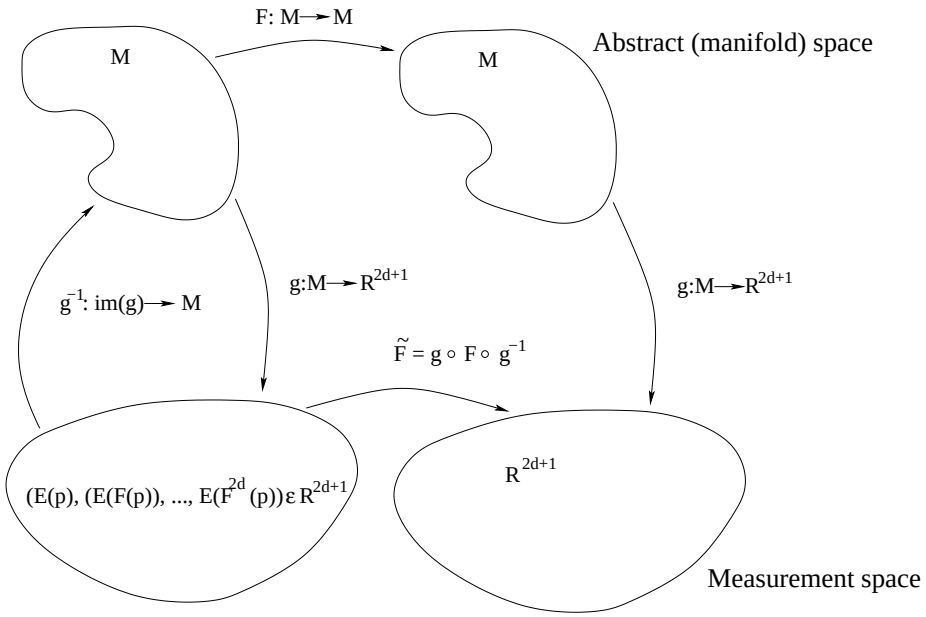
A potpourri of recent results in high-dimensional dynamical systems

D. J. Albers

September 22, 2005



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:

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

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:*

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

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

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

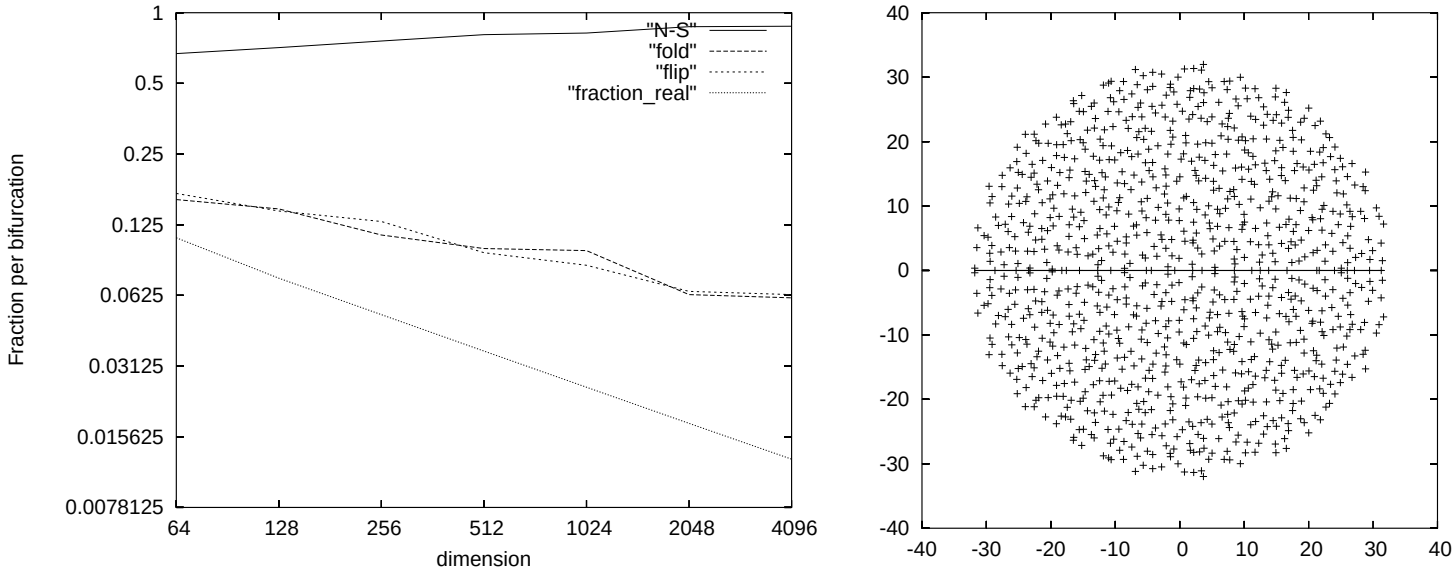
$\omega_{ij} \in N(0, s)$, β_i uniform on $[0, 1]$, $G \equiv \tanh()$, $d =$ number of inputs, $N =$ number of neurons.

Theorem 1 (Circular law (Bai)) *Suppose that the entries of a $n \times n$ matrix M have finite sixth moment and that the joint distribution of the real and imaginary part of the entries has a bounded density. Then, with probability* 1*, the empirical distribution $\mu_n(x, y)$ tends to the uniform distribution over the unit disk in two-dimensional space.*

