

Statistical mechanics of compressed sensing and memory through random matrices

Surya Ganguli

Dept. of Applied Physics

Stanford University

Joint work with:

Haim Sompolinsky, Harvard / Hebrew University

Ben Huh, UCSD / Gatsby Institute, UCL

Two Classical Dogmas

1) Nyquist-Shannon Sampling Theorem:

To accurately reconstruct a bandlimited signal whose maximal frequency component has frequency f , you must sample it at a rate greater than or equal to $2f$.

If only a small number of frequency components are present, no matter how high the largest frequency is, you need only sample the signal at a small number of random times to reconstruct the signal.

2) Solving linear equations:

If you have T unknowns and only N equations, where $N < T$, then give up.

If the true solution is sparse, you can still find it even if $N < T$.

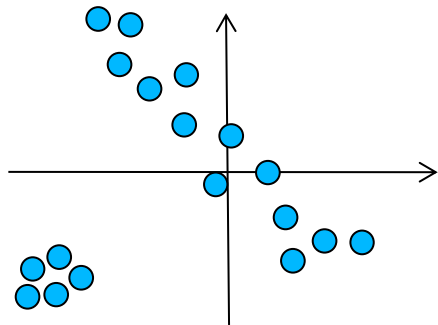
High dimensional data analysis

N = dimensionality of data
 M = number of data points

$$\alpha = N / M$$

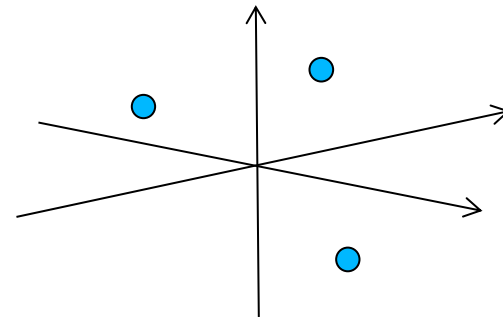
Classical Statistics

$$\begin{aligned} N &\sim O(1) \\ M &\rightarrow \infty \\ \alpha &\rightarrow 0 \end{aligned}$$



Modern Statistics

$$\begin{aligned} N &\rightarrow \infty \\ M &\rightarrow \infty \\ \alpha &\sim O(1) \end{aligned}$$



Curse of dimensionality:

How can we extract meaning from small amounts of high dim data?

How can neural systems acquire internal models of the world from small amounts of high dim stimuli?

Talk Outline

1) Intro to Compressed Sensing

2) Applications:

Aquiring temporal signals

Magnetic Resonance Imaging

DNA microarrays (biosensing)

Brain Connectomics

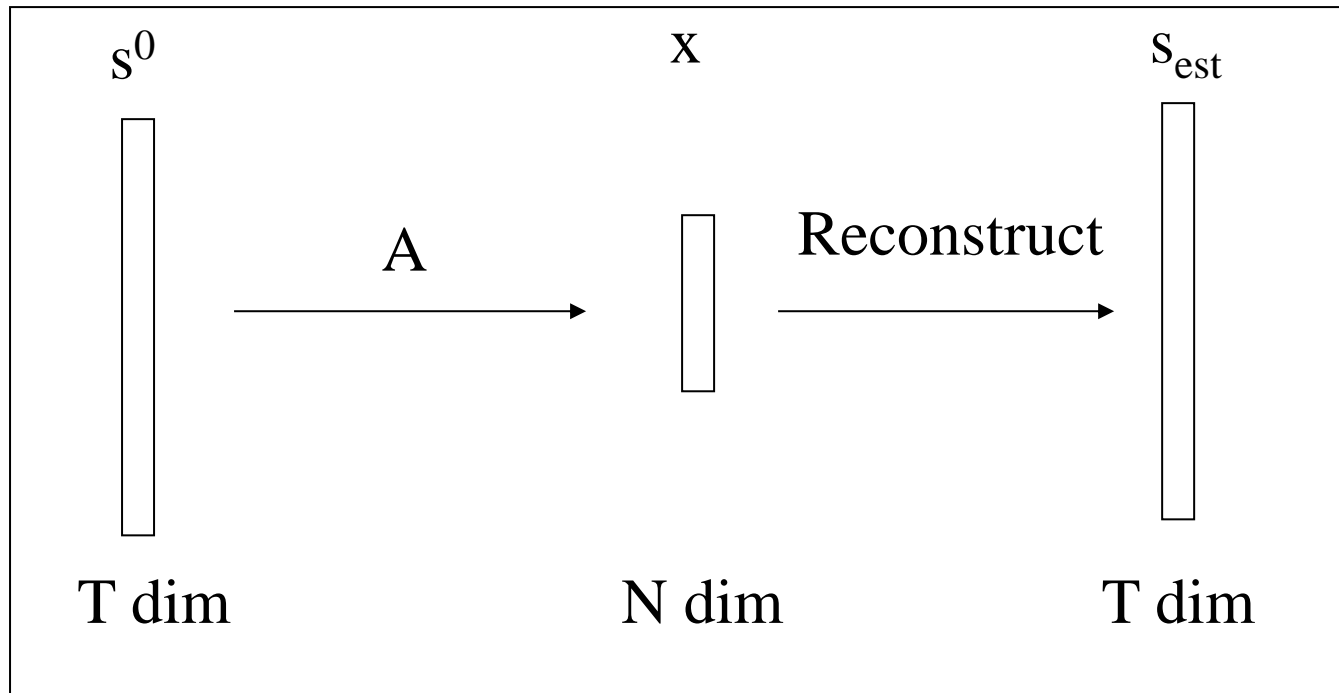
Genetic Regulatory Network Reconstruction

Compressed sensing in the brain?

3) Statistical mechanics approach to the analysis of compressed sensing.

4) Memory in neural networks through dynamical CS

Compressed Sensing



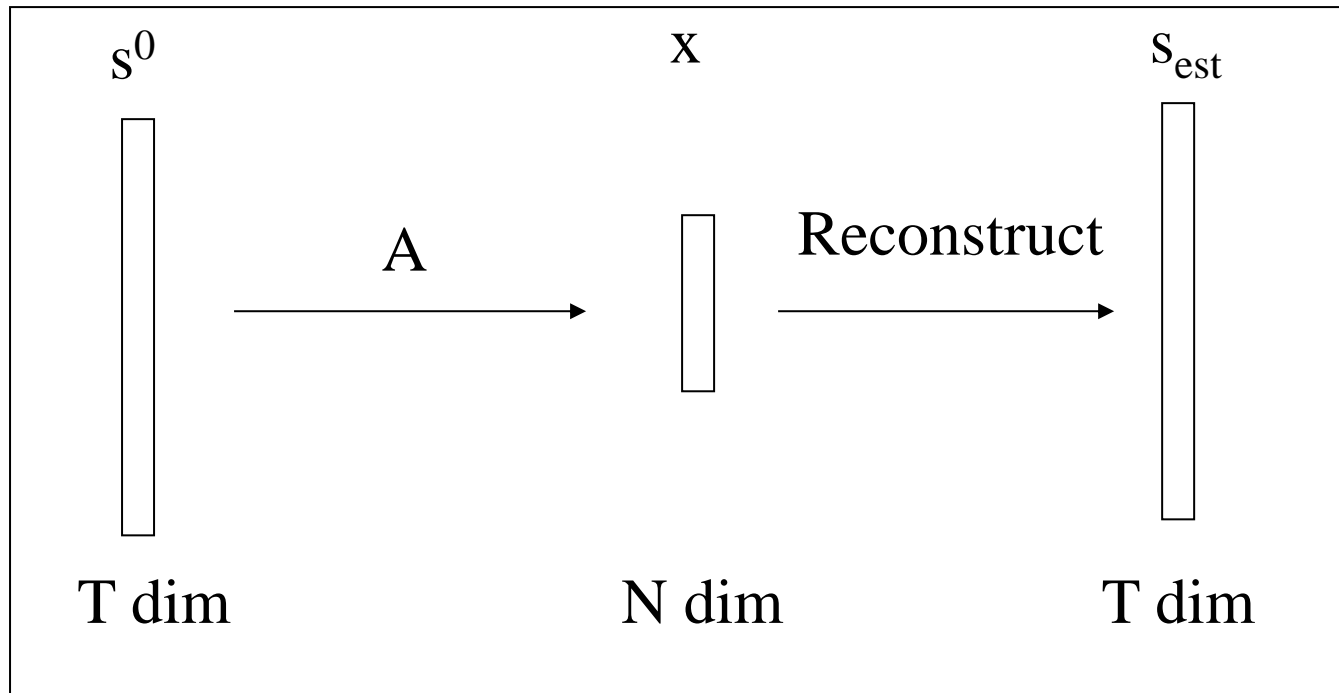
\mathbf{s}_0 : T dimensional signal with a fraction f elements nonzero

$\mathbf{x} = \mathbf{A}\mathbf{s}_0$: N dimensional measurement vector with $\alpha = N/T < 1$

In general, reconstructing \mathbf{s}_0 from \mathbf{x} is ill posed:

T-N dimensional space of possible signals \mathbf{s} consistent with measurement constraints.

Compressed Sensing



\mathbf{s}_0 : T dimensional signal with a fraction f elements nonzero

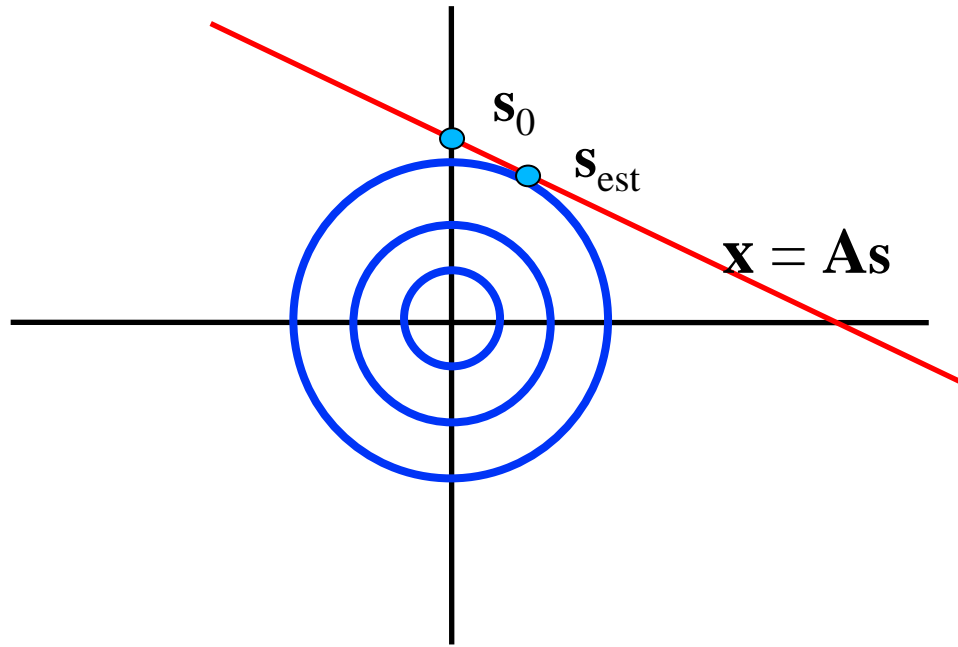
$\mathbf{x} = \mathbf{A}\mathbf{s}_0$: N dimensional measurement vector with $\square = N/T < 1$

