Persistent Chaos in High-Dimensional Neural Networks

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

April 15, 2005

Outline:

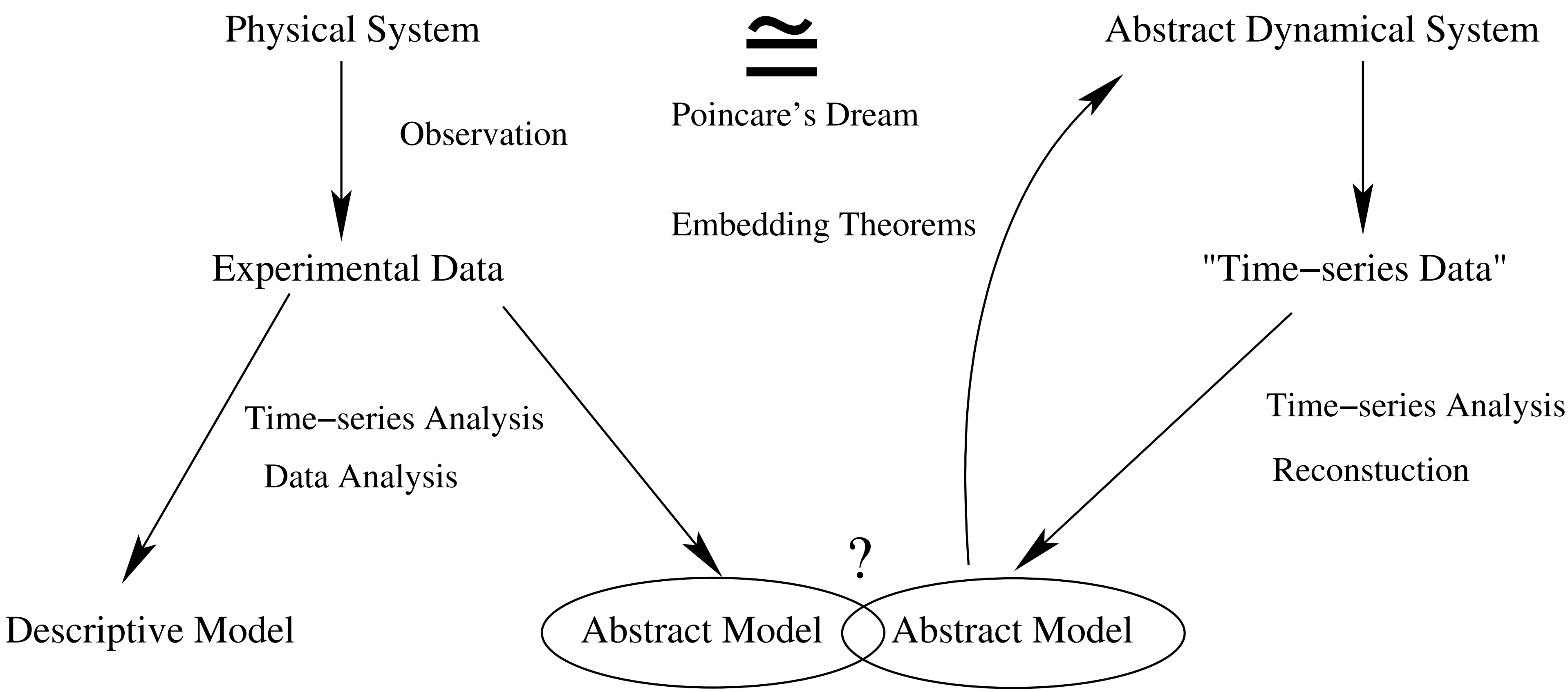
- Introduction and motivation
- Mathematical versus computational dynamics
- Neural networks: a "universal" function space
- Stratification of parameter space
- Persistence of chaos: Dynamics of the bifurcation chain region
- Future work

Poincaré 's Dream

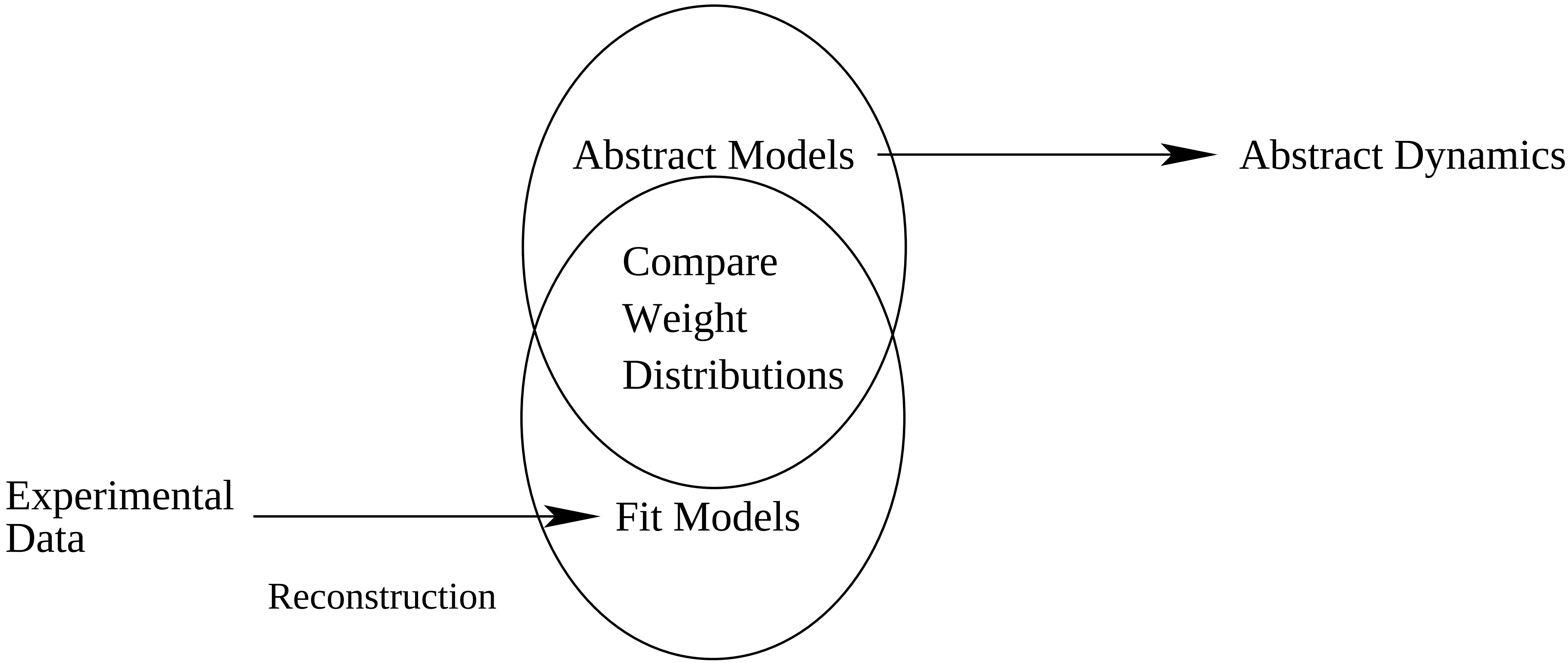
We want insight and understanding of the natural world.

We use ODE's, PDE's, maps and statistical models to reconstruct and model nature.

Split the world into dissipative and non-dissipative dynamics — then the study the space of models will yield insight into how nature behaves and is structured.