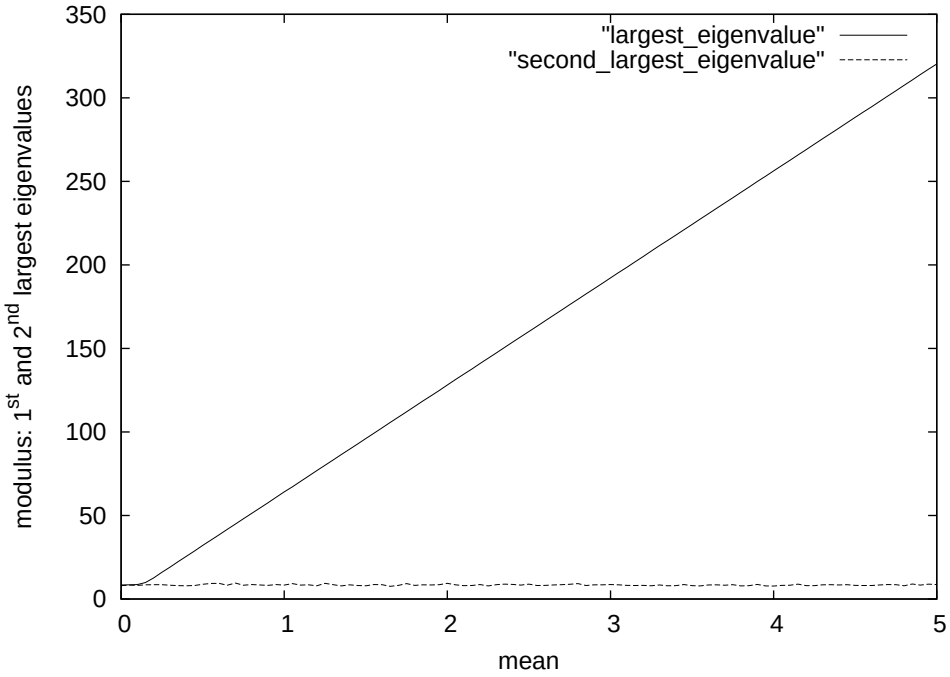
Theorem 2 (Theorem 6.3 **(Edelman))** *The density function $\widehat{\rho}$ converges pointwise to a very simple form as $n \to \infty$:*

$$\lim_{n\to\infty} \frac{1}{n}\widehat{\rho}(\widehat{x}, \widehat{y}) = \begin{cases} \frac{1}{\pi} & \widehat{x}^2 + \widehat{y^2} < 1 \\ 0 & \widehat{x}^2 + \widehat{y^2} > 1 \end{cases} \tag{4}$$

where $\widehat{\rho}_n$ is simply ρ as a function of $\widehat{x} = \frac{x}{\sqrt{n}}$ and $\widehat{y} = \frac{y}{\sqrt{n}}$. Note that $\frac{\widehat{\rho}(\widehat{x},\widehat{y})}{n}$ is a randomly chosen normalized eigenvalue in the upper half plane.



On the left, the observed probability of each bifurcation was recorded for 1000 matrices with i.i.d., mean zero, variance one, Gaussian elements for each d (in powers of 2) along with the fraction of eigenvalues that are real. On the right is the spectrum of eigenvalues in the complex plane that corresponds to a single 1024×1024 matrix ($d = 1024$)



This figure represents an ensemble of 1000 $d \times d$ matrices with $d = 64$. Depicted are the modulus of largest and second largest eigenvalues. The line representing the modulus of the largest eigenvalue is given by $64m$ while the line for the modulus of the second largest eigenvalue is given by $\sim \sqrt{d}$.

For $|m| > 0.1289$, $\lambda_d(m) = dm$, $\lambda_{d-1} =$ constant.

DF for time-delay dynamical systems

$$Df_x = \begin{bmatrix} a_1 & a_2 & a_3 & \cdots & a_{d-2} & a_{d-1} & a_d \\ 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \ddots & \vdots & & & & \\ 0 & 0 & 0 & \cdots & 0 & 1 & 0 \end{bmatrix}$$

The a_k's are given as:

$$a_k = \frac{\partial x_t}{\partial x_{t-k}} = \sum_{i=1}^{n} \beta_i s w_{ik} sech^2(s w_{i0} + s \sum_{j=1}^{d} w_{ij} x_{t-j}) \qquad (5)$$

Gaussian polynomials (and companion matrices)

$$a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \tag{6}$$

where the a_i coefficients are independent standard normals with mean zero. The expected number of real zeros, E_{real}, as $n \to \infty$ is given by the formula:

$$E_{real}(n) = \frac{2}{\pi}\log(n) + C_1 + \frac{2}{n\pi} + O(1/n^2) \tag{7}$$

where $C_1 = 0.6257358072$

Theorem 3 *Let $v(x) = (f_0(x), \ldots, f_n(x))^T$ be any collection of differentiable functions and $a_0, \ldots, a_n$ be the elements of a multivariate normal distribution with mean zero and covariance matrix C. The expected number of real zeros on an interval (or measurable set) I of the equation*

$$a_0 f_0(x) + a_1 f_1(x) + \cdots + a_n f_n(x) = 0 \tag{8}$$

is

$$\int_I \frac{1}{\pi} ||w'(x)|| dx, \tag{9}$$

where w is given by

$$w(x) = \frac{C^{1/2} v(x)}{||C^{1/2} v(x)||} \tag{10}$$

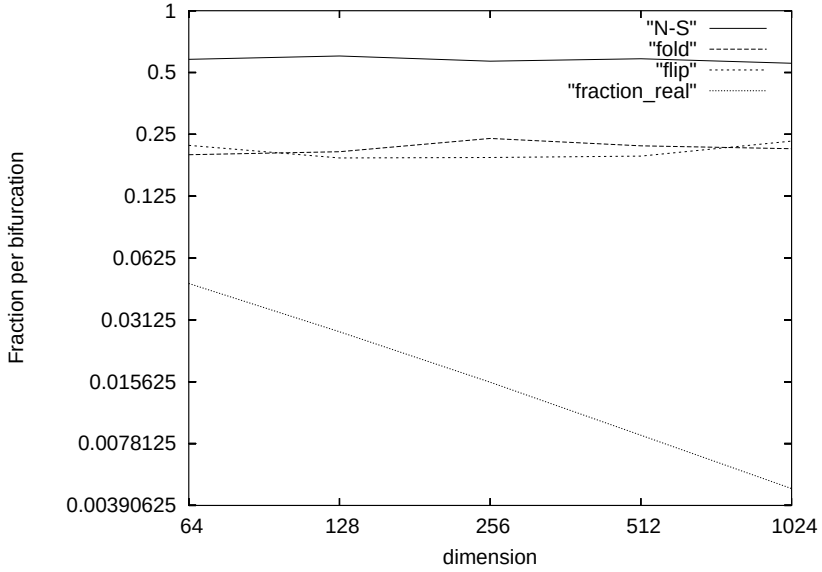
In logarithmic derivative notation this is

