Supplementary Material for "Stability and Memory-loss go hand-in-hand: Three Results in Dynamics & Computation"

G. Manjunath

Department of Mathematics & Applied Mathematics, University of Pretoria, Pretoria 0028

Email: manjunath.gandhi@up.ac.za

This supplementary material is organized as follows. In Section 1, we explain the intuition behind the notion of memory loss or the echo state property with respect to an input. In Section 2, we present essentially the same definitions used in the accompanying main paper [1] along with an additional explanation/figure. In Section 3 and Section 4 we present the same theorems presented in [1, Section 3] with an additional example or explanation that concern some mathematical aspects of the results. In Section 5, we present some additional details of the computations used in the computer simulations in [1, Section 2]. Lastly, in Section 6, we present additional numerical results that complement the results in [1, Section 2] that concern finding the input-specific echo state property threshold.

1 Memory-loss or Echo State Property

determines mass, composition etc

Consider a thought experiment whose set-up comprises a collection of containers, each containing an identical substance in mass and composition determined by a parameter α. Suppose X is an interval of real numbers, and for each value x in X, there is a container with a substance at temperature x. Say all of the containers have full thermal conductivity and are supplied heat simultaneously and equally by a source of heat. Suppose a quantum of heat u that feeds the containers updates a temperature (of the substance) x to $g(\alpha, u, x)$ in one "time-unit" through a function g. If we let heat to be varying in time, i.e., if we let u_k be the heat supplied to each of the containers at the kth time-unit, then the temperature at the $k+1$-th time unit is $x_{k+1} = g(\alpha, u_k, x_k)$. Since heat u_k fluctuates with time k, so would the temperatures of

$x \longmapsto g(\alpha, u, x)$ 1

$x_{k+1} = g(\alpha, u_k, x_k)$

heat varies with time

u = heat packet

Memory is lost when the initial state of a system is quickly forgotten. Conceptually, this can happen in two very different ways. The first is for trajectories to merge, so that in time, they evolve effectively as a single trajectory independent of their points of origin. This happens in systems that are contractive. Consider for example a system defined by the composition of a sequence of maps f_i of a compact metric space X to itself, and assume that all the f_i have a uniform Lipschitz constant $L < 1$, *i.e.*, for all $x, y \in X$, $d(f_i x, f_i y) \leqslant Ld(x, y)$. Since the diameter of the image of X decreases exponentially with time, all trajectories eventually coalesce into an exponentially small blob, which in general continues to evolve with time (except when all the f_i have the same fixed point). A similar phenomenon is known to occur in random dynamical systems. An SDE of the form

$$dx_t = a(x_t)\,dt + \sum_{i=1}^{n} b_i(x_t) \circ dW_t^i \tag{1.1}$$

gives rise to a stochastic flow of diffeomorphisms, in which almost every Brownian path defines a time-dependent flow (see *e.g.* [10]). When all of the Lyapunov exponents are strictly negative, trajectories are known to coalesce into *random sinks* (see [3, 13]). This phenomenon occurs naturally in applications, such as the Navier-Stokes system with sufficiently large viscosity (see *e.g.* [17, 18]), and in certain neural oscillator networks (see *e.g.* [14]).

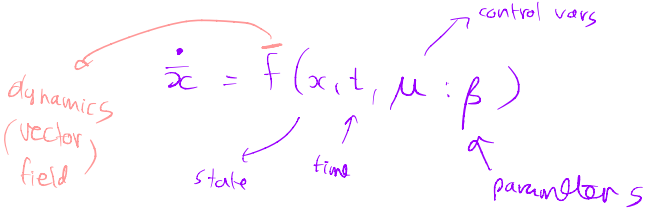
In chaotic systems (autonomous or not), memory is lost quickly not through the coalescing of trajectories but for a diametrically opposite reason, namely their sensitive dependence on initial conditions. Small errors multiply quickly with time, so that in practice it is virtually impossible to track a specific trajectory in a chaotic system. For this reason, a statistical approach is often taken. Let ρ_0 denote an initial probability density with respect to a reference measure m, and suppose its time evolution is given by ρ_t. As with individual trajectories, one may ask if these probability distributions retain memories of their pasts. We will say a system loses its memory in the statistical sense if for two initial distributions ρ_0 and $\hat{\rho}_0$, $\int |\rho_t - \hat{\rho}_t|\, dm \to 0$ as $t \to \infty$. It is this form of memory loss that is studied in the present paper. Of particular interest is when memory is lost quickly: we say a system has *exponential statistical loss of memory* if there is a number $\alpha > 0$ such that for any ρ_0 and $\hat{\rho}_0$, $\int |\rho_t - \hat{\rho}_t|\, dm < Ce^{-\alpha t}$. Such memory loss may happen over a finite time interval, *i.e.*, for $t \leqslant T$, or for all $t \geqslant 0$.

the substance in the containers, and hence there is no single equilibrium temperature that the substance in the containers approach.

Also, say that during this time-varying heating process, the container cools at the interfaces with the surrounding air when its temperature is not within a certain range. Now, suppose this experiment had been set-up "infinitely" long ago, and let the entire infinitely long history of the heat input be represented by $\bar{u} = (\ldots, u_{-1}, u_0, u_1, \ldots)$. Let $X_n(\alpha, \bar{u})$ be the collection or set of all temperature values of the substance in all the containers at time n. If $X_n(\alpha, \bar{u})$ is a single-value that is the temperature of the substance in all the containers evolve in unison, then there is memory-loss of all "initial" temperatures of the substances in the containers. Such forgetting may be intuitively attributed to the fact that the substance cannot be excited by the heat to raise its temperature beyond a certain value due to the dissipation of heat – this dissipation of heat out of a container may have reset its temperature and effectively ~~have~~ had it forget its initial temperature. On the other hand, if the set $X_n(\alpha, \bar{u})$ would contain more than one value, then there is at least one container that contains a substance that has not forgotten its initial temperature. In the case where $X_n(\alpha, \bar{u})$ has a single value, where substances change temperatures in unison, we expect the temporal input heat to be the sole factor in dictating the future evolution of the set $X_n(\alpha, \bar{u})$. In contrast, in the case $X_n(\alpha, \bar{u})$ has multiple values, the input heat pattern does not fully guide the evolution of temperatures since the substance in at least one of the containers is retaining the effect of its initial temperature. This intuitive idea of memory-loss or the lack of it can be extended to the dynamical systems theory, where, in addition, the input can be both amplified and attenuated. In such a setting, the consequences of such memory-loss are exciting that is demonstrated mathematically in this article. In particular, the evolution of the set $X_n(\alpha, \bar{u})$ as a function of $\bar{u}$ with and without memory-loss is surprisingly contrasting.

In dynamical systems theory, memory-loss appears in the context of two diametrically opposite reasons. When systems have sensitive dependence on initial conditions like in chaotic systems, small errors multiply within a short time gap, so that it is unfeasible to track a specific trajectory (e.g., [2, 3]). On the other hand, the diameter of the state space could asymptotically decrease to zero so that the trajectories tend to coalesce into a single trajectory. In this paper, we consider the latter case of memory-loss in which a dynamical system is driven exogenously.

A feature of memory-loss in driven systems is that a single bounded trajectory emerges to be a representative of the drive/input. To illustrate this idea, we consider a continuous-time driven system, in particular a scalar differential equation, $\dot{x}(t) = u(t) - x(t)$, where $u(t) = 2\cos t$. The solution of the differential equation with an initial condition $x(t_0) = x_0$ can be shown to be $x(t, t_0, x_0) = (x_0 - \sin t_0 - \cos t_0)e^{-(t-t_0)} + \rho(t)$, where $\rho(t) = \sin t + \cos t$. For a given t, $\lim_{t_0 \to -\infty} x(t, t_0, x_0) = \rho(t)$ regardless of x_0. Hence, in this case, or in cases similar to this (for a schematic



previously: derive from 1st principle

today: complex, lots of data.

Brain / Neuroscience

Degrees of freedom.

Linear: completely characterised.

Nonlinear: unknown f's
not even governing

challenges:

- High dimensional
- Chaos, transients
- Noise, stochasticity
- Multiscale (fractile-type)
- Uncertainty.

Using models

1. Future state prediction
2. Design & optimisation
3. Control // suppressing seizures.
4. interpretability + physical intuition

Techniques

- Regression
  - Sparse regression
- Neural nets
- Deep learning
- Genetic programs

see Fig. S1), for a fixed initial choice x_0, setting the starting time t_0 further back in time would get $x(t, t_0, x_0)$ closer to $\rho(t)$ for a given t. It follows that different solutions asymptotically forget or lose the memory of their initial condition, and a single bounded attractive solution $\rho(t)$ emerges as a "proxy" for the drive $u(t)$ (in the state space). This is the essence of memory-loss w.r.t. to the input $u(t) = 2\cos(t)$. In general, such memory-loss is not observed in all driven systems, and neither is it observed generically. Moreover, such a memory-loss phenomenon usually depends on the exogenous input $u(t)$ and also the equations governing the system. In cases, where there is no memory-loss w.r.t. to input, not a single trajectory like $\rho(t)$, but a set-valued function comprising multiple or a bunch of trajectories (solutions) would attract nearby solutions.