Physical System
≅
Abstract Dynamical System
Observation
Poincare's Dream
Embedding Theorems
Experimental Data
"Time−series Data"
Time−series Analysis
Data Analysis
Time−series Analysis
Reconstuction
?
Descriptive Model
Abstract Model
Abstract Model

The Link


Abstract Models
Abstract Dynamics
Compare
Weight
Distributions
Experimental
Data
Fit Models
Reconstruction

Ingredients:

Time-series analysis and reconstruction

Numerical stability of models

Embedding theorems

Statistical understanding of computational models

Abstract dynamics

Computational versus Mathematical Dynamical Systems: the approach

Abstract dynamics:

- Select a manifold (e.g. T^2)
- Impose dynamic types (e.g. Anosov diffeomorphims)
- Classify genericity or persistence of dynamics types

Computational dynamical systems:

- Select a dynamical system (with finitely many parameters)
- Vary parameters, observe geometry and dynamics types
- Classify behaviors based upon observations — common behaviors defined by statistics

Computational versus mathematical dynamical systems: results

Mathematical dynamical systems:

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

Computational dynamical systems

- The number of dimensions of the dynamical system matters; high and low-dimensional dynamical systems are very different.
- Parameters matter; the practical effects of parameters with respect to dynamical systems is significant.

Mathematical dynamics intuition

Everything is Anosov (cat map), “stacked” Anosovs, etc.

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

Periodic “windows” are likely rare for $d \geq 3$. (Pugh)

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

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

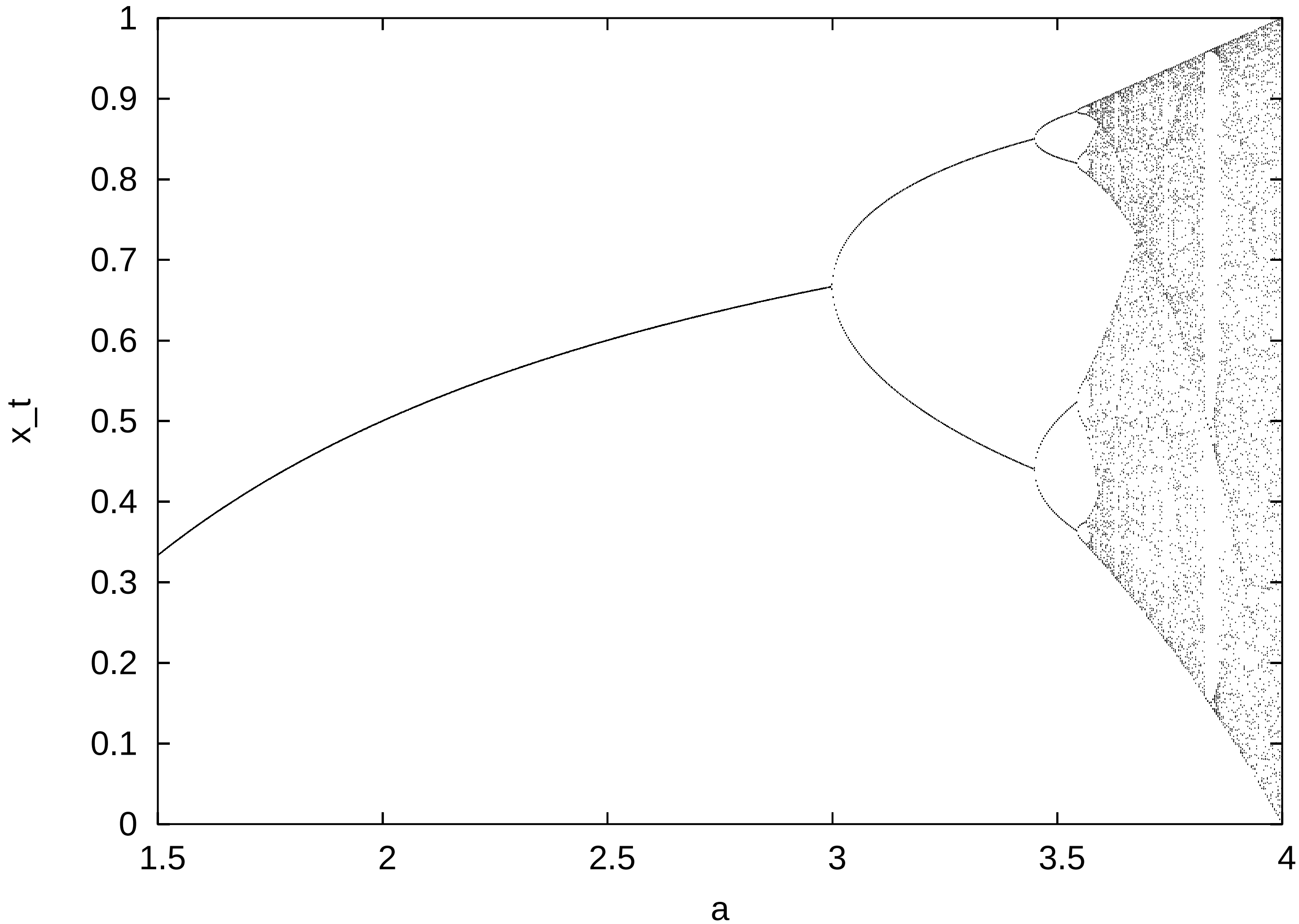
Computational and low-dimensional intuition

Logistic map: real Fatou lemma, Jakobson absolute continuity;

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

Windows of periodic orbits amidst windows of chaotic orbits are common in parameter space;

Standard logistic map



High-Dimensional observations (from our study):

Routes to chaos:

- As $d \to \infty$, the probability of the first bifurcation being Neimark-Sacker (bifurcation to quasi-periodic orbit) $\to 1$.

- As $d \to \infty$, the dominant route to chaos from a fixed point along a one-dimensional interval is the quasi-periodic route involving n-tori ($n < 3$);

Chaotic region:

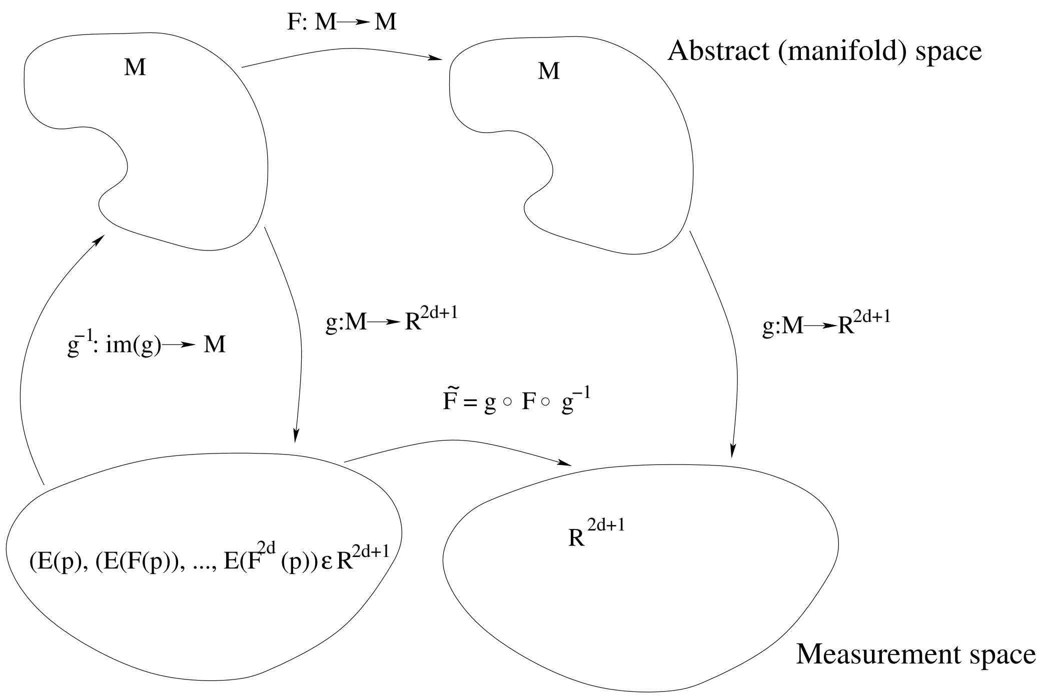
- As $d \to \infty$ chaos becomes persistent with respect to parameter variation.

