# Persistent Chaos in High-Dimensional Neural Networks

D. J. Albers
with J. C. Sprott and James P. Crutchfield

February 20, 2005

# Outline:

- Introduction and motivation
- Mathematical versus computational dynamics
- Neural networks: a "universal" function space
- Splitting of a parameter interval
- Qualitative results
- Dynamics of the bifurcation chain region
- Future work

# Introduction

Understanding the world *a la* Poincaré

Dissipative versus non-dissipative dynamics.

Problems: general qualitative understanding; invariance of measures (i.e. non-equilibrium physics); relationship with nature.

The goal is to present a framework for studying the above problems, provide links between mathematical dynamics and the real world.

First step, understand the dynamics of our construction.

# Difference in approaches

Mathematical dynamical systems:

- Little changes for $d \geq 3$
- Parameters are never an issue, the framework is with respect to $C^k$ perturbations in the $C^r$ Whitney topology

Computational dynamical systems

**The number of dimensions of the dynamical system matters;** *there is a stark difference between common dynamics in high and low-dimensional dynamical systems.*

**Parameters matter;** *the practical effects of parameters with respect to dynamical systems is significant. Instead of selecting a manifold to impose dynamics on about which one can classify a generic type behavior, the existence of parameters asks the opposite question: begin with a dynamical system, vary the parameters, and observe the types of manifolds that are present, and study the types of dynamics that are predominant on those manifolds.*

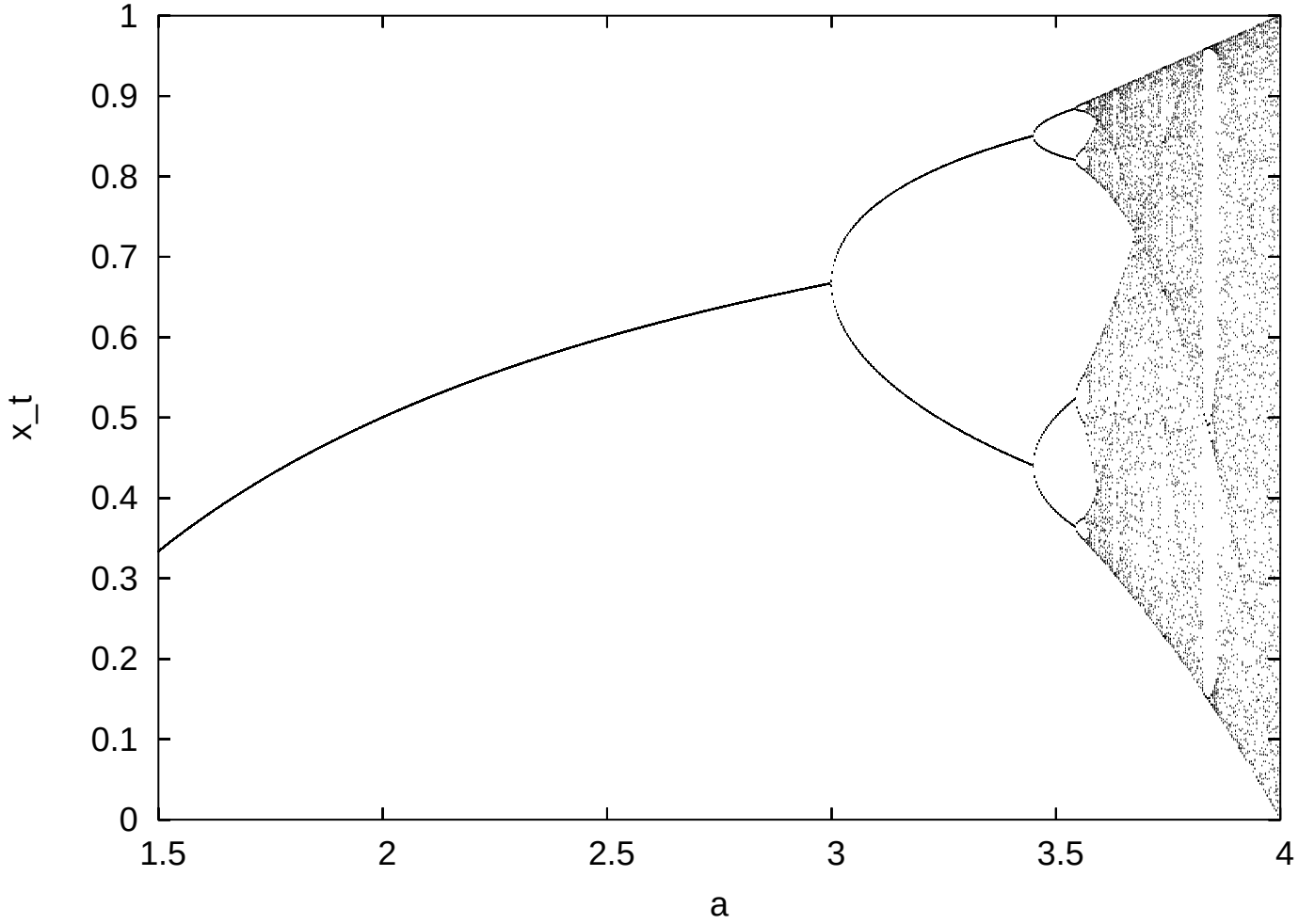
## Computational and low-dimensional intuition

Logistic map: real Fatou lemma, Jakobson absolute continuity;

Topological variation is dramatic, e.g. transitions from chaotic to periodic orbits;

Periodic orbits and windows amidst chaotic orbits are common;

Standard logistic map.



# Mathematical dynamics intuition

Everything is Anosov (cat map) like ("stacks" of Anosovs, etc).

Topological variation is relatively uncommon (but there are many conflicting stories). (Smale, Palis, Robbin, etc).

Periodic "windows" are likely rare. (Pugh)

Many (non-dissipative at the moment) dynamical systems are ergodic *a la* Boltzmann. (Pugh-Shub)

Dissipative dynamical systems are hard to handle (current hope — SRB measures).

## Artificial neural networks

**Definition 1** *A neural network is a $C^r$ mapping $\gamma : R^n \to R$. The set of feedforward networks with a single hidden layer, $\Sigma(G)$, can be written:*

$$\Sigma(G) \equiv \{\gamma : R^d \to R | \gamma(x) = \sum_{i=1}^{N} \beta_i G(\tilde{x}^T \omega_i)\} \tag{1}$$

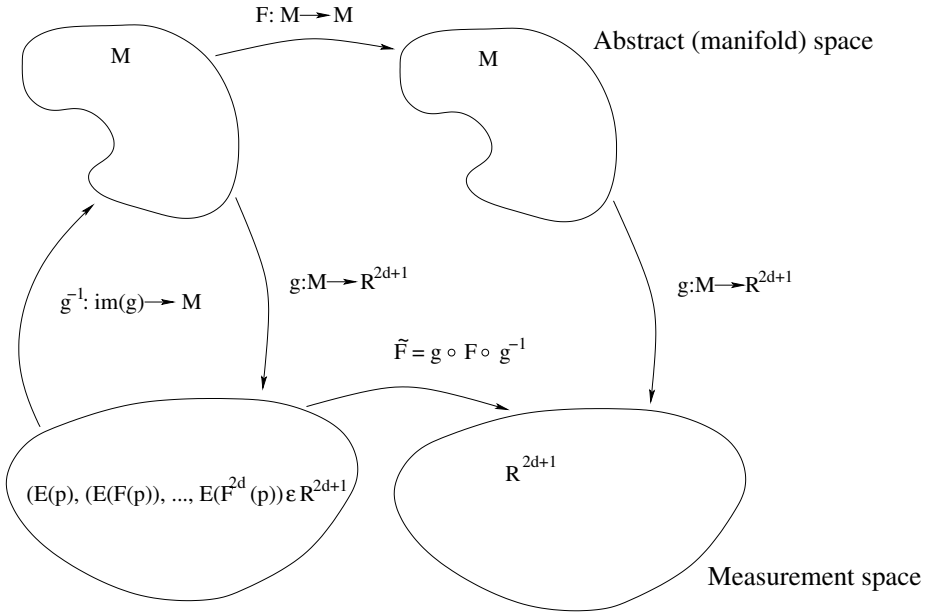
*where $x \in R^d$, is the $d-$vector of networks inputs, $\tilde{x}^T \equiv (1, x^T)$ (where $x^T$ is the transpose of $x$), $N$ is the number of hidden units (neurons), $\beta_1, \ldots, \beta_N \in R$ are the hidden-to-output layer weights, $\omega_1, \ldots, \omega_N \in R^{d+1}$ are the input-to-hidden layer weights, and $G : R^d \to R$ is the hidden layer activation function (or neuron).*

$$x_t = \beta_0 + \sum_{i=1}^{N} \beta_i G\left(s\omega_{i0} + s\sum_{j=1}^{d} \omega_{ij} x_{t-j}\right) \tag{2}$$

$\omega_{ij} \in N(0, s)$, $\beta_i$ uniform on [0, 1], $G \equiv \tanh()$, $d =$ number of inputs, $N =$ number of neurons.