Approaches to constructing an estimate \mathbf{s}_{est} of \mathbf{s}_0 from \mathbf{x} when \mathbf{s}_0 is sparse:

L_0 minimization: $\mathbf{s}_{\text{est}} = \arg \min_{\mathbf{s}} \sum_i |\mathbf{s}_i|^0$ subject to $\mathbf{x} = \mathbf{A}\mathbf{s}$ (hard)

L_p minimization: $\mathbf{s}_{\text{est}} = \arg \min_{\mathbf{s}} \sum_i |\mathbf{s}_i|^p$ subject to $\mathbf{x} = \mathbf{A}\mathbf{s}$ (convex for $p \geq 1$)

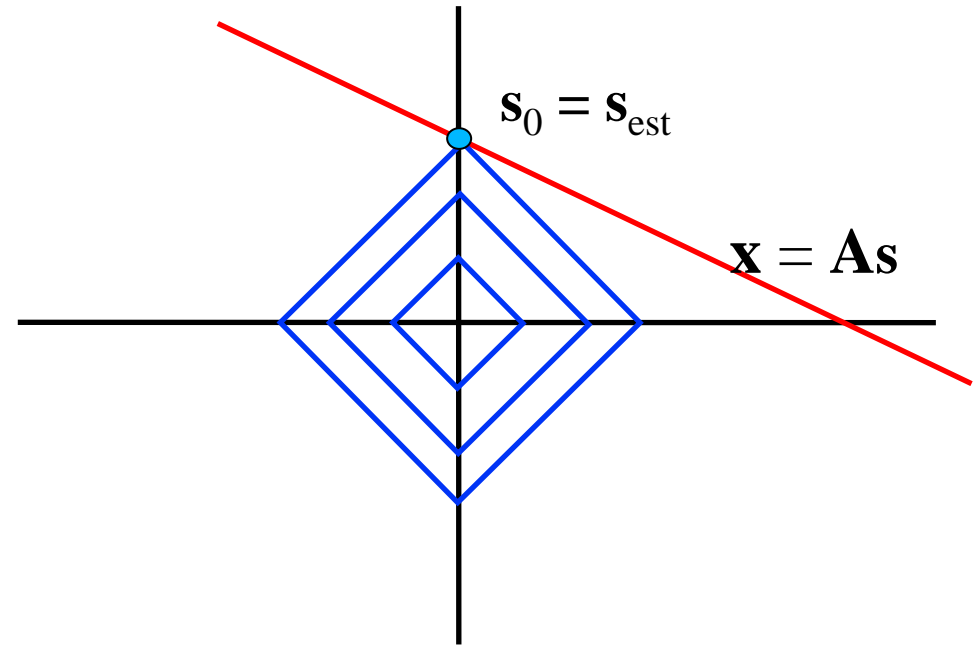
Why L1? Geometry behind compressed sensing



“Classical” L_2 minimization:

$$\mathbf{s}_{\text{est}} = \arg \min_{\mathbf{s}} \sum_i s_i^2$$

subject to $\mathbf{x} = \mathbf{A}\mathbf{s}$

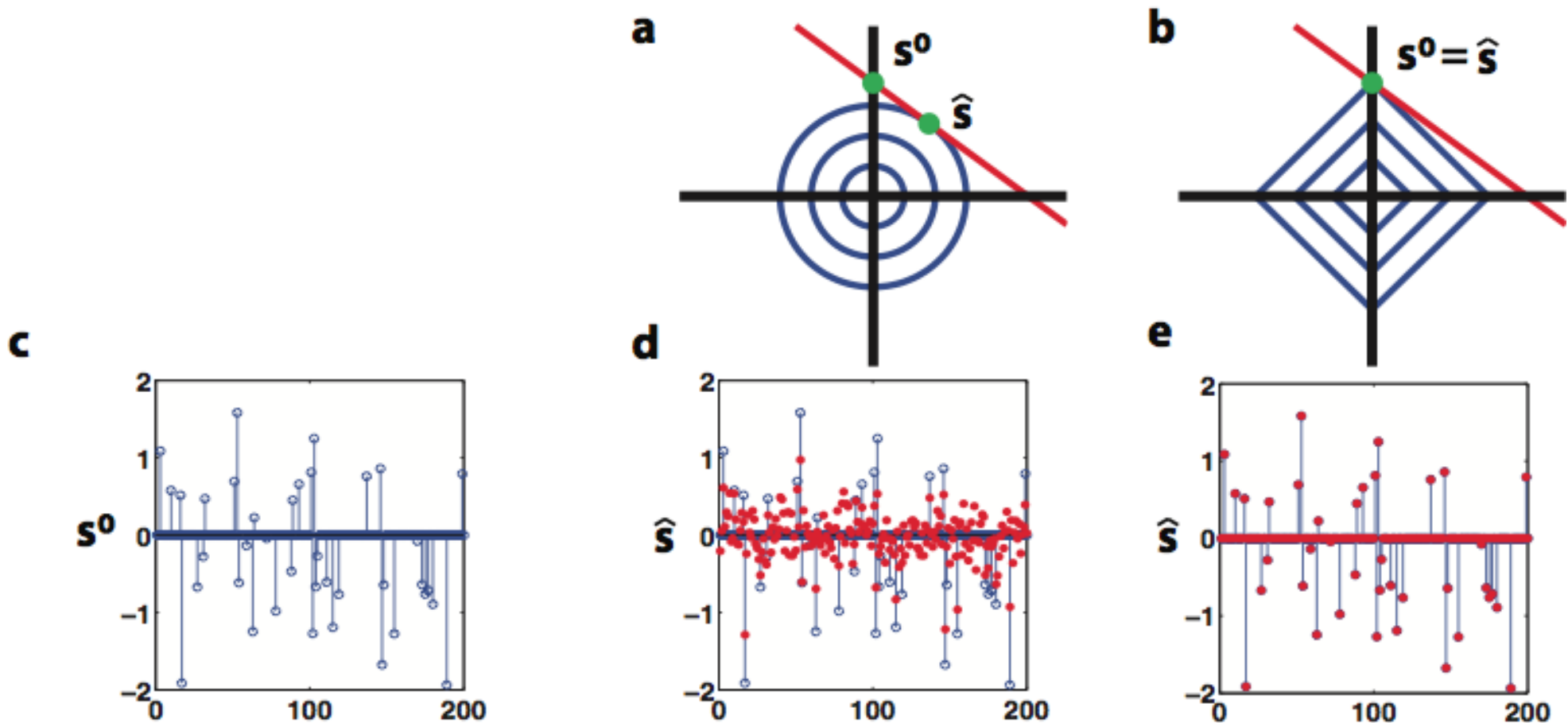


“Modern” L_1 minimization:

$$\mathbf{s}_{\text{est}} = \arg \min_{\mathbf{s}} \sum_i |s_i|$$

subject to $\mathbf{x} = \mathbf{A}\mathbf{s}$

Why L1? An example reconstruction



$T = 200$ dim signal

$K = 40$ nonzero elements

$N = 120$ random linear measurements

Key question: What is a good set of measurements?

- General scenario: assume the true signal \mathbf{s}^0 is sparse in an orthonormal basis given by the columns of the $T \times T$ matrix \mathbf{C} :

$$\mathbf{s}^0 = \mathbf{C} \mathbf{u}$$

\mathbf{u} is a sparse vector
with fT nonzeros

- Assume you take $N < T$ linear measurements \mathbf{x} of \mathbf{s}^0 obtained by choosing N orthonormal columns from the $T \times T$ measurement matrix \mathbf{B} :

$$\begin{aligned}\mathbf{x} &= \mathbf{R} \mathbf{B}^T \mathbf{s}^0 \\ &= \mathbf{R} \mathbf{B}^T \mathbf{C} \mathbf{u} \\ &= \mathbf{A} \mathbf{u}\end{aligned}$$

(\mathbf{R} is an $N \times T$ matrix which
picks off N rows of \mathbf{B}^T .)

- Reconstruction algorithm: L_1 min: $\mathbf{u}_{\text{est}} = \arg \min_{\mathbf{s}} \sum_i |\mathbf{u}_i|^1$ subject to $\mathbf{x} = \mathbf{A} \mathbf{u}$
Then $\mathbf{s}_{\text{est}} = \mathbf{C} \mathbf{u}_{\text{est}}$
- Question: Given a sparsity basis \mathbf{C} , what kind of measurement basis \mathbf{B} should you choose in order to still get perfect reconstruction ($\mathbf{s}_{\text{est}} = \mathbf{s}^0$) with a small value of $N < T$

Measurements should be incoherent w.r.t. sparsity basis

- Question: Given a sparsity basis \mathbf{C} , what kind of measurement basis \mathbf{B} should you choose in order to still get perfect reconstruction ($\mathbf{s}_{\text{est}} = \mathbf{s}^0$) with a small value of $N < T$

- Definition of coherence:

Let \mathbf{c}_i be the i 'th column of \mathbf{C} (a sparsity basis vector)

Let \mathbf{b}_j be the j 'th column of \mathbf{B} (a measurement basis vector)

Coherence between measurement and sparsity bases:

$$\text{Coh}(\mathbf{B}, \mathbf{C}) = \sqrt{T} \max_{k,l} \langle \mathbf{b}_k, \mathbf{c}_l \rangle \quad (\text{ranges from } 1 \text{ to } \sqrt{T})$$

- Theorem: Assume \mathbf{s}^0 has only fT nonzero active components in sparsity basis \mathbf{C} . Assume you chose a measurement basis \mathbf{B} .

Then if $N > \text{Coh}(\mathbf{B}, \mathbf{C})^2 fT \log T$ you will get perfect reconstruction.

Candes, Romberg 2007

- Moral: If your measurement basis \mathbf{B} is incoherent w.r.t to your sparsity basis \mathbf{C} (i.e. $\text{Coh}(\mathbf{B}, \mathbf{C}) \sim O(1)$) then you will only need $O(fT \log T)$ measurements.

Examples of incoherent bases

Measurement basis	Sparsity basis	Coherence	
■ Spikes (delta functions)	Sinusoids	1	
■ Noiselets	Haar Wavelets	$\sqrt{2}$	Coifman et. al. 01
■ Noiselets	Spikes	1	
■ Noiselets	Sinusoids	$O(1)$	
■ Random Basis	Any fixed basis	$\sqrt{2 \log T}$	

Application: Magnetic Resonance Imaging

Measurement basis

Sparsity basis

Fourier Modes

????

