \(v_{j,n-1}^a\), \(1 \leq j \leq k\). The corresponding background mean at step \(n\) is denoted by \(\bar{x}_n^b\) and the corresponding displacements by \(v_{j,n}^b\), where

\[
\bar{x}_n^b = \frac{1}{k} \sum_{j=1}^{k} f(\bar{x}_{n-1}^a + v_{j,n-1}^a),
\]

\[
v_{j,n}^b = f(\bar{x}_{n-1}^a + v_{j,n-1}^a) - \bar{x}_n^b.
\]

Note that the \(v_{j,n}^b\) are the columns of \(X_n\) in \(\S 2.1.3\). The average value of the measurements of \(h\) will be \(\overline{h}_n := \frac{1}{k} \sum_{j=1}^{k} h(\bar{x}_{n}^b + v_{j,n}^b)\). The measurement of \(h\) from the true trajectory \(x_n\) will be denoted by \(y_n = h(x_n)\). To define the analysis ensemble at step \(n\), we introduce some further notation. For \(1 \leq j \leq k\), let

\[
q_{j,n} := \frac{1}{\epsilon} (h(\bar{x}_{n}^b + v_{j,n}^b) - \overline{h}_n),
\]

\[
q_n^2 := \frac{1}{(k - 1) \sum_{j=1}^{k} q_{j,n}^2},
\]

\[
\gamma_n := \frac{1}{q_n^2} \left(1 - \frac{1}{\sqrt{1 + q_n^2}}\right) \text{ if } q_n \neq 0, \gamma_n = 0 \text{ otherwise, and}
\]

\[
v_{0,n} := \frac{1}{(k - 1)} \sum_{j=0}^{k} q_{j,n} v_{j,n}^b.
\]

The analysis ensemble is defined by:

\[
\bar{x}_n^a = \bar{x}_n^b + \frac{(y_n - \overline{h}_n) v_{0,n}}{1 + q_n^2},
\]

\[
v_{n,j}^a = v_{n,j}^b - \gamma_n q_{j,n} v_{0,n}, \text{ for } 1 \leq j \leq k.
\]
Relating to the notation in the previous section, the vector formed by the $q_{j,n}$ corresponds to $\epsilon^{-1}Hx_n$ in §2.1.3.

Let $f : M \ni x$ be as in §2.1.1, and let $A$ be a $k^u$ dimensionally unstable attractor for $f$, i.e. $\dim E^u = k^u$. Assume $k > k^u$ and $x_0 \in A$. Our main result is:

**Main Result.** For generic observation function $h$ there is a family $\{I_{\epsilon}\}_{\epsilon > 0}$ of open sets of initial ensembles such that whenever $E_{\epsilon} \subset I_{\epsilon}$, the kMEKF with initial ensembles $E_{\epsilon}$ and noiseless observation of $h$ is $O(\epsilon)$-reliable. The same conclusion holds if the measurement of $h$ has noise, provided its size is bounded by a small multiple of $\epsilon$, and also when the observations are generated by a pseudo-trajectory, provided its distance to a true trajectory is sufficiently small.

**Remark 2.** Sets of initial ensembles for which the kMEKF is $O(\epsilon)$-reliable are described explicitly in §4.3.1. Roughly speaking, they consist of ensembles for which $\|x_0 - p_0\| \lesssim \sqrt{\epsilon}$ and such that the corresponding perturbations are of adequate spread and lie sufficiently close to the unstable space $E^u_{x_0}$. The last condition is discussed and explained in §2.3.

This result is proved using an inductive scheme. In §3 it is established for the case $k = 2$ (see Proposition 1 and Corollary 1). The general case is deferred to §4 (see Proposition 4). The proofs follow a similar strategy, but the analysis at the linear level is straightforward in the former. Hence, we concentrate in controlling nonlinear terms in §3, and leave the complications at the linear level coming from higher dimensional unstable dynamics for §4.

### 2.3. Initialization of the ensemble

In this section we discuss how to identify an initial ensemble of trajectories that is appropriate for the EKF to be reliable. Proofs are left for subsequent sections. The most desirable characteristic of an initial ensemble is to well approximate the unstable space of the true trajectory. Even in the case of perfect model, which is the one treated here, this is a non-trivial task, as unstable spaces depend on the infinite future of the system.

An unstable cone at $x$, $K^u_x$, is a subset of $T_xM$ of the form $K^u_x = \{(v_u, v_s) \in E^u_x \oplus E^s_x = T_xM \|v_s\| \leq c\|v_u\|\}$ for some constant $c > 0$. The initialization we propose relies on the forward invariance of unstable cones for uniformly hyperbolic systems. This property ensures the existence of a family of cones $K^u_x \subset T_x(M)$ surrounding the unstable space $E^u_x \subset T_x(M)$ that is invariant under $Df$. Moreover, in the uniformly hyperbolic setting, the invariance is strict, in the sense that $Df_xK^u_x \subset intK^u_{f(x)} \cup \{0\}$. Thus, an unstable cone at a point $x$ gets mapped inside the interior the corresponding cone at $f(x)$ under the tangent dynamics $Df$. As we do not make use of the linearization, but of the map itself in the forecast step, strict invariance is essential to allow for the small errors associated to this difference to be negligible. This permits to ensure that when the displacements of ensemble vectors from the mean are small and lie inside the unstable cone, so do their corresponding images under $f$.

The above justifies the existence of an open set of ensembles having the desired property of remaining close to the unstable space under application of the dynamics. However, there is still something to be said about how to identify them. A reasonable approach is as follows. We may start with a cloud of points sufficiently dense in a sphere of small radius around the point $x$. By forecasting according to $f$, projecting back to a small sphere around $f(x)$, and repeating this procedure for a few steps, we could identify finite time unstable directions, which necessarily contain unstable cones. In fact, by performing a Gram-Schmidt orthogonalization procedure, we may be able to estimate the dimension of the unstable space. This estimate would dictate the number of ensemble members to keep track of during the data assimilation procedure. It is also possible to approximate positive Lyapunov exponents by keeping track of the total expansion or contraction gained along the forecast steps.
3. Properties of the EKF for hyperbolic systems with one-dimensional unstable spaces

Let \( f : \mathcal{M} \to \mathcal{M} \) be a diffeomorphism having a one-dimensional unstable hyperbolic attractor \( \mathcal{A} \), as in §2.1.1. We show that for generic \( C^2 \) function \( h \), as in §2.1.2, the 2-member ensemble Kalman filter (2MEKF) is \( O(\epsilon) \)—reliable, in the sense of Definition 1.

3.1. Evolution equations

Let \( \epsilon > 0 \). Starting from an initial ensemble of analysis vectors \( x_n^a \pm v_n^a \), we obtain the new background vectors at step \( n \) by forecasting according to \( f \). We denote these background ensemble vectors by \( \overline{x}_n^a \pm \overline{v}_n^a \). Thus,

\[
\overline{x}_n^a = \frac{1}{2} \left( f(x_{n-1}^a + v_{n-1}^a) + f(x_{n-1}^a - v_{n-1}^a) \right),
\]

\[
\overline{v}_n^a = f(\overline{x}_{n-1}^a + \overline{v}_{n-1}^a) - \overline{x}_n^a.
\]

The measurement of \( h \) from the true trajectory \( x_n \) will be denoted by \( y_n = h(x_n) \). Let \( q_n := \frac{1}{2}(h(\overline{x}_n^a + \overline{v}_n^a) - \overline{h}_n) \). The corresponding analysis vectors, obtained using an ensemble square root filter, are \( x_n^a \pm v_n^a \), with:

\[
\overline{x}_n^a = \overline{x}_n^a + \frac{2q_n}{1 + 2q_n^2} (y_n - \overline{h}_n)^{-1} \overline{v}_n^a,
\]

\[
\overline{v}_n^a = \frac{1}{\sqrt{1 + 2q_n^2}} \overline{v}_n^a.
\]

3.2. Basic definitions and notation

Under Takens’ genericity conditions (see Theorem 2.1.2), it is ensured that for every \( x \in \mathcal{M} \), the vectors

\[
\{ \nabla h(x), (Df_x)^T \nabla h(f(x)), (Df_x^2)^T \nabla h(f^2(x)), \ldots, (Df_x^{2N})^T \nabla h(f^{2N}(x)) \}
\]

span \( T_x \mathcal{M} \), for all \( x \in \mathcal{M} \).

Let

\[
\tilde{\gamma}(x) := \frac{1}{\| \nabla h(x) \|} \cos \angle (\nabla h(x), E_y^x) = \frac{1}{(v^y_x)^T \nabla h(x)},
\]

where \( v^y_x \in E^y_x \) is a unit length vector. We note that \( \tilde{\gamma}(x) < \infty \) whenever \( \nabla h(x) \) and \( v^y_x \) are not orthogonal. By compactness of \( \mathcal{A} \), there is some constant \( \tilde{\gamma} > 0 \) such that \( \tilde{\gamma} > \sup_{x \in \mathcal{A}} \min_{j=0,\ldots,2N} \{ \tilde{\gamma}(f^j(x)) \} \). We note that \( \tilde{\gamma} \) is finite by the non-degeneracy condition on \( h \). Whenever \( \tilde{\gamma}(x) < \tilde{\gamma}_0 \) we will say that the angle \( \angle (\nabla h(x), E^x_y) \) is good.

Using the Taylor expansion of \( h \) around \( x_n^a \), we know that for \( ||v^y_x|| \) small,

\[
h(x_n^a + v_n^a) = h(x_n^a) \pm \nabla h(x_n^a) \cdot v_n^a + O(||v_n^a||^2).
\]

Hence, \( \overline{h}_n = h(\overline{x}_n^a) + O(||\overline{v}_n^a||^2) \) and \( q_n = \nabla h(\overline{x}_n^a) \cdot \overline{v}_n^a + O(||\overline{v}_n^a||^2) \).

We have assumed the metric is adapted, and \( \lambda > 1 \) are strict lower and upper bounds on the expansion, respectively contraction, along unstable and stable spaces. Then, whenever \( x \) and \( y \) are sufficiently close we have, \( ||f(x) - f(y)||_u \geq \lambda ||x - y||_u \) and \( ||f(x) - f(y)||_s \leq \mu ||x - y||_s \), where \( ||u(x)|| \) denote distance along unstable (stable) spaces, to be defined precisely in the sequel. Let \( \overline{h} \) be a Lipschitz constant for \( f \), and \( L \) a Lipschitz constant for \( h \).
3.3. Outline of inductive estimates

Here we introduce some further notation aiming to outline the ideas behind the inductive arguments in the coming sections. The new notation in this section is not required in subsequent sections. For any \( n \), let

\[
A_n := \frac{\angle(v_n^a, E_{x_n}^u)}{\epsilon},
\]

\[
V_n := \frac{\|v_n^a\|}{\epsilon},
\]

\[
X_n := \frac{\|x_n - x_n^a\|}{\epsilon},
\]

\[
S_n := \frac{\|x_n - x_n^a\|}{\epsilon^2}.
\]

Relevant properties of the ensemble can be expressed in terms of the above quantities. For example the shadowing property is equivalent to \( \{X_n\}_{n\in \mathbb{N}} \) being bounded. The ensemble size is bounded provided \( \{V_n\}_{n\in \mathbb{N}} \) is bounded.