- As $d \to \infty$, topological variation is common but occurs on a zero Lebesgue measure set in parameter space.

Scalings (with relatively constant maximum entropy)

- As $d \to \infty$, the probability of a network being chaotic over some portion of its parameter space $\to 1$.
- The maximum $D_{KY} \sim \frac{d}{2}$.
- As $d \to \infty$, the probability of the existence of periodic windows in parameter space $\sim d^{-2}$.
- As $d \to \infty$, the maximum number of positive exponents $\sim \frac{d}{4}$.

Neural networks and the Takens embedding theorem



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:

$$g(x_t) = (E(x_t), E(F(x_t)), \ldots, E(F^{2d}(x_t))) \quad (1)$$

Artificial neural networks

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

$$\Sigma(G) \equiv \{\gamma : R^d \rightarrow R | \gamma(x) = \sum_{i=1}^{N} \beta_i G(\tilde{x}^T \omega_i)\} \tag{2}$$

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 $G : R^d \rightarrow R$ is the hidden layer activation function (or neuron).

$$x_t = \beta_0 + \sum_{i=1}^{N} \beta_i G\left(s\omega_{i0} + s\sum_{j=1}^{d} \omega_{ij} x_{t-j}\right) \tag{3}$$

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

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.

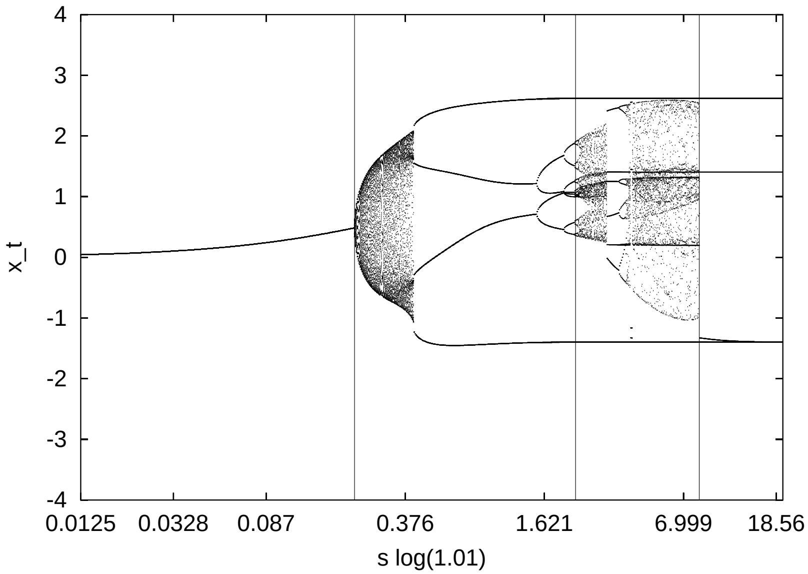
Stratification of the s parameter interval

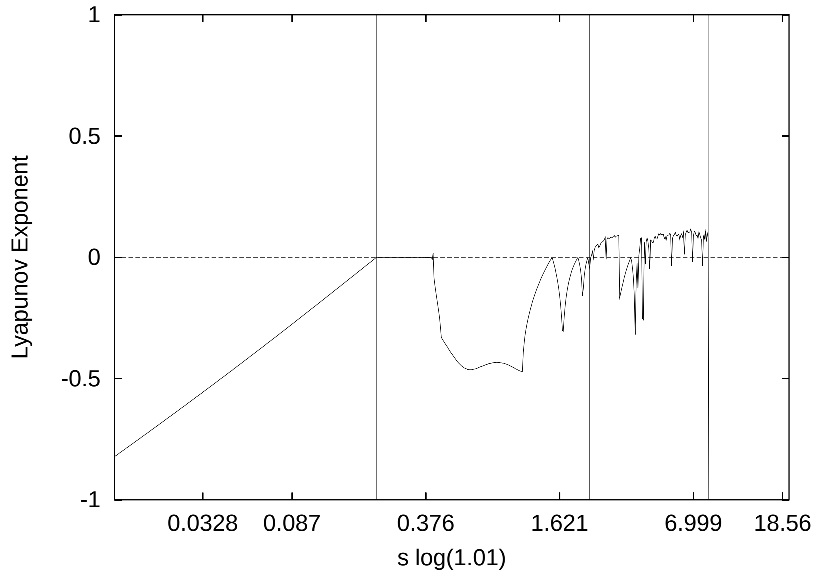
Bifurcation diagrams along an interval of the s parameter.

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: $\mathrm{I} \to \mathrm{II} \to \mathrm{IV} \to$ V.

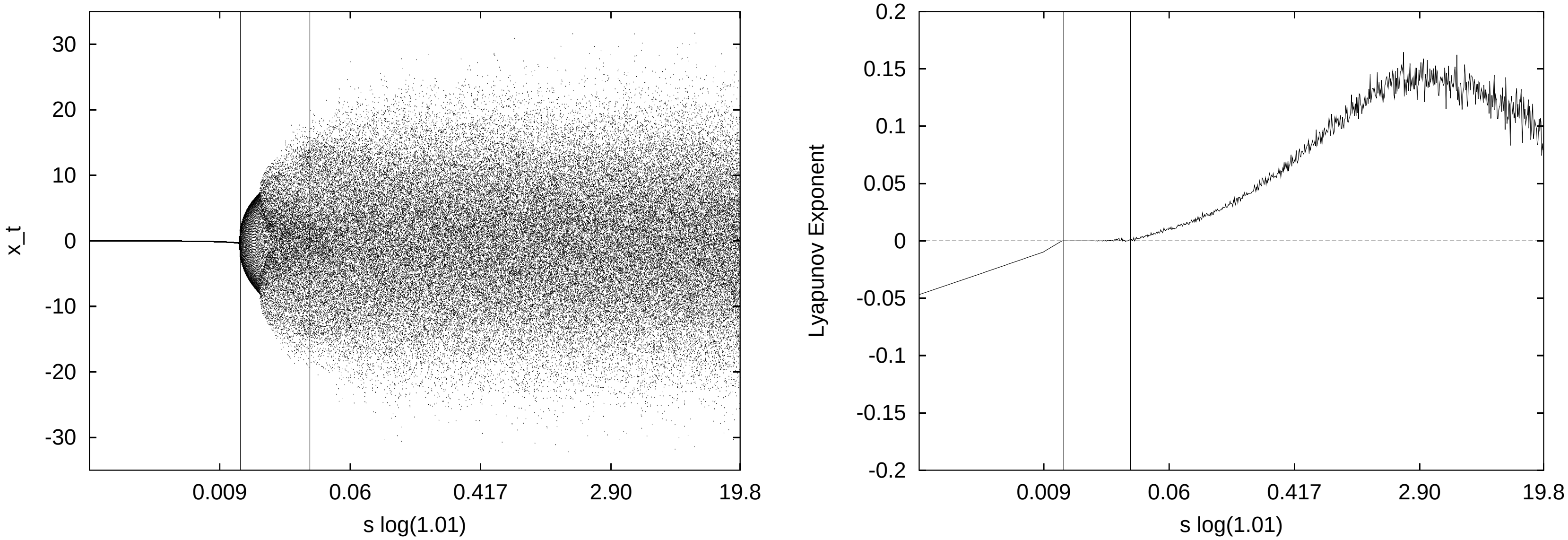
Bifurcation diagram with the largest Lyapunov exponent for $N = 4$ and $d = 4$ with regions I, II, III, and V — no region IV.