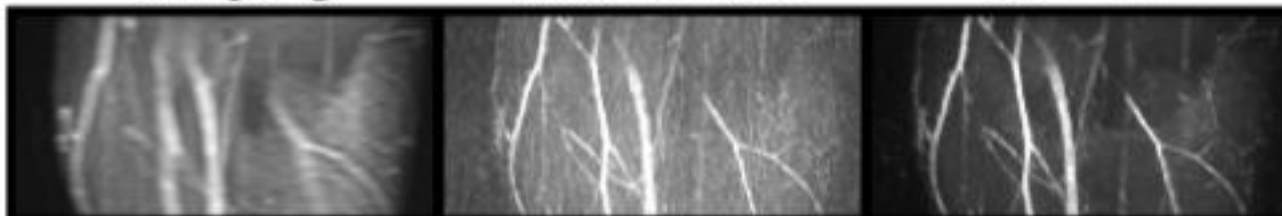
Due to physics of NMR, each RF pulse applied to image yields a Fourier component of the image.

Angiograms: sparse in pixel basis.
Other images: sparse in spatial differences of wavelet coefficients.

low resolution
Sampling

random undersampling
zero-fill w/dc

CS - TV



Sparse MRI: The Application
of Compressed Sensing for
Rapid MR Imaging

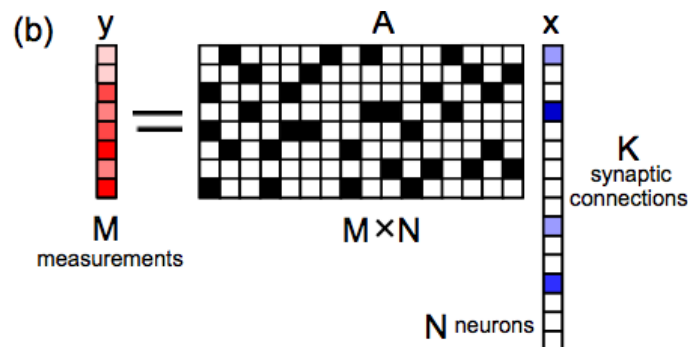
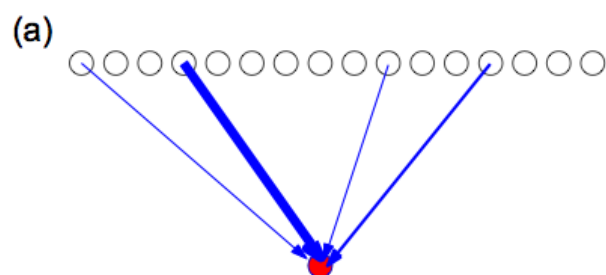
Michael Lustig, David Donoho
and John M. Pauly

Application: DNA Microarrays (biosensors)

- Goal: in a sample (i.e. soil, water), measure an unknown vector \mathbf{s}^0 of Concentrations of T DNA sequences, where each DNA sequence is in 1-1 correspondence with an organism of interest.
- Traditional approach: construct a DNA microarray with T spots, with each spot containing a single complementary DNA strand.
- BUT: in any given sample, we only expect a small number of organisms among the T organisms to be present; $\Rightarrow \mathbf{s}^0$ is sparse!
- Compressed sensing approach: construct a DNA microarray with $N \ll T$ spots, each spot containing a random subset of T possible complementary DNA sequences.
- $\mathbf{x} = \mathbf{A} \mathbf{s}^0$
 - \mathbf{x}_i = fluorescence level of spot i
 - \mathbf{A}_{ij} = binding probability of contents of spot i with sequence j
 - \mathbf{s}^0_j = concentration of sequence j in sample

Application: Brain connectomics

- Goal: measure an unknown vector \mathbf{s}^0 of T synaptic strengths onto a single postsynaptic cell.
- Naïve approach: excite each presynaptic cell one at a time, measure the postsynaptic response, and infer the corresponding synaptic strength
- BUT: for any given postsynaptic cell, we only expect a small number of synaptic strengths to be nonzero; $\Rightarrow \mathbf{s}^0$ is sparse!
- Compressed sensing approach: Stimulate a random subset of presynaptic cells, and measure the postsynaptic response. Repeat $N \ll T$ times.



Reconstruction of Sparse Circuits Using
Multi-neuronal Excitation (RESCUME)
Tao Hu and Dmitri Chklovskii, NIPS 2009

- $\mathbf{x} = \mathbf{A} \mathbf{s}^0$

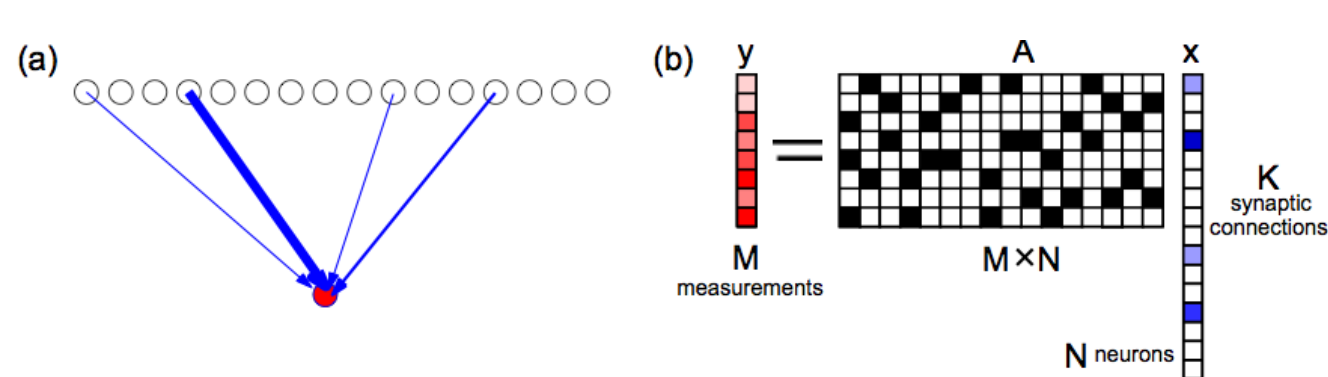
\mathbf{x}_i = response of postsynaptic cell on trial i

\mathbf{A}_{ij} = level of excitation of presynaptic cell j in trial i

\mathbf{s}^0_j = synaptic weight from presynaptic cell j

Application: Genetic Regulatory Network Reconstruction

- Goal: measure an unknown vector \mathbf{s}^0 of T regulatory weights governing the linear response of a candidate gene to the expression levels of T other genes.
- For any given candidate gene, we only expect a small number of the T genes to actually regulate it $\Rightarrow \mathbf{s}^0$ is sparse!
- Compressed sensing approach: Measure the expression levels of the T genes and the candidate gene under $N \ll T$ experimental conditions.

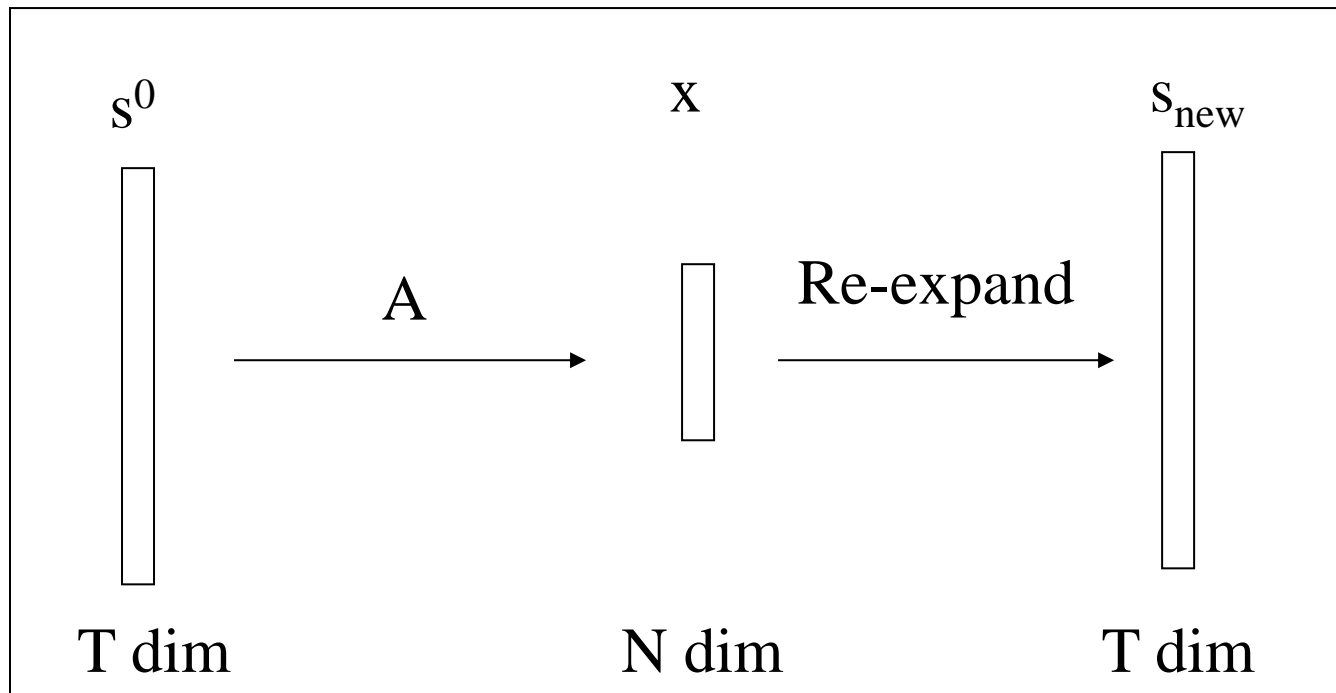


Reconstruction of Sparse Circuits Using
Multi-neuronal Excitation (RESCUME)
Tao Hu and Dmitri Chklovskii, NIPS 2009

- $\mathbf{x} = \mathbf{A} \mathbf{s}^0$

\mathbf{x}_i = expression level of candidate gene in condition i
 \mathbf{A}_{ij} = expression level of gene j in condition i
 \mathbf{s}^0_j = regulatory weight of gene j onto candidate gene

Compressed sensing as a principle of brain communication?



Isely, Hillar Sommer
NIPS 2010

s_0 = High dimensional sparse activity in one brain region

\mathbf{x} = Low dimensional dense activity in a set of axons s_0 communicating to another brain region.

s_{new} = Expansion of communicated signal \mathbf{x} in a downstream region for further processing / computation.

Examples: Hippocampal CA3/CA1 -> Subicular Output

Cerebellar granule cells -> Purkinje/ DCN Outputs

Layer 4 of sensory cortex -> Layer 5 outputs

Question: When does L1 minimization work?