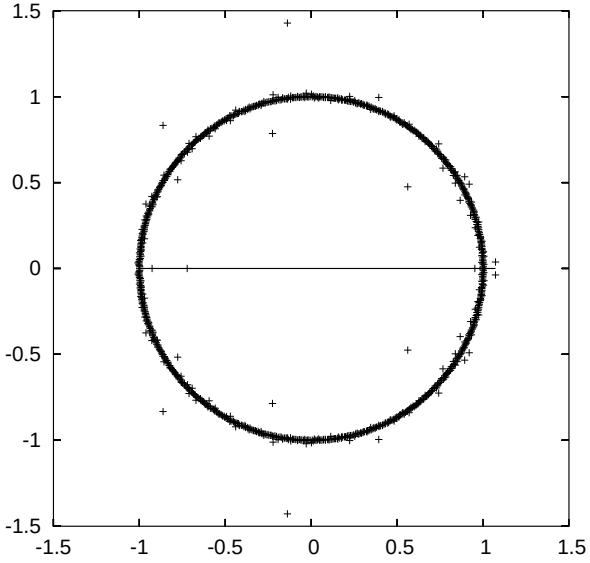
$$\frac{1}{\pi} \int_I (\frac{\partial^2}{\partial x \partial y} (\log(v(x)^T C v(y))|_{y=x=t}))^{1/2} dt \tag{11}$$

There are many applications of this profound theorem presented in (Edelman), one of particular interest is an application to a trigonometric series such as

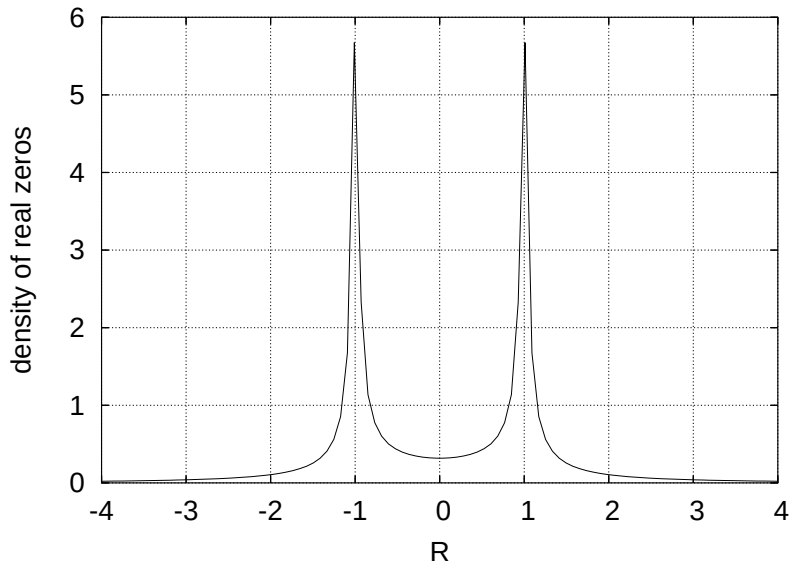
$$\sum_{k=0}^{\infty} a_k \cos(\nu_k \theta) + b_k \sin(\nu_k \theta) \tag{12}$$

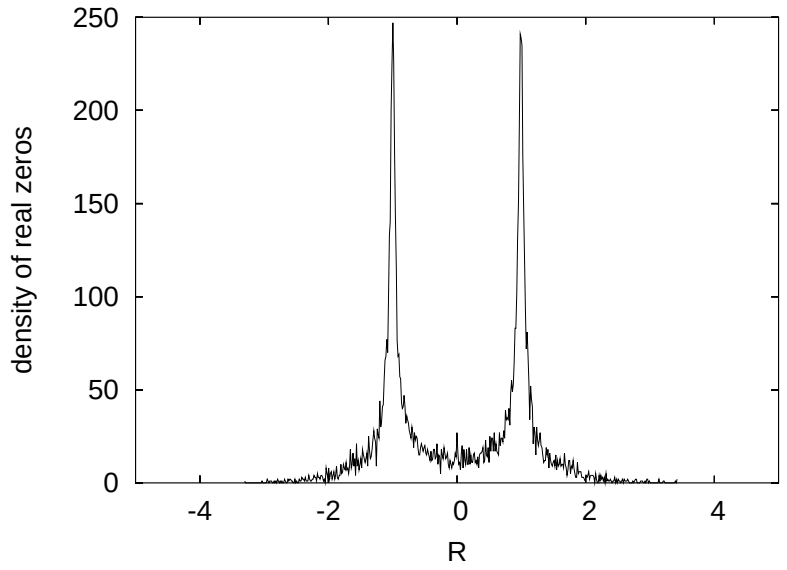
where a_k and b_k are independent normal random variables with mean zero and variance σ_k^2.