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

The prototypical scenario



Persistent chaos: Dynamics in region IV

Formulation of a conjecture along the 1-dimensional s-interval

List of properties we require for the conjecture

Evidence

High-dimensional generalization of parameter space

New definition: Persistent chaos of degree-p

Formulation of conjectures

Evidence

Lyapunov characteristic exponent (LCE) spectrum

LCEs: $\chi_1 \geq \chi_2 \geq \ldots \geq \chi_d \in R$, where indexing is chosen to give a monotonic ordering.

- χ_i represents the relative expansion ($\chi_i > 0$) or contraction ($\chi_i < 0$) in a particular direction (on a particular manifold).
- the number of positive exponents = the number of expanding (stretching) directions.
- the number of negative exponents = the number of contracting (squashing) directions.

LCE spectrum for a typical *individual* network with 32 neurons and 64 dimensions.

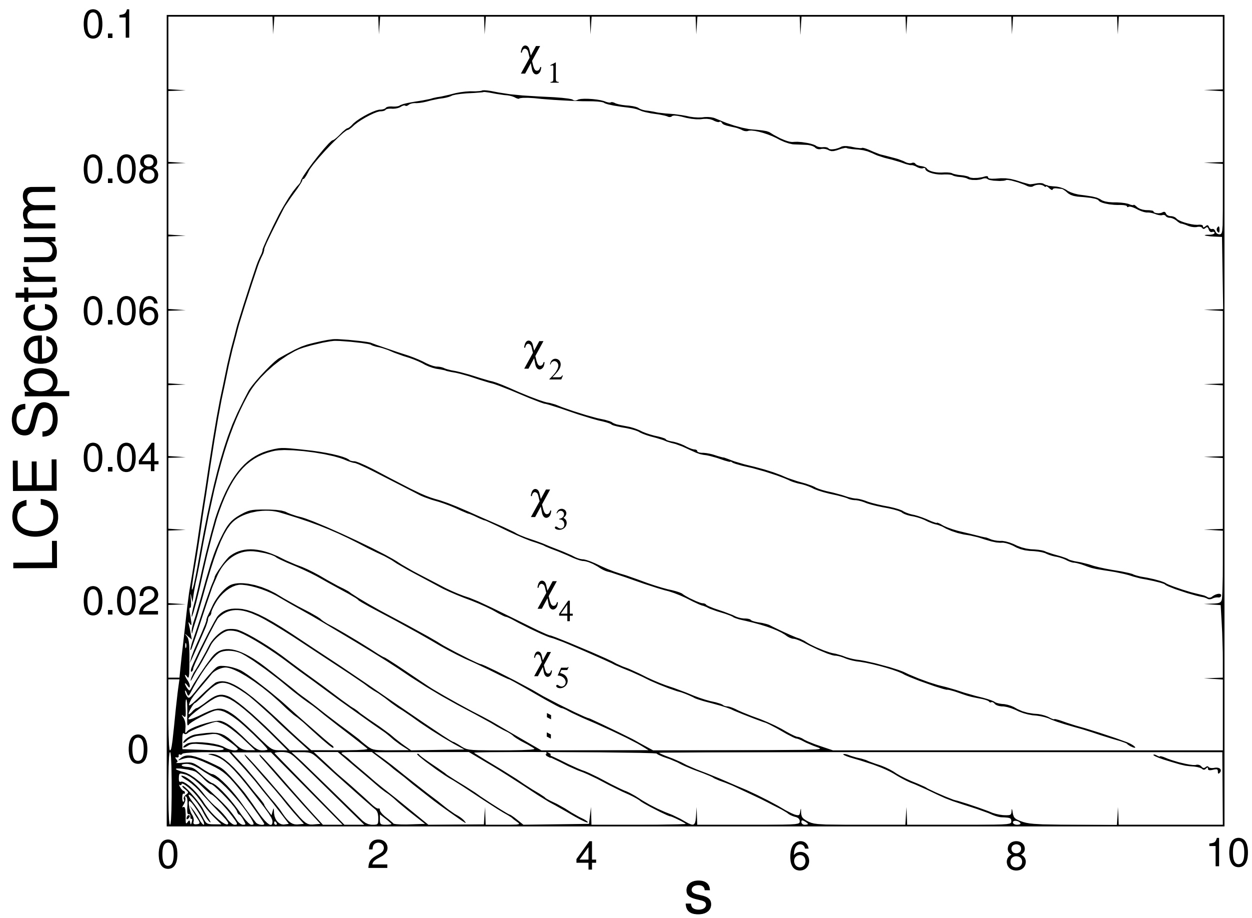
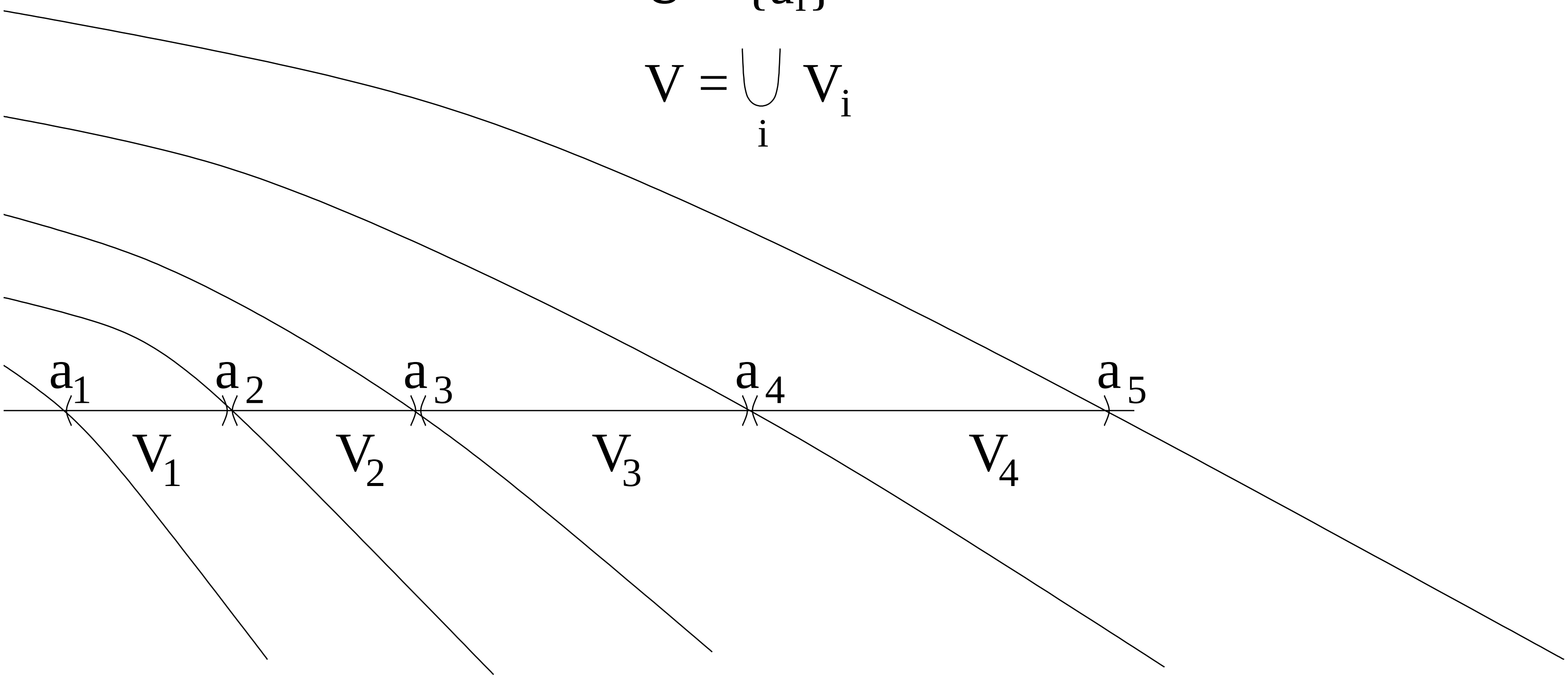