## Neural networks and the Takens embedding theorem

Schematic diagram of the Takens embedding theorem and how it applies to our construction.




$F$ is the dynamical system, $E : M \to R$ ($E$ is a $C^k$ map), where $E$ represents some empirical style measurement of $F$, and $g$ is the "Takens's" map:

$$g(x_t) = (E(x_t), E(F(x_t)), \ldots, E(F^{2d}(x_t))) \tag{3}$$

## What can we approximate with neural networks?

Any function from a Sobolev space (Lebesgue integrable functions with weak derivatives).

Any function in $C^r$, $r \geq 0$ (ODE's, maps, etc).

Piecewise smooth functions with properly chosen domains.

Most PDE's.

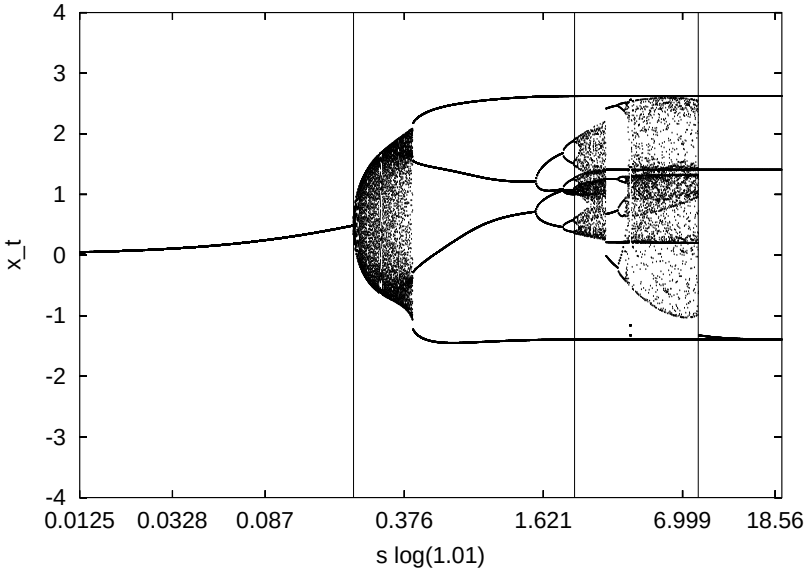
## Splitting of the $s$ parameter interval

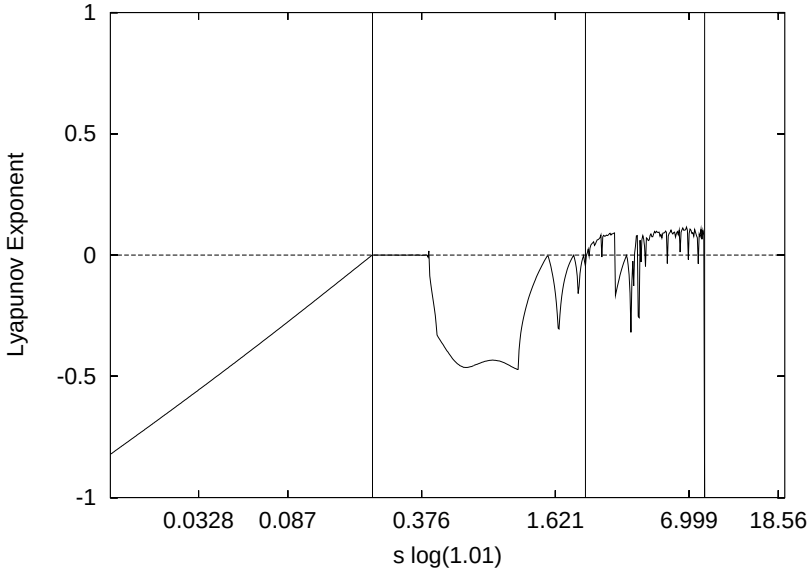
Bifurcation diagrams along an interval of the $s$ parameter.

There exist roughly five regions: the first bifurcation region (I), the routes to chaos region (II), chaos to bifurcation chains (III), bifurcation chain region (IV), bifurcation chains to finite state dynamics (V).

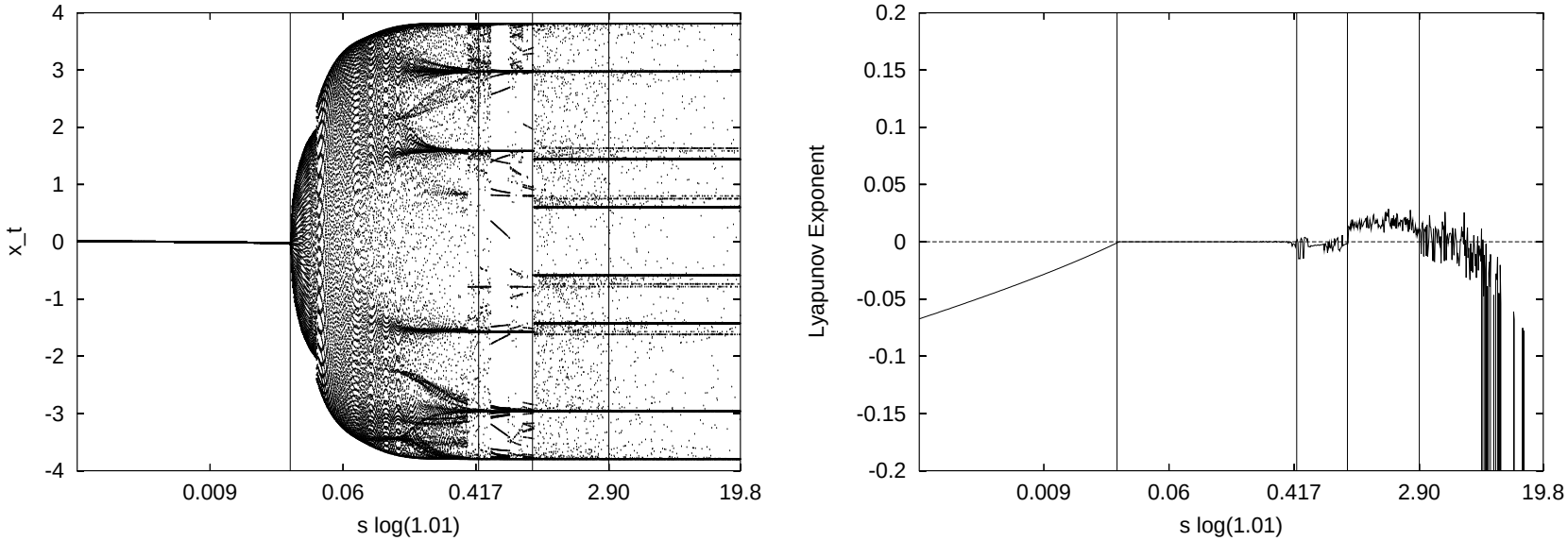
Prototypical high-dimensional case is: I $\rightarrow$ II $\rightarrow$ IV $\rightarrow$ V.

Bifurcation diagram with the largest Lyapunov exponent for $N = 4$ and $d = 4$ with regions I, II, III, and V — no region IV.