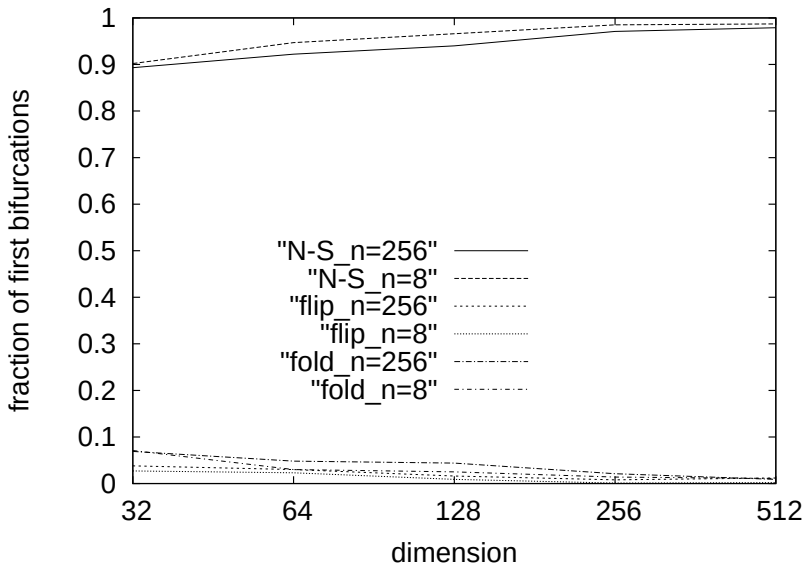


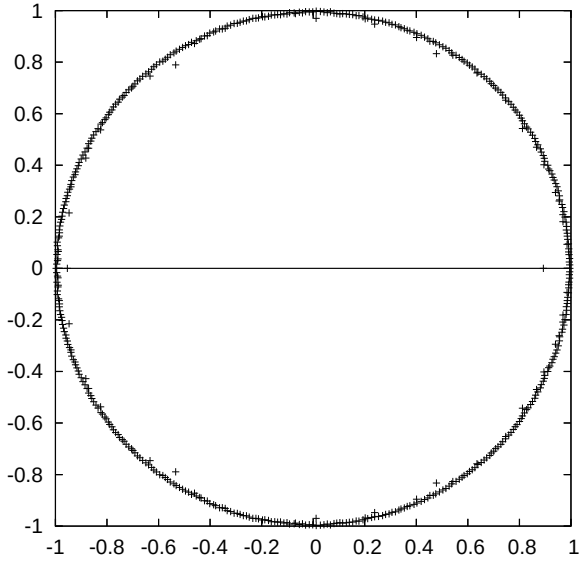
On the left, the observed probability of each bifurcation was recorded for 1000 matrices with i.i.d., mean zero, variance one, Gaussian a_k's for each d (in powers of 2) along with the fraction of eigenvalues that are real. On the right is the spectrum of eigenvalues in the complex plane that corresponds to a single 1024×1024 matrix ($d = 1024$).



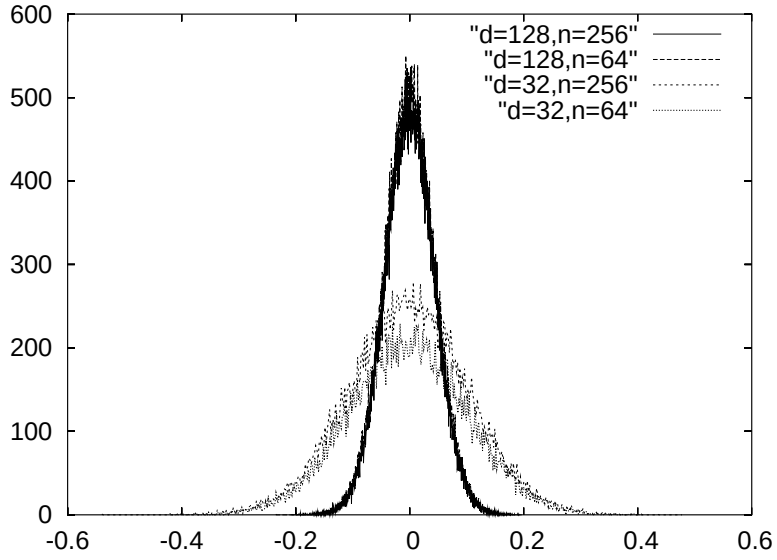


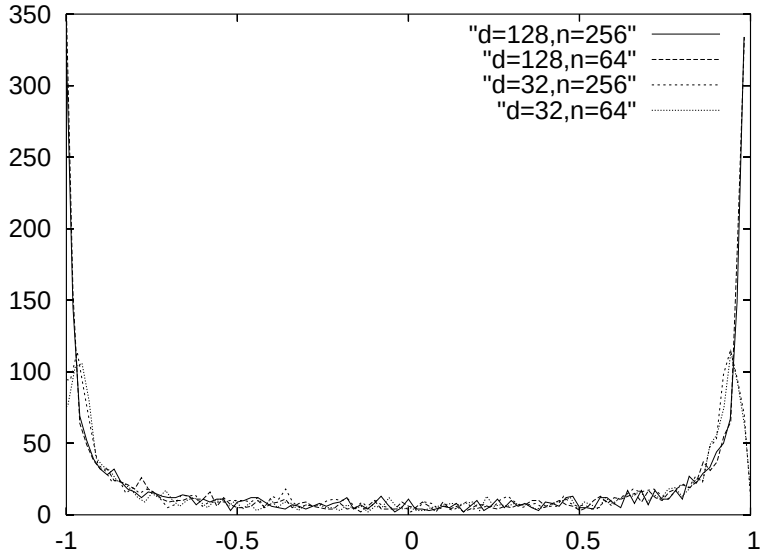
On the left is the theoretical real zero density for a 64-degree polynomial with random coefficients drawn from normals with mean zero and unit variance. On the right is the real zero density for a set of 3000 companion matrices with a_k's drawn from standard normals with mean zero and unit variance.





On the left, the observed probability of each bifurcation was recorded for 1000 neural networks for each d along with the fraction of eigenvalues that are real. On the right is the spectrum of eigenvalues in the complex plane that corresponds to a single neural network with $n = 256$ and $d = 512$.





The plot on the right is of the distribution of a_k's for 1000 neural networks with $n = 256, 64$ and $d = 128, 32$. The plot on the right is of the distribution of real eigenvalues along the real axis for the same set of neural networks.