We will later show that the following inequalities hold, provided \( \|x_n - x_n^a\|, \|v_n^a\| \) and \( \angle(v_n^a, E_{x_n}^u) \) are sufficiently small. First,

\[
V_{n+1} \leq \begin{cases} \frac{\hat{z}}{\sqrt{2}} & \text{if } \angle(\nabla h(x_{n+1}), E_{x_{n+1}}^u) \text{ is good}, \\ \Lambda V_n & \text{if } \angle(\nabla h(x_{n+1}), E_{x_{n+1}}^u) \text{ is bad}. \end{cases}
\]

Thus, \( \{V_n\} \) remains bounded if the number of consecutive bad angles is bounded above. Next, there exist some \( 0 < \nu < 1 \) and \( C, C' > 0 \), depending on \( f \) and \( h \), such that

\[
A_{n+1} \leq \nu A_n + C X_n + C' V_n.
\]

There exist some \( 0 < \mu < 1 \) and \( C, C' > 0 \), depending on \( f \) and \( h \), such that

\[
S_{n+1} \leq \mu S_n + C A_n V_n X_n + C' V_n^2.
\]

In general, for \( X_n \) we only have

\[
X_{n+1} \leq \tilde{X}(1 + CV_n)X_n.
\]

These estimates provide some insight on the evolution of the quantities \( A_n, V_n, S_n, X_n \) with respect to \( n \). However, showing that \( X_n \) is bounded requires some further considerations. It is in fact fruitful to study the quantities \( Q_n := \|x_n - x_n^a\| \) instead of \( X_n \). For appropriate choices of the initial ensemble, the size of the perturbation elements in the Kalman filter somehow keeps track of the the distance to the truth. Indeed, when good angles \( \angle(E_{x_{n+1}}^u, \nabla h(x_{n+1})) \) occur, \( Q_{n+1} \leq 1 + \frac{Q_n}{1 + 2C_f} \) provided the smallness assumptions above. In fact, contraction by a factor arbitrarily close to \( \frac{1}{1 + 2C_f} \) occurs provided \( Q_n \) is not too small. When bad angles occur, there may be exponential growth of the quotient \( \frac{Q_{n+1}}{Q_n} \), but if the number of consecutive bad angles is bounded above, this growth rate can be controlled in such a way that the expansion is compensated by the contraction gained by the occurrence of a good angle. These arguments are enough to show that \( \{Q_n\} \) is bounded, and furthermore, that it is eventually of order one. The same conclusion holds for \( \{A_n\} \) and \( \{S_n\} \).

In the next section, we present an inductive scheme making the above estimates rigorous. It is valid for \( \|v_0^a\| = O(\epsilon) \), and the quantities \( \|x_0 - x_0^a\| \) and \( \angle(v_0^a, E^u(x_0)) \) sufficiently small. As we will see in Proposition 1, in this setting, the shadowing property is guaranteed.
3.4. Inductive scheme

We will now establish the fundamental properties of the 2MEKF generated by \( f : M \circ \) using an inductive scheme. In short, the 2MEKF will be \( \mathcal{O}(\epsilon) \)-reliable provided the ensemble has been initialized in such a way that the displacement vector \( v_0 \) lies in a sufficiently narrow unstable cone and that the distance from the ensemble mean to the true trajectory is sufficiently small.

**Proposition 1** (Properties of 2-member ensemble Kalman filter).

Let \( x_0 \in \mathcal{A} \). Then, the following holds.

- For generic observation function \( h \), there is a family \( \{ \mathcal{J}_\epsilon \}_{\epsilon > 0} \) of open sets of initial ensembles such that whenever \( \mathcal{E}_\epsilon \in \mathcal{J}_\epsilon \), the 2MEKF with initial ensembles \( \{ \mathcal{E}_\epsilon \}_{\epsilon > 0} \) and noiseless observation of \( h \) is \( \mathcal{O}(\epsilon) \)-reliable. More precisely, this is the case for all 2MEKF initialized in such a way that the inductive hypothesis from §3.4.1 holds for suitable choice of constants \( C_1, \ldots, C_5 \). Moreover, the ensemble spread remains proportional to \( \epsilon \).

- The same conclusion holds if the measurement of \( h \) has noise, provided its size is bounded by a small multiple of \( \epsilon \), and also when the observations are generated by a pseudo-trajectory, provided its distance to a true trajectory is sufficiently small.

The proof of these results occupies the remainder of this subsection.

### 3.4.1. Inductive hypothesis \( IH(C_1, C_2, C_3, C_4, C_5, \epsilon) \)

**Definition 2.** Let \( \epsilon > 0 \). We say that the ensemble with mean \( \bar{x}_n \) and perturbations \( v_n \) (or concisely, the ensemble at time \( n \)) satisfies the inductive hypothesis \( IH_n(C_1, C_2, C_3, C_4, C_5, \epsilon) \) if the following holds:

1. **(i)_n** Lower bound on spread of ensemble.
   
   \[ C_1 \epsilon \leq \| v_n \| \]

2. **(ii)_n** Unstable cone.
   
   \[ \angle (v_n, E_{x_n}^u) \leq C_2 \epsilon \]

3. **(iii)_n** Upper bound on spread of ensemble.
   
   \[ \| v_n \| \leq C_3 \epsilon \]

4. **(iv)_n** Shadowing.
   
   \[ \| x_n - \bar{x}_n \| \leq C_4 \epsilon \]

5. **(v)_n** Bound on distance along stable direction.\(^4\)
   
   \[ \| x_n - \bar{x}_n \|_s \leq C_5 \epsilon^2 \]

\(^4\| x_n - \bar{x}_n \|_s \) is a shorthand for \( \sup_{v \in E_{x_n}^s \| v \|=1} |(x_n - \bar{x}_n) \cdot v| \)
3.4.2. Inductive step

In this section we show that if $\epsilon > 0$ is sufficiently small and $IH_0(C_1, C_2, C_3\bar{\Lambda}^{-2N}, C_4, C_5, \epsilon)$ is valid at the initial time, then $IH_n(C_1, C_2, C_3, C_4, C_5, \epsilon)$ will remain valid at all times $n \geq 0$, provided $\epsilon$ is sufficiently small and some relations between the constants $C_1, \ldots, C_5$ are satisfied. As for each value of $\epsilon$ the set of ensembles for which these conditions are valid contains a non-empty open set, this is enough to prove Proposition 1.

The letters $C, C'$ will denote positive constants independent of $\epsilon$ and $C_1, \ldots, C_5$, but may depend on $f$ and $h$, and are allowed to change from one appearance to the next. The letter $\nu$ will denote a constant between 0 and 1 with the same properties as $C$. The notation $\mathcal{O}_*$ is similar to the asymptotic $\mathcal{O}$ notation, but the constants involved are allowed to depend on $C_1, \ldots, C_5$ as well.

For the rest of this section we suppose the standing assumption

$$IH_0(C_1, C_2, C_3\bar{\Lambda}^{-2N}, C_4, C_5, \epsilon) \text{ and } IH_m(C_1, C_2, C_3, C_4, C_5, \epsilon) \text{ for all } m \leq n \quad (\text{IH})$$

is valid for some $n \geq 0$, and some (yet to be determined) constants $C_1, \ldots, C_5$. We will show that $IH_{n+1}(C_1, C_2, C_3, C_4, C_5, \epsilon)$ holds. The proof proceeds by induction provided $C_1, \ldots, C_5$ are chosen appropriately.

Proof of $(iii)_{n+1}$. For any $m \geq 0$ we have

$$\|v_m^a\| \leq \|v_{m+1}^b\| \leq \bar{\Lambda}\|v_m^a\|.$$  

Hence, by the standing assumption (IH), for all $0 \leq n < 2N$ we have that $(iii)_{n+1}$ of $IH_{n+1}(C_1, C_2, C_3, C_4, C_5, \epsilon)$ holds.

For $n \geq 2N$, we observe that when the angle $\angle(E_x^u, \nabla h(x_m))$ is good, $\epsilon$ is sufficiently small and $(iv)_{m-1}$ holds, then, recalling the definition of $\tilde{\gamma}$ from (8), we have

$$\|v_m^a\| = \frac{1}{\sqrt{1 + 2q_m^2}}\|v_m^b\| \leq \tilde{\gamma}\epsilon \frac{\epsilon}{\sqrt{2}}.$$  

By the genericity condition on $h$ a good angle will occur within any $2N + 1$ consecutive steps, so we can choose $m$ between $n - 2N + 1$ and $n + 1$. Then, choosing

$$C_3 \geq \frac{\bar{\Lambda}^{2N+1}}{\sqrt{2}} \quad (9)$$

together with the standing assumption (IH) imply that $(iii)_{n+1}$ holds for $\epsilon$ sufficiently small.

Proof of $(i)_{n+1}$. Recall that $\lambda > 1$ is a strict lower bound on the expansion of $f$ along $E^u$ and $L$ is a Lipschitz constant for $h$. Then, $q_{n+1} \leq L\|v_{n+1}^b\|$. Moreover, if $\|v_n^a\| \geq C_1\epsilon$, and $\epsilon$ is sufficiently small we have

$$\|v_{n+1}^a\| \geq \frac{\epsilon}{\sqrt{\epsilon^2 + 2L^2\|v_{n+1}^b\|^2}} \|v_{n+1}^b\| \geq \frac{\epsilon}{\sqrt{\lambda^2 + 2L^2}}.$$  

Then, $(i)_{n+1}$ is guaranteed by choosing $C_1$ such that

$$C_1 \leq \frac{1}{\sqrt{2L}}\sqrt{1 - \frac{1}{\bar{\Lambda}^2}} \quad (10)$$
Proof of \((ii)_{n+1}\). Let us assume \(C_2\epsilon\) is sufficiently small. Then, recalling from Equation (*) that 
\(v^a_{n+1}\) and \(v^b_{n+1}\) are collinear, we have that
\[
\angle(v^a_{n+1}, E^u_{x_{n+1}}) = \angle(v^b_{n+1}, E^u_{x_{n+1}})
\leq \angle(v^b_{n+1}, Df^{u}_{x_{n+1}}v^a_{n}) + \angle(Df^{u}_{x_{n+1}}v^a_{n}, Df_{x_{n+1}}v^a_{n}) + \angle(Df_{x_{n+1}}v^a_{n}, E^u_{x_{n+1}})
\leq CC_3\epsilon + C_2\mu\epsilon + O_*(\epsilon^2),
\]
where in the last inequality we have used conditions \((ii)_{n}\) and \((iv)_{n}\) of the inductive hypothesis. Hence, \((ii)_{n+1}\) holds, for \(\epsilon\) sufficiently small, as long as
\[
\frac{C_4}{C_2} < \frac{1 - \nu}{\epsilon}.
\]

Proof of \((v)_{n+1}\). For the background ensemble, we have
\[
\|x_{n+1} - \bar{x}^b_{n+1}\|_s \leq \|x_{n+1} - f(\bar{x}^a_{n})\|_s + \|f(\bar{x}^a_{n}) - \bar{x}^b_{n+1}\| \leq C_5\mu\epsilon^2 + CC_3^2\epsilon^2 + O_*(\epsilon^3).
\]
Therefore,
\[
\|x_{n+1} - \bar{x}^a_{n+1}\|_s \leq \|x_{n+1} - \bar{x}^b_{n+1}\|_s + \|\bar{x}^b_{n+1} - \bar{x}^a_{n+1}\|_s
\leq C_5\mu\epsilon^2 + CC_3^2\epsilon^2 + \|\bar{x}_{n+1} - \bar{x}_{n}\| \frac{2\eta_{n+1}}{1 + 2\eta_{n+1}} \frac{\|v^b_{n+1}\|}{\epsilon} + O_*(\epsilon^3)
\leq C_5\epsilon^2 + CC_3^2\epsilon^2 + C\epsilon \|\bar{x}_{n} - \bar{x}^a_{n}\| \|v^a_{n}\| C_2\epsilon + O_*(\epsilon^3)
\leq C_5\epsilon^2 \left(\frac{CC_3^2 + C\epsilon C_2 C_3 C_4}{C_5} + \mu\right) + O_*(\epsilon^3),
\]
where the next to last inequality follows from Equation (*) and the \(C^1\) dependence of \(E^u(x)\) with respect to \(x\). Hence, \((v)_{n+1}\) holds, for \(\epsilon\) sufficiently small, as long as
\[
\frac{CC_3^2 + C\epsilon C_2 C_3 C_4}{C_5} < 1 - \mu.
\]