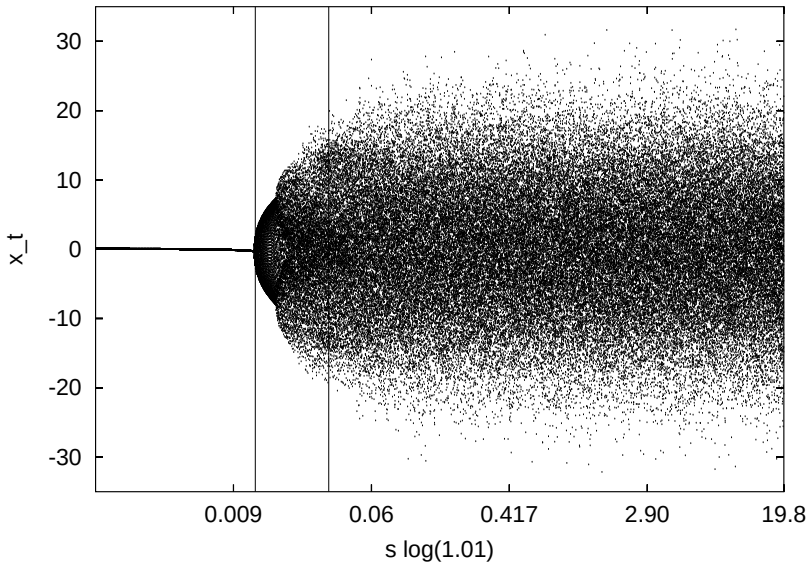



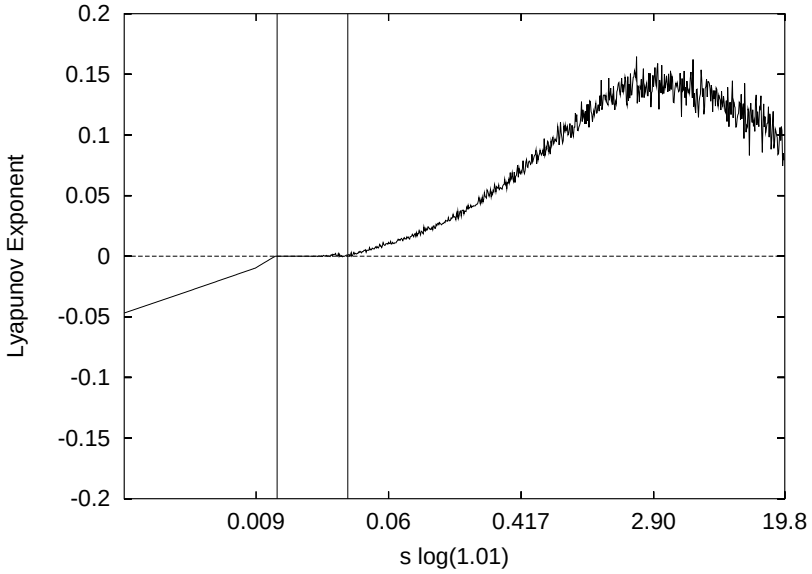
Bifurcation diagram with the largest Lyapunov exponent for $N = 4$ and $d = 64$ with regions I, II, III, IV, and V — all regions.



Bifurcation diagram with the largest Lyapunov exponent for $N = 64$ and $d = 64$ with regions I, II, IV, region V is not displayed.

## The prototypical scenario



## General lore over all regions:

1. As the dimension is increased, the probability of a network being chaotic over some portion of its parameter space goes to unity.

2. For the networks considered, at $d = 50$ there exists an $s$ value such that the maximum probability of chaos is about fifty percent.

3. As the dimension is increased, the Kaplan-Yorke dimension scales like $\sim \frac{d}{2}$.

4. As the dimension is increased, the probability of the first bifurcation being Neimark-Sacker (bifurcation to quasi-periodic orbit) approaches unity.

5. As the dimension is increased, the dominant route to chaos from a fixed point along a one-dimensional interval is the quasi-periodic route involving tori ($T^2$ for $d = 64$);

6. As the dimension is increased, the existence of periodic windows in parameter space scales like $\frac{1}{d}$.

7. As the dimension is increased, the number of positive exponents scales like $\sim \frac{d}{4}$.

8. As the dimension is increased, in the chaotic region of parameter space, hyperbolicity is violated on an increasingly countable, "dense," Lebesgue measure zero interval in parameter space.

9. As the dimension in increased chaos becomes more robust with respect to parameter changes.

## Outline for the numerical arguments for region IV

Along the 1-dimensional $s$-interval — formulation of conjecture

List of properties we require

New definitions

Formalization of conjectures

Evidence

High-dimensional generalization

New definitions

Formalization of conjectures

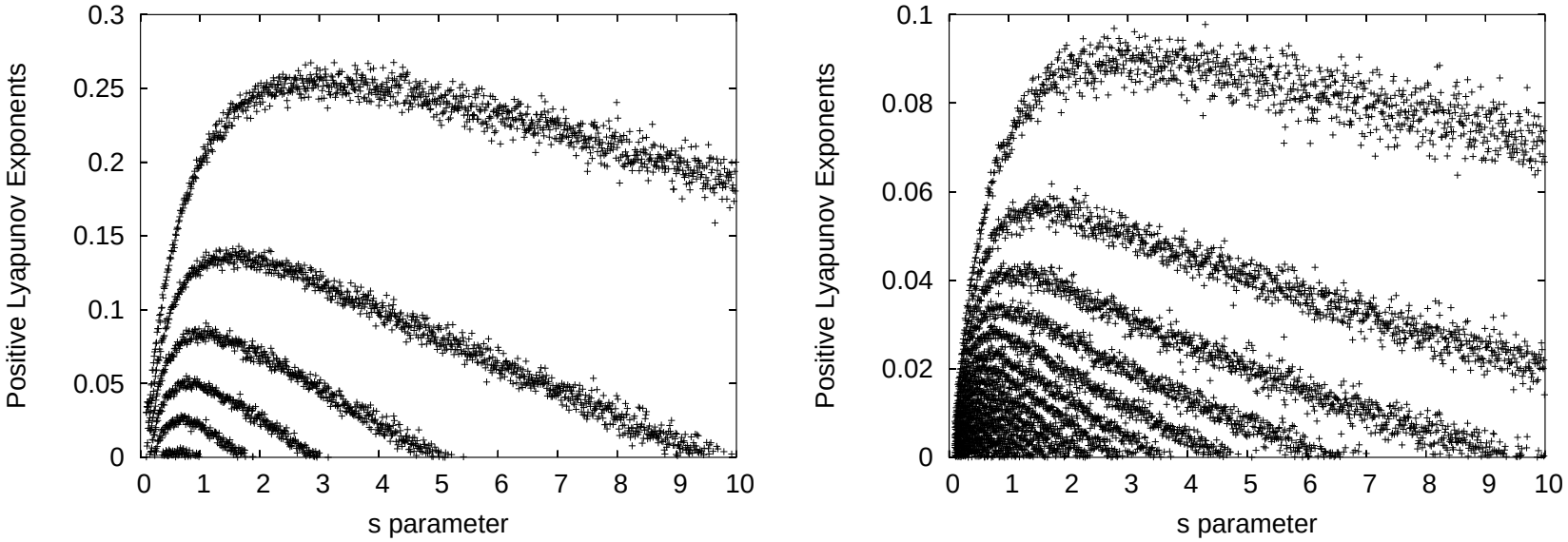
Evidence

## Lyapunov characteristic exponent (LCE) spectrum

1. the number of positive exponents = the number of expanding (stretching) directions.

2. the number of negative exponents = the number of contracting (squashing) directions.

Positive LE spectrum for typical *individual* networks with 32 neurons and 16 (left) and 64 (right) dimensions.




About the fuz

## List of properties for 1-dimensional parameter interval picture

The following properties must increase with dimension:

a. the number of positive exponents;

b. the continuity of the exponents relative to parameter change;

c. Asymptotic density of transversal Lyapunov exponent zero crossings;

## Numerical continuity

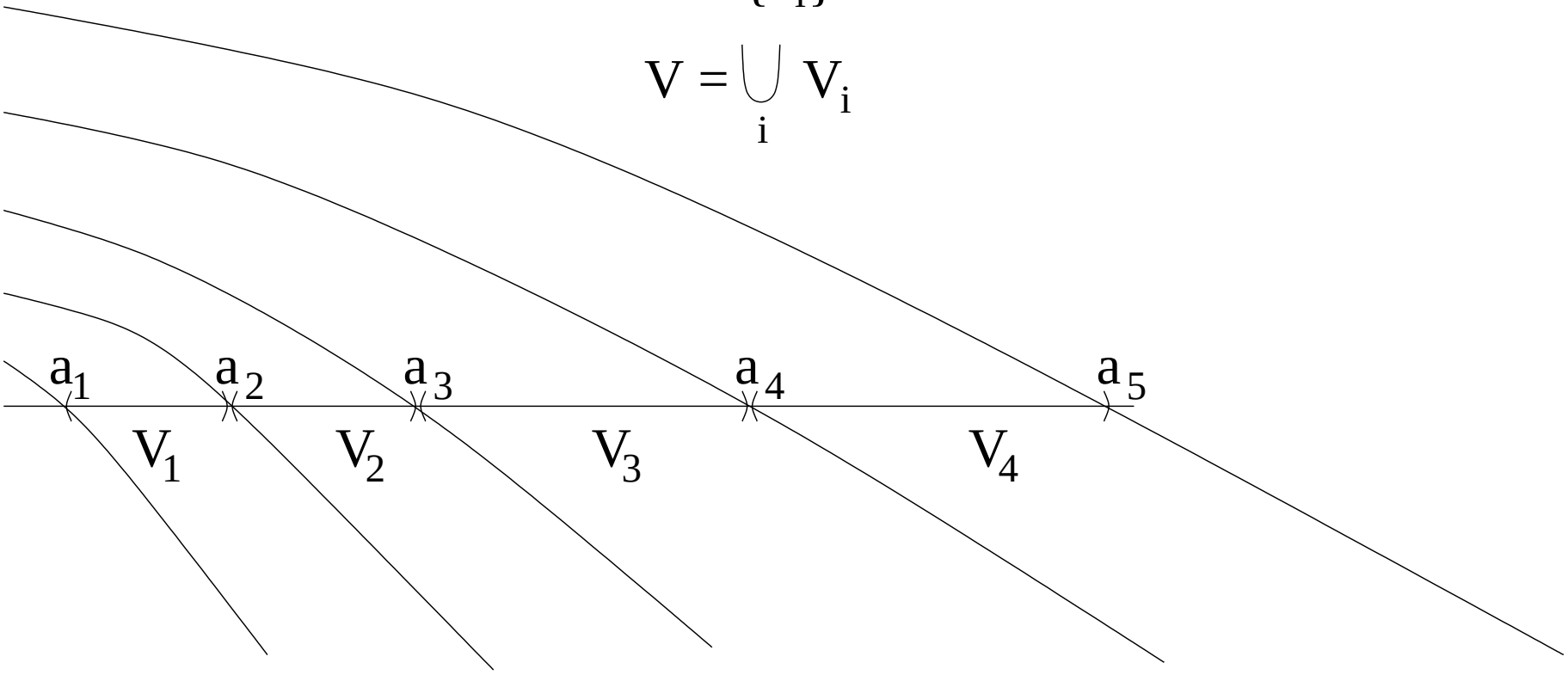
Small variation in the parameter leads to small variation in the LCEs.