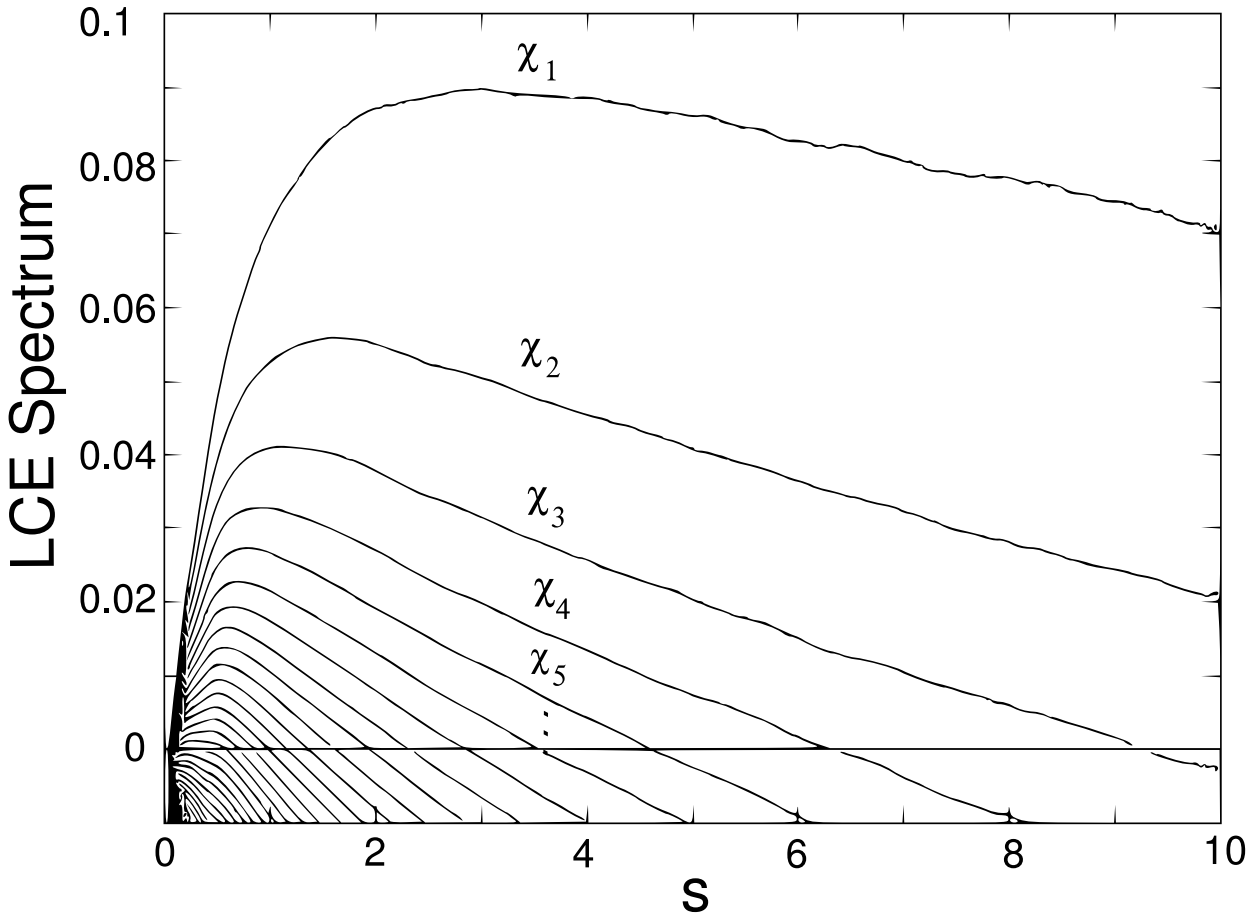
Three problems:

RMT: Why does perturbing the mean control the largest eigenvalue? (probably a simple answer)

Companion matrices: Use a Edelman's theorem to derive a connection between the weight distribution of with neural network and the real zeros of it's derivative matrix (this is the first step on a grand scheme to connect weight distributions of the neural networks to the LCE's)

Companion matrices and full matrices: Old problem, distribution on g_{ij} leads to what distribution on c_{1j}? Train the neural nets, find an empirical answer, use some of Edelman's machinery to attempt a proof.