Diagram for bifurcation chain sets

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.

$$U = \{a_i\}$$

$$V = \bigcup_i V_i$$



Two conjectures

Conjecture 1 (Existence of a Codimension ϵ 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 δ_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 $s + \delta$.*

Topological variation is inevitable

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) — periodic windows vanish

List of properties for 1-dimensional parameter interval picture

The following properties must increase with dimension:

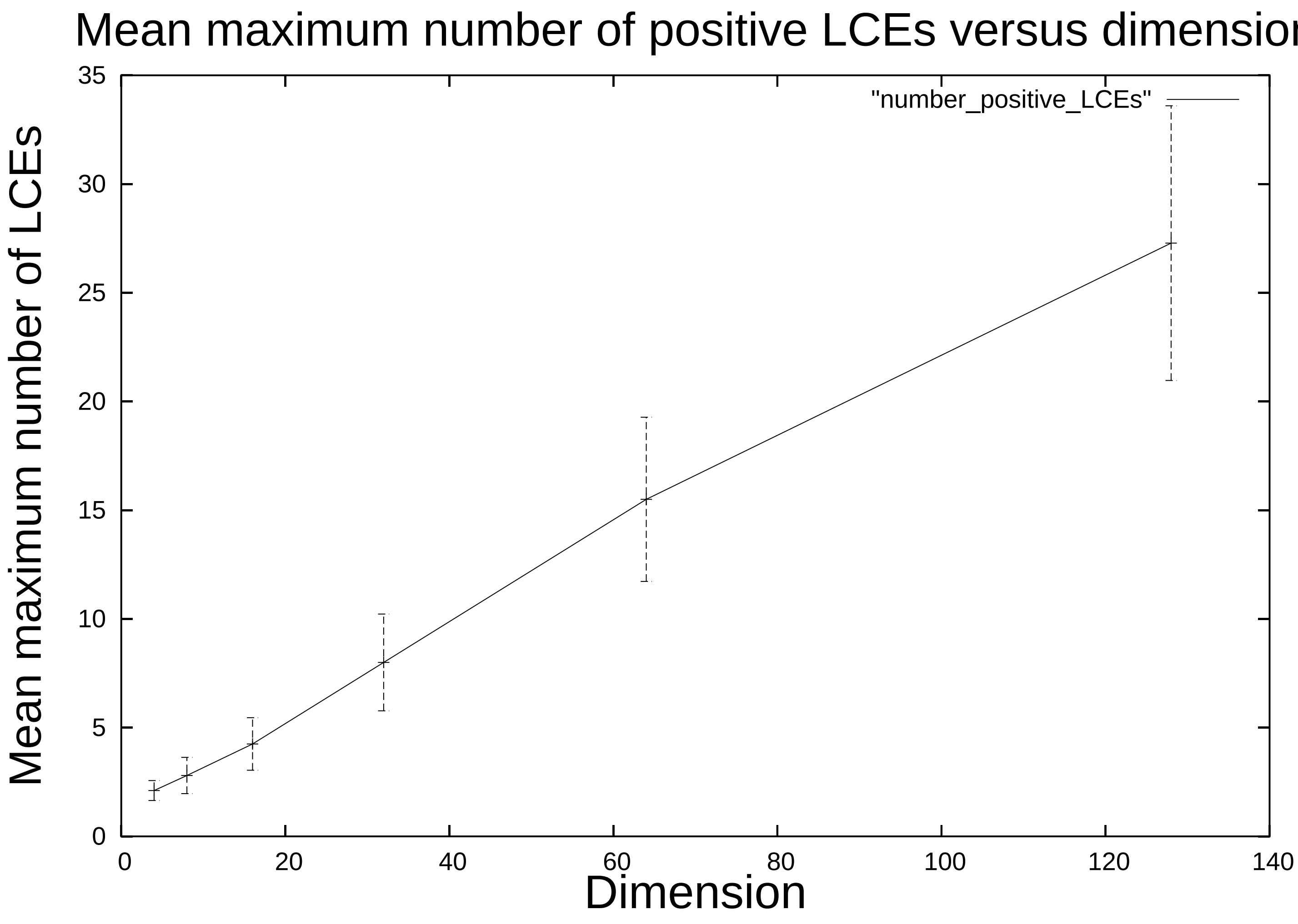
a. the number of positive exponents;

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

c. Asymptotic density of transversal Lyapunov exponent zero crossings;

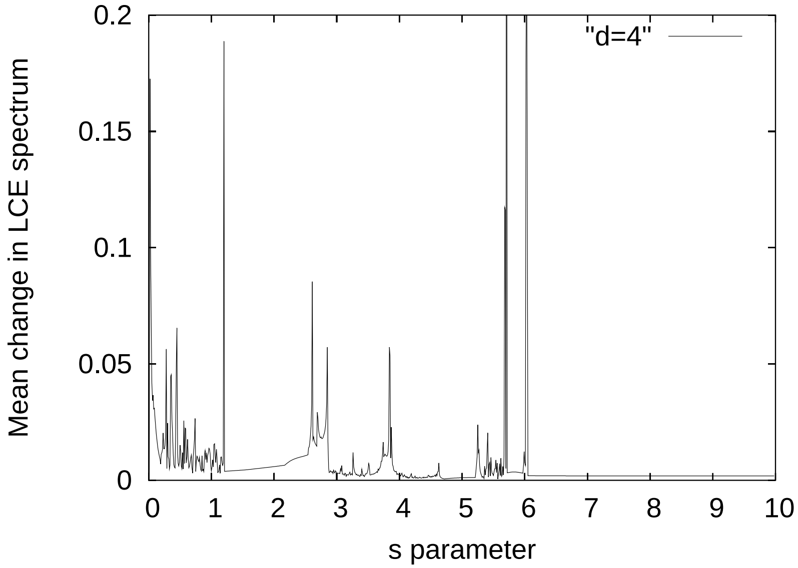
Mean max number of positive Lyapunov exponents

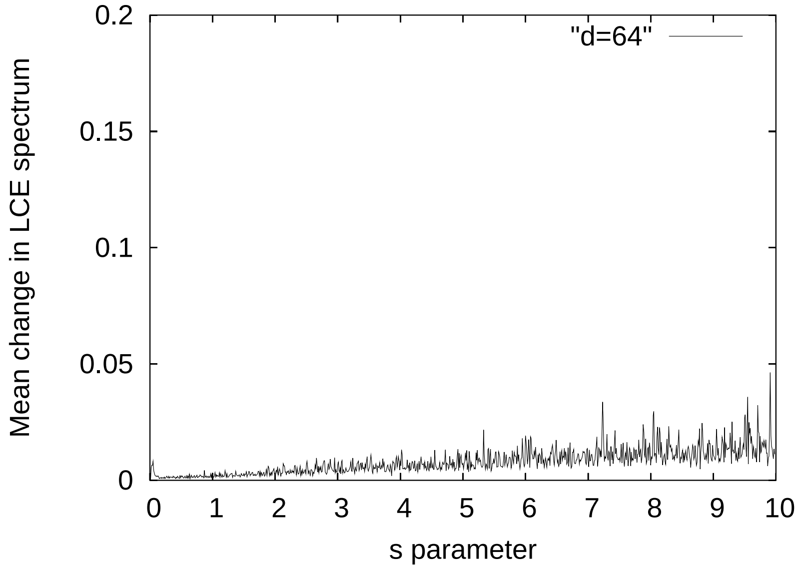
Mean maximum number of positive LE's versus dimension, all networks have 32 neurons (slope is approximately $\frac{1}{4}$).