## Diagram for bifurcation chain sets

An intuitive diagram for chain link sets, $V$, bifurcation link sets, $V_i$, and bifurcation chain sets, $U$. for an LCE decreasing chain link set $V$.

$$U = \{a_i\}$$

$$V = \bigcup_i V_i$$



**Definition 2 (Chain link set)** *Assume $f$ is a mapping (neural network) as previously defined. A* **chain link set** *is denoted:*

$$V = \{s \in R \mid \chi_j(s) \neq 0 \text{ for all } 0 < j \leq d \\ \text{and } \chi_j(s) > 0 \text{ for some } j > 0\}$$

Next, let $C_k$ be a connected component of the closure of $V$, $\overline{V}$. It can be shown that $C_k \cap V$ is a union of disjoint, adjacent open intervals of the form $\bigcup_i (a_i, a_{i+1})$.

**Definition 3 (Bifurcation link set)** *Assume $f$ is a mapping (neural network) as previously defined. Denote a* **bifurcation link set** *of $C_k \cap V$ as:*

$$V_i = (a_i, a_{i+1}) \tag{4}$$

**Definition 4 (Bifurcation chain subset)** *Let $V$ be a chain link set, and $C_k$ a connected component of $\overline{V}$. A* **bifurcation chain subset** *of $C_k \cap V$ is denoted:*

$$U_k = \{a_i\} \tag{5}$$

*or equivalently:*

$$U_k = \partial(C_k \cap V) \tag{6}$$

**Definition 5 ($\epsilon$-dense)** *Given an $\epsilon > 0$, an open interval $(a, b) \subset R$, and a sequence $\{c_1, \ldots, c_n\}$, $\{c_1, \ldots, c_n\}$ is $\epsilon$-dense in $(a, b)$ if there exists an $n$ such that for any $x \in (a, b)$, there is an $i$, $1 \leq i < n$, such that dist$(x, c_i) < \epsilon$.*

However, we really want a sequence of sets:

$$\begin{array}{ccc} c_1^1, & \cdots, & c_{n_1}^1 \\ c_1^2, & \cdots, & c_{n_2}^2 \\ \vdots & \vdots & \vdots \end{array}$$

where $n_{i+1} > n_i$

**Definition 6 (Asymptotically Dense ($a-$dense))** *A sequence $S_j = \{c_1^j, \ldots, c_{n_j}^j\} \subset (a, b)$ of finite subsets is asymptotically dense in $(a, b)$, if for any $\epsilon > 0$, there is an $N$ such that $S_j$ is $\epsilon$-dense if $j \geq N$.*

# Two conjectures

**Conjecture 1 (Existence of a Codimension $\epsilon$ bifurcation set)** *Assume $f$ is a mapping (neural network) as previously defined with a sufficiently high number of dimensions, $d$, and a bifurcation chain set $U$ as previously mentioned. The two following (equivalent) statements hold:*

i. *In the infinite-dimensional limit, the cardinality of $U$ will go to infinity, and the length $\max|a_{i+1} - a_i|$ for all $i$ will tend to zero on a one dimensional interval in parameter space. In other words, the bifurcation chain set $U$ will be $a-$dense in its closure, $\overline{U}$.*

ii. *In the asymptotic limit of high dimension, for all $s \in U$, and for all $f$ at $s$, an arbitrarily small perturbation $\delta_s$ of $s$ will produce a topological change. The topological change will correspond to a different number of global stable and unstable manifolds for $f$ at $s$ compared to $f$ at $s + \delta$.*

**Conjecture 2 (Periodic window probability decreasing)** *Assume $f$ is a mapping (neural network) as previously defined and a bifurcation chain set $U$ as previously defined. In the asymptotic limit of high dimension, the length of the bifurcation chain sets, $l = |a_n - a_1|$, increases such that the cardinality of $U \to m$ where $m$ is the maximum number of positive Lyapunov exponents for $f$. In other words, there will exist an interval in parameter space (e.g. $s \in (a_1, a_n) \sim (0.1, 4)$) where the probability of the existence of a periodic window will go to zero (with respect to Lebesgue measure on the interval) as the dimension becomes large.*

Or

$$\#U \to \infty \text{ as } d \to \infty$$

There exists only a single $V$ (chain link set)

## Reminder: List of properties for 1-dimensional parameter interval picture

The following properties must increase with dimension:

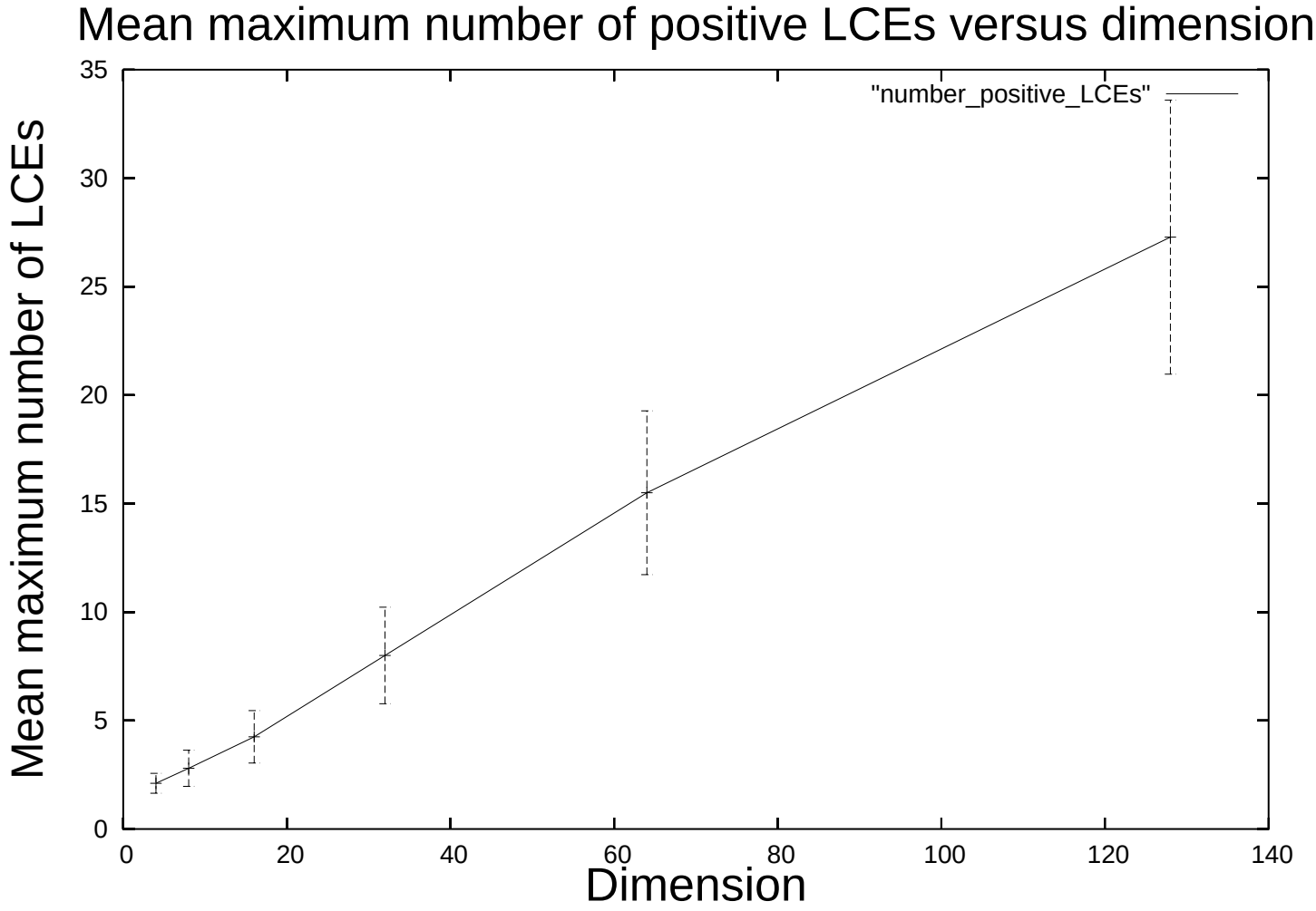
a. the number of positive exponents

b. the continuity of the exponents relative to parameter change

c. asymptotic density of transversal Lyapunov exponent zero crossings

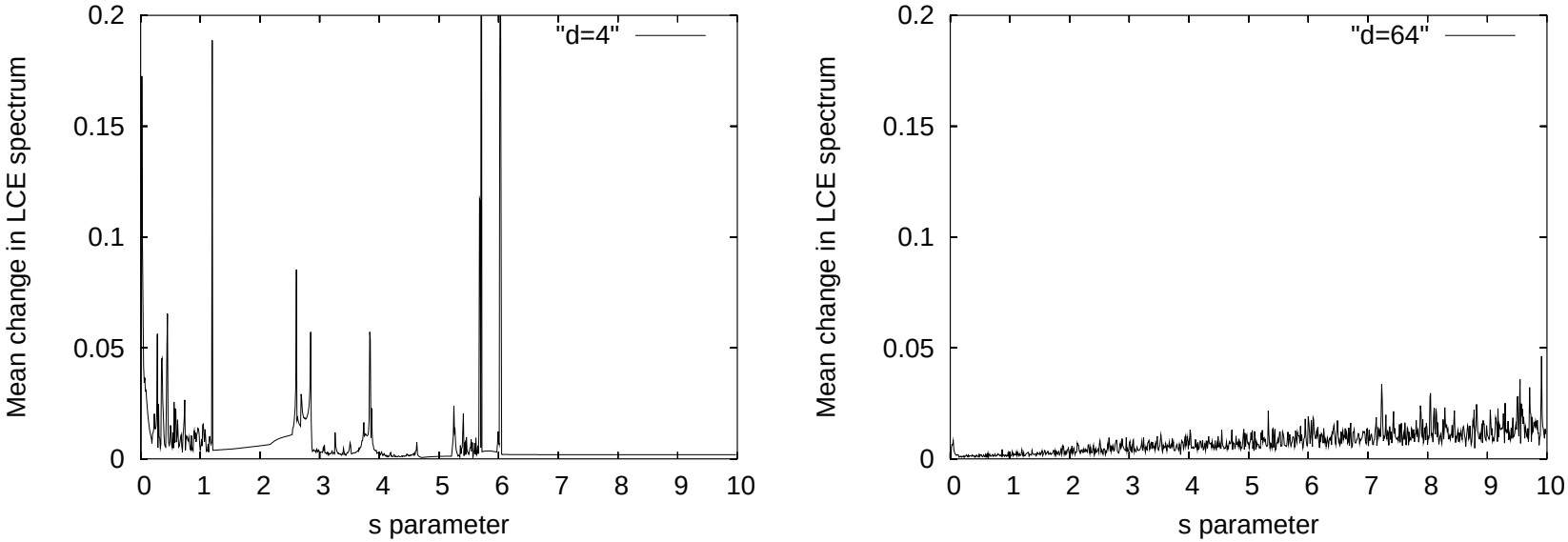
## Mean max number of positive Lyapunov exponents

Mean maximum number of positive LE's versus dimension, all networks have 32 neurons (slope is approximately $\frac{1}{4}$).



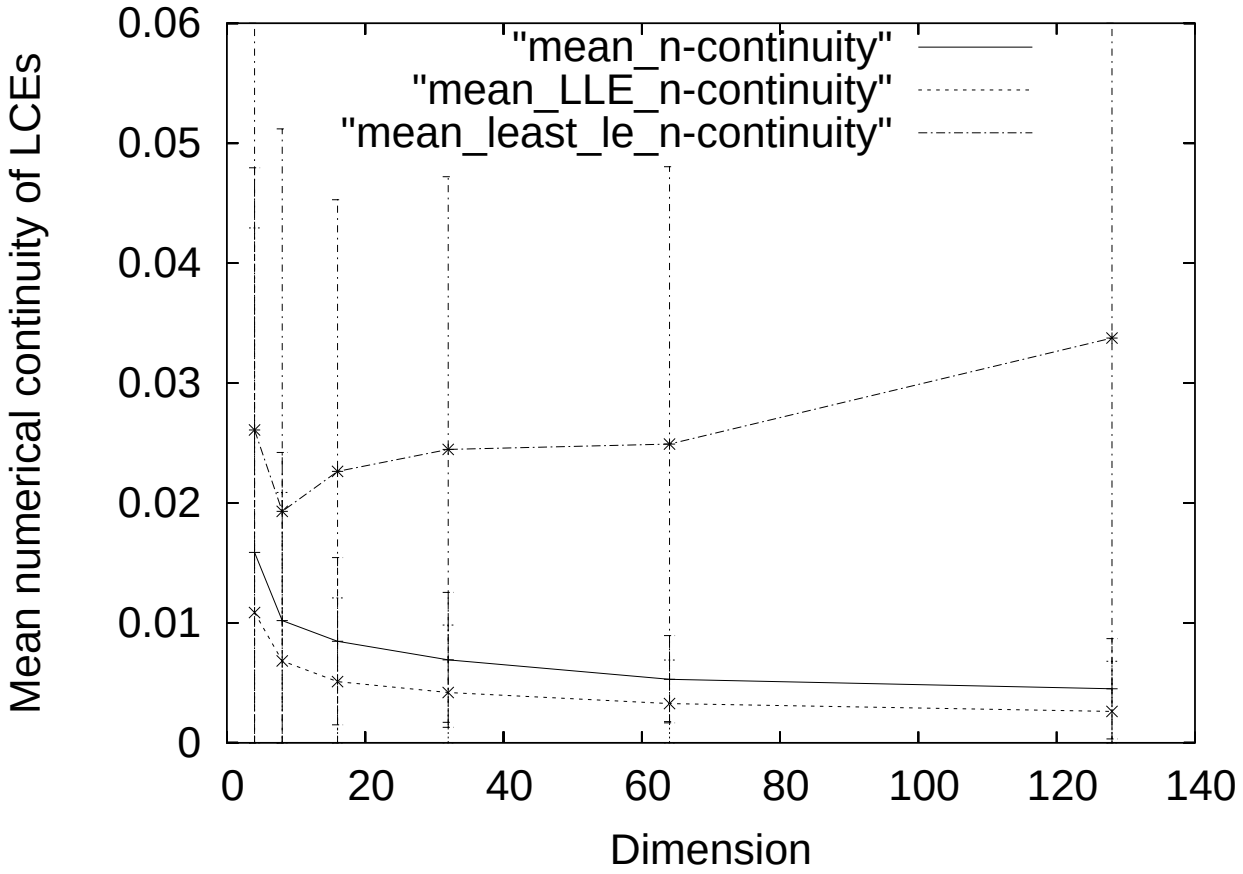
## $num$-continuity versus s

$num$-continuity (mean of $|\chi_i(s) - \chi_i(s + \delta s)|$ for each $i$) versus parameter variation: 32 neurons, 4 (left) and 64 (right) dimensions.