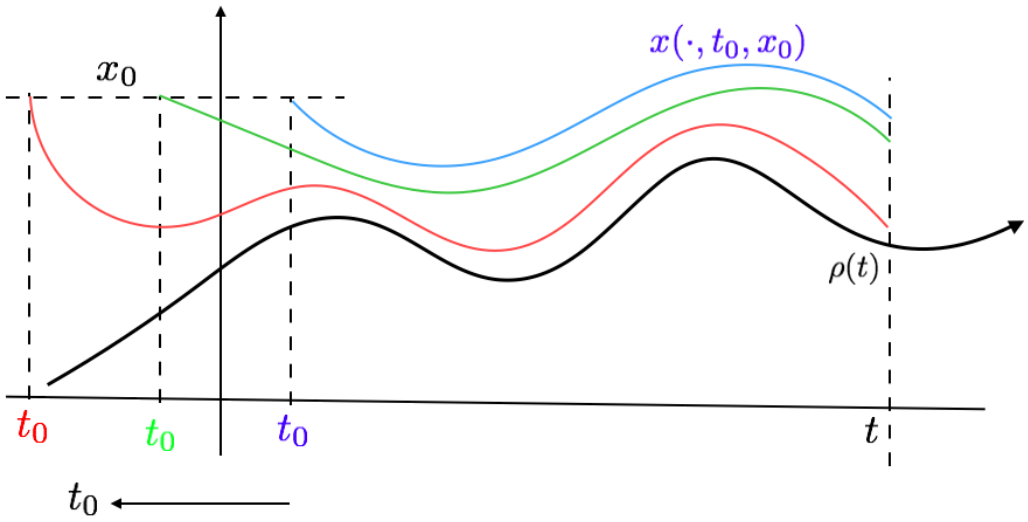


Figure S1: Schematic figure to explain memory-loss in continuous-time.

The idea of relating memory-loss to the presence of a single proxy corresponding to an input can be extended to the case of a discrete-time dynamical driven system (representative work is in [4]). The most prevalent exponent of memory-loss was in the field of reservoir computing (RC) (e.g., [5, 6, 7]). In most of the reservoir computing (RC) literature, the idea of memory-loss or fading memory popularly called the echo state property (ESP) is treated as a stability property of the system valid for all inputs. In such a case, the ESP guarantees that the entire past history of the input determines the state of the system precisely, i.e., there is no possibility of two or more reachable states at any given point in time [5] if the entire past values of the input have influenced the dynamics. The formal definition linking ESP intrinsically to an input is available in [4]. Here, we extend the results made in [4] and consider a general setup to analyze the effect of both an input and a design parameter. A design

parameter (not necessarily a scalar and includes the case of a vector or a matrix) α, an input value u, a state-variable x in a compact state space X and a continuous function $g(\alpha, u, x)$ that takes values in X would constitute a parametric driven system (we denote this system by g throughout, and other entities tacitly understood; for a formal definition see Section 2). The dynamics of a parametric driven system g is rendered through a temporal input $\bar{u} = (\ldots, u_{-1}, u_0, u_1, \ldots)$ via the equation $x_{n+1} = g(\alpha, u_n, x_n)$. Given an $\bar{u}$ and α, any sequence $\bar{x} = (\ldots, x_{-1}, x_0, x_1, \ldots)$ in the state space that satisfies the equation $x_{n+1} = g(\alpha, u_n, x_n)$ *for all* integers n is an "image" of $\bar{u}$ in the space X (and also called a solution of g). Having fixed α and $\bar{u}$ suppose we collect the n^{th} component of all images $\bar{x}$, and denote it by $X_n(\alpha, \bar{u})$, then we have a collection of sets $\{X_n(\alpha, \bar{u})\}$ which we call the representation of $\bar{u}$. It turns out the components of this representation satisfy $X_{n+1}(\alpha, \bar{u}) = g(\alpha, u_n, X_n(\alpha, \bar{u}))$ (see [4, Proposition 2] and [8] for details). Following [4] , we say that the driven system g has echo-state property (ESP) with respect to (w.r.t.) an input $\bar{u}$ at α if it has exactly one solution or one image $\bar{x}$ in the space X or (equivalently) if for each n, $X_n(\alpha, \bar{u})$ is single-valued, i.e., a singleton subset of X. The emergence of a single image or a proxy, in general, does not necessarily mean it captures the "information" content of the input. For instance, consider the discrete-time system, $g(\alpha, u, x) = ux$, where $U = [0, 1)$ and $X = [0, 1]$ with $g(\alpha, u, x) = ux$. It may be verified that for every input $\bar{u}$, the map $g(\alpha, u_k, \cdot)$ is a contraction on X for each $k \in \mathbb{Z}$, and it easily follows that there is exactly one solution in X (by applying Lemma 2 in [4]). However, every solution $\{x_n\}$ in X converges monotonically to 0 as $n \to \infty$. The aforementioned convergence means that any image of $\bar{u}$ would not reflect upon the temporal complexity of $\bar{u}$, and nor would there be any good separability of the representation of the inputs in the space X. Such a driven system is not useful for applications. On that account, in RC, one employs driven dynamical systems with a higher dimensional state space.

2 Preliminaries

A parametric driven system would comprise a parameter space Λ, an input metric space U and a compact metric space X and a continuous map $g : \Lambda \times U \times X \to X$. For brevity, we would often refer g to be the driven system with the other components silently understood. The reader may note henceforth, all additional hypotheses on a parametric driven system we assume are satisfied by the recurrent neural network $g(\alpha, u, x) = \overline{\tanh}(Au + \alpha Bx)$ cited frequently in [1].Throughout, $A \subset B$ stands for A is a subset of B and does not preclude the case that $A = B$. In our analysis, wherever a cartesian product of spaces is considered, the cartesian products are endowed with the most commonly used topology on it called the product topology (e.g., [9]). This implies that a sequence converges in an infinite cartesian product space if and only if the individual coordinate projections of the sequence also converges.

This section aims to create a set-valued function (called the encoding map) that represents all "attainable" states of the system if a left-infinite input was fed to the system at some scaling parameter, and then to describe the continuity notions of set-valued functions. In what is to follow, we describe the convergence of a sequence of sets through the Hausdorff metric, which we recall next.

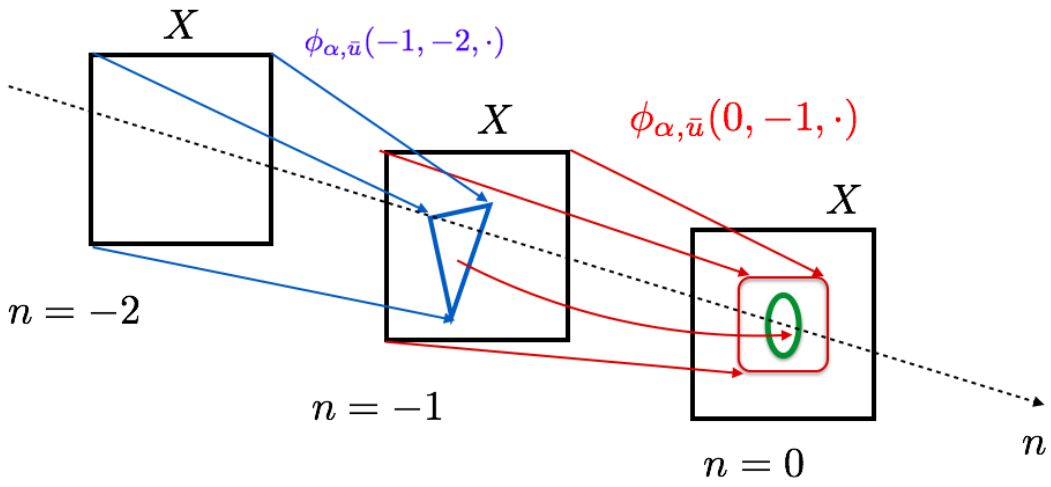


Figure S2: Schematic figure to explain the images of X under the map $\phi_{0,\bar{u}}$. The square X at time -2 is transformed to a triangle at time -1 and then into an oval at time 0, while the square at X at time -1 is transformed into another smaller square at time 0.

Hausdorff distance or Hausdorff metric: When X is a metric space with metric d, we denote by H_X the collection of all nonempty closed subsets of X. On this space we employ the Hausdorff metric defined by $d_H(A,B) := \max(dist(A,B), dist(B,A)) := \inf\{\epsilon : A \subset B_\epsilon(B)\ \&\ B \subset B_\epsilon(A)\}$, where $B_\epsilon(A) := \{x \in X : d(x,A) < \epsilon\}$ is the open ϵ-neighborhood of A. It is well known that whenever X is a compact metric space, H_X is also a compact metric space.

For the analysis to follow, we would adopt a composition-operator (called a process in [4, 10]). Given a parametric driven system g, we denote $g_{\alpha,u}(x) := g(\alpha,u,x)$ and $\{u_n\} \subset U$ by $\bar{u}$. Suppose a parametric driven system g has been fed input values $u_m, u_{m+1}, \ldots, u_{n-1}$ starting at time m. Then the map g transports a state-value $x \in X$ at time m to give a state-value $g_{\alpha,u_{n-1}} \circ \cdots \circ g_{\alpha,u_m}(x)$ at time n. Formally, for every choice of α, $\bar{u}$ and g, we define for all $m \leq n$, the function that 'transports' a system state at x at time m through the inputs $u_m, u_{m+1}, \ldots u_{n-1}$ to the state at time n given by an composition-operator $\phi_{\alpha,\bar{u}} : \mathbb{Z}^2_{\geq} \times X \to X$, where

$\mathbb{Z}^2_{\geq} := \{(n,m) : n \geq m, n, m \in \mathbb{Z}\}$ and

$$\phi_{\alpha,\bar{u}}(n,m,x) := \begin{cases} x & \text{if } n = m, \\ g_{\alpha,u_{n-1}} \circ \cdots \circ g_{\alpha,u_m}(x) & \text{if } m < n. \end{cases}$$