Proof of \((iv)_{n+1}\). Let \(\tau(n)\) be the number of iterates after the last good angle, minus one. We will show

Lemma 1. There exist some constants \(\sigma\) and \(\hat{C}_4\) such that
\[
\|x_n - \bar{x}^a_n\| \leq \hat{C}_4 \sigma^{\tau(n)} \|v^a_n\|.
\]

Remark 3. Lemma 1 and the already established property \((iii)_{n+1}\) combined with the fact that good angles occur at least once in every \(2N\) consecutive iterates guarantee the shadowing property
\[
\|x_n - \bar{x}^a_n\| \leq C_3 \hat{C}_4 \sigma^{2N} \epsilon =: C_4 \epsilon.
\]

Proof of Lemma 1. We proceed by induction. Assume the claim is true up to the \(n\)-th step. Applying the triangle inequality to Evolution Equations (*) gives
\[
\|x_{n+1} - \bar{x}^a_{n+1}\| \leq \Lambda \|x_n - \bar{x}^a_n\| (1 + \frac{C\|v^b_n\|}{\epsilon})
\leq \Lambda (1 + CC_3) \|x_n - \bar{x}^a_n\|.
\]
We consider two cases. Let us fix \(K < \frac{C_1}{\Lambda(1 + CC_3)}\).
Case I \[\|x_n - \tilde{x}_n\| \leq K\epsilon.\]

Then \[\|x_{n+1} - \tilde{x}_{n+1}\| \leq \bar{K}K\epsilon,\] and therefore \[\|x_{n+1} - \tilde{x}_{n+1}\| \leq \bar{K}(1 + CC_3)K\epsilon < C_1\epsilon \leq \|\tilde{v}_{n+1}\|.\] The only restriction imposed by this case is \(\hat{C}_4\sigma \geq 1.\)

Case II \[\|x_n - \tilde{x}_n\| > K\epsilon.\]

In this case, \((v)\) implies that \(\angle(x_n - \tilde{x}_n, E^u_{x_n}) = O_\ast(\epsilon).\) In view of \((ii)\), we also have \(\angle(x_{n} - \tilde{x}_n, v_n^a) = O_\ast(\epsilon).\) Let \(\beta_n = \|Df_{x_n}v^u(x_n)\|,\) where \(v^u(x_n) \in E^u_{x_n}\) is a unit length vector. Then,

\[
\|v_{n+1}^b\| = \beta_n\|v_n^a\| + O_\ast(\epsilon^2),
\]

\[
\|v_{n+1}^a\| = \frac{\beta_n}{\sqrt{1 + 2\gamma_n^2}}\|v_n^a\| + O_\ast(\epsilon^2),
\]

\[
\|x_{n+1} - \tilde{x}_{n+1}\| = \beta_n\|x_n - \tilde{x}_n\| + O_\ast(\epsilon^2).
\]

To estimate \(\|x_{n+1} - \tilde{x}_{n+1}\|,\) we consider two further subcases.

Case IIa \[\cos \angle(x_{n+1} - \tilde{x}_{n+1}, \nabla h(x_{n+1})) \geq \kappa,\] where \(\kappa > 0\) is a small constant, depending on \(f, h, C_1\) and \(C_3,\) to be specified later. The small angle conditions stated in the previous paragraph imply that \(x_{n+1} - \tilde{x}_{n+1}\) and \(v_{n+1}^b\) are nearly collinear. Thus,

\[
\|x_{n+1} - \tilde{x}_{n+1}\| = \frac{\beta_n}{1 + 2\gamma_n^2}\|x_n - \tilde{x}_n\| + \beta_n\|x_n - \tilde{x}_n\|(1 - \frac{\cos \angle(x_{n+1} - \tilde{x}_{n+1}, \nabla h(x_{n+1}))}{\cos \angle(v_{n+1}^a, \nabla h(x_{n+1}))}) + O_\ast(\epsilon^2)
\]

\[
= \frac{\beta_n}{1 + 2\gamma_n^2}\|x_n - \tilde{x}_n\| + O_\ast(\epsilon^2) \leq \frac{\hat{C}_4\sigma^\tau(n)}{\sqrt{1 + 2\gamma_n^2}}\|v_{n+1}^a\| + O_\ast(\epsilon^2),
\]

where the inequality follows from the inductive hypothesis. In particular, in this case \(\|x_{n+1} - \tilde{x}_{n+1}\| \leq \hat{C}_4\sigma^\tau(n)\|v_{n+1}^a\| + O_\ast(\epsilon^2).\) Furthermore, \(\|x_{n+1} - \tilde{x}_{n+1}\| \leq \frac{\hat{C}_4\sigma^\tau(n)}{\sqrt{1 + 2\gamma_n^2}}\|v_{n+1}^a\| + O_\ast(\epsilon^2)\) when the angle \(\angle(E_{x_{n+1}}^u, \nabla h(x_{n+1}))\) is good.

Case IIb \[\cos \angle(x_{n+1} - \tilde{x}_{n+1}, \nabla h(x_{n+1})) < \kappa.\] Then,

\[
\|x_{n+1} - \tilde{x}_{n+1}\| \leq \beta_n\|x_n - \tilde{x}_n\|(1 + \frac{C\kappa\|v_{n}^b\|}{\epsilon}) \leq \beta_n(1 + CC_3\kappa)\|x_n - \tilde{x}_n\|
\]

\[
\leq \sqrt{1 + 2C^2_3\|\nabla h\|_{\infty}^2\kappa^2(1 + CC_3\kappa)}\hat{C}_4\sigma^\tau(n)\|v_{n+1}^a\| + O_\ast(\epsilon^2).
\]

Choosing \(\kappa < \frac{1}{\sqrt{\|\nabla h\|_{\infty}^2}}\), ensures that Case IIb implies a bad angle \(\angle(E_{x_{n+1}}^u, \nabla h(x_{n+1})).\) Hence, it is guaranteed that at least once every \(N\) steps, either Case I or Case IIa occur. Whenever this happens, the possible exponential growth of the sequence \(\{\|x_j - \tilde{x}_j\|\}_{j \in \mathbb{N}}\) accumulated over the previous at most \(N - 1\) steps will be reset if we require that

\[
\left(\sqrt{1 + 2C^2_3\|\nabla h\|_{\infty}^2\kappa^2(1 + CC_3\kappa)}\right)^{4N} < 1 + 2\frac{C^2_1}{\gamma^2}.
\]

Indeed, this ensures that \((iv)\) holds, with \(\sigma \geq \sqrt{1 + 2C^2_3\|\nabla h\|_{\infty}^2\kappa^2(1 + CC_3\kappa)},\) provided \(\epsilon\) is sufficiently small and \(\hat{C}_4 \geq 1.\)
The last restriction on the constants \( C_1, \ldots, C_5 \) sufficient for the induction to move forward is therefore

\[
C_4 \geq C_3 \sigma^{2N} \tag{13}
\]

Hence, the induction can be carried on by choosing, in that order, constants \( C_1, C_3, C_4, C_2 \) and \( C_5 \), satisfying the boxed inequalities. The result holds for sufficiently small \( \epsilon \).

Finally, we extend the proof to the case of noisy observation. We remark that this case also covers the situation when the measurements do not come from a true trajectory, but from a pseudo-trajectory, provided its distance to a true trajectory is sufficiently small. Let us assume that the noise in the measurement of \( h \) from the true trajectory is bounded by \( B \epsilon \).

Consider the 2MEKF satisfies

\[
\sqrt{1 + 2C_3^2H}(1 + CC_3H) + \frac{C_3B}{C_1}4N < 1 + \frac{C_1^2}{\gamma^2}.
\]

3.5. Achieving the inductive hypothesis

The induction presented in \S3.4.2 motivates the following definition.

**Definition 3.** Given \( \epsilon, C_1, \ldots, C_5 > 0 \), we say that an initial ensemble \( E \) is attracted to \( IH(C_1, C_2, C_3, C_4, C_5, \epsilon) \) if for some \( n \), the \( n \)-th iterate of \( E \) under the 2MEKF satisfies \( IH_n(C_1, C_2, C_3, C_4, C_5, \epsilon) \). We say that the initial ensembles \( \{E_n\}_{n>0} \) are attracted to \( IH_n(C_1, C_2, C_3, C_4, C_5, \epsilon) \) if there exists \( n \) such that for all \( \epsilon \) sufficiently small, the \( n \)-th iterate of \( E_\epsilon \) under the 2MEKF satisfies \( IH_n(C_1, C_2, C_3, C_4, C_5, \epsilon) \).

**Remark 4.** When \( \epsilon > 0 \) is sufficiently small and \( C_1, \ldots, C_5 \) satisfy the boxed inequalities, the inductive arguments from \S3.4.2 imply that if an ensemble \( E \) is attracted to \( IH(C_1, C_2, C_3, C_4, C_5, \epsilon) \), then \( IH_n(C_1, C_2, C_3, C_4, C_5, \epsilon) \) holds for all sufficiently large \( n \).

Let \( C_1, \ldots, C_5 \) be constants for which the induction in \S3.4.2 is valid, with all boxed inequalities in the proof of Proposition 1 strict. Then, we have the following.

**Proposition 2.** Consider initial ensembles \( \{E_n\}_{n>0} \) satisfying \( IH(\tilde{C}_1, \tilde{C}_2, \tilde{C}_3, \tilde{C}_4, \tilde{C}_5, \epsilon) \) for constants that also satisfy the boxed inequalities. Then, for generic \( h \), the ensembles \( \{E_n\}_{n>0} \) are attracted to \( IH(C_1, C_2, C_3, C_4, C_5, \epsilon) \).

**Proof.** The proof generalizes that of Proposition 1. We observe that if \( \epsilon > 0 \) is sufficiently small, condition (iii) of the inductive hypothesis is attracting in the sense that if a good angle occurs at step \( n \) (this happens at least once within any \( 2N + 1 \) consecutive steps for generic \( h \)), then condition (iii) of \( IH_n(C_1, C_2, C_3, \epsilon, C_4, C_5, \epsilon) \) is satisfied, as the proof of Proposition 1 shows. The rest of the inductive argument remains applicable.

Conditions (i), (ii), (iv) and (v) are also attracting in some sense, but not as simply as (iii). The constants \( \tilde{C}_1, \tilde{C}_4, \tilde{C}_2 \) and \( \tilde{C}_5 \), in that order, can be improved until conditions (i), (iv),
(ii) and (v) of $IH_n(C_1, C_2, C_3\mathbb{X}^{−2N}, C_4, C_5, \epsilon)$ are achieved, and they are maintained thereafter. Here, we explain how to reach (iv) in detail, as is the most involved, and omit the details of the proofs for the other constants, which use similar ideas.

We go back to the cases presented to establish (iv) in the proof of Proposition 1. We observe that if the ensemble is in Case I, condition (iv) is valid at the next step. In Case IIb, the quotient $\frac{\parallel x_{n+1} - \bar{x}_{n+1}\parallel}{\parallel x_n\parallel}$ deteriorates with respect to the same quotient at time $n$ by a fixed multiplicative factor that can be controlled by the choice of $\kappa$, up to higher order terms in $\epsilon$. In Case IIa, this quotient gets reduced by a factor independent of $\kappa$, up to higher order terms in $\epsilon$. Again using the non-degeneracy condition on $h$, and choosing a sufficiently small value for $\kappa$, we can ensure exponentially fast decrease of the quotient $\frac{\parallel x_{n+1} - \bar{x}_{n+1}\parallel}{\parallel x_{n+1}\parallel}$, until it gets to order 1. (The exponential decrease occurs along times of good angles. In between, this quotient may deteriorate, but this deterioration is controlled by $\kappa$). In particular, condition (iv) of $IH_n(C_1, C_2, C_3\mathbb{X}^{−2N}, C_4, C_5, \epsilon)$ is achieved.