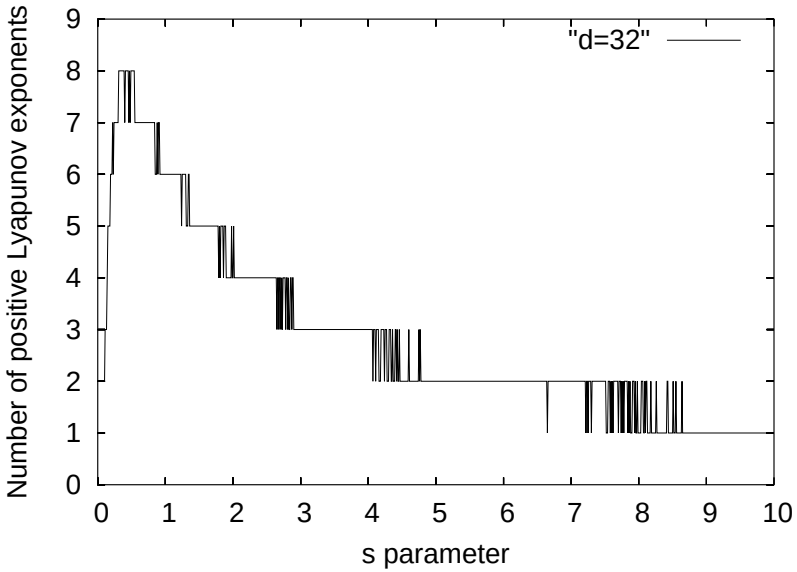
## $num$-continuity versus dimension

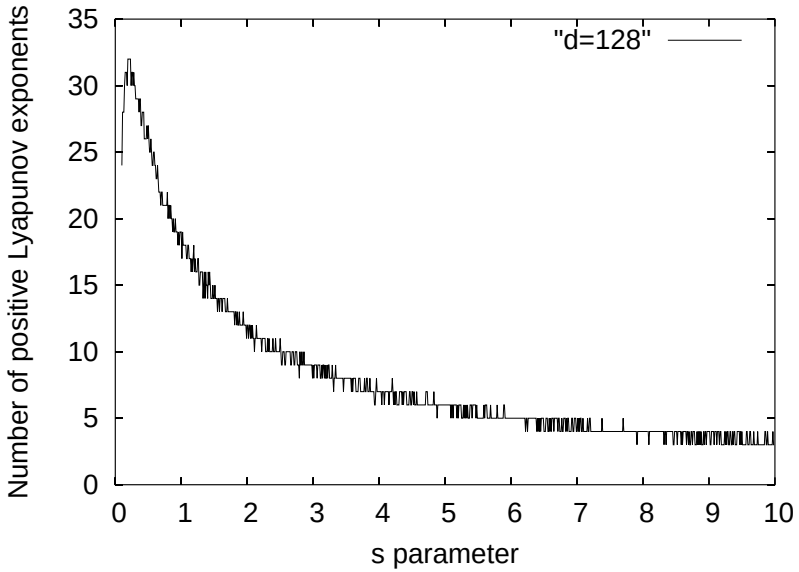
Mean $num$-continuity, $num$-continuity of the largest and the most negative Lyapunov exponent of many networks versus their dimension. The error bars are the standard deviation about the mean over the number of networks considered.



## Intuition for a-density of zero crossings

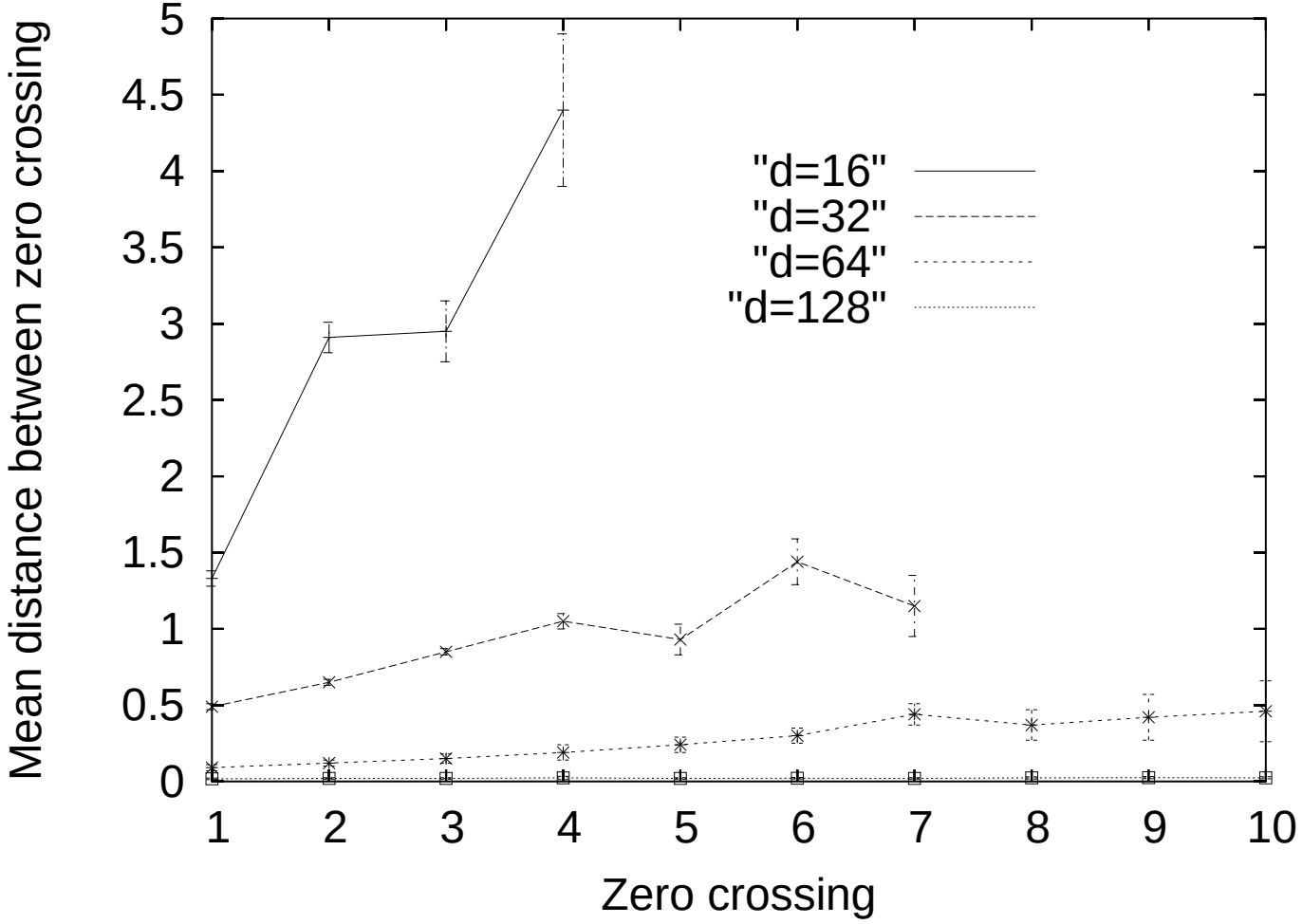
Number of positive LE's for typical individual networks with 32 neurons and 32 (left) and 128 (right) dimensions.



## a-density of the first 10 zero crossings

Mean distance ($\delta s$) between each of the first 10 zero crossings of LE's for many networks with 32 neurons and 16, 32, 64, and 128 dimensions.



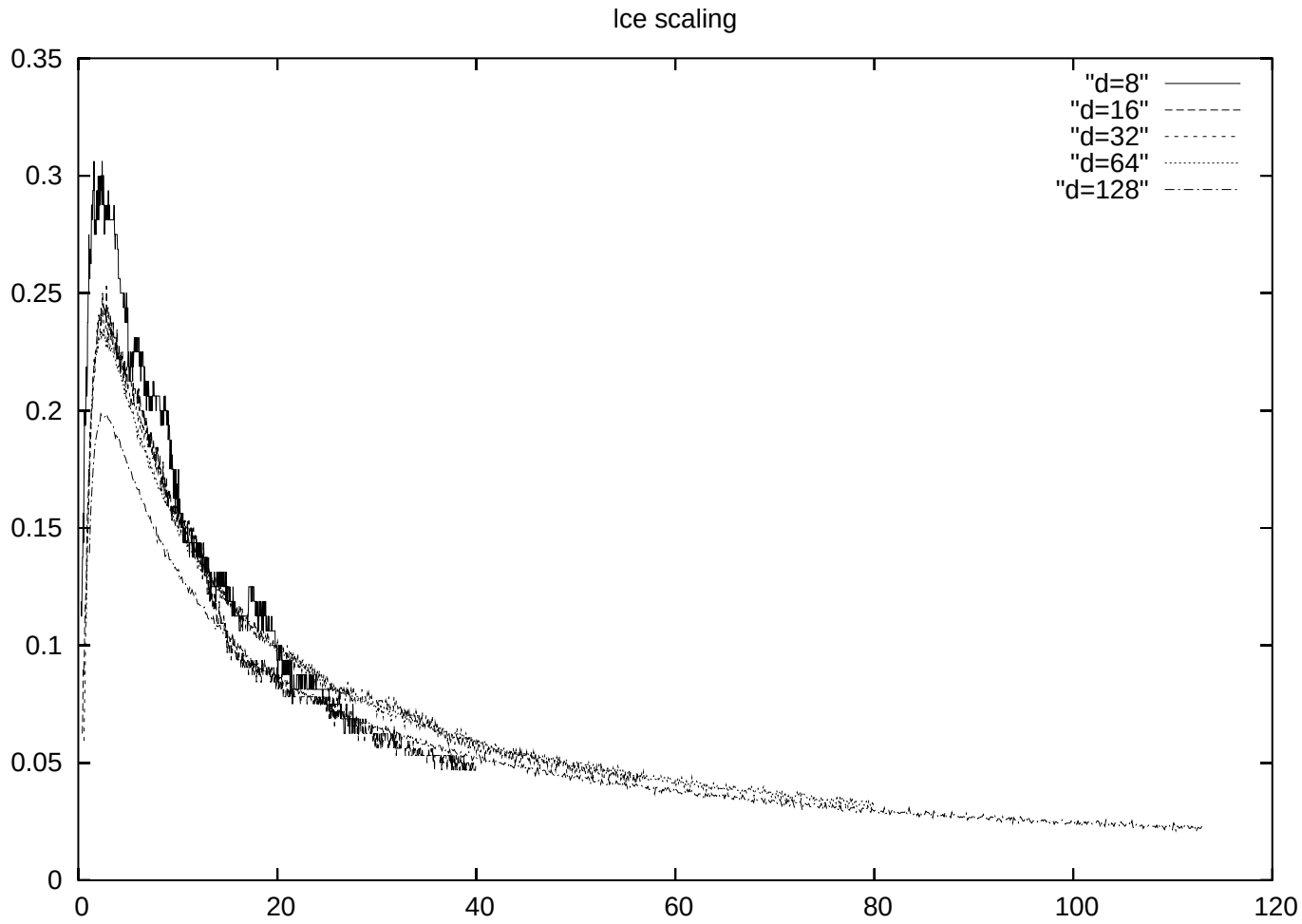
## Review 1-dimensional parameter interval conjectures