Thus, if the inputs $u_m, u_{m+1}, \ldots, u_{n-1}$ are fed in that order to every $x \in X$, the system would have evolved at time n to one of the states in $\phi_{\alpha,\bar{u}}(n,m,X)$. For a choice $(n,m) = (0,-1)$ and $(n,m) = (-1,-2)$, the action of a map $\phi_{\alpha,\bar{u}}$ is illustrated schematically in Fig. S2. A simple observation is that the set inclusion $\phi_{\alpha,\bar{u}}(m+2,m,X) \subset \phi_{\alpha,\bar{u}}(m+1,m,X)$ holds for all $m \in \mathbb{Z}$ since $g_{\alpha,u_{m+1}} \circ g_{\alpha,u_m}(X) \subset g_{\alpha,u_m}(X)$. Based on this observation, it follows that if the entire left-infinite input $\{u_m\}_{m<n}$ would have been fed to every $x \in X$, then $\phi_{\alpha,\bar{u}}(n,m,X)$ is a decreasing sequence of sets, i.e., $\phi_{\alpha,\bar{u}}(n,m-1,X) \subset \phi_{\alpha,\bar{u}}(n,m,X)$ for an m. Hence, if the entire left-infinite input $\{u_m\}_{m<n}$ had influenced the dynamics of the parametric driven system g, the system would have evolved at time n to an intersection of a decreasing sequence of sets:

$$X_n(\alpha,\bar{u}) := \bigcap_{m<n} \phi_{\alpha,\bar{u}}(n,m,X). \tag{1}$$

When X is compact, each set $\phi_{\alpha,\bar{u}}(n,m,X)$ is a closed subset of X and hence $X_n(\alpha,\bar{u})$ is a nonempty closed subset of X (a proof is available in [4, Lemma 2.1]). Note that $X_n(\alpha,\bar{u})$ is an element of the space H_X. Since H_X is a compact metric space, every sequence of sets has a convergent subsequence. Whenever a sequence $\{A_n\}$ converges in H_X, it means that $d_H(A_n,A) \to 0$ and we denote it by $d_H\text{-}\lim_{n\to\infty} A_n = A$. It may be verified that (see for e.g., [11]) when $\{A_n\}$ is a sequence of decreasing sets, i.e., $A_{n+1} \subset A_n$, then $A = \bigcap_{n=1}^{\infty} A_n$ if and only if $d_H\text{-}\lim_{n\to\infty} A_n = A$, i.e., the limit in the compact space H_X always exists for a decreasing sequence of sets and is equal to the nested intersection of the sequence of sets. Hence an alternate definition of (1) would be

$$X_n(\alpha,\bar{u}) := d_H\text{-}\lim_{m\to\infty} \phi_{\alpha,\bar{u}}(n,-m,X). \tag{2}$$

When the input originates from an invertible dynamical system, one could treat the function $\phi_{\alpha,\bar{u}}(n,-m,x)$ to be just dependent on a single value u in U since u then would have determined its unique past. In such a special case the notion of the composition-operator above coincides with that of the echo state mappings defined in [12].

Given an input $\bar{u}$, a solution (called an entire-solution in [4, 10]) of g is a bi-infinite sequence $\{x_n\}$ that satisfies $x_{n+1} = g(\alpha,u_n,x_n)$ for all $n \in \mathbb{Z}$. One can easily show that "a sequence $\{x_n\}$ is a solution of g obtained through an input $\bar{u}$ if and only if x_k belongs to $X_k(\alpha,\bar{u})$ for all $k \in \mathbb{Z}$" (see [4, Lemma 2.1] for a proof). We remark that $X_n(\alpha,\bar{u})$ was defined initially through a solution in [1] and it is of no consequence since the definitions are equivalent. We say that a parametric driven system g has the ESP w.r.t. to $\bar{u}$ at the parameter α if there is exactly one single solution of g

(when the input is $\bar{u}$, and the scaling parameter is α). We leave out the phrase "at the parameter α" whenever it is obvious. Putting all these together, it follows that a parametric driven system g has the ESP w.r.t. to its input $\{u_n\}$ at the parameter α if and only if $X_n(\alpha, \bar{u})$ is a singleton subset of X for each n.

We say a parametric driven system g to be an *open-parametric* driven system if X is connected and the mapping $g(u, \alpha, \cdot)$ does not map any uncountable set contained in X to a single value in X. Given a driven system g, a parameter α and input $\bar{u}$, if X is connected, then it follows that $X_n(\alpha, \bar{u})$ is connected for all n since $X_n(\alpha, \bar{u})$ is a nested intersection of connected sets. A nonempty connected set is either a singleton or uncountably infinite subset of X, when X is a metric space. Suppose $X_{n-1}(\alpha, \bar{u})$ is uncountable then $X_n(\alpha, \bar{u})$ cannot be a singleton subset of X if g were to be an open parametric system since $X_n(\alpha, \bar{u}) = g(\alpha, u_n, X_{n-1}(\alpha, \bar{u}))$. It follows that when $X_n(\alpha, \bar{u})$ is a singleton subset of X, $X_{n+k}(\alpha, \bar{u})$ for any integer k is forced to be a singleton subset of X as well when g is a open-parametric driven system. Thus, an open-parametric driven system g has the ESP w.r.t. $\overleftarrow{u}$ at α if and only if $\mathcal{E}(\alpha, \overleftarrow{u})$ is a singleton subset of X.

In particular, we know that an open-parametric driven system g has the ESP w.r.t. to an input $\bar{u}$ if and only if $\mathcal{E}(\alpha, \overleftarrow{u}) := X_0(\alpha, \bar{u})$ is a singleton subset of X. Note that $\mathcal{E}(\alpha, \overleftarrow{u}) = d_H\text{-}\lim_{n\to\infty} \phi_{\alpha,\bar{u}}(0, -n, X)$. Next, we define the encoding family to be $\{\mathcal{E}_n(\alpha, \overleftarrow{u})\}_{n\geq 1} \subset \mathsf{H}_X$, where $\mathcal{E}_n(\alpha, \overleftarrow{u}) := \phi_{\alpha,\bar{u}}(0, -n, X)$. Hence, we also have $\mathcal{E}(\alpha, \overleftarrow{u}) = d_H\text{-}\lim_{n\to\infty} \mathcal{E}_n(\alpha, \overleftarrow{u})$.

An example of an open-parametric driven system is a recurrent neural network(RNN) of the form $g(\alpha, u, x) = \overline{\tanh}(Au + \alpha Bx)$, where $\overline{\tanh}(*)$ is (the nonlinear activation) tanh performed component-wise on $*$, α, a real-valued parameter would correspond to the scaling of the reservoir (invertible) matrix B of dimension N and A is a matrix with input connections and $X = [-1, 1]^N$ (the cartesian product of N copies of $[-1, 1]$). We observe that whenever $\alpha \neq 0$ and B is invertible, $g(\alpha, u, \cdot)$ is always invertible since $\overline{\tanh}$ is invertible and hence g is an open-parametric driven system when $\alpha = 0$ is not in the parameter space Λ.

Given a parametric driven system g we define the following set-valued maps associated with the encoding map:

1. For every given $\alpha \in \Lambda$, the map $\mathcal{E}(\alpha, \cdot) : \overleftarrow{u} \mapsto \mathcal{E}(\alpha, \overleftarrow{u})$ is said to be an *input-encoding map.*
2. For every left-infinite input $\overleftarrow{u}$, the map $\mathcal{E}(\cdot, \overleftarrow{u}) : \alpha \mapsto \mathcal{E}(\alpha, \overleftarrow{u})$ is said to be a *parameter-encoding map.*
3. For every given $\alpha \in \Lambda$, the collection of mappings $\{\mathcal{E}_n(\alpha, \cdot)\}_{n\geq 1}$ where $\mathcal{E}_n(\alpha, \star) :=$

$\phi_{\alpha,\bar{\star}}(0,-n,X)$ is called the *input-encoding family*. Similarly, for a given $\bar{u}$, the collection of mappings $\{\mathcal{E}_n(\cdot,\bar{u})\}_{n\geq 1}$ where $\mathcal{E}_n(\star,\bar{u}) := \phi_{\star,\bar{u}}(0,-n,X)$ is called the *parameter-encoding family*.

For set-valued maps, different notions of continuity exist. These notions are schematically illustrated in Fig. S3. Intuitively, the notion of upper semicontinuity permits sudden explosions (Fig. S3 (b) & (e)) while the notion of lower semicontinuity permits sudden implosions (Fig. S3 (c) & (f)), and the notion of continuity does not permit both explosions and implosions. Formally, let X be a topological space and Y_X be a subspace of the power set of X, and let Z be another topological space. A map $f : Z \to Y_X$ is said to be upper semicontinuous at z if for every open set V in X containing $f(z)$ there exists an open neighborhood U containing z such that $f(U) \subset V$. A map $f : Z \to Y_X$ is said to be lower semicontinuous at z if for every open set V in X such that $f(z) \cap V \neq \emptyset$, then $\{z : f(z) \cap V \neq \emptyset\}$ is an open neighborhood of x. A map f that is upper-semicontinuous for all $z \in Z$, then it is called upper semicontinuous (u.s.c). A map f that is lower-semicontinuous for all $z \in Z$, then it is called lower semicontinuous (l.s.c). A map is upper semicontinuous only (usc-only) if it is u.s.c and not l.s.c at some point $z \in Z$. A map is lower semicontinuous only (lsc-only) if it is l.s.c and u.s.c at some point $z \in Z$. A map that is both upper semicontinuous and lower semicontinuous at a point $z \in Z$ is said to be continuous at z. We use the fact that if a map is u.s.c at z and is single-valued, i.e., $f(z)$ is single-valued, i.e., a singleton subset of X, then f is continuous at z.

In addition to the cases illustrated in Fig. S3, the "graphs" of set-valued maps could behave wildly with explosions. For instance, the input-encoding map for a class of driven systems is not l.s.c when there is no ESP Theorem 2. It is not possible to explicitly describe such maps where the arguments are infinite-dimensional.

The continuity of the function $(\alpha,\bar{u}) \mapsto X_n(\alpha,\bar{u})$ can be also be defined by considering the topology induced on H_X by the Hausdorff metric d_H. Such continuity defined using the Hausdorff metric is equivalent [11] to the continuity of $X_n(\alpha,\bar{u})$ when $X_n(\alpha,\bar{u})$ is treated as a set-valued function of the variable $(\alpha,\bar{u})$. We make use of this equivalence without further remarks.