num-**continuity versus s** (individual networks)

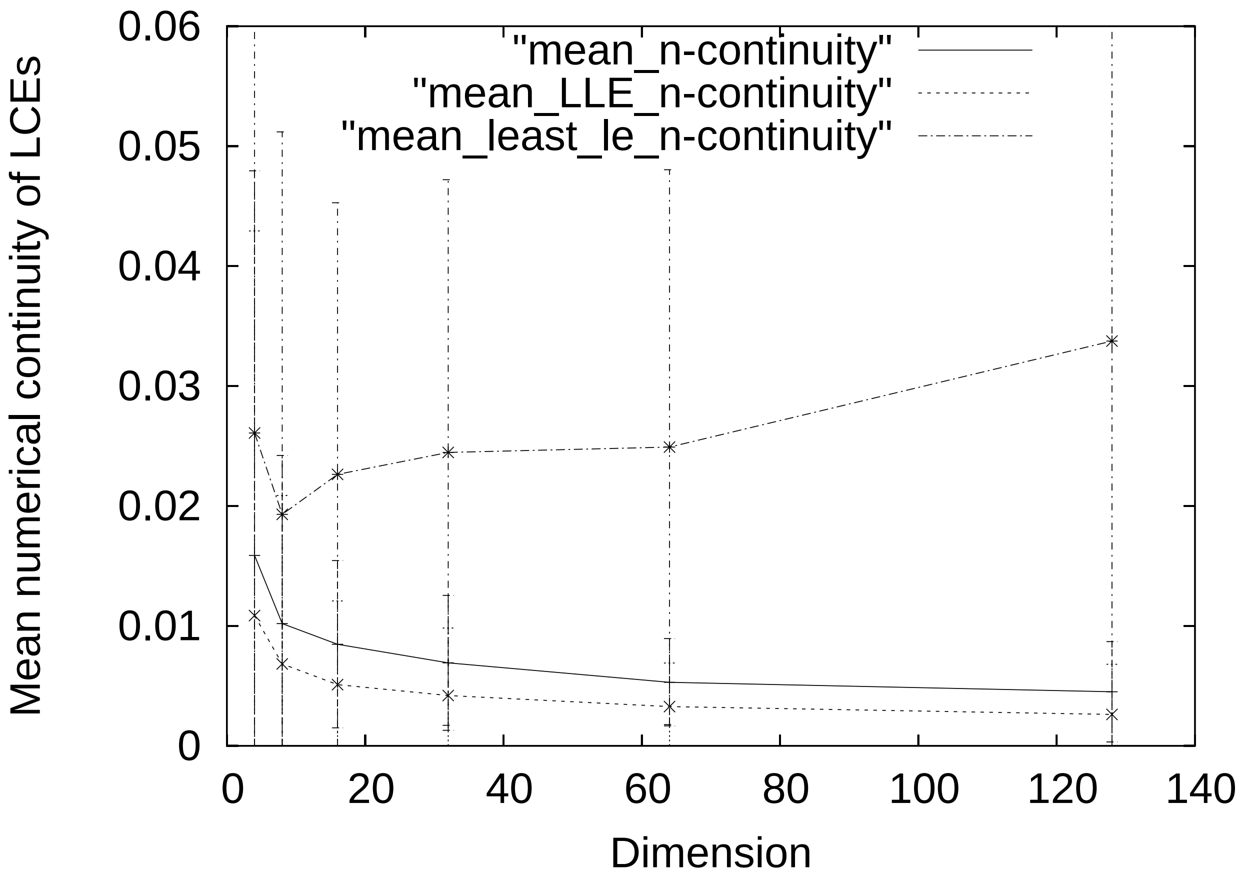
num-continuity (mean of $|\chi_i(s) - \chi_i(s + \delta s)|$ for each i) versus parameter variation: 32 neurons, 4 (left) and 64 (right) dimensions.





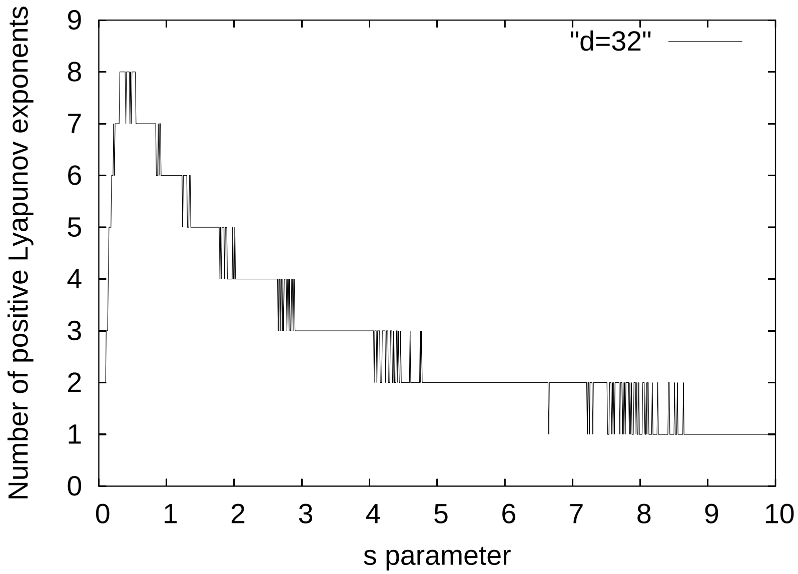
num-**continuity versus dimension** (ensemble of networks)

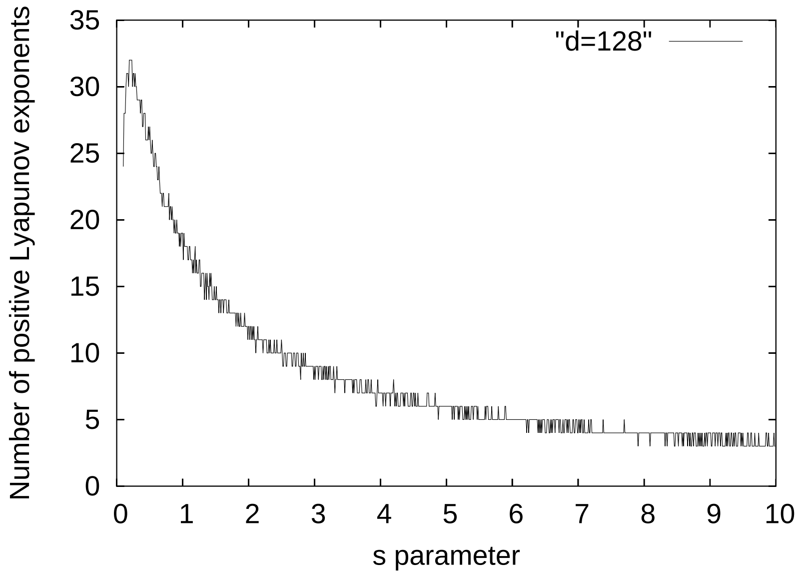
Mean num-continuity, num-continuity of the largest and the most negative Lyapunov exponent of many networks versus their dimension. The error bars are the standard deviation about the mean over the number of networks considered.