\mathbf{s}_0 : T dimensional signal with a fraction f elements nonzero

$\mathbf{x} = \mathbf{A}\mathbf{s}_0$: N dimensional measurement vector with $\alpha = N/T < 1$

L_1 minimization: $\mathbf{s}_{\text{est}} = \arg \min_{\mathbf{s}} \sum_i |s_i|^1$ subject to $\mathbf{x} = \mathbf{A}\mathbf{s}$

- When is perfect recovery possible: i.e. when is \mathbf{s}_{est} equal to \mathbf{s}_0 ?
- Traditional approach: What are sufficient conditions on \mathbf{A} such that perfect recovery is guaranteed? (Donoho, Tao, Candes).
- Problem: many large random measurement matrices which violate such sufficient conditions nevertheless yield good signal reconstruction.
- **Statistical mechanics approach:** compute the typical performance of L_1 minimization as a function of α and f for large random measurement matrices.

Statistical mechanics approach

\mathbf{s}_0 : T dimensional signal with a fraction f elements nonzero

$\mathbf{x} = \mathbf{A}\mathbf{s}_0$: N dimensional measurement vector with $\alpha = N/T < 1$

L_1 minimization: $\mathbf{s}_{\text{est}} = \arg \min_{\mathbf{s}} \sum_i |\mathbf{s}_i|^1$ subject to $\mathbf{x} = \mathbf{A}\mathbf{s}$

- Define an energy function on the space of candidate signals whose global minimum is the solution to L_1 minimization:

$$E(\mathbf{s}) = \lambda/2 \|\mathbf{A}\mathbf{s} - \mathbf{A}\mathbf{s}_0\|^2 + \sum_i |\mathbf{s}_i| \quad \text{later will take } \lambda \rightarrow \text{infinity}$$

- This yields a Gibbs distribution

$$P_G(\mathbf{s}) = 1/Z \exp(-\beta E(\mathbf{s})) \quad \text{later will take } \beta \rightarrow \text{infinity}$$

- Now compute the typical error as a function of α and f :

$$\langle\langle 1/T \int D\mathbf{s} \|\mathbf{s} - \mathbf{s}_0\|^2 P_G(\mathbf{s}) \rangle\rangle_{\mathbf{A}, \mathbf{s}_0}$$

Mean field theory of compressed sensing

- The full theory:

$$\mathbf{u} = \mathbf{S} - \mathbf{S}_0$$

$$P_G(\mathbf{u}_i, \{\mathbf{u}_j\}_{j \neq i}) \propto e^{-\beta \sum_{a=1}^N \frac{\lambda}{2} (\mathbf{A}_{ai} \mathbf{u}_i + \sum_{j \neq i} \mathbf{A}_{aj} \mathbf{u}_j)^2 - \beta |\mathbf{s}_i| - \beta \sum_{j \neq i} |\mathbf{s}_j|}$$

$$P_G(\mathbf{u}_i, \{\mathbf{u}_j\}_{j \neq i}) \propto e^{-\frac{\beta \lambda}{2} \left(\sum_{a=1}^N \mathbf{A}_{ai}^2 \right) \mathbf{u}_i^2 - \beta \lambda \mathbf{h}_i \mathbf{u}_i - \beta |\mathbf{s}_i| + F(\{\mathbf{u}_j\}_{j \neq i})}$$

$$\downarrow \quad \mathbf{h}_i = \sum_{a=1}^N \mathbf{A}_{ai} \sum_{j \neq i} \mathbf{A}_{aj} \mathbf{u}_j$$

- The residual \mathbf{u}_i couples to all other residuals only through the field \mathbf{h}_i
Idea: approximate the distribution of \mathbf{h}_i with a Gaussian.

- Mean field theory: $P_{MF}(\mathbf{u}_i | \mathbf{h}_i) \propto e^{-\frac{1}{2} \frac{\alpha \beta \lambda}{1 + \beta \lambda \Delta Q} (\mathbf{u}_i - \mathbf{h}_i)^2 - \beta |\mathbf{s}_i|}$

- Statistics of \mathbf{h}_i : zero mean Gaussian with variance Q_0 / α

- Order parameters : $Q_0 = \frac{1}{T} \sum_{i=1}^T \langle \mathbf{u}_i \rangle_{P_G}^2 \quad \Delta Q = \frac{1}{T} \sum_{i=1}^T \langle (\delta \mathbf{u}_i)^2 \rangle_{P_G}$

Self-consistency of the mean field approximation

- Mean field theory: $P_{MF}(\mathbf{u}_i | \mathbf{h}_i) \propto e^{-\frac{1}{2} \frac{\alpha \beta \lambda}{1 + \beta \lambda \Delta Q} (\mathbf{u}_i - \mathbf{h}_i)^2 - \beta |\mathbf{s}_i|}$
- Statistics of \mathbf{h}_i : zero mean Gaussian with variance Q_0 / α
- Order parameters : $Q_0 = \frac{1}{T} \sum_{i=1}^T \langle \mathbf{u}_i \rangle_{P_G}^2 \quad \Delta Q = \frac{1}{T} \sum_{i=1}^T \langle (\delta \mathbf{u}_i)^2 \rangle_{P_G}$
- Self consistency: Averaging \mathbf{u}_i over P_{MF} , and the distribution of \mathbf{h}_i and The true signal \mathbf{s}_i^0 , should recover the same order parameters above, obtained for a fixed realization of the measurements A_{ai}
- Self consistent equations for the order parameters:

$$Q_0 = \langle\langle \langle \mathbf{u}_i \rangle_{MF}^2 \rangle\rangle_{\mathbf{h}_i, \mathbf{s}_i^0} \quad \Delta Q = \langle\langle \langle (\delta \mathbf{u}_i)^2 \rangle_{MF} \rangle\rangle_{\mathbf{h}_i, \mathbf{s}_i^0}$$

A brief tour of the literature of MFT

Physics

Spin glass theory: replica
method, cavity method, Bethe
Approximation
Mezard et.al. Spin Glass Theory and Beyond

Statistics

Belief propagation (BP): exact
marginals on tree graphical models

Fixed points of BP on a loopy graph =
Extremal points of Bethe Free Energy

Yedida,
Freeman, Weiss, 2004

Cavity Method

~

BP

Cavity Method at
Next Level of Approximation

~

Survey Propagation

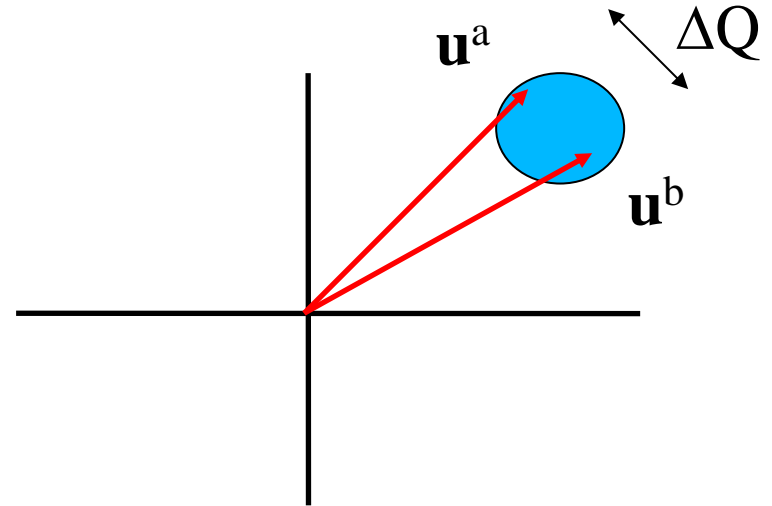
Mezard and Montanari: Physics Information and Computation

Donoho, Maleki, Montanari, PNAS 2010, Message passing approach to compressed sensing

Order parameters and the geometry of low energy (high probability) configurations

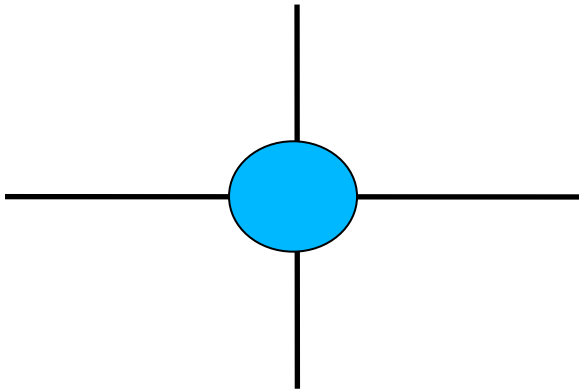
$$Q_0 = \frac{1}{T} \sum_{i=1}^T \langle \mathbf{u}_i \rangle_{P_G}^2$$

$$\Delta Q = \frac{1}{T} \sum_{i=1}^T \langle (\delta \mathbf{u}_i)^2 \rangle_{P_G}$$