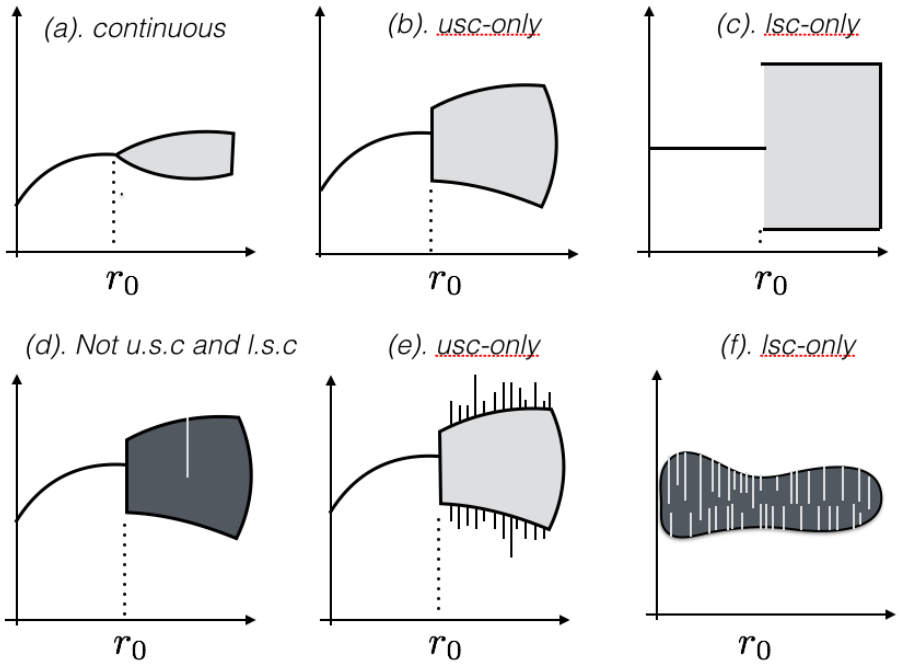


Figure S3: Schematic of graphs of set-valued maps to explain the notions of upper semicontinuity (u.s.c) and lower semicontinuity (l.s.c). The lighter grey values indicate the interior, while the darker values indicate the boundaries of the set-valued function's graph, and the comb-like outward projections refer to the explosions due to the absence of l.s.c; $S(r_0)$ is a singleton subset in (a) and (c), while it is multi-valued (subset of X with multiple elements) in all other figures. For an open set V containing $S(r_0)$ there is a neighborhood of S_0 that maps into V in (b) but not in (c). Also note that there is an explosion of S at r_0 in (b) while there is an implosion in (c). S is upper semicontinuous only (usc-only) if it is u.s.c and not l.s.c at some point $z \in Z$. S is lower semicontinuous only (lsc-only) if it is l.s.c and u.s.c at some point $z \in Z$.

3 Continuity of the Input-Encoding Map and the ESP

Given a parametric dynamical system g, for every given $\alpha \in \Lambda$, we call the set-valued mapping $\bar{u} \mapsto \{X_n(\alpha, \bar{u})\}_{n\in\mathbb{Z}}$ to be the *input-representation map*. The main purpose of the analysis in this section is to establish the fundamental connection between the echo state property and continuity of the input-representation map in Theorem 1.

First, we adopt a terminology to ensure the driven system has an input that causes a contraction in a subdomain of X. Given an open-parametric driven system g and an α, if there exists a $\overleftarrow{v}$ such that g has the ESP w.r.t. $\overleftarrow{v}$ at α, then we say g is

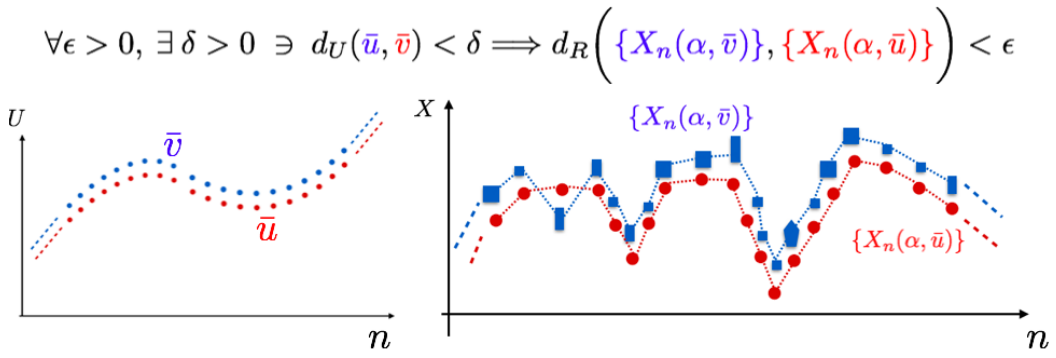


Figure S4: Schematic to show slightly different inputs $\bar{u}$ and $\bar{v}$ in the input space could lead to mostly similar responses $\{X_n(\alpha, \bar{u})\}$ and $\{X_n(\alpha, \bar{v})\}$ in the state space that are measured by a metric d_R (see Remark 1).

contractible at α. For example, the RNNs that we consider are always contractible except in the case of the input weight matrix A having a row with all zeroes. It is a result in [4, (ii) of Theorem 4.1] that for an RNN with an input-weight matrix A having non-zero rows and any norm of the reservoir matrix B, there always exists an input $\overline{v}$ so that it has the ESP w.r.t. $\overline{v}$, and hence such systems are contractible. The idea behind contractibility is that if the input $\overline{v}$ is such that if the entity $(Av_n + \alpha Bx_n)$ is thrown into the saturation region of the activation function σ sufficiently often, one can always witness ESP w.r.t. $\overline{v}$. Given any $\overline{u}$, an $\overline{v}$ is readily obtained by scaling $\overline{u}$ to have a sufficiently large amplitude.

Theorem 1 establishes that for an open-parametric driven system that is contractible, slightly different inputs lead to mostly similar responses if and only if the ESP is satisfied (see Fig. S4).

Theorem 1. (Continuity of the input-representation map and the input-encoding map) *An open-parametric driven system g that is contractible at α has the ESP w.r.t. $\bar{u}$ at $\alpha \iff$ the input-representation map $\bar{u} \mapsto \{X_n(\alpha, \bar{u})\}$ is continuous at $\bar{u} \iff$ the input-encoding map $\bar{u} \mapsto \mathcal{E}(\alpha, \overline{u})$ is continuous at $\bar{u}$.*

The proof of Theorem 1 would be presented after establishing the relationship of the ESP w.r.t. $\bar{u}$ and the input-encoding map as stated in Lemma 1. The pedantic reader may note that Theorem 1 would not be true without the contractiblility condition.

For example, if $X = [0,1]$ and $g(\alpha, u, x) = x$, for all n, $X_n(\alpha, \bar{u}) = X$ regardless of the input $\bar{u}$ and hence the input-representation map $\bar{u} \mapsto \{X_n(\alpha, \bar{u})\}$ is continuous (i.e., both usc and lsc), but there is no ESP to all inputs since $X_n(\alpha, \bar{u})$ is not a singleton subset of X for every n.

Lemma 1. (Continuity of the input-encoding map) *An open-parametric-driven system g that is contractible at α has the ESP w.r.t. $\overline{u}$ at α if and only if the input-encoding map $\mathcal{E}(\alpha, \cdot)$ is continuous at $\overline{u}$.*

Proof. ($\Longrightarrow$) The open-parametric driven system g has the ESP w.r.t. $\overline{u}$ at α is equivalent to saying that $\mathcal{E}(\alpha, \overline{u})$ is a singleton subset of X. By (ii) of Lemma 2, $\mathcal{E}(\alpha, \cdot)$ is u.s.c at $\overline{u}$. We use the fact that a map that is u.s.c and is a singleton subset of the space X is continuous to conclude $\mathcal{E}(\alpha, \cdot)$ is a continuous at $\overline{u}$.

($\Longleftarrow$) Let $\mathcal{E}(\alpha, \cdot)$ be continuous at $\overline{u}$ and suppose that g does not have the ESP w.r.t. $\overline{u}$ at α. This means $\mathcal{E}(\alpha, \overline{u})$ contains a subset of X that has at least two elements of X. Hence the encoding map evaluated at $(\alpha, \overline{u})$ is bounded away from all singleton subsets of X, i.e., $r := \inf_{\{x\} \in X} d_H(\mathcal{E}(\alpha, \overline{u}), \{x\}) > 0$.

Since g is contractible at α, let $\overline{v}$ be an input for which $\mathcal{E}(\alpha, \overline{v})$ is a singleton subset of X. We define the sequence of left-infinite sequences $\{\overline{w}_n\}_{n>1}$ by $\overline{w}_n := (\ldots, v_{-2}, v_{-1}, u_{-n}, \ldots, u_{-2}, u_{-1})$. Since, the first n elements of $\overline{w}_n$ are identical to that of $\overline{u}$, it follows that the sequence $\overline{w}_n \to \overline{u}$ in the product topology on $U^{(-\infty,-1)}$, where $U^{(-\infty,-1)}$ is the infinite cartesian product space $\prod_{i=-1}^{\infty} U_i$ with $U_i = U$.

Since $\mathcal{E}(\alpha, \overline{v})$ is a singleton subset of X, we find that $\mathcal{E}(\alpha, \overline{w}_n)$ is a singleton subset of X since

$$\mathcal{E}(\alpha, \overline{w}_n) = \bigcap_{j>0} \phi_{\alpha, *_n}(0, -j, X) = \bigcap_{j>0} \phi_{\alpha, \dagger}(0, -j, X) = \mathcal{E}(\alpha, \overline{v}),$$

where $*_n = \overline{w}_n$ and $\dagger = \overline{v}$ (notations to avoid obscurity in the subscripts). Since $\mathcal{E}(\alpha, \overline{w}_n)$ is a singleton subset of X, it follows that $d_H(\mathcal{E}(\alpha, \overline{u}), \mathcal{E}(\alpha, \overline{w}_n)) \geq r > 0$ for all n. This implies $\mathcal{E}_n(\alpha, \overline{w}_n)$ does not converge to $\mathcal{E}(\alpha, \overline{u})$ although we had $\overline{w}_n \to \overline{u}$. This contradicts our assumption that $\mathcal{E}(\alpha, \overline{u})$ is continuous at $\overline{u}$. ■.