Intuition for a-density of zero crossings

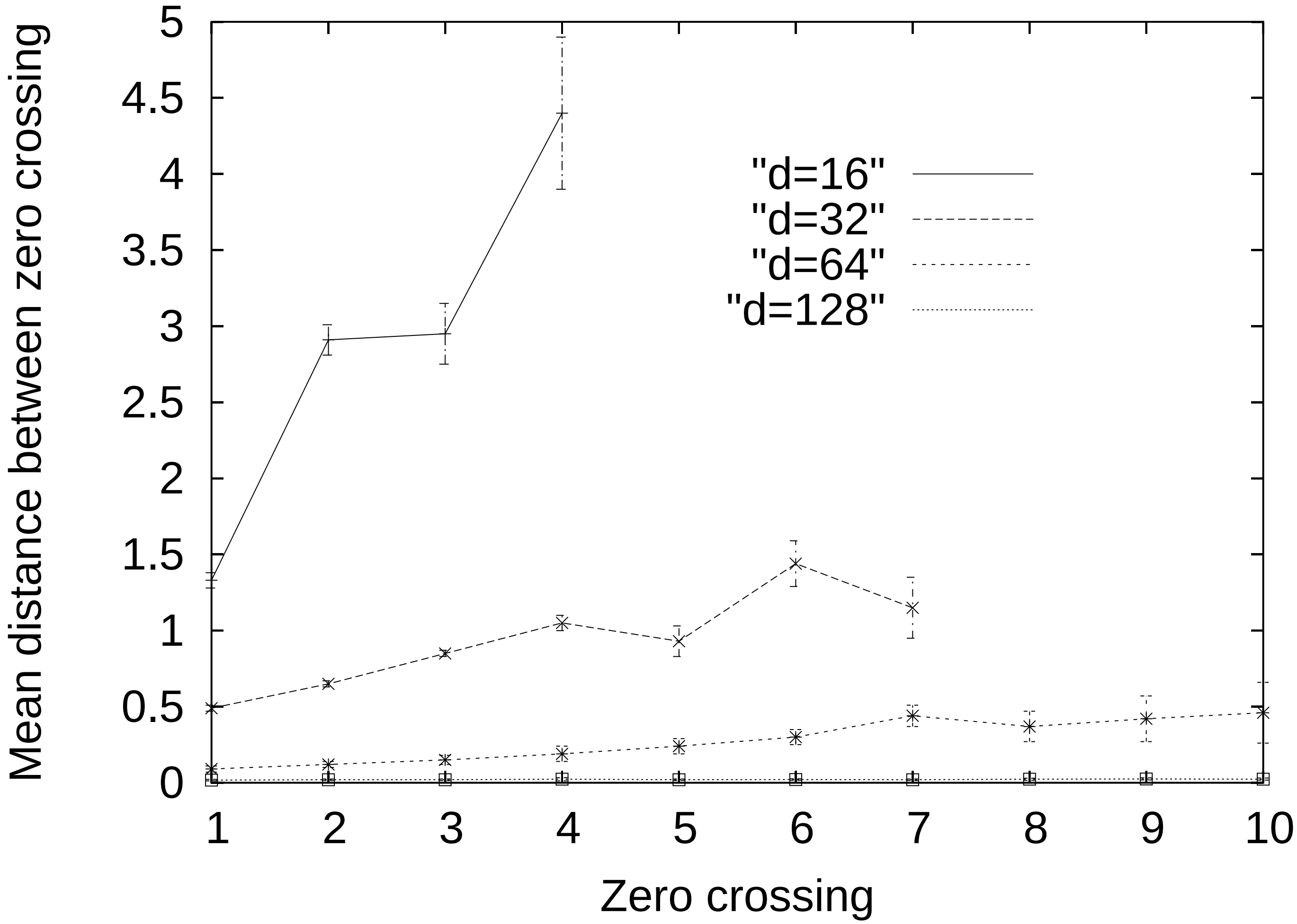
Number of positive LE's for typical individual networks with 32 neurons and 32 (left) and 128 (right) dimensions.





a-density of the first 10 zero crossings

Mean distance (δs) 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. **Number of positive exponents increases with d** — With arbitrarily large dimension, there will be arbitrarily many positive Lyapunov exponents to provide our needed exponent zero crossings.

b. **Continuity increase with d increase** — 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. **Dense zero crossings of LCEs** — The Lyapunov exponent zero crossings become more tightly packed, i.e. the bifurcation chain set is becoming a-dense in it's closure.

High-dimensional parameter set generalization

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:

$$\phi : R^{N(d+2)+1} \to \Sigma(G) \tag{4}$$

Things to be concerned about with respect to ϕ: continuity, affine structure, differentiability (e.g. is ϕ 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) .

p-degree Persistent Chaos

Definition 2 (Degree-p Persistent Chaos) *Assume a discrete-time map f that takes a compact set to itself. The map has persistent chaos of degree-p if there exists an open subset U of parameter space, such that, for all $\xi \in U$ and a given open set $\mathcal{O}$ of initial conditions, $f|_\xi$ retains $p \geq 1$ positive Lyapunov exponents.*

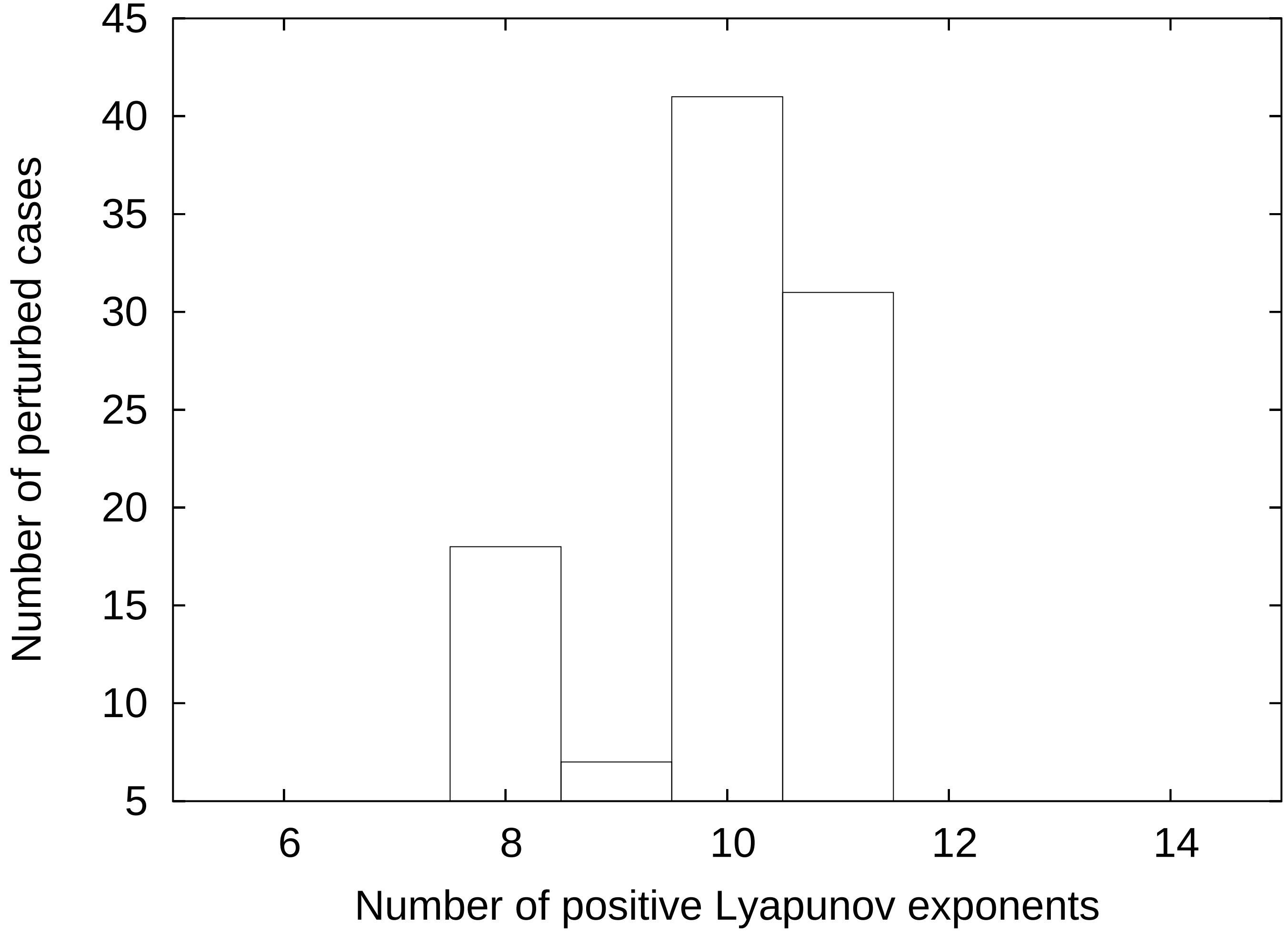
Conjectures

Conjecture 3 (Persistence of chaos) *Assume f is a network with a sufficiently large number d of dimensions and number of parameters $k = n(d+2)+1$. There exists an open set of significant Lebesgue measure in parameter space R^k for which chaos will be degree-p persistent with $p \to \infty$ as $d \to \infty$.*