The upper bound on $\epsilon$ for which this argument applies is determined by higher order terms ignored in the above estimates. It depends on $f, h$ and the values of $\tilde{C}_1, \ldots, \tilde{C}_5$.

**Remark 5.** In fact, when the quantities $\parallel x_0 - \bar{x}_0\parallel$ and $\angle(v_n^u, E_{\bar{x}_n}^u)$ are small, but much larger than $\epsilon$, the 2MEKF algorithm is still useful. Indeed, the inductive procedure from §3.4.2 remains applicable when $IH(C_1, C_2, C_3, C_4, C_5, \epsilon)$ is replaced by $IH(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)$ with $\epsilon \leq \delta \leq c_v \epsilon$, for some $c > 0$, where $IH(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)$ defined as follows.

**Definition 4.** Let $\epsilon, \delta > 0$. We say that the ensemble satisfies the inductive hypothesis $IH_n(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)$ at time $n$ if the following holds:

(i) $\delta_n^L$ Lower bound on spread of ensemble: $C_1 \epsilon \leq \parallel v_n^a\parallel$.

(ii) $\delta_n^U$ Unstable cone: $\angle(v_n^a, E_{\bar{x}_n}^u) \leq C_2 \delta$.

(iii) $\delta_n^U$ Upper bound on spread of ensemble: $\parallel v_n^a\parallel \leq C_3 \delta$.

(iv) $\delta_n^U$ Shadowing: $\parallel x_n - \bar{x}_n\parallel \leq C_4 \delta$.

(v) $\delta_n^U$ Bound on distance along stable direction: $\parallel x_n - \bar{x}_n\parallel_s \leq C_5 \delta^2$.

In this case, the proof of Proposition 2 remains applicable and yields the following.

**Corollary 1.** Let $\epsilon > 0$ be sufficiently small. Assume that $IH_0(C_1, C_2, C_3\mathbb{X}^{−2N}, C_4, C_5, \epsilon, \delta)$ holds for some initial ensembles $\{\tilde{E}_\epsilon\}_{n>0}$, with constants $C_1, \ldots, C_5$ satisfying the boxed inequalities in §3.4.2. Then, there exists some $c > 0$ independent of $\epsilon$ such that whenever $\epsilon \leq \delta \leq c_v \epsilon$, the forward evolution of $\tilde{E}_\epsilon$ under the 2MEKF satisfies $IH_n(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)$ for all $n \geq 0$. Moreover, $\tilde{E}_\epsilon$ is attracted to $IH(C_1, C_2, C_3, C_4, C_5, \epsilon)$. In other words, the 2MEKF with initial ensembles $\{\tilde{E}_\epsilon\}_{n>0}$ is $O(\epsilon)$-reliable.

### 3.6. Lyapunov exponent

A consequence of the properties of 2MEKF presented above is that the positive Lyapunov exponent of the true trajectory can be well approximated using the ensemble.

Let $x_0$ be any initial condition. Let $x_j$ denote its trajectory, and $z_j := \bar{x}_j^*, \tilde{z}_j := z_j + v_j^a$, with $\bar{x}_j^*, v_j^a$ are as in Equations (*). Let $v^u$ be the unstable direction of $x_0$, $\chi_n = \frac{1}{n} \log \parallel Df_{x_0}^n v^u\parallel$ be the $n$-th step approximation to the positive Lyapunov exponent of $x_0$, and $\chi = \lim_{n \to \infty} \chi_n$ the corresponding Lyapunov exponent.
**Proposition 3.** Let \( \hat{\chi}_n := \frac{1}{n} \sum_{j=0}^{n-1} \log \frac{\| f(z_j) - f(z_0) \|}{\| z_j - z_0 \|} \). Assume that the initial ensemble satisfies \( IH_0(C_1, C_2, C_3X^{-2N}, C_4, C_5, \epsilon) \), for constants \( C_1, \ldots, C_5 \) satisfying the boxed inequalities (9)-(13). Then, for generic \( h \), we have

\[
\| \hat{\chi}_n - \chi_n \| = \mathcal{O}_*(\epsilon)
\]

uniformly on \( n \). In particular, \( \lim_{n \to \infty} \hat{\chi}_n - \chi = \mathcal{O}_*(\epsilon) \).

**Proof.** This follows from the shadowing property of 2MEKF (iv), invariance of unstable cones (ii), and the boundedness of 2MEKF (iii), showed in Proposition 1.

**Remark 6.** In fact, any data assimilation algorithm for which the mean of analysis members shadows the true trajectory, and the displacement vectors have uniformly bounded spread and lie close to the unstable direction also gives a good approximation of the positive Lyapunov exponent of the true trajectory.

4. Properties of the EKF with higher dimensional unstable spaces

In this section we generalize the properties of the 2MEKF presented in §3 to systems with higher dimensional unstable spaces. For the rest of the section, we let \( f : M \to \mathbb{R} \) be a diffeomorphism having a \( k_u \) dimensionally unstable hyperbolic attractor \( \mathcal{A} \), as in §2.1.1 and \( h : M \to \mathbb{R} \) be generic in the sense of §2.1.2. In this case, we consider an ensemble Kalman filter with \( k > k_u \) members, which is denoted \( k\)MEKF.

4.1. Simplification in analysis step

Here we show a simplification of the analysis step, presented in §2.1.3 (Equation (7)), in the case of scalar observation function. This allows to reduce the computational complexity of the algorithm to quadratic order in \( k \), instead of cubic. The simplification is a consequence of the following simple lemmas, whose content can be traced back, at least, to Potter’s work in 1964 [16, 3].

**Lemma 2.** Let \( Q \) be a \( k \)-dimensional row vector, and let \( q^2 = QQ^T \). Then, the symmetric square root of \( (I + Q^T Q)^{-1} \) is given by

\[
(I + Q^T Q)^{-\frac{1}{2}} = I - \gamma(Q)Q^T Q, \quad \text{where} \quad \gamma(Q) = \frac{1}{q^2} \left( 1 \pm \frac{1}{\sqrt{1 + q^2}} \right)^5.
\]

**Proof.** We drop the \( Q \) dependence of \( \gamma(Q) \) for brevity. Now, we verify the claim directly. First, we note that \( I - \gamma Q^T Q \) is symmetric. We let \( M = Q^T Q \), and observe that \( M^2 = q^2 M \). Thus,

\[
(I - \gamma Q^T Q)^2 (I + Q^T Q) = (I - 2\gamma M + \gamma^2 M^2) + (M - 2\gamma M^2 + \gamma^2 M^3)
\]

\[
= I + (-2\gamma + q^2 \gamma^2 + 1 - 2q^2 \gamma + q^4 \gamma^2) M.
\]

The choice of \( \gamma \) ensures that the second term vanishes.

**Lemma 3.** The matrix \( W = I - \gamma Q^T Q \) has an orthonormal basis of eigenvectors, with \( \frac{1}{\sqrt{1 + q^2}} \) as a simple eigenvalue of corresponding eigenvector \( Q^T \) and 1 as an eigenvalue with multiplicity \( k - 1 \).

**Proof.** The first claim follows from symmetry of \( W \). The second claim can be checked directly. The last claim follows from the fact that for every \( v \in \mathbb{R}^k \) such that \( v^T Q^T = 0 \), \( \gamma Q^T Q v = 0 \) and thus \( W v = v \).

\[\text{For } Q = 0, \gamma(Q) := 0. \text{ To ensure that } I - \gamma(Q)Q^T Q \text{ is positive definite, the minus sign must be chosen in the definition of } \gamma(Q).\]
4.2. Evolution equations

Let $\epsilon > 0$. The evolution equations of the kMEKF are as follows. At time $n-1$ we start with an initial ensemble of vectors with mean $\overline{x}_{n-1}^b$ and displacement vectors $v_{j,n-1}^a$, $1 \leq j \leq k$. To obtain the corresponding background vectors at step $n$, we forecast according to $f$. Let us denote the mean of these background ensemble vectors by $\overline{x}_{n}^b$ and the corresponding displacements by $v_{j,n}^b$. Thus,

$$\overline{x}_{n}^b = \frac{1}{k} \sum_{j=1}^{k} f(\overline{x}_{n-1}^a + v_{j,n-1}^a),$$

$$v_{j,n}^b = f(\overline{x}_{n-1}^a + v_{j,n-1}^a) - \overline{x}_{n}^b.$$

The average value of the corresponding observation will be $\overline{y}_{n} := \frac{1}{k} \sum_{j=1}^{k} h(\overline{x}_{n}^b + v_{j,n}^b)$. The measurement of $h$ from the true trajectory $x_n$ will be denoted by $y_n = h(x_n)$.

For $1 \leq j \leq k$, let $q_{j,n} := \frac{1}{\epsilon} (h(\overline{x}_{n}^b + v_{j,n}^b) - \overline{y}_{n})$. To be consistent with the notation used in [11], we let $X_n$ be the matrix whose columns are the displacement vectors $v_{j,n}^b$ and $Y_n$ the row vector with entries $\epsilon q_{j,n}$. To make use of the simplification presented in §4.1, we let $Q_n = \sqrt{\epsilon} = 1 Y_n$ and $q_n^2 = \frac{1}{(k-1)} \sum_{j=1}^{k} q_{j,n}^2 = Q_n Y_n^T$. The corresponding mean and displacement analysis vectors, obtained using an ensemble square root filter, are given by:

$$\overline{x}_{n}^a = \overline{x}_{n}^b + X_n^b ( (k-1) \epsilon^2 I + (Y_n^b Y_n^b)^{-1} (Y_n^b)^T (y_n - \overline{y}_{n}) = \overline{x}_{n}^b + X_n^b (I - \frac{1}{1+q_n^2} Q_n Y_n^T) (y_n - \overline{y}_{n}) = \overline{x}_{n}^b + \frac{1}{(k-1) \epsilon^2} (y_n - \overline{y}_{n}) X_n^b (Y_n^b)^T, \quad X_n^a = X_n^b (I + Q_n Y_n^T) = X_n^b (I - \gamma(Q_n) Q_n^T Q_n) = X_n^b W_n,$$

with $\gamma(Q_n) = \frac{1}{q_n^2} (1 - \frac{1}{\sqrt{1+q_n^2}})$ for $Q_n \neq 0$ and $\gamma(0) = 0$, as in Lemma 2. Let

$$v_{0,n} := \frac{1}{(k-1) \epsilon} X_n^b (Y_n^b)^T = \frac{1}{(k-1)} \sum_{j=0}^{k} q_{j,n} v_{j,n}^b.$$

Then, the equations above simplify to:

$$\overline{x}_{n}^a = \overline{x}_{n}^b + \frac{(y_n - \overline{y}_{n}) v_{0,n}}{1+q_n^2}, \quad (***)$$

$$v_{0,n}^a = v_{0,n}^b - \gamma(Q_n) q_{j,n} v_{0,n}, \quad \text{for } 1 \leq j \leq k.$$

In words, the coordinates of the displacements of analysis ensemble members from the mean in the ordered basis formed by the background ensemble, i.e. the columns of $X_n^b$, are given by the columns of $W_n$, and the transformation from background to analysis ensemble is a contraction by a factor of $(1 + q_n^2)^{-\frac{3}{2}}$ in the direction determined by $Q_n$ (equivalently, by $Y_n^b$); in model space, this contraction is achieved by a displacement in the direction of $v_{0,n}$.