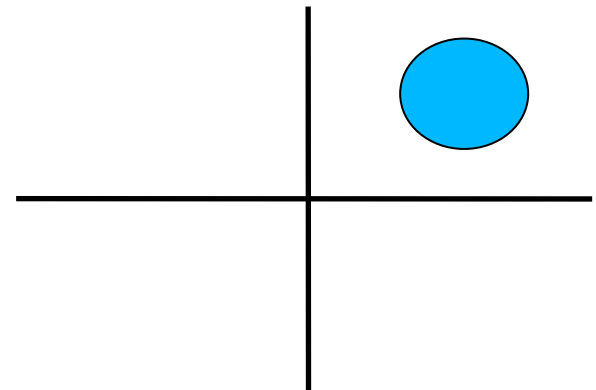
Perfect Reconstruction Solutions

$$\Delta Q \sim O(1/\beta^2)$$
$$Q_0 \sim O(1/\beta^2)$$



Error Solutions

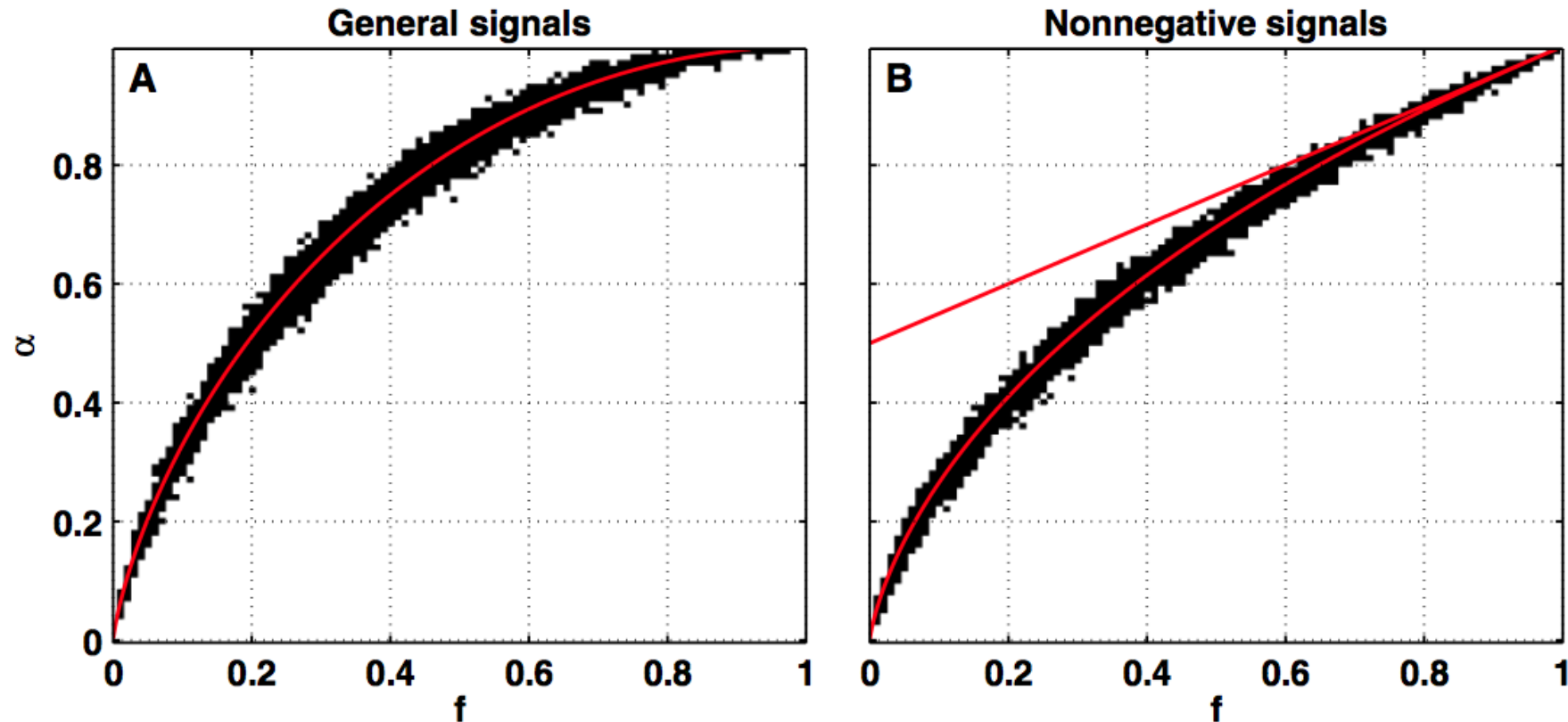
$$\Delta Q \sim O(1/\beta)$$
$$Q_0 \sim O(1)$$



Phase transitions in compressed sensing

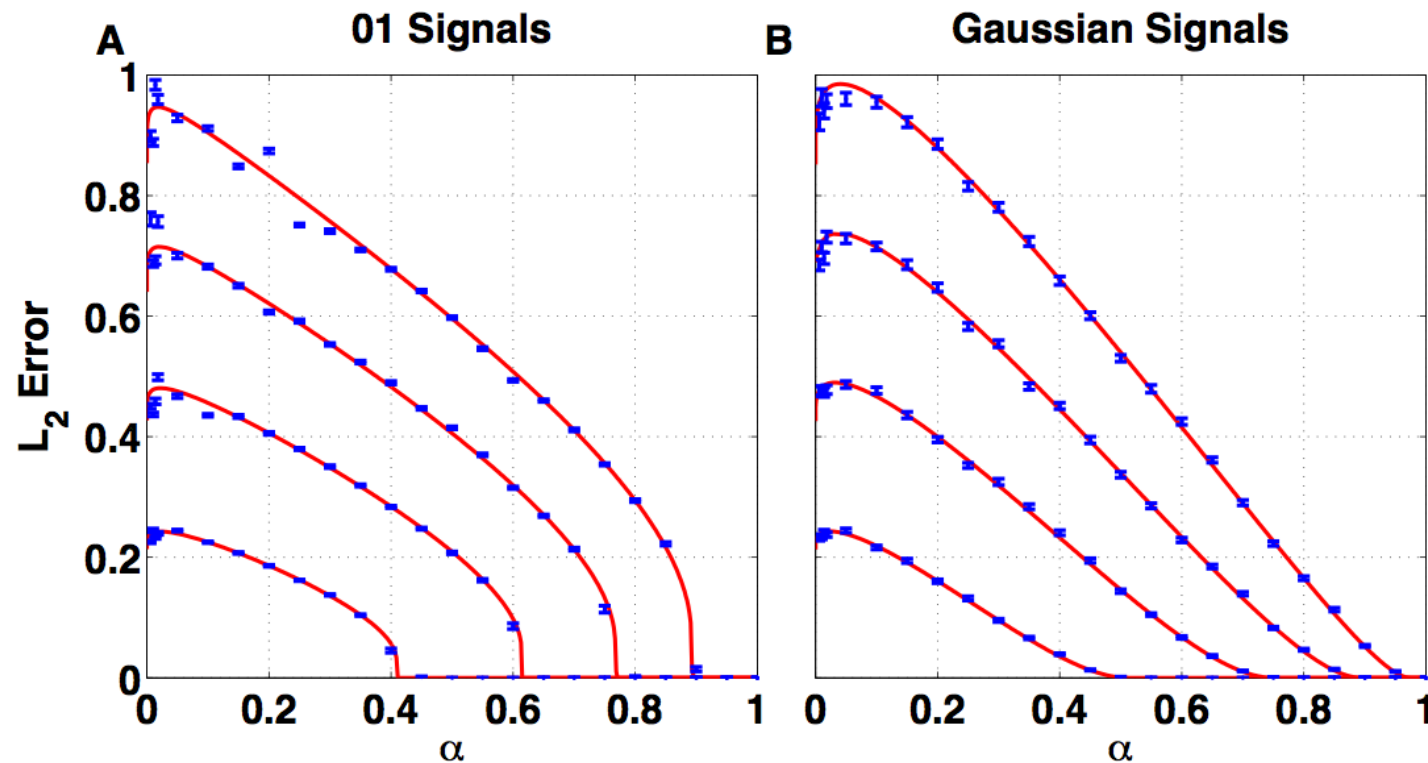
$\alpha > \alpha_c(f)$: perfect reconstruction possible
 $\alpha < \alpha_c(f)$: perfect reconstruction not possible

See also Donoho et.al. 2006



As $f \rightarrow 0$ $\alpha_c(f) \rightarrow f \log 1/f$ (expected from entropic arguments)

Compressed sensing in the error regime



Rise of the error near the phase transition depends only on the distribution of nonzero elements near the origin. Let $\delta\alpha$ be distance into error phase:

A gap in this distribution \Rightarrow Error rises sharply as $1/\log(1/\delta\alpha)$

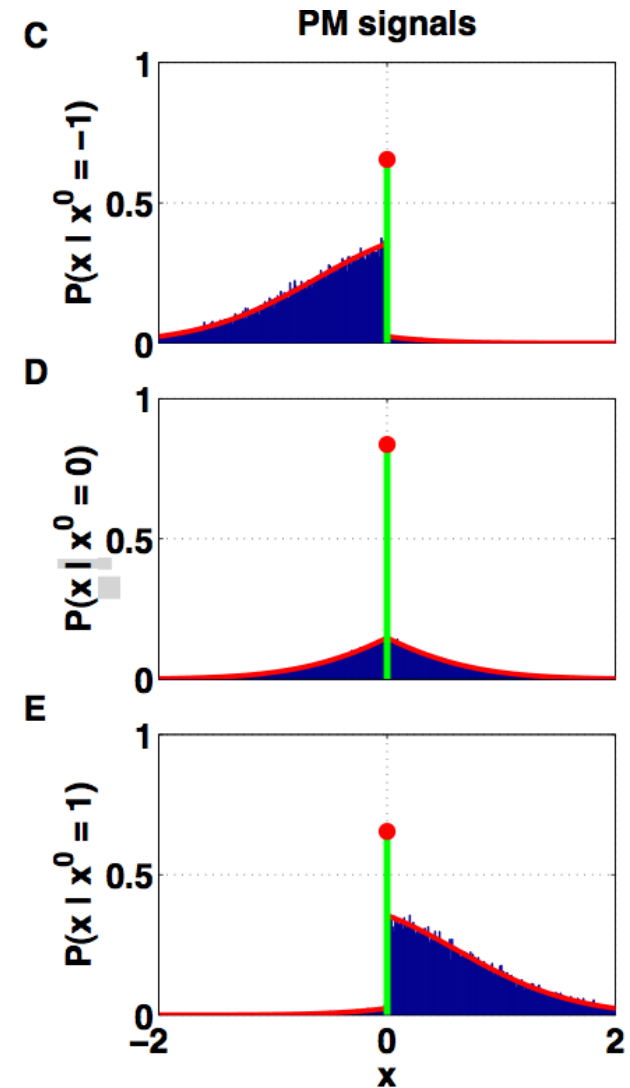
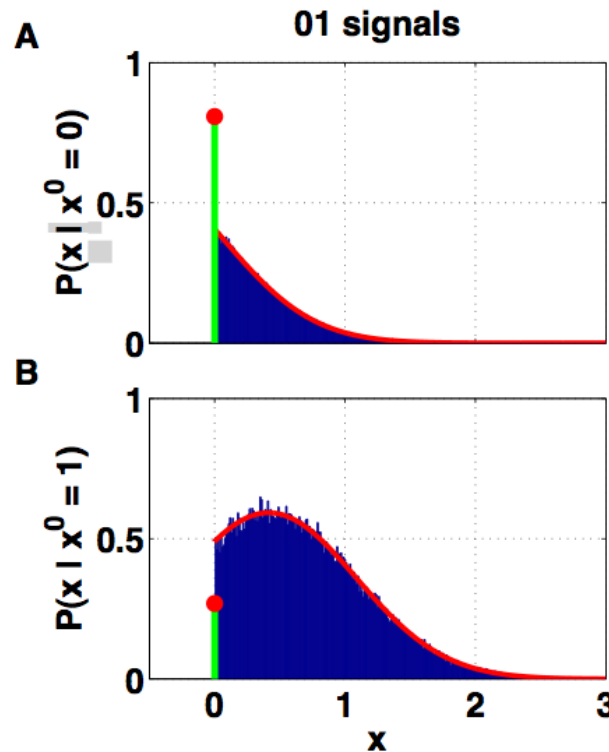
Power law behavior (s^ν) \Rightarrow Error rises as $(\delta\alpha)^{2/(1+\nu)}$

Sharper confinement of nonzeros to origin (smaller ν) \Rightarrow shallower rise of error