a. With arbitrarily large dimension, there will be arbitrarily many positive Lyapunov exponents to provide our needed exponent zero crossings.

b. The exponents become more continuous with respect to variation of the $s$ parameter as the dimension of the dynamical system is increased which helps to guarantee "smooth," transversal zero crossings of Lyapunov exponents as well as no abrupt topological change.

c. The Lyapunov exponent zero crossings become more tightly packed, i.e. the bifurcation chain set is becoming $a$-dense in it's closure.

# Alternative argument with positive LCE scaling



## High-dimensional parameter set generalization

Definitions, conjectures, and an outline of evidence.

Basic results:

- General stability of instability, chaos reigns, periodic windows disappear or become unobservable in portions of parameter space.

- The geometry of the dynamics is not drastically changed upon perturbations of parameters.

## The abstraction of parameter perturbation

Consideration of the map:

$$\phi : R^{N(d+2)+1} \rightarrow \Sigma(G) \tag{7}$$

Things to be concerned about with respect to $\phi$: continuity, affine structure, differentiability (e.g. is $\phi$ $C^0$, $C^r$ $r > 1$, etc).

For now, we will assume that an open ball in $R^{N(d+2)+1}$ yields an open ball in $\Sigma(G)$ (however, $\Sigma(G)$ is infinite dimensional where as $R^{N(d+2)+1}$ is not) .

## Robustness of a property with respect to a space

A property $X$ of some operation or morphism (e.g. a mapping) $Y$ is robust if $X$ is a persistent property in an open neighborhood of $Y$.

To specify robustness I need:

- the property;
- the mapping;
- the open set upon which the property is persistent;

## $k$-degree LCE stability

**Definition 7 (degree $k$ Lyapunov exponent equivalent)** *Assume two discrete-time maps $f$ and $g$ (networks) as previously defined from a compact set to itself. The mappings $f$ and $g$ are called $k$ Lyapunov equivalent if $f$ and $g$ have the same number of Lyapunov exponents and if $f$ and $g$ have the same number of positive Lyapunov exponents.*

**Definition 8 (Robust chaos of degree $k$)** *Assume a discrete time map $f$ (network) as previous defined mapping a compact set to itself. The map $f$ has robust chaos of degree $k$ if there exists a $p$-dimensional subset $U \in R^p$, $p \in N$, such that, for all $\xi \in U$, and for all initial conditions of the map $f$ at $\xi$, $f$ retains $k$ positive Lyapunov exponents. In other words, $f$ is $k$-Lyapunov equivalent on the subset $U$.*

## Conjectures

**Conjecture 3 (Robust chaos with respect to parameter perturbation in $R^p$)** *Assume $f$ is a mapping (neural network) as previously defined with a sufficiently high number of dimensions, $d$, and let $p = N(d+2)+1$. There will exist a open set $V$ in $R^p$ (with significant Lebesgue measure) of parameter space for which chaos will be a robust dynamic.*

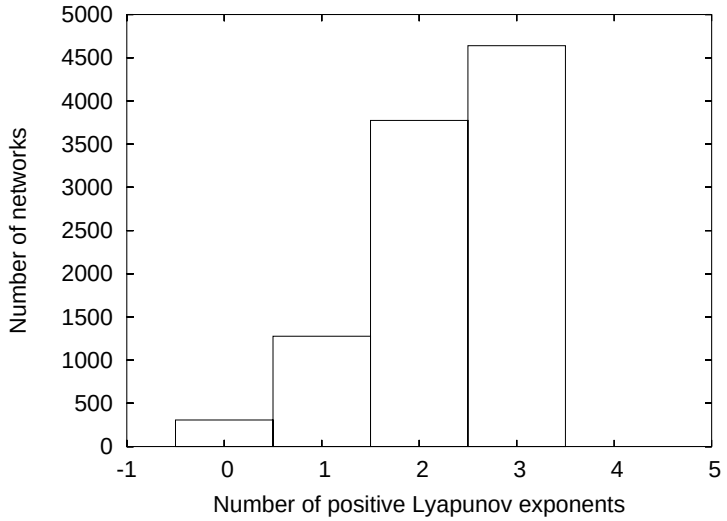
**Conjecture 4 ($k-$robust chaos in large dynamical systems)** *Assume $f$ is a mapping (neural network) as previously defined with a sufficiently high number of dimensions, $d$, and an arbitrarily high number of parameters (neurons) $p = N(d+2)+1$. There will exist an open interval (with significant Lebesgue measure) in parameter space ($R^p$) for which chaos will be robust of degree $k$ with $k \to \infty$ as $d \to \infty$.*

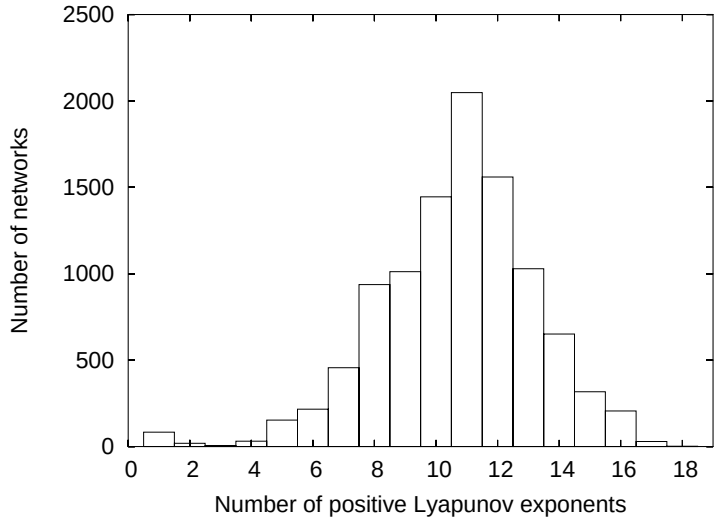
**Conjecture 5 (Periodic window probability diminishing)** *Assume $f$ is a mapping (neural network) as previously defined with a sufficiently high number of dimensions, $d$. There will exist a set $V \in R^p$ (again let $p = N(d+2)+1$) of parameter space such that there will not exist periodic windows on a positive Lebesgue measure set within $V$.*

# $k$-degree LCE stability

Point: with respect to the distribution of positive exponents, the mean must increase, the variance must not explode.

Histograms of the number of positive Lyapunov exponents for $d = 8$ and $d = 64$.