Region IV update

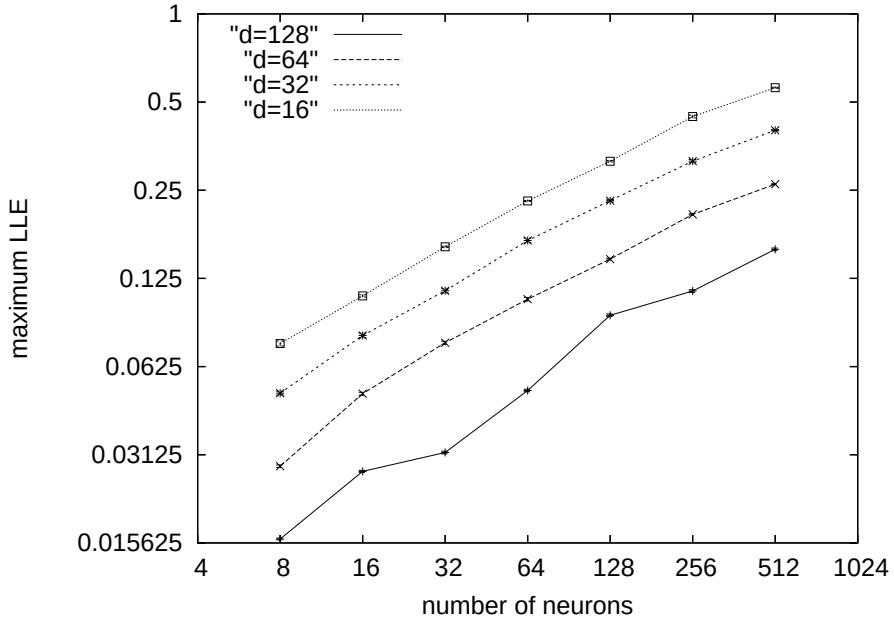


LCE spectrum: 32 neurons, 64 dimensions

Scalings — factors to consider

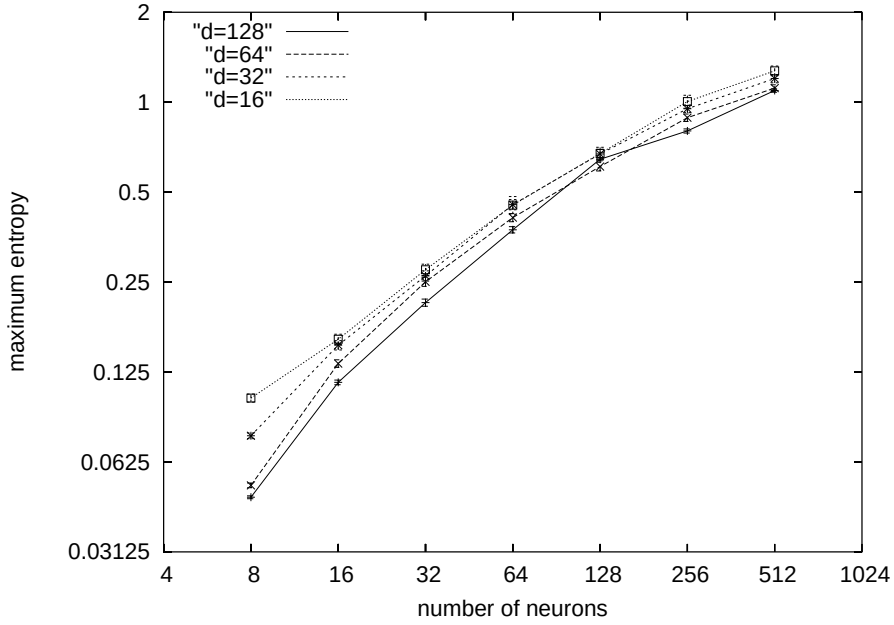
Effects of adding delays

Effects of adding parameters: functional analysis, RMT, training



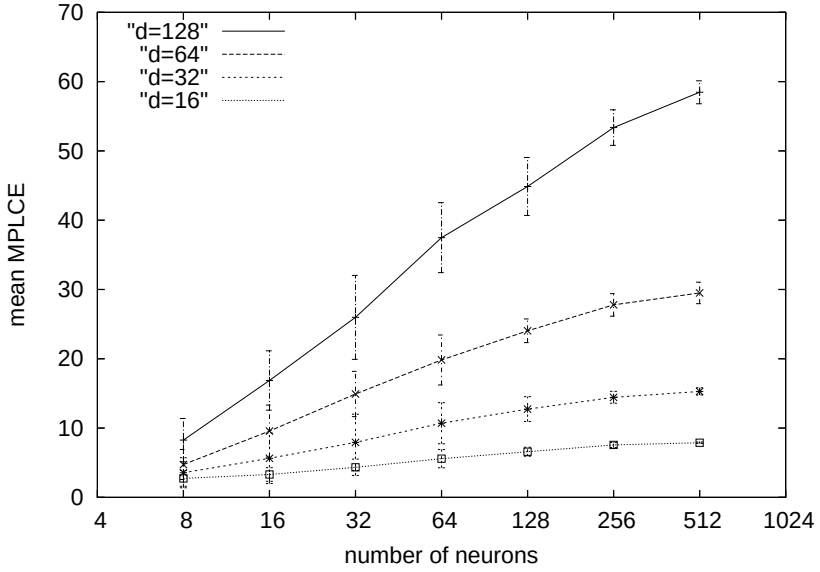
Maximum LLE (over s) versus n for various values of d. Each point on each curve represents an average over at least 100 networks.

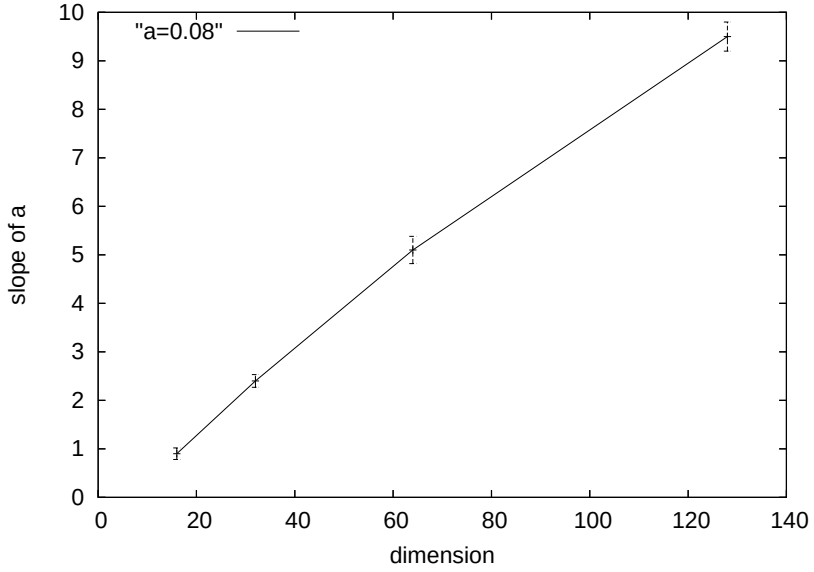
Scalings: $d = 16$, $\chi_{max} \sim n^{0.49}$; $d = 128$ $\chi_{max} \sim n^{0.5548}$.



Maximum entropy (over s) versus n for various values of d. Each point on each curve represents an average over at least 100 networks.

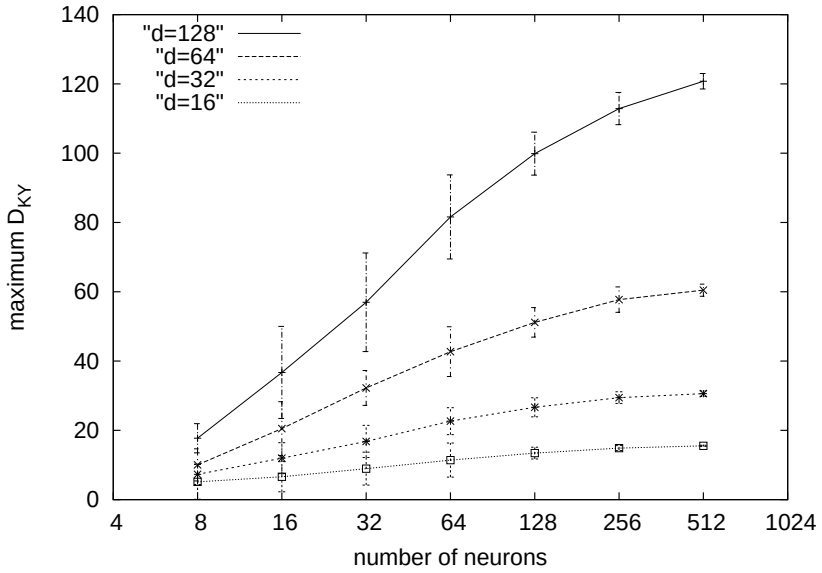
Scalings: $h_\mu \sim n^{0.625}$ at $d = 16$; $h_\mu \sim n^{0.73}$ at $d = 128$.