The nature of errors in compressed sensing

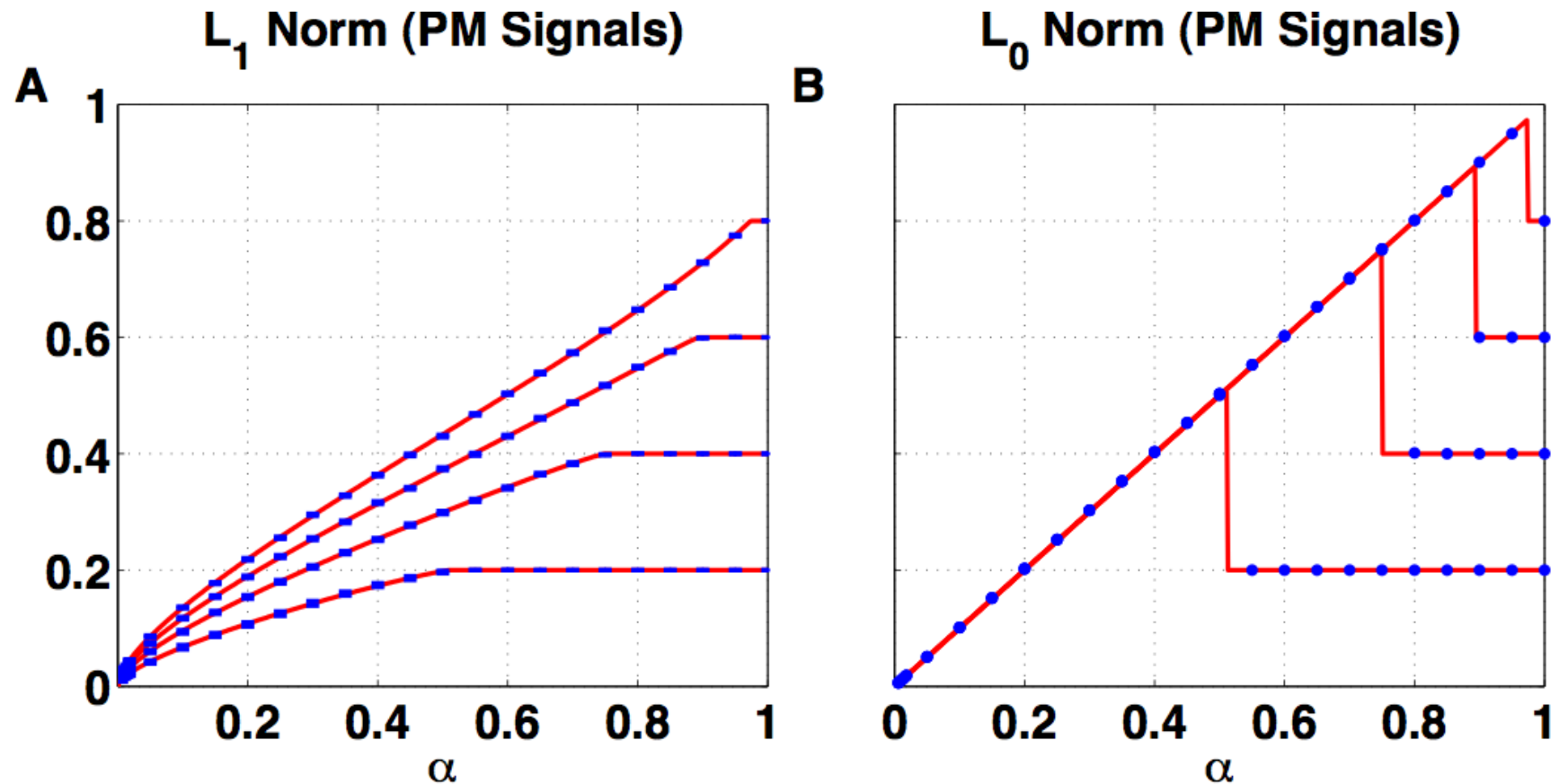
$$P(x|x^0) = \frac{1}{\sqrt{2\pi q_0}} \exp\left(-\frac{(x - x_0 + \Delta q)^2}{2q_0}\right) + H(-z^+) \delta(x).$$

$$z^\pm = \frac{-x^0 \pm \Delta q}{\sqrt{q_0}}$$



$$P(x|x^0) = \frac{1}{\sqrt{2\pi q_0}} \exp\left(-\frac{(x - x_0 + \text{sgn}(x)\Delta q)^2}{2q_0}\right) + (H(z^-) - H(z^+))\delta(x)$$

Behavior of L_p norms under L_1 minimization



A procedure to detect successful reconstruction even when you do not know the true signal: if the number of nonzeros in your reconstruction is less than the number of measurements, with overwhelming probability, you have found the true signal.

High dimensional data analysis: a null model for sparse regression.

A_{nk} = n'th T dimensional "input" data $n = 1..N$
 y_n = n'th scalar "output" measurement $k = 1..T$

We wish to explain the relation between inputs and outputs via a sparse rule x : I.e. $y_n = \sum_k A_{nk} x_k$ for each n

Suppose we do L1 regularized regression and we
Get a candidate rule x_{est} .

Is x_{est} sparse? Need a null model for sparsity in high dimensional data analysis. Analyze random data: independent gaussian y and A

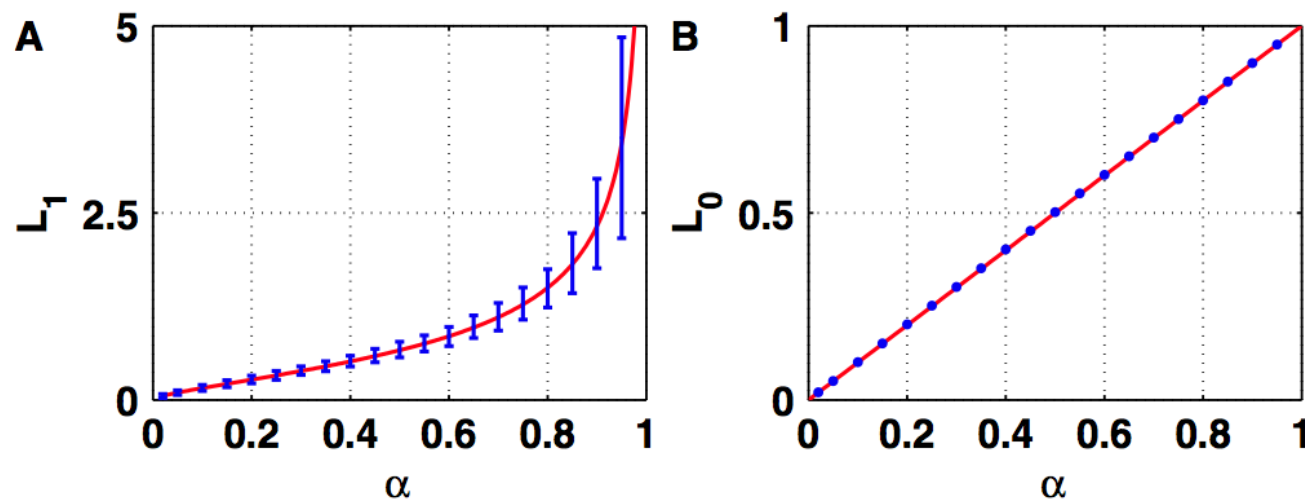
$$E(\mathbf{x}) = \frac{\lambda}{2T} (\mathbf{y} - \mathbf{A}\mathbf{x})^T (\mathbf{y} - \mathbf{A}\mathbf{x}) + \sum_{i=1}^T |x_i|.$$

High dimensional data analysis: a null model for sparse regression.

A_{nk} = n'th T dimensional "input" data $n = 1..N$

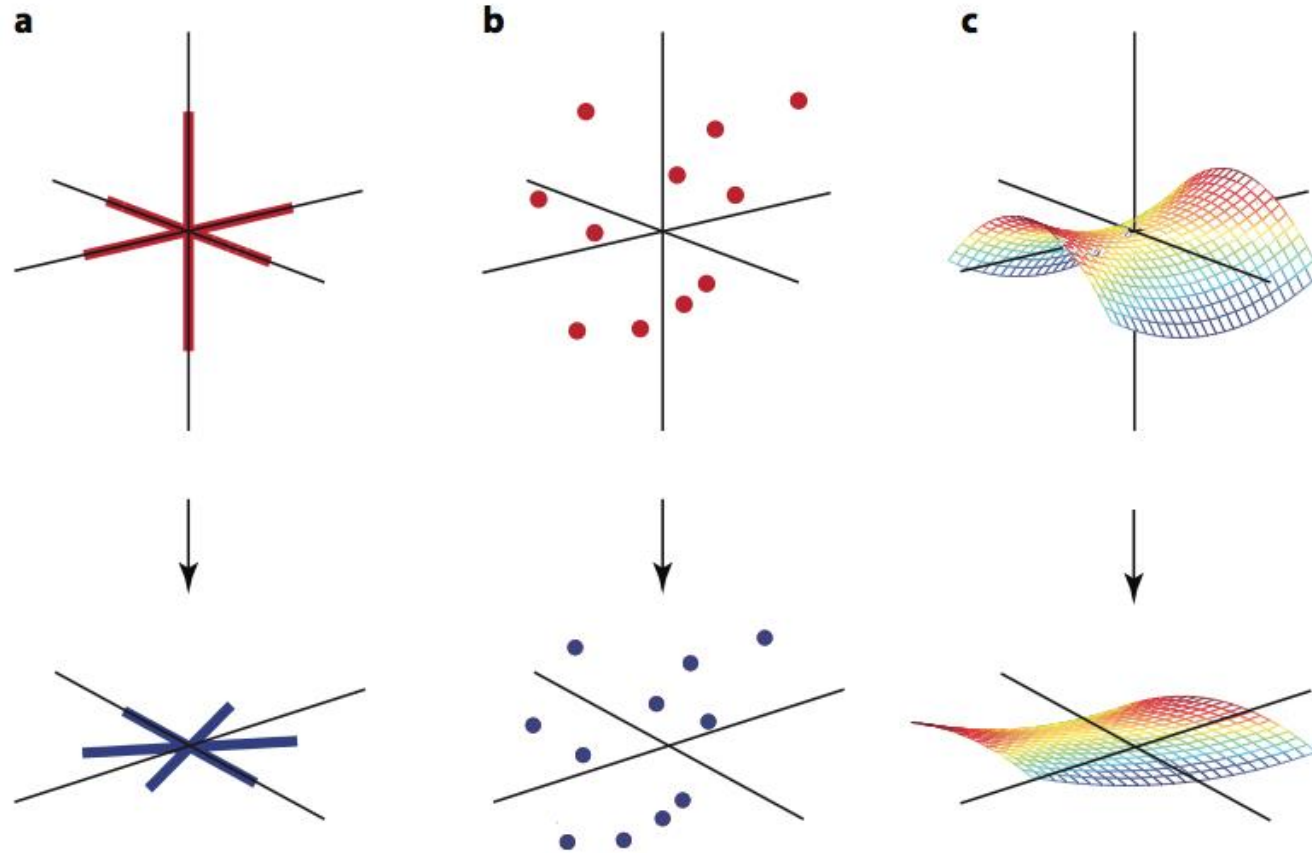
y_n = n'th scalar "output" measurement $k = 1..T$

We wish to explain the relation between inputs and outputs via a sparse rule x : I.e. $y_n = \sum_k A_{nk} x_k$ for each n



Expected sparsity (L1 and L0 norms) of a candidate rule that Explains random data (gaussian inputs and outputs).