---

*When $k = 2$, $q_{1,n} = -q_{2,n}$ and therefore $q_n^2 = 2q_{1,n}^2$. Moreover $v_{0,n} = 2q_{1,n} v_{1,n}$. This yields Equations (**).*
4.3. Inductive scheme

We now generalize the proof of reliability of the 2MEKF to higher dimensions. The main differences between the two cases arise at the linear level. While for systems with one-dimensional unstable spaces the linear analysis is straightforward, the lack of conformality in the forecast step and the fact that at each analysis step there is contraction along (at most) one direction make the inductive step somewhat more challenging in the higher dimensional case.

Adopting a strategy similar to that of §3.4 and relying on Takens’ embedding theorem proves to be fruitful. In fact, properties (iii) and (v), generalize rather directly. Maintaining a lower bound on the spread of the ensemble in all unstable directions, corresponding to (i), and establishing the shadowing property, corresponding to (iv), require some further work. Property (ii) would remain valid in the setting of $k = k^u + 1$ ensemble members. Here, it is slightly modified to allow for larger ensemble, $k > k^u + 1$.

The main result of the one-dimensional unstable setting, Proposition 1, is extended to the case of $k^u \geq 1$ unstable directions and $k > k^u$ ensemble members in Proposition 4.

4.3.1. Inductive hypothesis $IH^+(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)$

**Definition 5.** Given $\epsilon, \delta > 0$, we say that the ensemble satisfies the inductive hypothesis $IH^+_n(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)$ at time $n$ if the following holds:

**(i+)** Lower bound on spread of ensemble:

$$C_1^2 \epsilon^2 \leq \sum_{j=1}^{k} (v \cdot v_{j,n}^a)^2 \quad \forall \ v \in E_{x_n}^s \text{ with } \|v\| = 1.$$  

**(ii+)** Closeness of ensemble perturbations to unstable directions:

$$\sum_{j=1}^{k} (v \cdot v_{j,n}^a)^2 \leq C_2^2 \delta^4 \quad \forall \ v \in E_{x_n}^{u} \text{ with } \|v\| = 1.$$  

**(iii+)** Upper bound on spread of ensemble:

$$\|v_{j,n}^a\| \leq C_3 \delta, \quad \forall \ 1 \leq j \leq k.$$  

**(iv+)** Shadowing:

$$\|x_n - \overline{x}_n^a\| \leq C_4 \delta.$$  

**(v+)** Bound on distance along stable directions:

$$|\langle x_n - \overline{x}_n^a, v \rangle| \leq C_5 \delta^2 \quad \forall \ v \in E_{x_n}^{u} \text{ with } \|v\| = 1.$$  

**Remark 7.** For $k = 2$, the existence of a constant $C_1$ satisfying (i+) implies the existence of a (possibly different) constant satisfying (i) of §3.4.1. Also, (i+) and (ii+) combined yield (ii) of §3.4.1. See Remark 13 for another implication of (i+).

4.3.2. Inductive step

Now, we will extend the main result of §3, Proposition 1, to this setting. Assuming the genericity conditions on $f$ and $h$ stated in the beginning of the section, we have the following.
Proposition 4 (Properties of k-member ensemble Kalman filter). Let \( x_0 \in A \). Then, there is a family \( \{ \mathcal{E}_\epsilon \}_{\epsilon > 0} \) of open sets of initial ensembles such that whenever \( \mathcal{E}_\epsilon \in \mathcal{L}_\epsilon \) for all \( \epsilon > 0 \), the kMEKF with initial ensembles \( \{ \mathcal{E}_\epsilon \}_{\epsilon > 0} \) and noiseless observations of \( h \) is \( O(\epsilon) \)-reliable.

The same conclusion holds if the measurement of \( h \) has sufficiently small noise of order \( \epsilon \), and also when the observations are generated by a pseudo-trajectory, provided its distance to a true trajectory is sufficiently small.

Strategy of the proof. As in §3.4.2, the proof of Proposition 4 follows from an inductive procedure. We will show that there exist constants \( C_1, \ldots, C_5 \), \( c > 0 \) such that whenever \( \epsilon > 0 \) is sufficiently small, \( \epsilon \leq \delta \leq c \sqrt{\epsilon} \) and \( IH_0^+ (C_1, C_2, C_3, \mathbb{T}^{-2N}, C_4, C_5, \epsilon, \delta) \) holds for some ensembles \( \{ \mathcal{E}_\epsilon \}_{\epsilon > 0} \), then, the forward evolution of \( \mathcal{E}_\epsilon \) under the kMEKF with noiseless measurements of \( h \) satisfies \( IH_n^+ (C_1, C_2, C_3, C_4, C_5, \epsilon, \delta) \) for all \( n \geq 0 \). The \( O(\epsilon) \) reliability of kMEKF follows as in the case of 2MEKF. We restrict ourselves to the noiseless case, as the extension to the noisy setting is also similar to that of the 2MEKF. To this end, we divide the proof of Proposition 4 in several paragraphs, that give the conditions on the constants \( c, C_1, \ldots, C_5 \) for the induction to follow.

Remark 8. Constants \( C_3 \) and \( C_1 \) are independent of the other ones and may be made explicit from our arguments. The remaining relations among constants \( C_1, \ldots, C_5 \) in this setting are similar to those obtained in §3.4.2. In the coming paragraphs we show how to perform the inductive step without obtaining these relations explicitly.

Standing assumption and notation. For the rest of this section we suppose the standing assumption

\[
IH_0^+(C_1, C_2, C_3, \mathbb{T}^{-2N}, C_4, C_5, \epsilon, \delta) \quad \text{and} \quad IH_m^+(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta)
\]

for all \( m \leq n \) \((\text{IH}+)\)

is valid for some \( n \geq 0 \), and some (yet to be determined) constants \( C_1, \ldots, C_5 \). We will show that \( IH_{n+1}^+(C_1, C_2, C_3, C_4, C_5, \epsilon, \delta) \) holds. The proof proceeds by induction provided \( \epsilon \) is sufficiently small, some relation between \( \epsilon \) and \( \delta \) holds, and \( C_1, \ldots, C_5 \) are chosen appropriately.

Before presenting the proof of the inductive step, we introduce some notation and useful remarks.

Definition 6. For each \( x \in A \), let

\[
\Gamma(x) := \max_{\|w\| = 1} \min_{n' \in \{0, \ldots, 2N\}} \frac{1}{w^T (Df_x^{-n'})^T \nabla h(f^{-n'}(x))}, \quad \text{and} \quad \Gamma := \sup_{x \in A} \Gamma(x).
\]

Remark 9. We note that for generic \( h \), for each \( x \in M \), \( \Gamma(x) < \infty \) because the set \( \{ w : \|w\| = 1 \} \) is compact and by Takens’ theorem, \( \{ (Df_x^{-n'})^T \nabla h(f^{-n'}(x)) \}_{0 \leq n' \leq 2N} \) span \( T_x M \). Moreover, \( \Gamma < \infty \) because \( \Gamma(x) \) is continuous in \( x \) and \( A \) is compact.

Definition 7. We say that \( h \) makes a good angle with \( E_x^u \) at time \( n' \) if

\[
\max_{\|w\| = 1, w \in E_x^u} \frac{1}{w^T (Df_x^{-n'})^T \nabla h(f^{-n'}(x))} \leq \Gamma.
\]

Remark 10. Since \( \dim E_x^u = k^u \), the definition of \( \Gamma \) combined with elementary orthogonality considerations implies that for any \( x \), there exists a subset of \( k^u \) numbers,

\[
\{ g_t(x) < \cdots < g_k(x) \} \subset \{ 0, \ldots, 2N \}
\]

such that for each \( 1 \leq i \leq k^u \), \( h \) makes a good angle with \( E_x^u \) at time \( g_i(x) \).
Remark 11. Recall that $f$ is a diffeomorphism and for all $n \in \mathbb{Z}$, $Df^n_x E^u_x = E^u_{f^n(x)}$. Because the norm of $Df^n_x$ is uniformly bounded for $x \in \mathcal{A}$ and $|n| \leq 2N$, there is some $\tilde{\Gamma} > 0$ independent of $x \in \mathcal{A}$ such that whenever $h$ makes a good angle with $E^u_x$ at time $0 \leq n' \leq 2N$, we have that

$$\max_{\|w\|=1, w \in E^u_{f^{-n'}(x)}} \frac{1}{|w^T \nabla h(f^{-n'}(x))|} \leq \tilde{\Gamma}. \quad (\text{Bounded ensemble})$$

Proposition 5. Let $\mathcal{A} \subset \mathbb{R}^N$ to be $\mathcal{A} := \{w \in \mathbb{R}^N \mid \angle(E^u_x, \nabla h(x)) \text{ is good}\}$ exactly when $\frac{1}{\|\nabla h(x)\| \cos \angle(E^u_x, \nabla h(x))} \leq \Gamma$. \quad (I)

Remark 12. By Remarks (10) and (11), in each sequence of $2N + 1$ consecutive iterates $f^{-2N}(x), \ldots, f^{-1}(x), x$ there are at least $k^* > 0$ good angles, say at times $0 \leq g_1(x) < \cdots < g_4(x) \leq 2N$. Furthermore, the vectors $\{(Df_x^{-g_i(x)})^T \nabla h(f^{-g_i(x)}(x))\}_{1 \leq i \leq 4}$ may be chosen to be linearly independent.

Proof of $(iii)_n$. Specifically, we prove the following.

Proposition 5 (Bounded ensemble). Let $C_3 \geq \sqrt{\Gamma}k^2$, where $k$ is the number of ensemble members and $\Gamma$ is given by Definition 6. Then, whenever $c$ and $\epsilon$ are sufficiently small, independently of $n$, and $\epsilon \leq \delta < c\sqrt{\epsilon}$, then $\|v^a_{j,n+1}\| \leq C_3\delta$ for all $1 \leq j \leq k$, i.e. $(iii)_n$ holds.

The proof of the Proposition 5 relies on the following lemma.

Lemma 4. Let $\epsilon$ be sufficiently small, and $m \leq n$. Then, there exists $c > 0$ independent of $m$ and $n$ such that whenever $\epsilon \leq \delta < c\sqrt{\epsilon}$ and $\|v^a_{j,m-2N}\| \leq C_3\delta$ for some $m \geq 0$ and all $1 \leq j \leq k$, we have the following. For all $m - 2N \leq n' \leq m$,

1. $\|(X^a_{m})^T \nabla h_{n'}\| < \sqrt{\epsilon}$,
2. $\|(Df^{m-2N}X_{m-2N} W_{m-2N} \ldots W_n)^T \frac{(Df^{n'-m})^T \nabla h_{n'}}{\|(Df^{n'-m})^T \nabla h_{n'}\|} \leq \frac{k\epsilon}{\|(Df^{n'-m})^T \nabla h_{n'}\|}$, where $\nabla h_{n'}$ is a shorthand for $\nabla h(f^{n'}(x))$. Furthermore,
3. For any $m - 2N \leq n' \leq m$,
4. $\|(X^a_{m})^T \frac{(Df^{n'-m})^T \nabla h_{n'}}{\|(Df^{n'-m})^T \nabla h_{n'}\|} \leq \frac{k\epsilon}{\|(Df^{n'-m})^T \nabla h_{n'}\|}$. \quad (IV)

Proof of Lemma 4. Assume $c > 0$ sufficiently small, $\epsilon \leq \delta < c\sqrt{\epsilon}$ and $\|v_{j,m-2N}\| \leq C_3\delta$ for some $m \leq n$ and all $1 \leq j \leq k$.