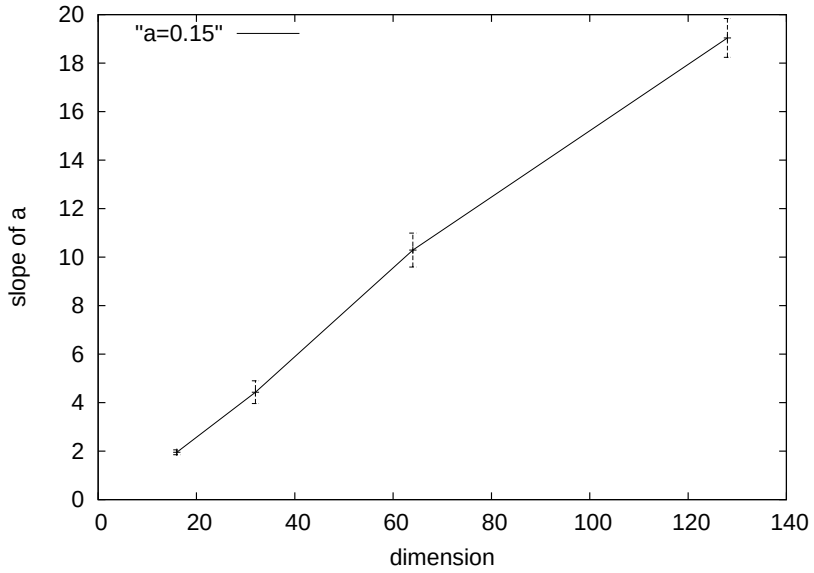




Maximum number of positive LCE's (over s) versus n for various values of d. Each point on each curve represents an average over at least 100 networks.

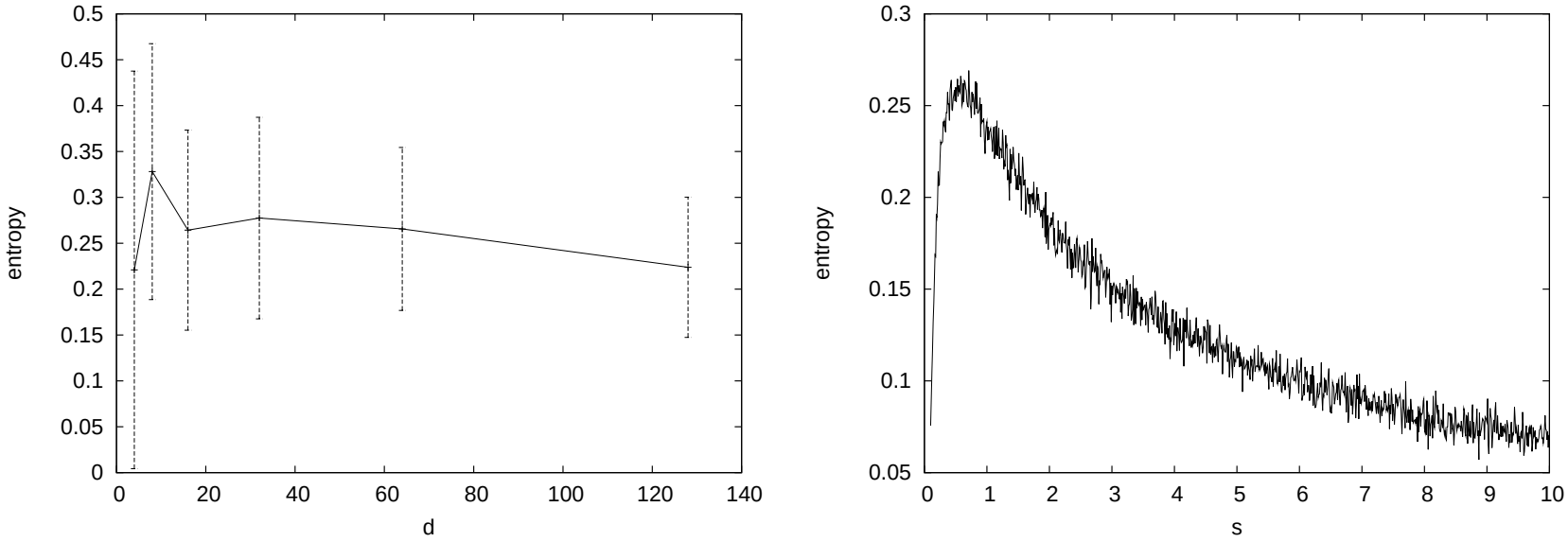
Scaling: MPLCE $\propto a\log(n)$ where a is a function of d — $a(d) = 0.08d$, yielding $MPLCE \approx 0.8dlog(n)$.





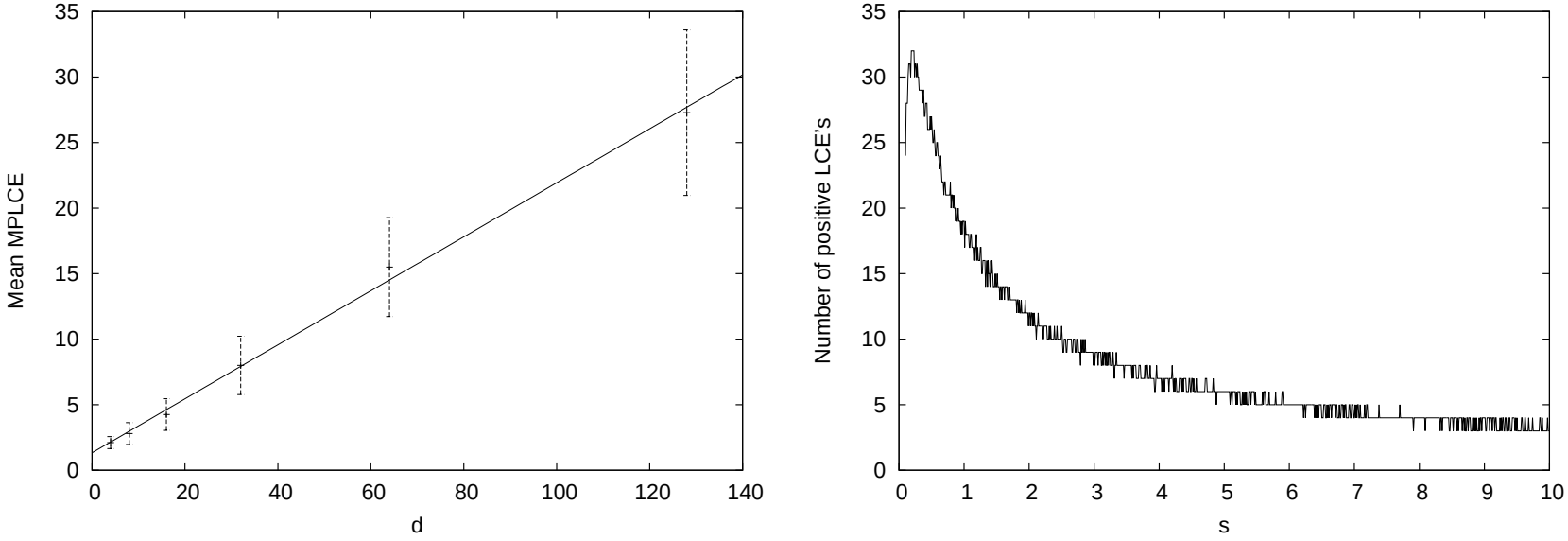
Maximum D_{KY} (over s) versus n for various values of d. Each point on each curve represents an average over at least 100 networks.