A larger context: random projections



$\mathbf{x} = \mathbf{A}\mathbf{s}$ is a random projection from a T dim space down to an N dim space

Data / interesting signals live on a K -dim submanifold in T -dim space

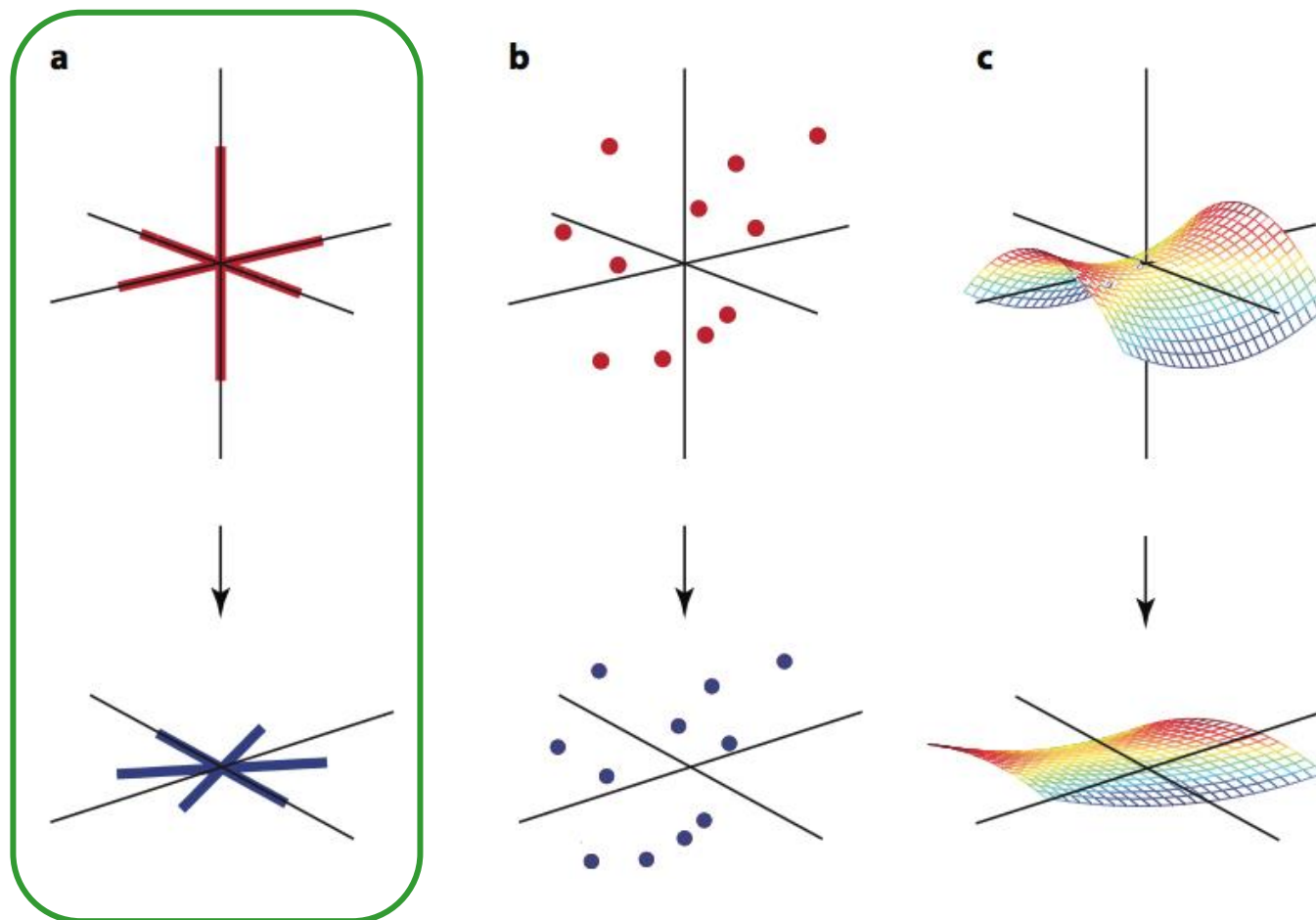
When will the geometry of this manifold be preserved under a random proj. ?

Distortion: $D_{ab} = (\| \mathbf{A}\mathbf{s}^a - \mathbf{A}\mathbf{s}^b \|^2 - \| \mathbf{s}^a - \mathbf{s}^b \|^2) / \| \mathbf{s}^a - \mathbf{s}^b \|^2$

A larger context: random projections

K-dim manifold
T-dim space

Random proj
to N-dim space



Manifold of K-sparse signals = Union of N choose K K-dim hyperplanes

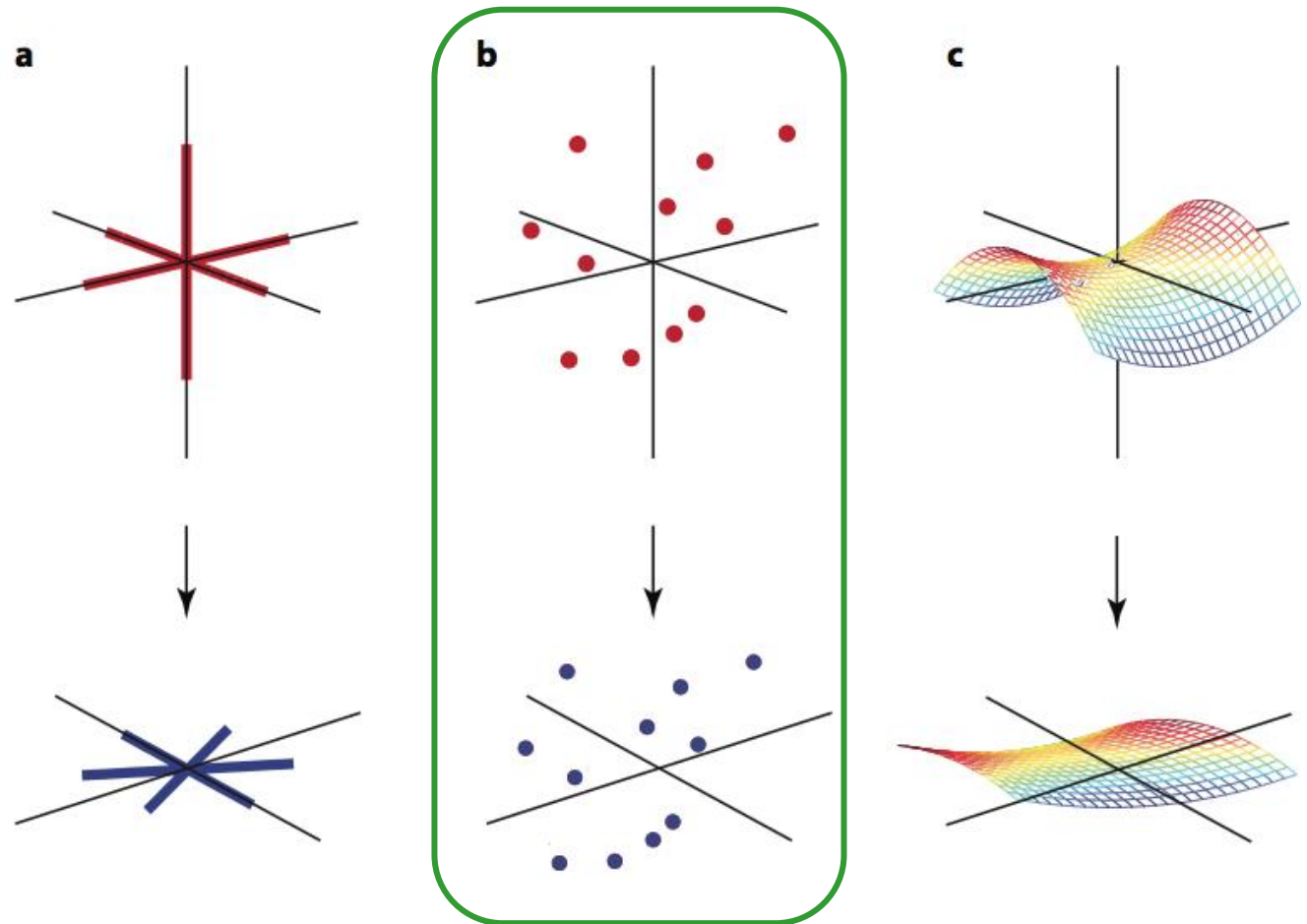
As long as $N > O(K \log T/K)$, then $\max_{ab} |D_{ab}| = O(1)$ with high prob over random choice of projection \mathbf{A} Baraniuk et. al. 2008

Deterministic result: for any projection \mathbf{A} with small distortion, one can reconstruct sparse signal from its projection (i.e. compute its pre-image)

A larger context: random projections

(K-dim) manifold
T-dim space

Random proj
to N-dim space



Point cloud = Union of P points in T-dim space choose K K-dim hyperplanes

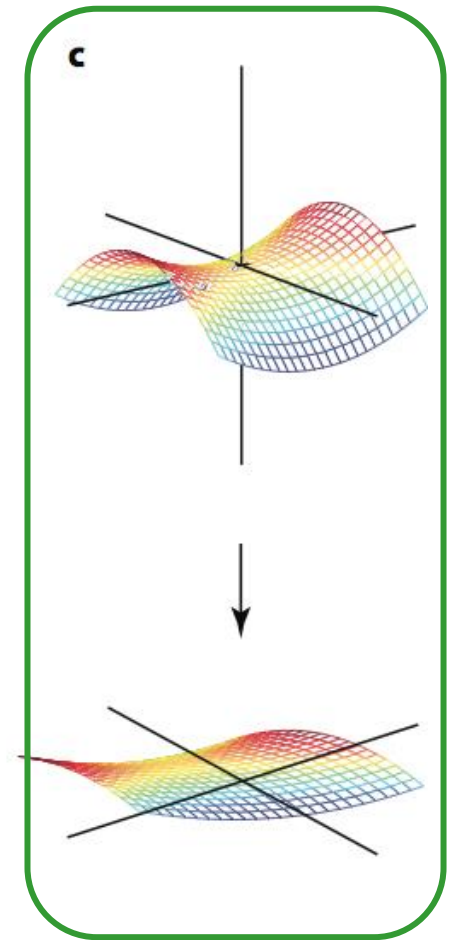
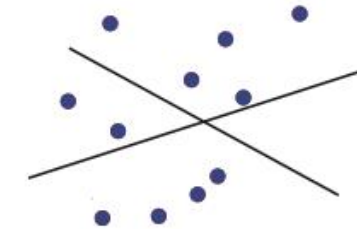
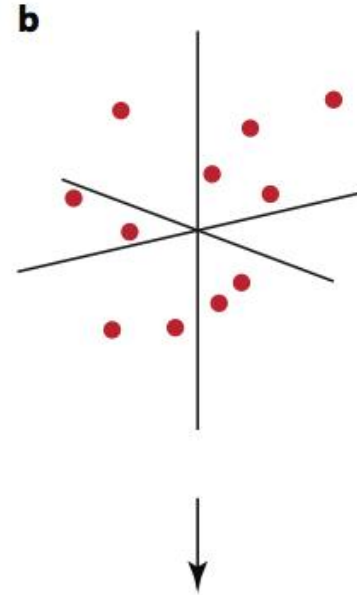
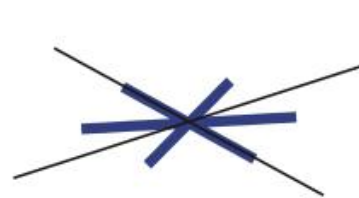
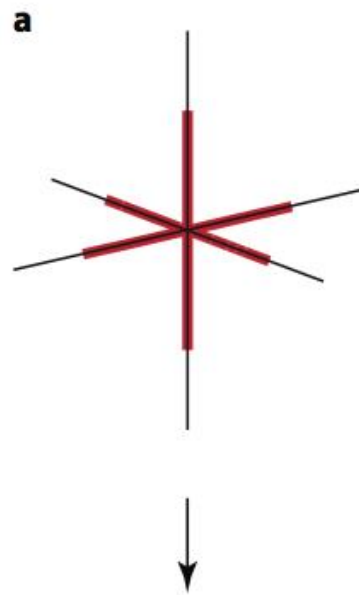
As long as $N > O(\log P)$, then $\max_{ab} |\mathbf{D}_{ab}| = O(1)$ with high prob over random choice of projection \mathbf{A}