Corollary 1. *If an open-parametric-driven system g that is not necessarily contractible at α has the ESP w.r.t. $\overline{u}$ at α then the input-encoding map $\mathcal{E}(\alpha, \cdot)$ is continuous at $\overline{u}$.*

Proof of Corollary 1. The first paragraph of the proof of Lemma 1.

Proof of Theorem 1. From Lemma 1, we have: an open-parametric driven system g that is contractible at α has the ESP w.r.t. $\bar{u}$ at α if and only if the

input-encoding map $\bar{u} \mapsto \mathcal{E}(\alpha, \overleftarrow{u})$ is continuous at $\bar{u}$. It remains to be proven that the input-encoding map $\bar{u} \mapsto \mathcal{E}(\alpha, \overleftarrow{u})$ is continuous at $\bar{u}$ if and only if the input-representation map $\bar{u} \mapsto \{X_n(\alpha, \bar{u})\}$ is continuous at $\bar{u}$. Equip the representation space $R := \prod_{i=-\infty}^{\infty} Y_i$, where $Y_i = \mathsf{H}_X$ for all $i \in \mathbb{Z}$ with the product topology. Let $\overleftarrow{v}_n := (\ldots, u_{n-2}, u_{n-1})$. Clearly, $X_n(\alpha, \bar{u}) = \mathcal{E}(\alpha, \overleftarrow{v}_n)$ and the mapping $\bar{u} \mapsto X_n(\alpha, \bar{u})$ is identical to $\overleftarrow{v}_n \to \mathcal{E}(\alpha, \overleftarrow{v}_n)$. The following equivalent statements (i)–(vi) lead to the proof: (i) $\bar{u} \mapsto \{X_n(\alpha, \bar{u})\}$ is continuous if and only if (iff) the coordinate mappings $\bar{u} \mapsto X_n(\alpha, \bar{u})$ is continuous for all $n \in \mathbb{Z}$ (by definition of convergence in product topology) (ii). $\bar{u} \mapsto X_n(\alpha, \bar{u})$ is continuous for all $n \in \mathbb{Z}$ iff $\overleftarrow{v}_n \mapsto \mathcal{E}(\alpha, \overleftarrow{v}_n)$ is continuous for all $n \in \mathbb{Z}$ (by definition of $\overleftarrow{v}_n$). (iii). $\overleftarrow{v}_n \mapsto \mathcal{E}(\alpha, \overleftarrow{v}_n)$ is continuous for all $n \in \mathbb{Z}$ iff g has the ESP w.r.t. $\overleftarrow{v}_n$ for all $n \in \mathbb{Z}$ (by Lemma 1). (iv). g has the ESP w.r.t. $\overleftarrow{v}_n$ for all $n \in \mathbb{Z}$ iff g has the ESP w.r.t. $\overleftarrow{v}_0$ (since g is open-parametric driven system) (v). g has the ESP w.r.t. $\overleftarrow{v}_0$ iff g has the ESP w.r.t. $\overleftarrow{u}$ (by definition of $\overleftarrow{v}_n$). (vi.) g has the ESP w.r.t. $\overleftarrow{u}$ iff the encoding map $\overleftarrow{u} \mapsto \mathcal{E}(\alpha, \overleftarrow{u})$ is continuous (by Lemma 1). ■

Remark 1. We remark that the topology on representation space R can be obtained through a metric d_R [9]. An example is $d_R(A, B) := \sum_{i=-\infty}^{\infty} d_H(A, B)/2^{|i|}$. An analogously defined metric d_U can be used as a metric for the input sequence space $U^{(-\infty,+\infty)}$.

Lemma 2. *Let g be a driven dynamical system. Then*

(i). The parameter-encoding map $\mathcal{E}(\cdot, \overleftarrow{u})$ is u.s.c for a given input $\overleftarrow{u}$.

(ii). The input-encoding map $\mathcal{E}(\alpha, \cdot)$ is u.s.c for a given α.

(iii). Every function in the input-encoding family $\{\mathcal{E}_n(\cdot, \overleftarrow{u})\}$ or in the parameter-encoding family $\{\mathcal{E}_n(\cdot, \overleftarrow{u})\}$ is continuous.

Proof. Fix a left-infinite $\overleftarrow{u}$, and fix a choice of bi-infinite input $\bar{u}$ whose elements $\{\ldots, u_{-2}, u_{-1}\}$ is identical to $\overleftarrow{u}$. Note that by definition, the encoding map $\mathcal{E}(\alpha, \overleftarrow{u}) := X_0(\alpha, \bar{u})$. Given an $(\alpha, \bar{u})$, let $X_0(\alpha, \bar{u}) \subset V$ where V is open in X. We know $X_0(\alpha, \bar{u}) = d_H\text{-}\lim_{j\to\infty} \phi_{\alpha,\bar{u}}(0, -j, X)$. Hence by the definition of the limit, there exists an n such that $\phi_{\alpha,\bar{u}}(0, -n, X) \subset V$. Fix such an n. Since by letting $n \to \infty$ in $\phi_{\alpha,\bar{u}}(0, -n, X)$ we get a decreasing sequence in X, to prove (i), it is sufficient to show that there exists a neigbhorhood W_1 of α such that $\phi_{r,\bar{u}}(0, -n, X) \subset V$ for all $r \in W_1$. In a similar vein, to prove (ii) it is sufficient to show that there exists a neighborhood W of $\bar{u}$ so that $\phi_{\alpha,\bar{v}}(0, -n, X) \subset V$ for all $\bar{v} \in W$.

Proof of (i). We will verify the definition of u.s.c of the parameter-encoding map at each α. The set inclusion $\phi_{\alpha,\bar{u}}(0, -n, X) \subset V$ means that $\phi_{\alpha,\bar{u}}(0, -n, x) \in V$ for all $x \in X$. Denote $U^{(-n,-1)}$ to be the product space $\prod_{i=-1}^{-n} Z_i$, where Z_i is the input space

U. Since $\phi_{\alpha,\bar{u}}(0,-n,x)$ actually depends only on $u_{(-1,-n)} := \{u_{-1}, u_{-2}, \ldots, u_{-n}\}$ and g depends on u continuously, and the composition of continuous functions is continuous, it follows that there is a continuous function

$$h : \Lambda \times U^{(-n,-1)} \times X \to X \text{ so that } h(\alpha, u_{(-1,-n)}, x) = \phi_{\alpha,\bar{u}}(0,-n,x) \tag{3}$$

for all $\alpha \in \Lambda$, $u_{(-1,-n)} \in U^{(-n,-1)}$ and $x \in X$. Since h is continuous and X is compact it follows that there exists a basis element $W_1 \times W_2$ of the product space $\Lambda \times U^{(-n,-1)}$ containing $(\alpha, u_{(-1,-n)})$ so that $h(W_1, W_2, X) \subset V$ which implies $h(r, u_{(-1,-n)}, X) \subset V$ for $r \in W_1$ or equivalently, $\phi_{r,\bar{u}}(0,-n,X) \subset V$ for all $r \in W_1$.

Proof of (ii). Let $h : \Lambda \times U^{(-n,1)} \times X \to X$ be as in (3), and W_1 and W_2 be the basis element as in the Proof of (i). Hence $h(W_1, W_2, X) \subset V$ implies that $h(\alpha, v_{(-1,-n)}, X) \subset V$ for $v_{(-1,-n)} \in W_2$. Define a basis element W in the space $U^{(-\infty,+\infty)} := \prod_{i=-\infty}^{\infty} Z_i$, where Z_i for $i = -1, -2, \ldots, -n$ satisfies $\prod_{i=-1}^{-n} Z_i = W$ and $Z_i = U$ for all other $i \in \mathbb{Z}$. Clearly by definition of a basis element in the infinite product space $U^{(-\infty,+\infty)}$, W is a neighborhood of $\bar{u}$ in the space $U^{(-\infty,+\infty)}$. Now $h(\alpha, v_{(-1,-n)}, X) \subset V$ for $v_{(-1,-n)} \in W_2$ implies $\phi_{\alpha,\bar{v}}(0,-n,X) \subset V$ for all $\bar{v} \in W$ where $\bar{v}$ is such that $v_{(-1,-n)} \in W_2$.

Proof of (iii). The statements follow the fact that for a fixed n, the function h is continuous in all its arguments. ■

4 The Parameter-Encoding map and the ESP

Theorem 2. (Continuity of the parameter-encoding and the parameter-representation map) *Fix an $\bar{u}$ and hence $\overleftarrow{u}$. If an open-parametric driven system g has the ESP w.r.t. $\overleftarrow{u}$ at α then the parameter-encoding map $\mathcal{E}(\cdot, \overleftarrow{u})$ is continuous at α. More generally, if g has the ESP w.r.t. $\bar{u}$ at α then the parameter-representation map $\alpha \mapsto \{X_n(\alpha, \bar{u})\}$ is continuous at α.*

Proof. Fix an $\bar{u}$ and hence the corresponding $\overleftarrow{u}$. We note that from (i) of Lemma 2 a parameter-encoding map is u.s.c. Further, whenever it is single-valued it is continuous. Hence $\mathcal{E}(\cdot, \overleftarrow{u})$ is continuous at α.