\footnote{If $n' < 0$ the content of (I)-(IV) is meaningless.}
• Proof of (I). Let $m - 2N \leq n' \leq m$. Then,
\[
\|(X^a_{n'})^T \nabla h_{n'}\| = \|((X^b_{n'})^T - \gamma(Q_{n'})Q^T_{n'}Q_{n'}(X^b_{n'})^T) \nabla h_{n'}\|
\]
\[
= \|((I - \gamma(Q_{n'})Q^T_{n'}Q_{n'})Y_{n'}^T) + O(\max_{1 \leq j \leq k} \|v^b_{j,n'}\|^2)\|
\]
\[
= \frac{c\sqrt{k} - 1}{\sqrt{1 + q^2_{n'}}} \|Q^T_{n'}\| + O(\max_{1 \leq j \leq k} \|v^b_{j,n'}\|^2) < \sqrt{k}\epsilon,
\]
where the last inequality is valid for sufficiently small $c > 0$.

• Proof of (II). Let $m - 2N \leq n' \leq m$. Recall that
\[
X^a_{n'} = Df_{n'-1}X^a_{n'-1} + O(\max_{1 \leq j \leq k} \|v^b_{j,n'}\|^2), \quad \text{and} \quad X^a_{n'} = X^b_{n'}W_{n'},
\]
where $Df_{n'}$ is the linearization of $f$ at the point $x^a_{n'}$ and $W_{n'}$ was introduced in §4.2. Then, in view of the standing assumption (III+),
\[
X^a_{n'} = Df^{n'-m-2N}_{m-2N}X^b_{m-2N}W_{m-2N} \ldots W_{n'} + O(\max_{1 \leq j \leq k} \|v^b_{j,n'}\|^2).
\]
Then, if $c > 0$ is sufficiently small, the following holds for all sufficiently small $\epsilon > 0$.
\[
\|(X^a_{n'})^T \nabla h_{n'}\| < \sqrt{k}\epsilon \quad \Rightarrow \quad \|(Df^{n'-m-2N}_{m-2N}X^b_{m-2N}W_{m-2N} \ldots W_{n'})^T \nabla h_{n'}\| < k\epsilon \quad \Rightarrow \quad \|(Df^{2N}_{m-2N}X^b_{m-2N}W_{m-2N} \ldots W_{n'})^T \nabla h_{n'}\| < \frac{k\epsilon}{k\epsilon}
\]
\[
\|(Df^{2N}_{m-2N}X^b_{m-2N}W_{m-2N} \ldots W_{n'})^T \nabla h_{n'}\| < \frac{k\epsilon}{k\epsilon}.
\]

• Proof of (III). Since $W_{n'}$ is a (non-strict) contraction for every $n'$, we have that
\[
\|(X^a_{m})^T (Df^{n'-m}_{m})^T \nabla h_{n'}\| \leq \|(Df^{2N}_{m}X^b_{m-2N}W_{m-2N} \ldots W_{n'})^T (Df^{n'-m}_{m})^T \nabla h_{n'}\| + O(\max_{1 \leq j \leq k} \|v^b_{j,n'}\|^2).
\]
Hence, if $c > 0$ is sufficiently small, for all sufficiently small $\epsilon > 0$ and all $m - 2N < n' < m$ we have that
\[
\|(X^a_{m})^T (Df^{n'-m}_{m})^T \nabla h_{n'}\| \leq \frac{k\epsilon}{k\epsilon}.
\]

• Proof of (IV). First, note that for all $v, w \in \mathbb{R}^N$ such that $w^Tv \neq 0$ we have that $\|v\| = \frac{|v^T w|}{|\cos\angle(v,w)|}$. Using (III) and the definition of $\Gamma$ we have,
\[
\|(X^a_{m})^T\| \leq k \max_{0 \leq j \leq k} \|v^a_{j,m}\|
\]
\[
\leq k \max_{0 \leq j \leq k} \min_{m-2N \leq n' \leq m} \frac{\|(X^a_{m})^T (Df^{n'-m}_{m})^T \nabla h_{n'}\|}{|\cos\angle(v^a_{j,m}, (Df^{n'-m}_{m})^T \nabla h_{n'})|}
\]
\[
\leq \max_{\|w\|=1} \min_{m-2N \leq n' \leq m} \frac{k^2\epsilon}{|(Df^{n'-m}_{m})^T \nabla h_{n'}\| |\cos\angle(w, (Df^{n'-m}_{m})^T \nabla h_{n'})|}
\]
\[
= \max_{\|w\|=1} \min_{m-2N \leq n' \leq m} \frac{k^2\epsilon}{|w^T (Df^{n'-m}_{m})^T \nabla h_{n'}\|} \leq \Gamma k^2\epsilon.
\]

22
Proof of Proposition 5. Let $C_3 \geq \Gamma k^2$, and assume $\epsilon, c > 0$ are sufficiently small. By the standing assumption (IH+), $\|v_{j,n}\| \leq C_3^{2N-\delta}$ for all $1 \leq j \leq k$, then $\|v_{j,n}\| \leq C_3^{2N-\delta}$ for all $1 \leq j \leq k$ and $0 \leq n \leq 2N$.

Furthermore, it follows from Lemma 4 and the choice of $C_3$ that whenever $\|v_{j,m}\| \leq C_3^{2N-\delta}$ for some $m$ and all $1 \leq j \leq k$, then $\|v_{j,m}\| \leq C_3\epsilon$, for all $1 \leq j \leq k$. Hence, when $n \geq 2N$, using the standing assumption (IH+) we have that $\|v_{j,n+1}\| \leq C_3\epsilon \leq C_3\delta$ for all $1 \leq j \leq k$.

Proof of (i+)$n+1$. Before presenting the main result of this part, we introduce some notation. For each $m \geq 0$ and $\psi \in T^*_m M$, the dual space of $T^*_m M$, let

$$\phi^a_m(\psi) = \sum_{j=1}^k \psi(v_{j,m})^2 \frac{1}{\|\psi\|^2},$$

where $\|\psi\|^2 := \sup_{\|v\|=1} |\psi(v)|$. Given a basis $\{w_1, \ldots, w_N\}$ of $T^*_m M$ for which $w_j \in E^a$ for $1 \leq j \leq k^a = \dim E^a$ and $w_j \in E^s$ for $k^a \leq j \leq N$, we consider the dual basis $\{w'_1, \ldots, w'_N\}$ of $T^*_m M$, defined by $w'_j(w_j) = \delta_{ij}$. The spaces $E' = \langle w'_1, \ldots, w'_k \rangle$, $E'' = \langle w'_k, \ldots, w'_N \rangle$, are independent of the particular choice of the vectors $w_1, \ldots, w_N$. In fact, the one-to-one correspondence between $T^* M$ and $TM$ induced by the Riemannian metric on $M$ defines a one-to-one correspondence between $E'$ and $E''$.

We let

$$z^a_m = \inf_{\psi \in E^a_m \setminus \{0\}} \phi^a_m(\psi) = \min_{\psi \in E^a_m / \{0\}} \phi^a_m(\psi) = \min_{\psi \in E^a_m / \{0\}} \phi^a_m(\psi).$$

Let $\phi^b_m(\psi)$ and $z^b_m$ be defined analogously.

The main result of this part is the following.

**Proposition 6** (Spread of ensemble). Let $C_1 \leq \min\{\left(\frac{\lambda^2-1}{M}\right)^{\frac{3}{2}}, \frac{3}{(1+k\lambda)^2}c^3\}$, where $\lambda > 1$, is a strict lower bound on the expansion along unstable directions, $\Lambda$ is a Lipschitz constant for $f$, $k$ is the number of ensemble members and the constants $L$ and $M$ are defined in the course of the proof. Then, whenever $c$ and $\epsilon$ are sufficiently small, independently of $n$, and $\epsilon \leq \delta < c\sqrt{\epsilon}$, then $z^a_{m+1} \geq C_1^2 \epsilon^2$, i.e. (i+)$n+1$ holds.

Before the proof, we show two auxiliary estimates.

**Lemma 5.** For every $m > 0$,

(1) $z^b_m \geq \lambda^2 z^a_{m-1}$.

(2) For all $\psi \in T^*_m M$, $\phi^a_m(\psi) \geq \frac{1}{1+q_m} \phi^b_m(\psi)$.

**Proof of (1).** For any $\psi \in T^*_m M$, linear approximation yields

$$\psi(v_{m,j}) = Df^*\psi(v_{m-1,j}) + O_4(\delta^2),$$

where $Df^* : T^* M \otimes$ is defined by $Df^*\psi(v) = \psi(Df v)$, for $\psi \in T^*_f M, v \in T_x M$. Therefore, when $\psi \neq 0$,

$$\phi^b_m(\psi) = \frac{\|Df^*\psi\|^2_{\psi}}{\|\psi\|^2_{\psi}} \phi^a_m(\psi) + O_4(\delta^3).$$

23
Let \( D'f = (Df^*)^{-1} \), the so-called the co-differential of \( f \). The spaces \( E^u \) and \( E^v \) are invariant under \( D'f \), by the corresponding invariance of \( E^u \) and \( E^v \) under \( Df \). Moreover, for \( \psi \in E^v_1 \), we have that

\[
\|D'f\psi\|^* = \sup_{w \in E^v_1 \setminus \{0\}} \frac{|D'f\psi(w)|}{\|w\|} = \sup_{v \in E^v_1 \setminus \{0\}} \frac{|D'f\psi(Dfv)|}{\|Dfv\|} = \sup_{v \in E^v_1 \setminus \{0\}} \frac{|\psi(v)|}{\|Dfv\|} < \lambda^{-1} \sup_{v \in E^v_1 \setminus \{0\}} \|\psi(v)\| = \lambda^{-1} \|\psi\|^*.
\]

Thus,

\[
\phi^b_m(D'f\psi) = \frac{\|\psi\|^*}{\|D'f\psi\|^*} \phi^a_{m-1}(\psi) + \mathcal{O}_* (\delta^3) > \lambda^2 \phi^a_{m-1}(\psi) + \mathcal{O}_* (\delta^3).
\]

Thus, if \( \delta \) is sufficiently small, \( \phi_m^b > \lambda^2 \phi^a_{m-1} \).

**Proof of (2).** We work at step \( m \), and to simplify the notation, we drop the explicit dependence on \( m \). Let \( \psi \in T^*_x M \), and let \( w \in T_x M \) be the vector such that \( \psi(v) = v \cdot w \). Using Cauchy-Schwartz inequality,

\[
(v_0 \cdot w)^2 = \frac{1}{(k-1)^2} \left( \sum_{j=1}^k q_j v_j^b \cdot w \right)^2 \\
\leq \frac{1}{(k-1)^2} \left( \sum_{j=1}^k q_j^2 \right) \left( \sum_{j=1}^k (v_j^b \cdot w)^2 \right).
\]

Then, \( \gamma(v_0 \cdot w)^2 \leq \frac{1}{(k-1)^2} \left( 1 - \frac{1}{\sqrt{1+q^2}} \right) \sum_{j=1}^k (v_j^b \cdot w)^2 \).

Equations (**) yield

\[
\sum_{j=1}^k (v_j^b \cdot w)^2 = \sum_{j=1}^k (v_j^b \cdot w)^2 + \gamma^2 \sum_{j=1}^k (q_j v_0 \cdot w)^2 - 2 \gamma \sum_{j=1}^k (v_j^b \cdot w)(q_j v_0 \cdot w) \\
= \sum_{j=1}^k (v_j^b \cdot w)^2 + \frac{k(k-1)}{4} \gamma^2 (v_0 \cdot w)^2 - 2 \gamma (v_0 \cdot w)^2 \\
= \sum_{j=1}^k (v_j^b \cdot w)^2 - (k-1) \gamma (v_0 \cdot w)^2 (\gamma q^2 - 2) \\
\geq \frac{1}{1+q^2} \sum_{j=1}^k (v_j^b \cdot w)^2,
\]

where the last inequality follows from the calculation above. Thus, \( \phi^a(\psi) \geq \frac{1}{1+q^2} \phi^b(\psi) \).

Letting \( w = \nabla h \), it is straightforward to get the following.