Scaling: $D_{KY} \sim a \log(n)$, where $a \sim 0.15d$, yielding $D_{KY} \sim 0.15d \log(n)$.

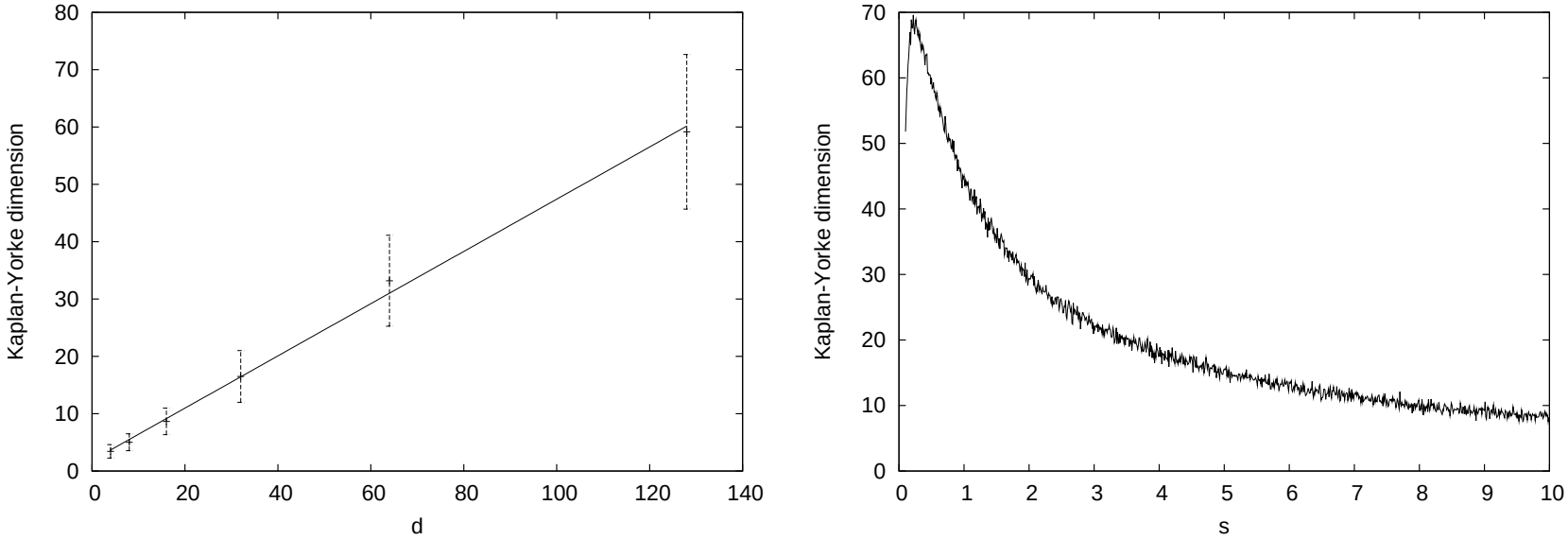


Mean maximum entropy versus dimension, d (left). Variation in the entropy of a single network with s (right). The network considered has 32 neurons and 128 dimensions.

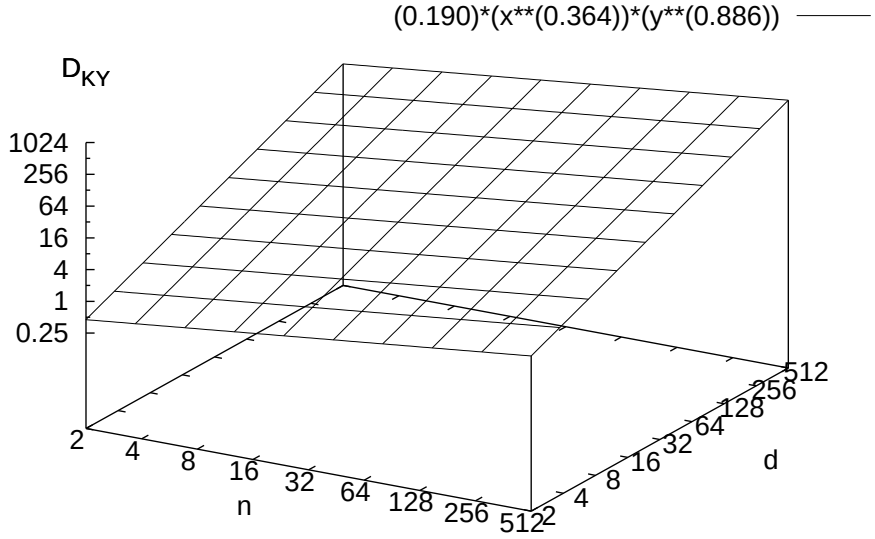
Since $h_\mu(n) \sim n^l$ where $l \leq 0.75$, it is likely that over a fixed range of d, for fixed n, the entropy will be fairly constant.



Mean maximum number of positive LE's versus dimension, all networks have 32 neurons (slope is approximately $\frac{1}{4}$) (left). Number of positive LCE's for a typical individual network with 32 neurons and 128 dimensions (right).



Mean maximum Kaplan-Yorke dimension versus dimension, d. For the set of networks analyzed, $D_{KY} \sim 0.46d$ (left). Variation of Kaplan-Yorke dimension versus s for a single network with $N = 32$ and $d = 128$ (right).



$$D_{ky} = 0.190 n^{0.364} d^{0.886} \tag{13}$$

which has $R^2 = 0.952$

$$\text{MPLCE} = 0.105 n^{0.367} d^{0.843} \tag{14}$$

which has $R^2 = 0.943$.

Topological variation and scaling laws.