Conjecture 4 (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.*

Example: 64-dimensional network

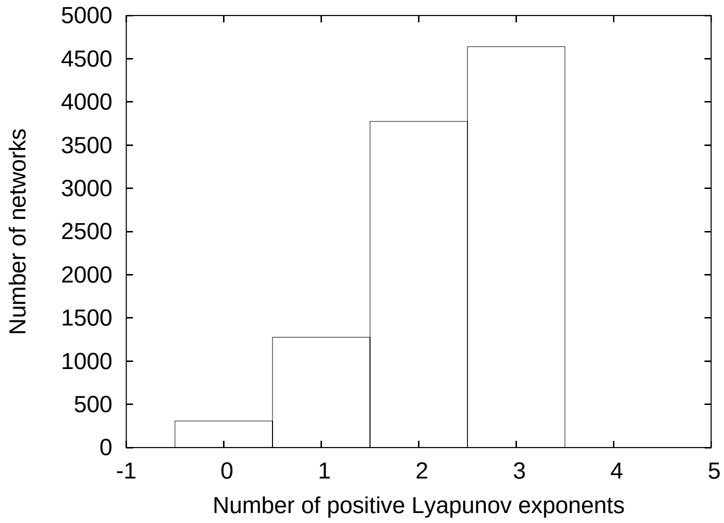
The histogram of the number of positive Lyapunov exponents for a typical 64-dimensional network perturbed 100 times (order: 10^{-3}).

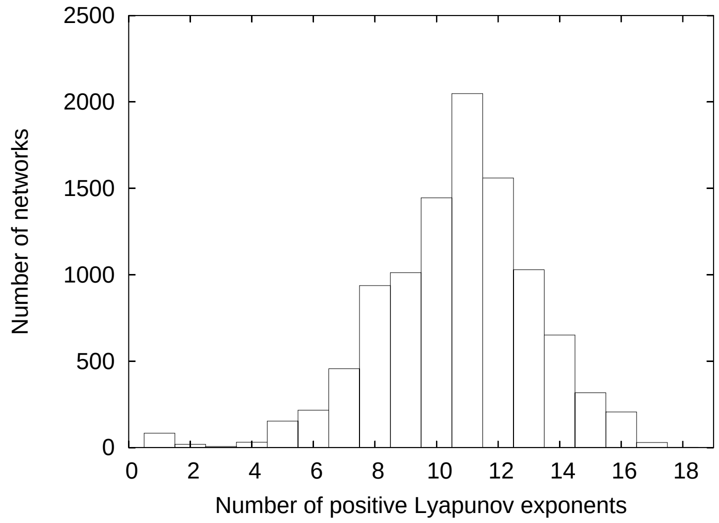


p-degree LCE stability

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 p

The degree of p-degree LCE stability for the **ensemble** could be defined as

i. the minumum number of positive exponents for an ensemble of perturbed networks. $p = 1$ at $d = 64$ at $s = 3$

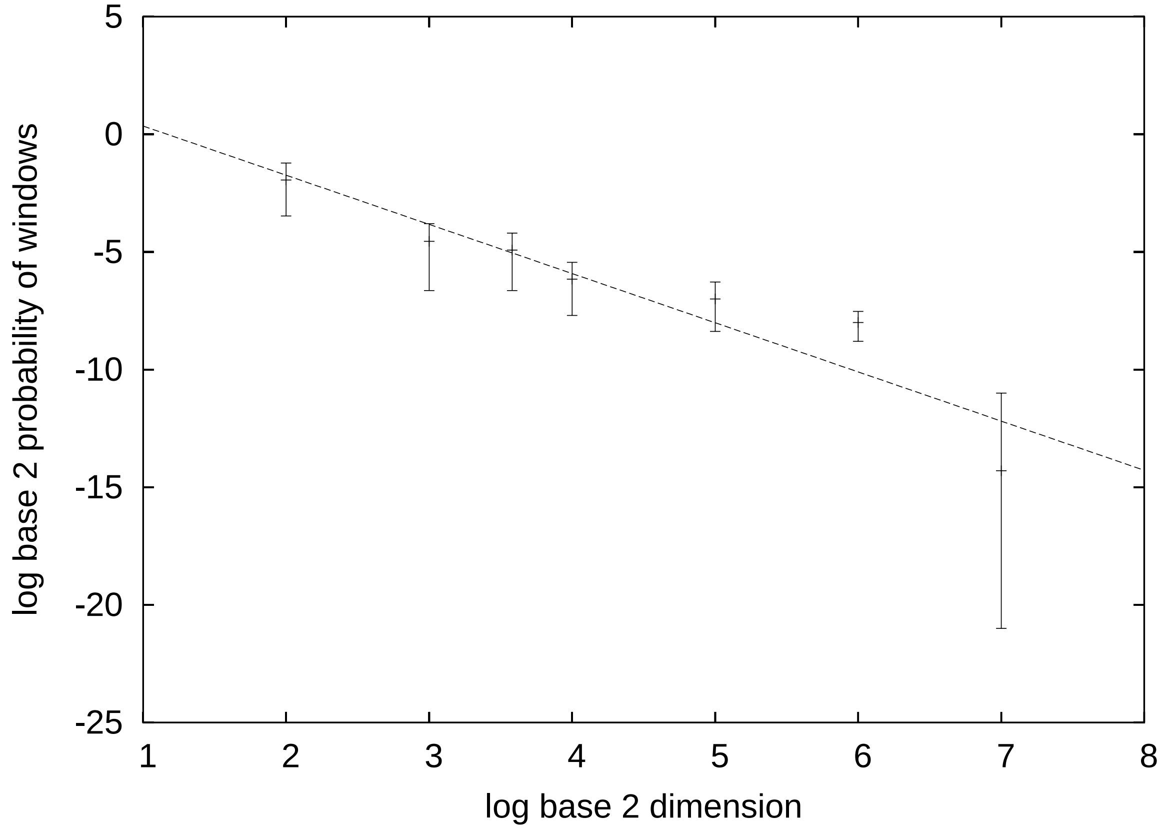
ii. the mean number of positive Lyapunov exponents minus the standard deviation about the mean. $p = 1$ at $d = 4$, $p = 8$ for $d = 64$ at $s = 3$

iii. 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. $p = 1$ at $d = 32$, and $p = 5$ at $d = 64$ for $s = 3$

p is increasing with d

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 $\sim \frac{1}{d^2}$.



Conclusions with respect to region IV:

Chaos becomes persistent with respect to parameter change as $d \to \infty$

Topological change is inevitable with arbitrary parameter variation as $d \to \infty$
(e.g. many precise mathematical notions of dynamic stability are violated)

Period windows vanish as $d \to \infty$

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. Scalings with respect to the LCE spectrum.
4. Scalings with respect to the number of positive LCEs versus s.
5. Scalings with respect to the number of neurons
6. Analysis of the center-bunching hypothesis of Pugh and Shub

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;