**Corollary 2.**

\[
\sum_{j=1}^k (v_j^b \cdot \nabla h)^2 = \frac{1}{1+q^2} \sum_{j=1}^k (v_j^b \cdot \nabla h)^2. \tag{14}
\]

**Proof of Proposition 6.** Let \( \psi \in E^v_1 \), \( M \), such that \( \|\psi\|^* = 1 \). We think of \( \psi \) as a horizontal vector, so \( \psi(v) = \psi v \) for all \( v \in T^*_x M \); thus \( \psi \psi^T = 1 \). Let \( Dh = Dh^u + Dh^s \) with \( Dh^u \in \)
$E^u$ and $Dh^s \in E^{s'}$. Let us decompose $\psi = \psi_h + \psi_0$, where $\psi_h \in (Dh^u)$, and $\psi_0$ such that $\psi_0 X_{n+1}^b (X_{n+1}^b)^T (Dh^v)^T = 0$. Then,

$$\psi X_{n+1}^b (X_{n+1}^b)^T \psi^T = \psi_h X_{n+1}^b (X_{n+1}^b)^T \psi_h^T + \psi_0 X_{n+1}^b (X_{n+1}^b)^T \psi_0^T.$$

Also, by Lemma 5(2),

$$\psi X_{n+1}^a (X_{n+1}^a)^T \psi^T \geq \frac{1}{1 + q_{n+1}^2} \psi_h X_{n+1}^b (X_{n+1}^b)^T \psi_h^T + \psi_0 X_{n+1}^b (X_{n+1}^b)^T \psi_0^T + O_s(\delta^3).$$

Fix $K \geq 1$ to be determined later.

**Case I** $\psi_0 \psi_0^T \leq K \psi_h \psi_h^T$.

Then,

$$\phi_n^{a+1} (\psi) \geq \frac{\phi_n^{b+1} (\psi)}{1 + q_{n+1}^2} = \frac{\phi_n^{b+1} (\psi)}{1 + \frac{Dh(Dh)^T}{(k-1)\epsilon^2} \phi_n^{b+1} (\psi)} + O_s(\delta).$$

Thus,

$$\phi_n^{b+1} (\psi) \geq \frac{\phi_n^{b+1} (\psi)}{1 + \frac{Dh(Dh)^T}{(k-1)\epsilon^2} \phi_n^{b+1} (\psi)} + O_s(\delta).$$

Let $\hat{L} = \frac{4L^2h^2}{(k-1)}$. Then, Lemma 5(1) yields $\phi_n^{a+1} (\psi) \geq \frac{\lambda^2 z^a_n}{1 + \frac{1}{K} \delta} (z^a_n)$, for $\delta$ sufficiently small.

**Case II** $\psi_0 \psi_0^T > K \psi_h \psi_h^T$.

By Cauchy-Schwartz inequality,

$$2\psi_0 \psi_0^T \leq \frac{1}{\sqrt{K}} \psi_h \psi_h^T + \sqrt{K} \psi_h \psi_h^T \leq \frac{2}{\sqrt{K}} \psi_0 \psi_0^T.$$}

Hence,

$$1 = \psi \psi^T < (1 + \frac{1}{\sqrt{K}})^2 \psi_0 \psi_0^T.$$}

Then,

$$\phi_n^{a+1} (\psi) \geq \psi_0 \psi_0^T \phi_n^{b+1} (\psi) + O_s(\delta^3) \geq \frac{1}{1 + \frac{1}{\sqrt{K}}^2} \phi_n^{b+1} (\psi) + O_s(\delta^3) \geq \frac{\phi_n^{b+1} (\psi)}{1 + \frac{1}{\sqrt{K}}} + O_s(\delta^3).$$}

Hence, for $\delta$ sufficiently small, Lemma 5(1) yields $\phi_n^{a+1} (\psi) \geq \frac{\lambda^2 z^a_n}{1 + \frac{1}{\sqrt{K}}}$.

Choosing $K = \left( \frac{3e^2}{2} \frac{L_{n}}{z_{n-1}} \right)^2$ and $\hat{M} = (9\hat{L})^{\frac{1}{4}}$ yields $\phi_n^{a+1} (\psi) \geq \frac{\lambda^2 z^a_n}{1 + \frac{1}{M} \left( \frac{\hat{M}}{2} \right)}$ whenever $K \geq 1$.

When $K < 1$, $\frac{3e^2}{L_{n-1}^2} < 1$ and therefore $z^a_{n+1} \geq \frac{1}{1 + q_{n+1}^2} z^a_n \geq \frac{3e^2}{L(1 + k\hat{M})C_3^1}$. Thus, $z^a_n \geq \min\{z_0, (\frac{\lambda^2}{M})^2, \frac{3e^2}{L(1 + k\hat{M})C_3^1}\}$.

By the standing assumption (IH+) and the choice of $C_1$, the proof is complete.

**Remark 13.** Condition $(i+)_n$ implies that there is a constant $\hat{C} > 0$ such that when the angle $\angle(\nabla h_{z_{n}}, E_{z_{n}}^2)$ is good and $\delta$ is sufficiently small, $q_n \geq \hat{C}$. 

25
Lemma 5(2). displacement of the analysis (background) ensemble mean to the truth. Compared to \( C \) of the \( O \) \( \phi \) of the proof of (1) is analogous to that of Lemma 5(1), we note that the constant in front of \( \sum_{j=1}^{k} \| D h_{x_n}^{u} \|_{*}^{2} \phi_{m}^{b}(D h_{x_n}^{u}) + O_{*}(\delta) \) \( > \frac{1}{\epsilon^{2}(k-1)} C^{2} C_{f}^{2} \epsilon^{2} + O_{*}(\delta) \).

Letting \( \hat{C} = \frac{C^{2} C_{f}^{2}}{k-1} \) yields the claim, for \( \delta \) sufficiently small. \( \square \)

Proof of \((ii+)_n+1\). The proof of \((ii+)\) follows similarly to that of Proposition 6. For each \( m \geq 0 \), we let

\[
\hat{z}_{m}^{a} = \sup_{\psi \in E_{m}^{f} \setminus \{ 0 \}} \phi_{m}^{a}(\psi) = \max_{\psi \in E_{m}^{f}} \phi_{m}^{a}(\psi) = \max_{\psi \in E_{m}^{f}} \sum_{j=1}^{k} (v \cdot u_{j,m}^{a})^{2}.
\]

We define \( \hat{z}_{m}^{b} \) analogously.

Then, we have the following.

**Proposition 7** (Closeness to unstable directions). There exists some \( C_{2} > 0 \) independent of \( n \) such that whenever \( c \) and \( \epsilon \) are sufficiently small, independently of \( n \), and \( \epsilon \leq \delta < c \sqrt{\epsilon} \), then \( \hat{z}_{n+1}^{a} \leq C_{2}^{2} \delta^{4} \), i.e. \((ii+)_n+1\) holds.

As above, the proof relies on two auxiliary estimates.

**Lemma 6.** For all \( m > 0 \),

1. \( \hat{z}_{m}^{b} \leq \mu^{2} \hat{z}_{m-1}^{a} + O(C_{3}^{2} \delta^{4}) \).
2. For all \( \psi \in T_{s}^{*} M \), \( \phi_{m}^{a}(\psi) \leq \phi_{m}^{b}(\psi) \).

**Proof.** The proof of (1) is analogous to that of Lemma 5(1), we note that the constant in front of the \( O(C_{3}^{2} \delta^{4}) \) error term depends on \( f \) and \( h \) only. Part (2) follows directly from the proof of Lemma 5(2). \( \square \)

**Proof of Proposition 7.** Follows directly from Lemma 6, by choosing \( C_{2} \) sufficiently large compared to \( C_{3} \). \( \square \)

**Proof of \((v+)_n+1\).** The proof of \((v+)\) is entirely analogous to that of \((v)\).

**Proof of \((iv+)_n+1\).** To shorten notation, for each \( m \geq 0 \), we let \( \delta_{m}^{a(b)} = x_{m} - \bar{x}_{m}^{a(b)} \) be the displacement of the analysis (background) ensemble mean to the truth.

The goal of this part is to show the following.

**Proposition 8** (Shadowing). There is a constant \( C_{4} \) depending on \( f, h, C_{1} \) and \( C_{3} \) and independent of \( n \) such that whenever \( c \) and \( \epsilon \) are sufficiently small, independently of \( n \), and \( \epsilon \leq \delta < c \sqrt{\epsilon} \), then \( \| \delta_{n+1}^{a(b)} \| \leq C_{4} \delta \), i.e. \((iv+)_n+1\) holds.
Before presenting the proof we establish some properties of \( d_{m}^{(b)} \). We will write \( d_{m}^{(b)} \) in coordinates with respect to the ensemble displacements \( v_{m,k}^{(a)} \), up to small error terms. Thus, let \( e_{m}^{a} \) be such that \( d_{m}^{(b)} = X_{m}^{(b)} e_{m}^{a} + \mathcal{O}_{\ast}(\delta^{2}) \). (Note that \( e = \mathcal{O}_{\ast}(\frac{\epsilon}{\epsilon}) \).

From \( \S 4.2 \), we know that \( X_{m+1}^{b} = Df_{x_{m}}X_{m}^{a} + \mathcal{O}_{\ast}(\delta^{2}) \) and \( d_{m+1}^{b} = Df_{x_{m}}d_{m}^{a} + \mathcal{O}_{\ast}(\delta^{2}) \). Then, we get

\[
d_{m+1}^{b} = Df_{x_{m}}X_{m}^{a} e_{m}^{a} + \mathcal{O}_{\ast}(\delta^{2}) = X_{m+1}^{b} e_{m}^{a} + \mathcal{O}_{\ast}(\frac{\delta^{3}}{\epsilon}).
\]

For the analysis step, from equations (**) we get

\[
d_{m+1}^{a} = d_{m+1}^{b} + \frac{(y_{m+1} - \bar{h}_{m+1})}{1 + q_{m+1}^{2}} e_{m+1}.
\]

For convenience of notation, we now drop the \( m+1 \) indices for the reminder of this paragraph, writing indices only when necessary. Recall that \( y - \bar{h} = \nabla h \cdot d^{b} + \mathcal{O}_{\ast}(\delta^{2}) = Dh^{b} + \mathcal{O}_{\ast}(\delta^{2}) \) and \( v_{0} = \frac{1}{(k-1)\epsilon} X^{b}(Y^{b})^{T} = \frac{1}{(k-1)\epsilon} X^{b}(X^{b})^{T} Dh^{T} + \mathcal{O}_{\ast}(\frac{\delta^{3}}{\epsilon}) \). Then,

\[
d^{a} = d^{b} - \frac{1}{1 + q^{2}} \left( \frac{1}{(k-1)\epsilon} X^{b}(X^{b})^{T} Dh^{T} Dh^{b} + \mathcal{O}_{\ast}(\frac{\delta^{4}}{\epsilon^{2}}) \right) = X^{b}(I - \frac{1}{1 + q^{2}} \left( \frac{1}{(k-1)\epsilon} X^{b}(X^{b})^{T} Dh^{T} Dh^{b} \right)) e^{b} + \mathcal{O}_{\ast}(\frac{\delta^{4}}{\epsilon^{2}}) = X^{a} W^{-1}(I - \frac{1}{1 + q^{2}} \left( \frac{1}{(k-1)\epsilon} X^{b}(X^{b})^{T} Dh^{T} Dh^{b} \right)) e^{b} + \mathcal{O}_{\ast}(\frac{\delta^{4}}{\epsilon^{2}}).
\]

Since \( Y = Dh^{b} + \mathcal{O}_{\ast}(\delta^{2}) \) and \( Q = \frac{1}{\sqrt{k-1}\epsilon} Y^{b} \), we can write

\[
d^{a} = X^{a} W^{-1}(I - \frac{1}{1 + q^{2}} (Q)Q) e^{b} + \mathcal{O}_{\ast}(\frac{\delta^{4}}{\epsilon^{2}}).
\]

We can simplify the last expression by recalling that \( W = I - \gamma(Q)Q^{T}Q \). Straightforward algebra shows that

\[
d^{a} = X^{a} (I - \frac{1}{1 + q^{2}} (1 - \frac{1}{\sqrt{1 + q^{2}}})(Q)Q) e^{b} + \mathcal{O}_{\ast}(\frac{\delta^{4}}{\epsilon^{2}}).
\]

Since \( q^{2} = QQ^{T} \), up to an error proportional to \( \frac{\delta^{4}}{\epsilon^{2}} \leq C_{4} \), \( e_{m+1}^{a} \) is obtained from \( e_{m}^{a} \) by contracting in the direction of \( Q^{T} \) by a factor of \( \frac{1}{1 + q^{2}} \), and leaving all orthogonally complementary directions unchanged. From Corollary 2, we know that in model space, the analysis step contracts the total projection onto \( \nabla h \) by a factor of \( \frac{1}{1 + q^{2}} \) by means of adjusting in the direction of \( X^{b}(Y^{b})^{T} \) (equivalently, \( v_{0} \)).