Note that $\{X_n(\alpha, \bar{u})\} \in R := \prod_{i=-\infty}^{\infty} Y_i$, where $Y_i = \mathsf{H}_X$ for all i and (the representation space) R is equipped with the product topology. We know that $\alpha \mapsto \{X_n(\alpha, \bar{u})\}$

is continuous if and only if the coordinate mappings $\alpha \mapsto X_n(\alpha, \bar{u})$ is continuous for all $n \in \mathbb{Z}$. Let $\overleftarrow{v}_n := (\ldots, u_{n-2}, u_{n-1})$. Clearly, $X_n(\alpha, \bar{u}) = \mathcal{E}(\alpha, \overleftarrow{v}_n)$ and the mapping $\alpha \mapsto X_n(\alpha, \bar{u})$ is identical to the parameter-encoding map $\alpha \to \mathcal{E}(\alpha, \overleftarrow{v}_n)$. From Lemma 2, a parameter-encoding map is always u.s.c, and when it is single-valued it is continuous. Since, g is an open-parametric dynamical system, $\mathcal{E}(\alpha, \overleftarrow{v}_n)$ is a singleton subset of X for all n if and only if $\mathcal{E}(\alpha, \overleftarrow{u})$ is a singleton subset of X. When g has the ESP w.r.t. $\overleftarrow{u}$ at α it follows that $\alpha \mapsto \{X_n(\alpha, \bar{u})\}$ is continuous. ■

The bifurcation in which an attractive solution (like a stable equilibrium solution) loses "stability" and gives rise to attractive nonautonomous sets that comprise multiple solutions is called a shovel bifurcation in the nonautonomous dynamical systems literature, e.g., [10]. In [10], asymptotically constant inputs are used to study the shovel bifurcation. However, here through the definitions of soft-ESP and hard-ESP thresholds, we can obtain results for any general input (see Theorem 3) by relating a bifurcation to the equicontinuity of the parameter-encoding family $\{\mathcal{E}_n(\cdot, \overleftarrow{u})\}$.

First, we consider very elementary examples. To understand the transition across a soft-ESP threshold, consider the driven system family described by $g(\alpha, u, x) = u \tanh(\alpha x)$, where $X = [-1, 1]$. Further, consider the input $\overleftarrow{u}$ that is identically equal to 1, i.e., the entire left infinite sequence has elements equal to 1 and $\alpha \in (0, \infty)$. For all $\alpha \in [0, 1]$, the set $g(\alpha, u, \cdot)$ is a contraction and from it, it follows that $\mathcal{E}(\alpha, \overleftarrow{u})$ is $\{0\}$. When $\alpha > 1$, it may be easily verified that there are two attractive fixed points $p_\alpha > 0$ and $-p_\alpha < 0$ for the map $\tanh(\alpha x)$ in addition to the fixed point at 0. From the fact $\tanh(\alpha x) < x$ for $x > p_\alpha$ and $\tanh(\alpha x) > x$ for $x < -p_\alpha$, it follows that $\mathcal{E}(\alpha, \overleftarrow{u}) = [-p_\alpha, p_\alpha]$. Also, since p_α moves continuously away from 0 as α is varied beyond 1, it follows that the set $\mathcal{E}(\alpha, \overleftarrow{u})$ as a set varies continuously (in the space H_X) as α is varied beyond 1. Thus there is a "jump-less" transition of the set $X_0(\alpha, \overleftarrow{u})$ from being single-valued for to be multivalued, i.e., to be a subset of X with multiple elements as α crosses 1. Such a transition is continuous and falls into the case of a transition in Fig. S3 (a).

To understand the transition across a hard-ESP threshold, consider the example $g(\alpha, u, x) = u\,\alpha\,x$ where $X = [-1, 1]$. Again, consider the input $\overleftarrow{u}$ that is identically equal to 1 with $\alpha \in (0, \infty)$. For all $\alpha < 1$, it can be easily verified that $g(\alpha, u, \cdot)$ is a contraction on X and hence $\mathcal{E}(\alpha, \overleftarrow{u})$ is $\{0\}$. When $\alpha = 1$, $\mathcal{E}(\alpha, \overleftarrow{u})$ is $[-1, 1]$ since $g(1, 1, \cdot)$ is the identity map on X. Thus unlike a jump-less transition of $\mathcal{E}(\alpha, \overleftarrow{u})$ past α_t in the preceding example, there is a discontinuous transition of the set $X_0(\alpha, \overleftarrow{u})$ from being single-valued to be multivalued at $\alpha = 1$ (falls into the case of Fig. S3 (b)). We would have a similar discontinuous change in $\mathcal{E}(\alpha, \overleftarrow{u})$, when instead $\overleftarrow{u}$ is a left-infinite sequence that has 1's except for finitely many non-zero elements between 0 and !1. In such a case, at the bifurcation value α_t, the input $\overleftarrow{u}$ would be such that infinite number of elements u in $\overleftarrow{u}$ render the maps $g(\alpha_t, u, \cdot)$ to be non-contracting on some subset of X and finitely many elements u in $\overleftarrow{u}$ render the maps $g(\alpha_t, u, \cdot)$ to

be contracting, and the result is a hard-ESP threshold at α_t.

We conjecture that for inputs $\overline{u}$ that are not asymptotically periodic or that does not converge to a limit, α_t would only be a hard-ESP threshold and that in an interval $[\alpha_t, b]$, the set-valued function $\mathcal{E}(\cdot, \overline{u})$ would be not l.s.c and hence discontinuous.

Theorem 3 distinguishes the two ESP thresholds defined in [1] and, more generally, the continuous or discontinuous points of the parameter-encoding map through the notion of equicontinuity of the parameter-encoding family. Intuitively, equicontinuity is meant to deal with continuity of the entire set of functions at once. Formally, a family of functions $\{f_i\}_{i\in I}$ defined between two metric spaces X and Y is equicontinuous at $x \in X$ if for every $\epsilon > 0$, there exists a $\delta > 0$ such that $d_X(x, y) < \delta$ would imply $d_Y(f_i(x), f_i(y)) < \epsilon$ for all $i \in I$. The family $\{f_i\}_{i\in I}$ is said to be *equicontinuous* on a set $A \subset X$ if it is equicontinuous at every point $x \in A$. A very simple example, is family of functions defined by $f_n(x) = x^n$, $n = 1, 2, \ldots$ on $[0, 1]$ is not equicontinuous at 1 since for instance when $\epsilon = 1/2$, for every $\delta > 0$, there is an n large enough so that $|f_n(y) - f_n(y)| = |1 - y^n| > \epsilon$ while $y \in (1 - \delta, 1)$. The family of functions is equicontinuous for $x \in [0, 1)$ though. Similar to this behavior, we witness a discontinuity in the parameter-encoding map when the parametric encoding family $\{\mathcal{E}_n(\cdot, \overline{u})\}$ fails to be equicontinuous as Theorem 3 asserts.

Theorem 3. *Consider an open-parametric-driven system g and an input $\overline{u}$. The parameter-encoding map $\mathcal{E}(\cdot, \overline{u})$ is continuous at β if and only if $\{\mathcal{E}_n(\cdot, \overline{u})\}$ is equicontinuous at β. In particular, when α is a real-parameter in the interval $[a, b]$ and α_t is the ESP threshold, then α_t is a soft-ESP threshold when $\{\mathcal{E}_n(\cdot, \overline{u})\}$ is equicontinuous at β and a hard-threshold otherwise.*

Proof. We first show that $\{\mathcal{E}_n(\cdot, \overline{u})\}$ is equicontinuous at β implies $\mathcal{E}(\cdot, \overline{u})$ is continuous at β. To prove this it is sufficient to show that given a sequence, $\alpha_k \to \beta$ we have $\mathcal{E}(\alpha_k, \overline{u}) \to \mathcal{E}(\beta, \overline{u})$. Fix a sequence $\alpha_k \to \beta$. We know by definition of the sets $\mathcal{E}_n(\alpha, \overline{u})$, $\mathcal{E}_n(\alpha, \overline{u}) \to \mathcal{E}(\alpha, \overline{u})$ for every α. Hence for a given $\epsilon > 0$, we can define $N(\alpha)$ to be an integer such that $d_H(\mathcal{E}(\alpha, \overline{u}), \mathcal{E}_n(\alpha, \overline{u})) < \epsilon/3$ for all $n > N(\alpha)$. Thus for every $k \in \mathbb{N}$, we deduce

$$\begin{aligned} d_H(\mathcal{E}(\alpha_k, \overline{u}), \mathcal{E}(\beta, \overline{u})) &\leq d_H(\mathcal{E}(\alpha_k, \overline{u}), \mathcal{E}_n(\alpha_k, \overline{u})) + d_H(\mathcal{E}_n(\alpha_k, \overline{u}), \mathcal{E}_n(\beta, \overline{u})) + \\ &\qquad d_H(\mathcal{E}_n(\beta, \overline{u}), \mathcal{E}(\beta, \overline{u})) \\ &\leq \quad \epsilon/3 \quad + \quad d_H(\mathcal{E}_n(\alpha_k, \overline{u}), \mathcal{E}_n(\beta, \overline{u})) \quad + \quad \epsilon/3, \end{aligned}$$

whenever $n > max(N(\alpha_k), N(\beta))$. Since $\{\mathcal{E}_n(\cdot, \overline{u})\}$ is equicontinuous at β, there exists an integer K such that for $k > K$, $d_H(\mathcal{E}_n(\alpha_k, \overline{u}), \mathcal{E}_n(\beta, \overline{u})) \leq \epsilon/3$. Hence, $\mathcal{E}(\alpha_k, \overline{u}) \to \mathcal{E}(\beta, \overline{u})$.