## Characterizing $k$

Characterization the degree of $k$ LCE stability:

i. the degree of $k$-degree LCE stability for the ensemble could be defined as the minumum number of positive exponents for an ensemble of perturbed networks. $k = 1$ at $d = 64$ at $s = 3$

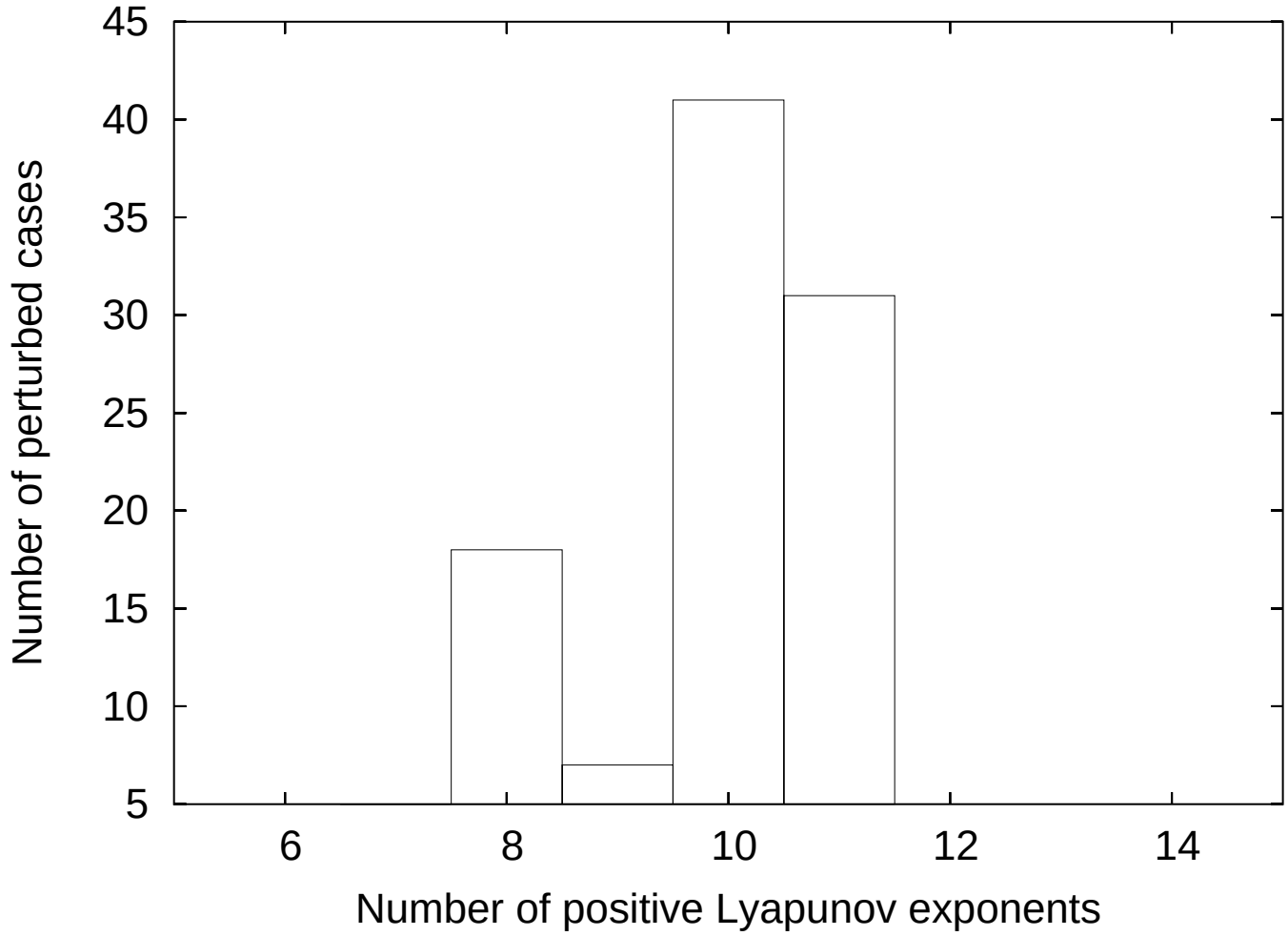
ii. the degree of $k$-degree LCE stability for the ensemble could be defined as the mean number of postive Lyapunov exponents minus the standard deviation about the mean. $k = 1$ at $d = 4$, $k = 8$ for $d = 64$ at $s = 3$

iii. the degree of $k$-degree LCE stability for the ensemble could be defined as the lower boundary of the curve under which 99 percent of the area of the distribution of the number of positive Lyapunov exponents is contained. $k = 1$ at $d = 32$, and $k = 5$ at $d = 64$ for $s = 3$

**$k$ is increasing with $d$**

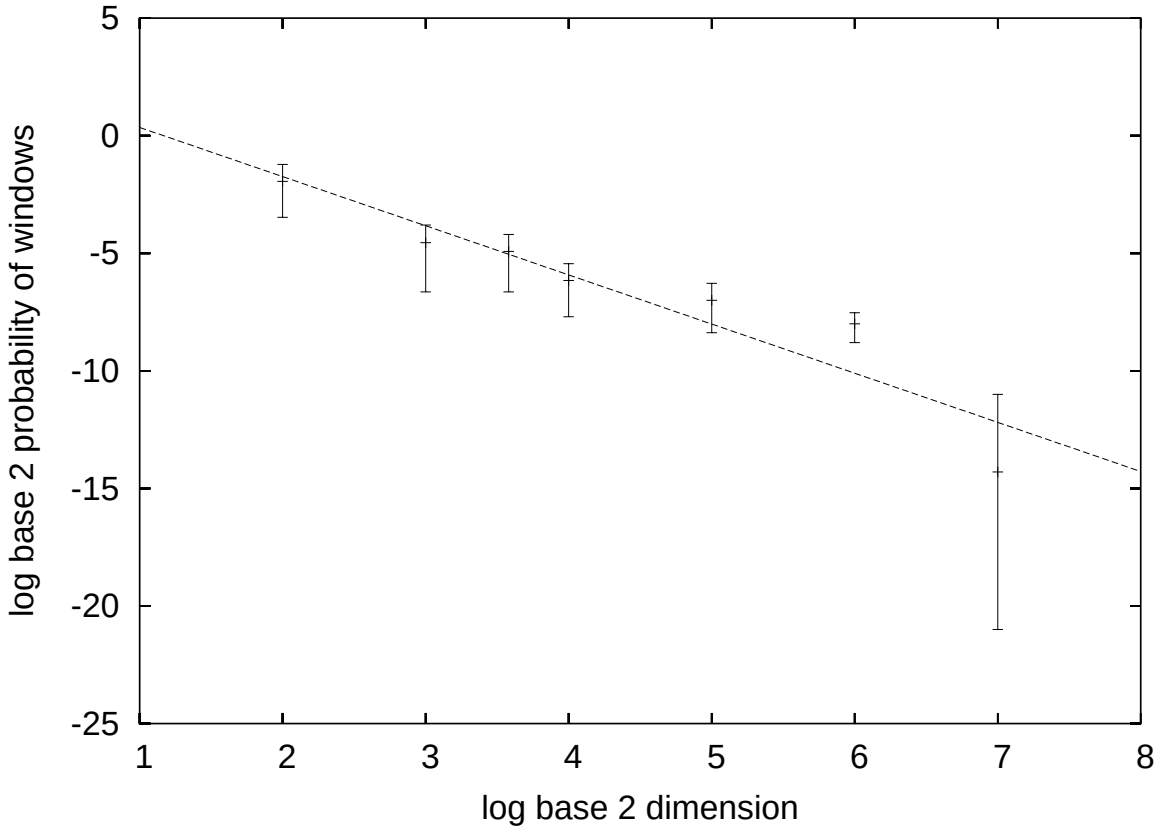
## Example: 64-dimensional network

The histogram of the number of positive Lyapunov exponents for a typical 64-dimensional network perturbed 100 times.



# Probability of periodic windows decreases with dimension like $\sim \frac{1}{d^2}$

Log probability of the existence of periodic versus log of dimension for 500 cases per $d$. Each case has all the weights perturbed on the order of $10^{-3}$ 100 times per case. The line of best is $\sim \frac{1}{d^2}$.



## Periodic windows summary:

Probability that a network has a window is decreasing much more slowly than the probability of observed periodic orbits in networks with observed periodic orbits.

As the dimension of a dynamical system is increased, few windows are observed, and those windows are concentrated in networks with many windows.

## Conclusions with respect to region IV:

As the dimension is increased, chaos becomes dynamic type which is robust with respect to parameter change.

Topological change, i.e. a change in the number of positive Lyapunov exponents and hence a change in the number of global stable and unstable manifolds, is much less drastic versus parameter variation as the dimension of a dynamical system is increased.

Periodic orbits are much less common in high-dimensional, chaotic dynamical systems, and observable periodic orbits concentrated in a smaller set of networks as the dimension of the network is increased.

## Future and work:

Current projects

1. Stability of the LCE algorithm.

2. Analysis of region II — the routes to chaos region (analytical and numeric).

3. Scaling with respect to the LCE spectrum.

4. Scaling with respect to the number of positive LCEs versus s.

5 Linking Takens embedding theorem to the neural network approximation theorems.

Future projects

1. Uniform and non-uniform partial hyperbolicity of high-dimensional neural networks. Compare this with results of both Pesin and Bonatti and Viana regarding SRB measures and Lyapunov exponents.

2. Basins of attractors, existence of Milnor attactors, relation with the finite SRB measure conjecture of Palis and Milnor attractor results of Kaneko;

3. A symbolic dynamics, anti-integrable limit study of region $V$. In other words, a detailed study of the transition from chaos to finite state orbits.

4. Direct connection to nature: train networks on high-dimensional experimental and numerical data sets, study weight distributions.

5. Forever transient — a generalized notion of dynamic stability in systems never allowed time to converge to an ergodic-type limit: e.g. time-evolution of weight distribution;

6. Robustness with respect to weight distributions: increases generality; useful for comparison with fitted networks;