The argument above shows that, in the linear approximation, the map \( e_{m}^{a} \mapsto e_{m+1}^{a} \) is a (non-strict) contraction. Furthermore, using Remark 13, we have that each time a good angle \( \angle (\nabla h_{x_{m}}, E_{x_{m}}^{u}) \) occurs, the contraction is by a factor

\[
\nu_{m} \leq (1 + \tilde{C})^{-\frac{3}{2}} := \nu.
\]

By Remark 12, we know that the composition of the \( 2N + 1 \) consecutive contractions \( e_{m}^{a} \mapsto e_{m+2N+1}^{a} \), includes at least \( k^{a} = \dim E^{a} \) contractions by at least \( \nu \) in linearly independent directions. This implies that the composition is a contraction on \( E^{a} \), though the contraction factor may be close to 1 if the contraction directions are close to being linearly dependent. But for generic \( h \), we can bound the contraction factor away from 1 by compactness, as in Remark 9.
Proof of Proposition 8. Let $C_{n+1}$ be a space spanned by $k^n$ linearly independent directions of contraction of strength at least $\nu = (1 + \tilde{C})^{-\frac{1}{2}}$. The existence of such a space is guaranteed by the previous paragraph. Furthermore, the choice can be made in such a way that there exists a (sufficiently large) $K$ independent of $x$ and $n$ for which, up to higher order terms, all vectors inside a cone

$$K_{n+1} := \{ v = v_{C_{n+1}} + v_{C_{n+1}}^\perp | v_{C_{n+1}} \in C_{n+1}, v_{C_{n+1}}^\perp \in C_{n+1}^C, \frac{\|v_{C_{n+1}}\|}{\|v_{C_{n+1}}^\perp\|} \leq K \}$$

are contracted by at least $(1 - \frac{1-\epsilon^2}{1+K^2})^{\frac{1}{2}}$. In particular, that is the case for all $v \in C_{n+1}$.

When $\epsilon, \delta$ are sufficiently small, we may incorporate higher order terms to get that the orthogonal projections of $e_{n+1}^a$ to $C_{n+1}$ are uniformly bounded, say by $C \epsilon$ independent of $n$. In Proposition 5 we proved that the columns of the matrix $X_{n+1}^a$ are bounded by $C \delta$. Therefore, the orthogonal projections of $d_{n+1}^a$ to $X_{n+1}^aC_{n+1}$ are bounded by $\tilde{C} \delta$, with $\tilde{C}$ independent of $n$.

By the choice of $C_{n+1}$, $\mathcal{L}(X_{n+1}^aC_{n+1}, E_{n+1}^a) = O_{\epsilon}(\frac{\delta}{\epsilon^2})$, and the multiplicative constant is controlled by the choice of $c$. Hence, if $c > 0$ is sufficiently small, $\mathcal{L}(X_{n+1}^aC_{n+1}, E_{n+1}^a)$ is bounded away from zero independently of $n$. Combining this with the already established property ($+\nu$) yields an upper bound on $\|d_{n+1}^a\|$ proportional to $\delta$ and depending on $f, h, C_1, C_3$ and smallness of $\epsilon$ and $c$. \hfill $\Box$

4.3.3. Improvement to $O(\epsilon)$ reliability

Let us assume that $IH_0^+(\tilde{C}_1, \tilde{C}_2, \tilde{C}_3 \tilde{X}^{-2N}, \tilde{C}_4, \tilde{C}_5, \epsilon, \delta)$ holds for some sufficiently small $c$ and $\epsilon$ with $\epsilon \leq \delta < c\sqrt{\epsilon}$, and suitable constants $\tilde{C}_1, \ldots, \tilde{C}_5$. In §4.3.2, we have just proved that $IH_0^+(\tilde{C}_1, \tilde{C}_2, \tilde{C}_3, \tilde{C}_4, \tilde{C}_5, \epsilon, \delta)$ remains valid for all $n \geq 0$.

As in the case of one-dimensionally unstable direction, we can improve this result to the following.

Corollary 3. Let $C'_1, \ldots, C'_5$ be any constants for which the induction of §4.3.2 applies. Then, there exist $(C_1, \ldots, C_5)$ arbitrarily close to $(C'_1, \ldots, C'_5)$ such that if $\epsilon$ is sufficiently small, $IH_0^+(C_1, C_2, C_3, C_4, C_5, \epsilon, \epsilon)$ holds for all sufficiently large $n$, i.e. $IH_0^+(C_1, C_2, C_3, C_4, C_5, \epsilon, \epsilon)$ is attracting.

Proof. It follows from the induction of §4.3.2 and the proof of Proposition 5 that, for all $n \geq 2N + 1, \|v_{a,n}\| \leq C_3 \epsilon$ for all $1 \leq j \leq k$, so that (iii+) of $IH_0^+(C_1, C_2, C_3 \tilde{X}^{-2N}, C_4, C_5, \epsilon, \epsilon)$ is attracting.

For (i+), we refer to the proof of Proposition 6. Since the fixed point $z_* := (\frac{\lambda^2 - 1}{M})^3 \epsilon^2$ of $F(z) := \frac{\lambda z}{1 + M(\frac{z}{M})^3}$ is a global attractor, we can also conclude that (i+) is attracting.

The remaining properties may be established in a similar manner; see also the proof of Proposition 2. \hfill $\Box$

4.4. Lyapunov exponents

The KMEK also allows to approximate the maximal Lyapunov exponent of $f|_A$, as in §3.6. For an initial ensemble with mean $\mathbb{E}_0^\epsilon \in A$ and perturbations $X_0^\delta$ satisfying the standing assumption (IH+), with constants $C_1, \ldots, C_5$ for which the inductive procedure of §4.3.2 holds, we have the following.

Proposition 9.

$\chi_{\max} := \lim_{n \to \infty} \frac{1}{n} \log \|Df_{x_0}^n\| = \underline{\lim}_{n \to \infty} \frac{1}{n} \log \|X_n^\delta(W_0 \ldots W_n)^{-1}\| + O(\epsilon)$. 

28
Moreover, when \( k = k^u + 1 \) we have

\[
\chi_{\text{max}} = \lim_{n \to \infty} \frac{1}{n} \log \| (\Pi_{j=0}^{n} W_j)^{-1} \| + O(\epsilon).
\]

**Proof.** Let us consider matrices \( \tilde{D} f_j \) such that for large \( j \), \( \tilde{D} f_j \) is \( O(\epsilon) \) close to \( D f_{x_j} \) and

\[
X^a_n = (\Pi_{j=0}^{n} \tilde{D} f_j) X^a_0 W_0 \ldots W_n.
\]

The shadowing condition \( (iv) \) of §4.3.1 yields

\[
\chi_{\text{max}} = \lim_{n \to \infty} \frac{1}{n} \log \| (\Pi_{j=0}^{n} \tilde{D} f_j) \| + O(\epsilon).
\]

Moreover, since

\[
\Pi_{j=0}^{n} \tilde{D} f_j X^a_0 = X^a_n (W_0 \ldots W_n)^{-1},
\]

and the span of the columns of \( X^a_0 \) contains a \( k^u \) dimensional space inside an unstable cone around \( E_{x_0}^u \),

\[
\lim_{n \to \infty} \frac{1}{n} \log \| \Pi_{j=0}^{n} \tilde{D} f_j \| = \lim_{n \to \infty} \frac{1}{n} \log \| X^a_n (W_0 \ldots W_n)^{-1} \|.
\]

For the second part, let us assume \( k = k^u + 1 \). We observe that the upper and lower bounds on the ensemble spread \( (i) \) and \( (iii) \) of §4.3.1 yield

\[
C \epsilon \| (W_0 \ldots W_n)^{-1} \| \leq \| X^a_n (W_0 \ldots W_n)^{-1} \| \leq \sqrt{k} C \epsilon \| (W_0 \ldots W_n)^{-1} \|.
\]

(The upper bound is straightforward. For the lower bound, we use that for each \( l \), (1) \( W_l \) has an orthonormal set of eigenvectors, and all but one of them have eigenvalue 1. The remaining one is \( Q_l^T \), whose corresponding eigenvalue is smaller, (2) \( (1, \ldots, 1) Q_l^T = 0 \), and (3) Zero is a singular value of \( X^a_l \) of multiplicity one, corresponding to the fact that \( \sum_{j=1}^{k} v^a_{j,l} = 0 \); the other \( k^u \) singular values of \( X^a_l \) are greater than or equal to \( C \epsilon \), for some constant \( C \) independent of \( l \), due to Proposition 6.)

Hence,

\[
\lim_{n \to \infty} \frac{1}{n} \log \| \Pi_{j=0}^{n} \tilde{D} f_j \| = \lim_{n \to \infty} \frac{1}{n} \log \| (W_0 \ldots W_n)^{-1} \|.
\]

Combining, we obtain the result,

\[
\chi_{\text{max}} = \lim_{n \to \infty} \frac{1}{n} \log \| (W_0 \ldots W_n)^{-1} \| + O(\epsilon).
\]

\[ \square \]

**Remark 14.** In §3.6 we gave a simpler approximation to \( \chi_{\text{max}} \) based on the expansion of the ensemble during the forecast steps being essentially a scalar process. The approximation above is based instead on accounting for the contraction of the ensemble during the analysis steps. For a multidimensionally unstable system, the latter approach is computationally simpler, making use of the already computed \( W_j \)'s.

**Acknowledgements**

This research was partially supported by NSF Award DMS0616585, and the first author acknowledges CONACyT (Mexico) for financial support.

URL http://dx.doi.org/10.1016/S1874-575X(06)80027-5


URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0493%281998%29126%3C1719%3AASITEK%3E2.0.CO%3B2


URL http://dx.doi.org/10.1007/978-3-642-03711-5


URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0493%281998%29126%3C0796%3ADAUAEK%3E2.0.CO%3B2

URL http://dx.doi.org/10.1016/j.physd.2006.11.008


URL http://www.sciencedirect.com/science/article/B6TVK-3WWCDFX-4/2/f51f885545aae414b9dfdc931bc0c34

URL http://www.sciencedirect.com/science/article/B6TVK-46X4M29-B/2/bf50634d9fe6315e334610a128df5de4


URL http://dx.doi.org/10.1103/PhysRevLett.64.821


URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0493%282003%29131%3C1485%3AESRF%3E2.0.CO%3B2

URL http://journals.ametsoc.org/doi/abs/10.1175/1520-0493%282004%29132%3C1590%3AWIBAE0%3E2.0.CO%3B2


31