Next, to show $\mathcal{E}(\cdot,\overline{u})$ is continuous at β implies $\{\mathcal{E}_n(\cdot,\overline{u})\}$ is equicontinuous at β, we prove its contrapositive. Let $\{\mathcal{E}_n(\cdot,\overline{u})\}$ be not equicontinuous at some fixed β. Then, there exists an $\epsilon > 0$, a sequence $\alpha_k \to \beta$ and a sequence of integers $n_k \to \infty$ such that $d_H(\mathcal{E}_{n_k}(\alpha_k,\overline{u}),\mathcal{E}_{n_k}(\beta,\overline{u})) > 3\epsilon_0 > 0$. Hence, by triangle inequality,

$$\begin{aligned} 3\epsilon_0 \;<\; d_H(\mathcal{E}_{n_k}(\alpha_k,\overline{u}),\mathcal{E}_{n_k}(\beta,\overline{u})) \;\leq\; & d_H(\mathcal{E}_{n_k}(\alpha_k,\overline{u}),\mathcal{E}(\alpha_k,\overline{u})) \quad + \\ & d_H(\mathcal{E}(\alpha_k,\overline{u}),\mathcal{E}(\beta,\overline{u})) + d_H(\mathcal{E}(\beta,\overline{u}),\mathcal{E}_{n_k}(\beta,\overline{u})). \end{aligned} \tag{4}$$

Since $\mathcal{E}_n(\beta,\overline{u}) \to \mathcal{E}(\beta,\overline{u})$, given $\epsilon_0 > 0$, we can find an integer M such that the inequality $d_H(\mathcal{E}_M(\beta,\overline{u}),\mathcal{E}(\beta,\overline{u})) < \epsilon_0$ holds. Note $\mathcal{E}_M(\cdot,\overline{u})$ is continuous for any finite M by (iii) of Lemma 2. Assume $\mathcal{E}(\cdot,\overline{u}))$ is continuous at β. Hence, by continuity the set $V_\beta := \{\alpha \in \Lambda : d_H(\mathcal{E}_M(\alpha,\overline{u}),\mathcal{E}(\alpha,\overline{u})) < \epsilon_0\}$ contains a neighborhood of β in the space Λ. Since $\mathcal{E}_n(\beta,\overline{u})$ is monotonically decreasing, $d_H(\mathcal{E}_n(\beta,\overline{u}),\mathcal{E}(\beta,\overline{u})) \to 0$ monotonically. Hence $V_\beta := \{\alpha \in \Lambda : d_H(\mathcal{E}_n(\alpha,\overline{u}),\mathcal{E}(\alpha,\overline{u})) < \epsilon_0 \;\forall\; n \geq M\}$. There exists an integer K_1 so that for all $k \geq K$, we have $\alpha_k \in V_\beta$ and $n_k \geq M$. Hence the term $d_H(\mathcal{E}_{n_k}(\alpha_k,\overline{u}),\mathcal{E}(\alpha_k,\overline{u}))$ in (4) can be made less than ϵ_0 whenever $k \geq K_1$. Thus, for $k \geq K_1$ we can rewrite (4) as

$$\begin{aligned} 3\epsilon_0 \;<\; d_H(\mathcal{E}_{n_k}(\alpha_k,\overline{u}),\mathcal{E}_{n_k}(\beta,\overline{u})) \;\leq\; & \epsilon_0 \quad + \\ & d_H(\mathcal{E}(\alpha_k,\overline{u}),\mathcal{E}(\beta,\overline{u})) + d_H(\mathcal{E}(\beta,\overline{u}),\mathcal{E}_{n_k}(\beta,\overline{u})). \end{aligned}$$

Now, by the assumption of continuity of $\mathcal{E}(\cdot,\overline{u})$ at β, there exists an integer K_2 so that $d_H(\mathcal{E}(\alpha_k,\overline{u}),\mathcal{E}(\beta,\overline{u})) < \epsilon_0$. Also since $\mathcal{E}_{n_k}(\beta,\overline{u})) \to \mathcal{E}(\beta,\overline{u})$, there exists an integer K_3 so that $d_H(\mathcal{E}(\beta,\overline{u}),\mathcal{E}_{n_k}(\beta,\overline{u})) < \epsilon_0$. For $k \geq \max(K_1,K_2,K_3)$, we have $3\epsilon_0 < \epsilon_0 + \epsilon_0 + \epsilon_0$ which is absurd. Hence our assumption $\mathcal{E}(\cdot,\overline{u}))$ is continuous at β is incorrect. When $\Lambda = [a,b]$ and α_t is the ESP threshold, setting $\beta = \alpha_t$ in the above proof, we obtain α_t to be the soft-ESP threshold and the hard-ESP threshold when $\{\mathcal{E}_n(\cdot,\overline{u})\}$ is equicontinuous and not-equicontinuous respectively at α_t. ■

5 Data used in the Figures of [1, Section 2]

We furnish details used in the computations involved in the figures in [1]. All figures referred here are from [1].

Details used in Fig. 2. The dynamics of a RNN using a sinusuoidal input of length 5000 with a maximum amplitude of 1 is used to obtain two sequences $x_{n+1} = \overline{\tanh}(Au + \alpha Bx)$ by setting $\alpha = 1.02$ and $\alpha = 1.05$ in two separate computations; here A and B (with unit spectral radius) and the initial condition x_0 are all chosen randomly, and they are identical in both the computations.

Details used in Fig. 3. We consider $g(\alpha, u, x) = \overline{\tanh}(Au + \alpha Bx)$, where A is randomly generated while the matrix B with unit spectral radius is chosen to be

$$B = \begin{pmatrix} 0 & 2.0425 & 3.4566 & 1.6340 \\ 0 & -0.4923 & -1.4455 & 1.3617 \\ 0 & 0 & 0 & 2.0425 \\ 0.2409 & 0.7018 & -1.1836 & 0 \end{pmatrix}.$$

Details used in Fig. 4. We consider $g(\alpha, u, x) = \overline{\tanh}(Au + \alpha Bx)$ with 2 neurons and with 250 neurons (in two separate computations), a randomly generated input of length 500, a input matrix A and a reservoir matrix B (chosen randomly, but with unit spectral radius). For the plot on the left in Fig. 4, we use 1000 initial conditions Y_0 chosen on the boundary of $X = [-1, 1]^2$ (to exploit the fact that when viewed as $g(\alpha, u, \cdot) : \mathbb{R}^n \to \mathbb{R}^n$ is an open mapping [9] and hence maps the boundary of a set to its boundary). For the plot on the right, there were only 50 randomly chosen samples in the space $X = [-1, 1]^{250}$. Clearly, in both plots of Fig. 4, the clustering towards a single-point for values of α is obvious. Hence, one does not need to be worried about simulating the parameter-encoding map accurately if one's task is only to identify if $\mathcal{E}(\cdot, \bar{u})$ has a single point-cluster or not. On the other hand, if one's task is to identify instabilities due to the parameter while there is no ESP, one needs to analyze the parameter-encoding map further, as in Fig. 5.

Details used in Fig. 5. We use the parameter-stability plot to find the hard-ESP threshold of an RNN. We employ the same RNN used in the plot on the right in Fig. 4 (the reservoir matrix B has a unit spectra radius) to obtain the parameter-stability plot shown in (a) of Fig. 5. The smallest α_k for which this plot turns conspicuously positive can be identified as the hard-ESP threshold. Further, wherever the parameter-encoding map is discontinuous, it fails to be lower-semicontinuous in view of [8, Lemma 2], and hence it has wildly behaving explosions. Thus, the parameter-stability plot in (a) of Fig. 5 is wiggly in addition to being positive for $\alpha > \alpha_t$. We remark that the scenario at a point of discontinuity of the parameter-encoding map would be similar to the behavior at a point of discontinuity of the example in Equation (4) in [1]). We note during the simulations, a very accurate approximation of $\mathcal{E}(\alpha_k, \bar{u})$ turns out to be not essential in determining the continuity and discontinuity points of the parameter-encoding map. Even with few (here, 50) initial conditions, we can identify the discontinuous points in the parameter-stability plot. All numerical simulations are computationally inexpensive and quick (< 3 minutes).

6 Cross-checking the inferences from the parameter-stability plot in [1, Fig. 4]

.

We reproduce a plot from [1, Fig. 4 (a)] in Fig. S5. Several cross-checks can be made to ensure that the idea of determining the ESP threshold via the parameter-stability plot in Fig. S5 is robust in the sense that the finite approximations of the underlying parameter-encoding map have not obscured the real value of α_t. Firstly, for the RNN employed in obtaining Fig. S5, we set two other additional step sizes $|\alpha_k - \alpha_{k-1}|$, and observe similar behavior in the parameter-stability plot (see Fig. S6).

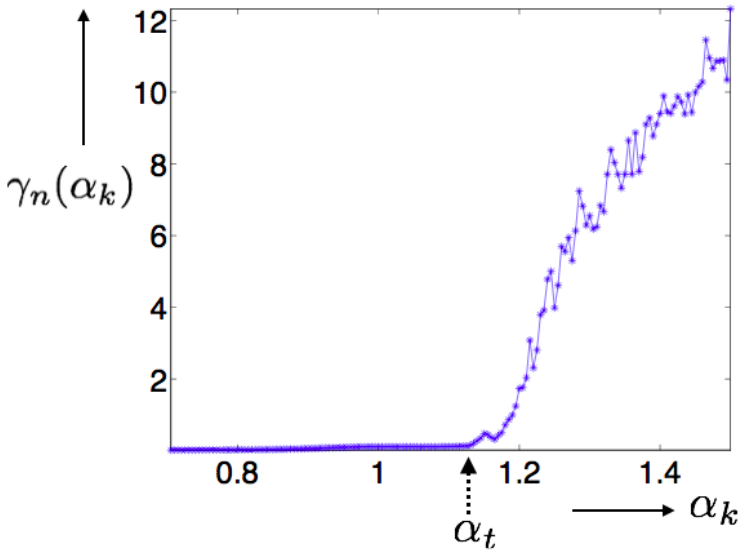


Figure S5: Parameter-Stability plot: $\gamma_n(\alpha_k) = d_H(\mathcal{E}_n(\alpha_k, \bar{u}), \mathcal{E}_n(\alpha_{k-1}, \bar{u}))$ plotted against α_k with $|\alpha_k - \alpha_{k-1}| = 0.005$. The smallest value of α_k where the plot turns positive is the edge-of-criticality or the hard-ESP threshold α_t in the interval $[0.7, 1.5]$.

Secondly, in Fig. S7, we verify the ESP threshold evaluated in Fig. S5 is nearly synchronized with $\mathcal{E}(\alpha_k, \bar{u})$ turning from being single-valued in X to being multi-valued. Towards, this end, we plot one coordinate of $\mathcal{E}(\alpha_k, \bar{u})$ against α_k in the top panel of Fig. S7. We reproduce Fig. S5 in the middle panel of Fig. S7. In the last panel of Fig. S7, we place a plot of a heuristic measure of the deviation from a single point-cluster $C(\alpha_k, \bar{u})$ of $\mathcal{E}(\alpha_k, \bar{u})$ against α_k. Since $\mathcal{E}(\alpha, \bar{u})$ comprises elements of

X, one can define their centroid when X is a subset of $\mathbb{R}^N$. If $\mathcal{E}(\alpha, \bar{u})$ is single-valued in X then it is identical to its centroid $C(\alpha, \bar{u})$ evaluated in X, and when it is multivalued, there is an element of X that is different from the centroid of $\mathcal{E}(\alpha, \bar{u})$, and hence the cosine of the angle between such an element and the centroid falls away from 1. With this idea, we define the clustering coefficient $C_n(\alpha, \bar{u})$ of the set $\mathcal{E}_n(\alpha, \bar{u})$ (that takes a value closer to 1 while it is close to single point-cluster and decreases as it moves away from being a point-cluster) by

$$C_n(\alpha, \bar{u}) := \min \left(\frac{M_\alpha \cdot y}{||M_\alpha|| ||y||} : y \text{ belongs to } \mathcal{E}_n(\alpha, \bar{u}) \right), \tag{5}$$

where M_α is the centroid of the set $\mathcal{E}_n(\alpha, \bar{u})$ and $M_\alpha \cdot y$ represents the dot product of the two vectors M_α and y. We observe the collated plots corroborate with the estimate α_t.

Thirdly, we observe the effect of scaling the input on the ESP threshold α_t obtained in Fig. S5 or Fig. S7. The RNN we consider is contractible at any $\alpha > 0$ since when the input's amplitude is sufficiently large, the dynamics of the RNN is driven to a subdomain of the state-space where there is contraction. This phenomenon also means that for an RNN, increasing the amplitude of the input should increase the ESP threshold (for a precise result, see [4, (ii) of Theorem 2]). We scale the input $\bar{u}$ used in the computations in Fig. S7 by a factor of 1.5, and then as expected, we observe in Fig. S8 that the hard-ESP threshold increases as well.

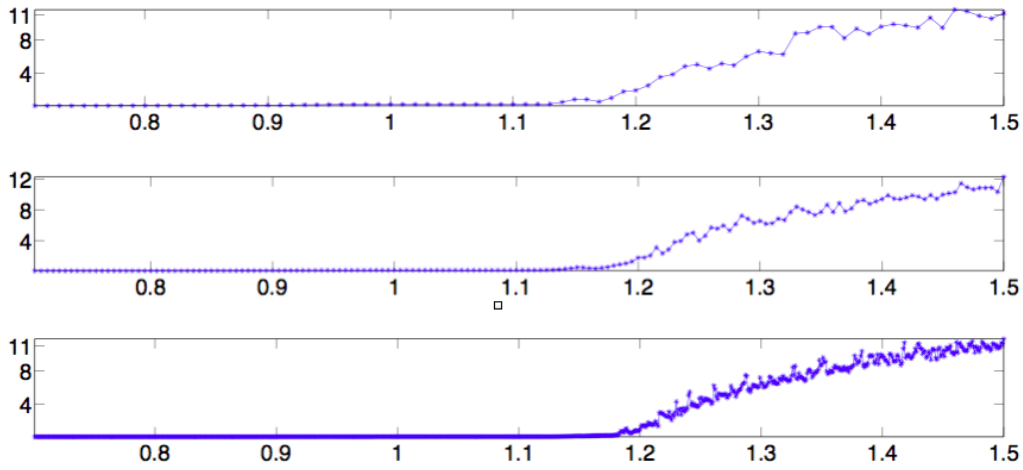


Figure S6: Parameter-stability plot in all three panels; Top panel: $|\alpha_k - \alpha_{k-1}| = 0.01$, Middle Panel: $|\alpha_k - \alpha_{k-1}| = 0.005$ (parameter-stability plot – Fig. S5 that is reproduced), Bottom Panel: $|\alpha_k - \alpha_{k-1}| = 0.001$.

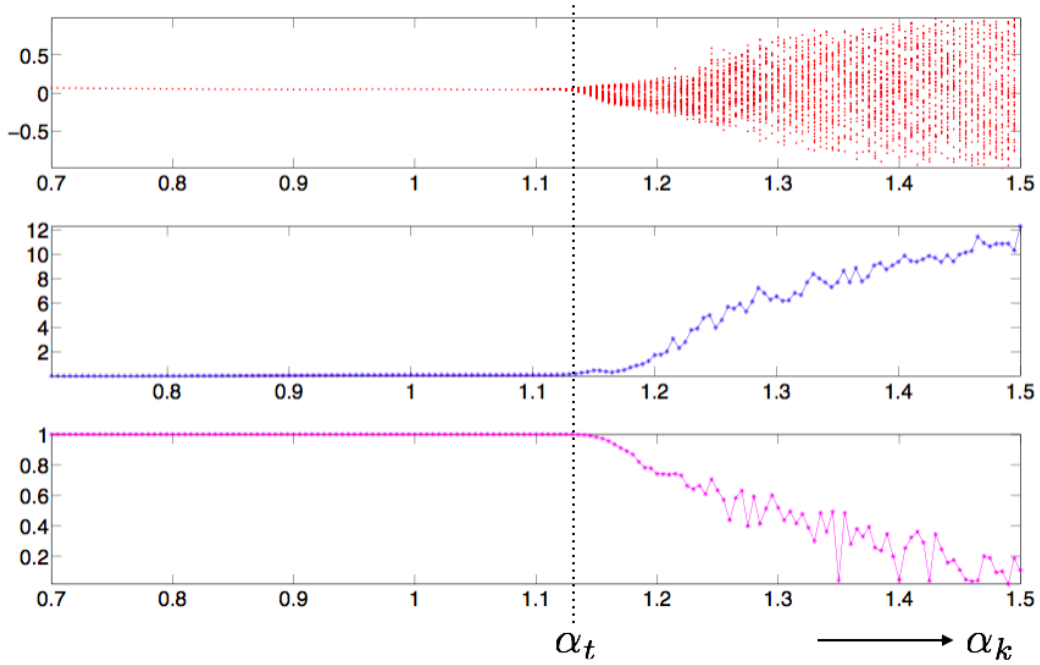


Figure S7: Complementary plots. Top Panel: Single coordinate of the parameter-encoding map (or states in a single neuron) against α_k. Middle Panel : parameter-stability plot – Fig. S5 that is reproduced. Bottom Panel: Clustering coefficient defined in (5) against α_k.

References

[1] Manjunath G (2020) Accompanying main paper: Stability and memory-loss go hand-in-hand: Three results in dynamics & computation.

[2] Bunimovich L (1985) Decay of correlations in dynamical systems with chaotic behavior. *Zh. Eksp. Teor. Fiz* 89:1452–1471.

[3] Kaneko K, Tsuda I (2003) Chaotic itinerancy. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 13(3):926–936.

[4] Manjunath G, Jaeger H (2013) Echo state property linked to an input: Exploring a fundamental characteristic of recurrent neural networks. *Neural computation* 25(3):671–696.

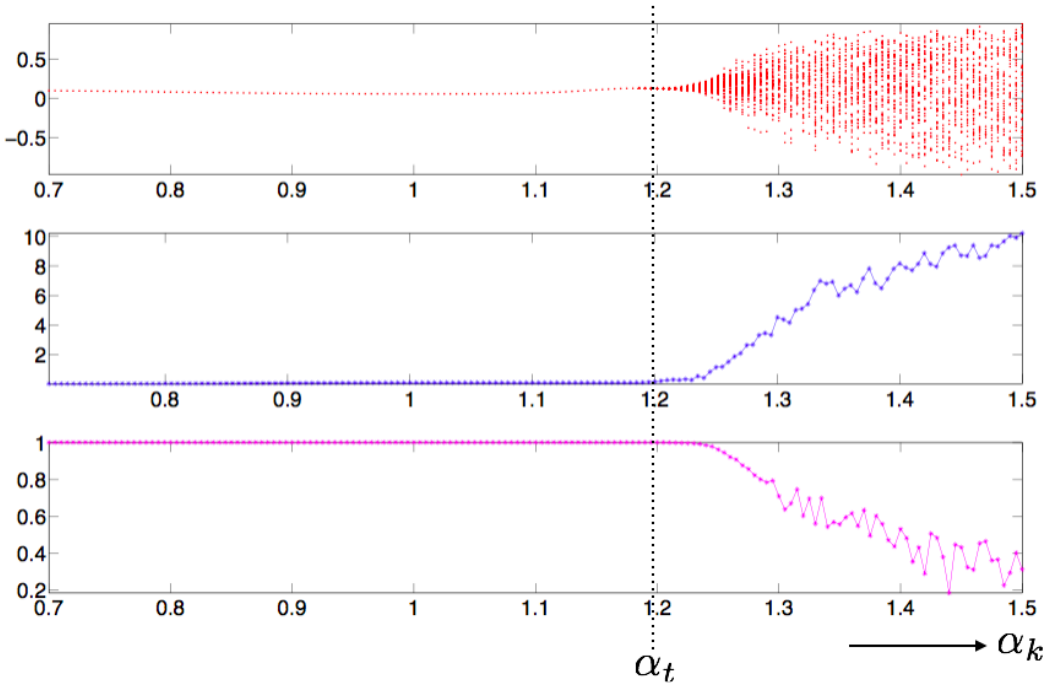


Figure S8: The simulation in panels of Fig. S7 carried out after having scaled the input $\bar{u}$ by a factor 1.5.

[5] Jaeger H (2001) The echo state approach to analysing and training recurrent neural networks-with an erratum note. *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report* 148(34):13.

[6] Jaeger H, Haas H (2004) Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *science* 304(5667):78–80.

[7] Lukoševičius M, Jaeger H (2009) Reservoir computing approaches to recurrent neural network training. *Computer Science Review* 3(3):127–149.

[8] Manjunath G (2020). *Supplementary Material that is available online along with this publication.*

[9] Munkres J (2014) *Topology*. (Pearson Education).

[10] Kloeden PE, Rasmussen M (2011) *Nonautonomous dynamical systems*. (American Mathematical Soc.) No. 176.

[11] Aubin JP, Frankowska H (2009) *Set-valued analysis*. (Springer Science & Business Media).

[12] Hart A, Hook J, Dawes J (2020) Embedding and approximation theorems for echo state networks. *Neural Networks*.