Johnson-Lindenstrauss Lemma

Compressed computation: with so few measurements, one cannot recover high-dim points, but any algorithm which depends on pairwise distances can be applied in low dim space

A larger context: random projections

(K-dim) manifold
T-dim space



Random proj
to N-dim space

Arbitrary K-dim manifold in T dim space

As long as $N > O(C K \log T)$, where C is related to curvature, then $\max_{ab} |D_{ab}| = O(1)$ with high prob over random choice of projection \mathbf{A}

Recovery is difficult (nonconvex) but **compressed computation** still applies

The fundamental problem of short term memory.

We can remember multiple stimuli over the time course of seconds.
(e.g. speech, phone numbers...)

Isolated neurons forget synaptic inputs on the time course of milliseconds.

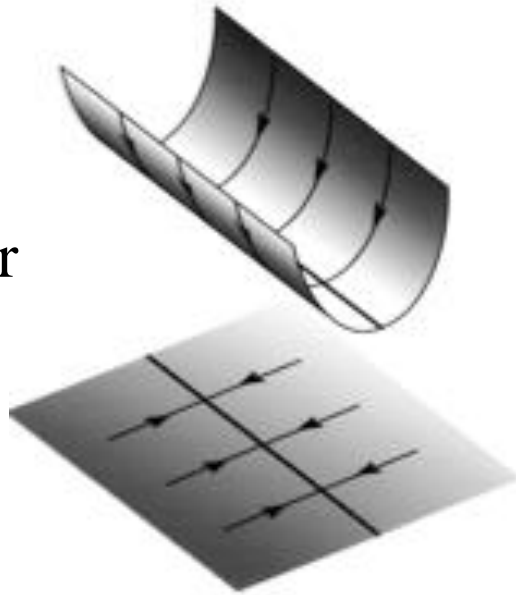
So to mediate short-term memory, networks of neurons must interact with each other to keep our memories alive.

But what kind of interactions are capable of extending single neuron memory to the cognitive timescale?

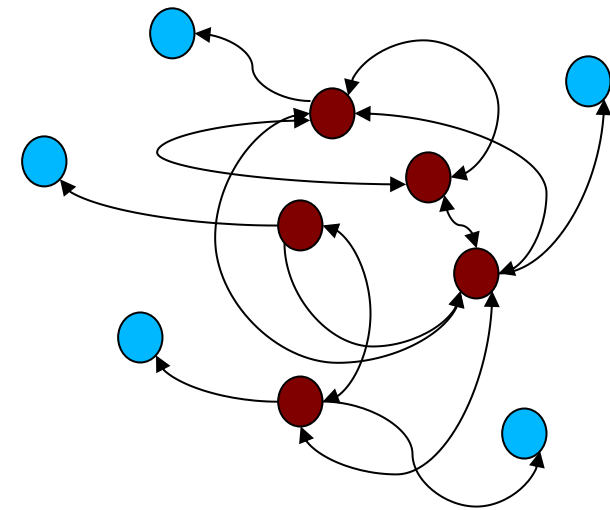
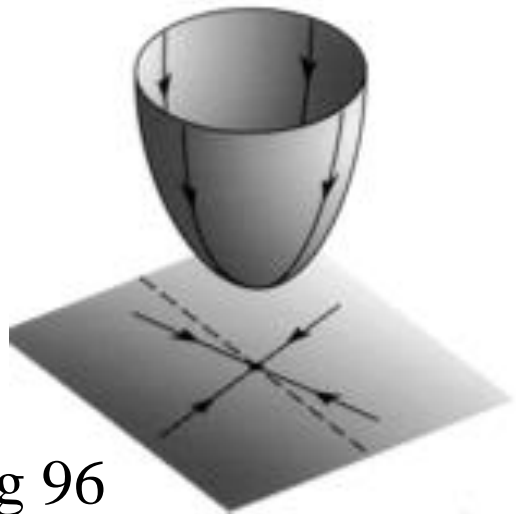
And how can networks store multiple items in a temporal sequence?

An Old Paradigm: Persistent Activity Stabilized by Attractor Dynamics in Recurrent Networks

Line Attractor

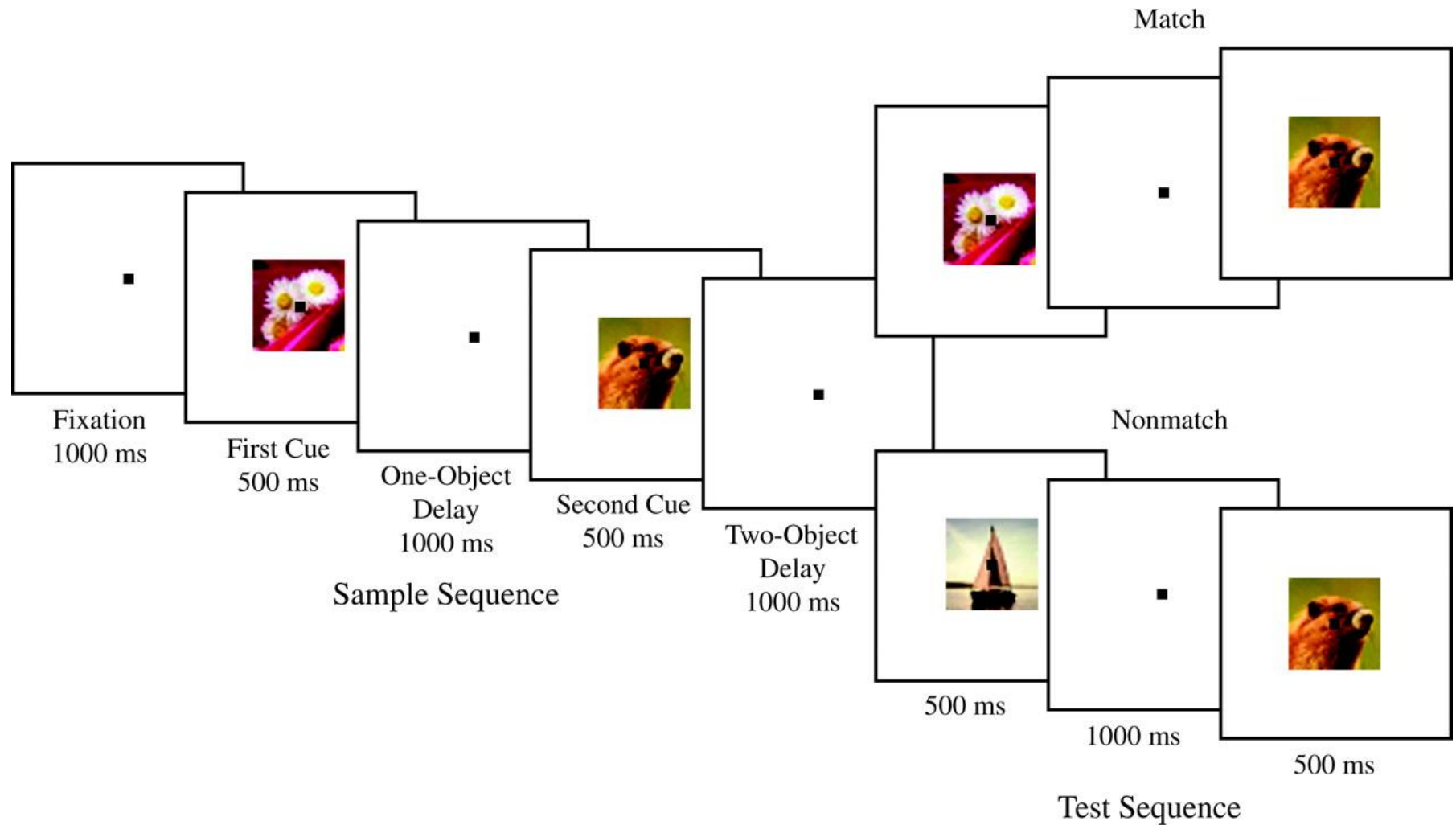


Fixed Point
Attractor

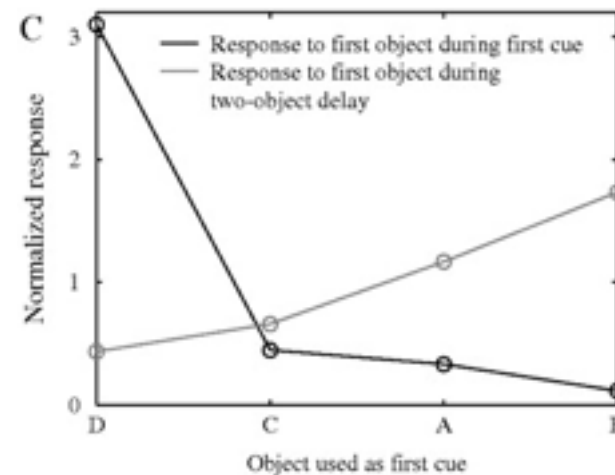
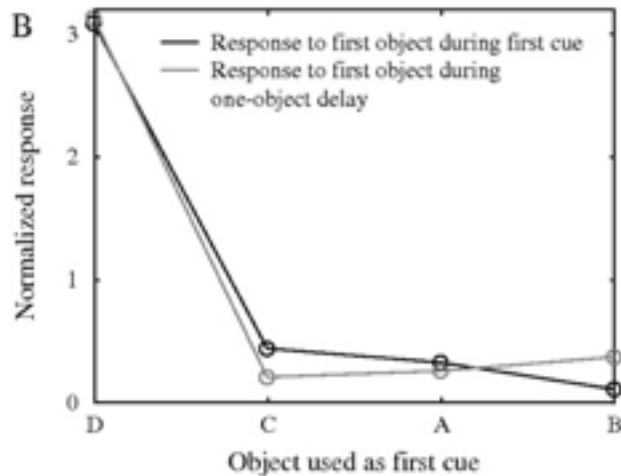
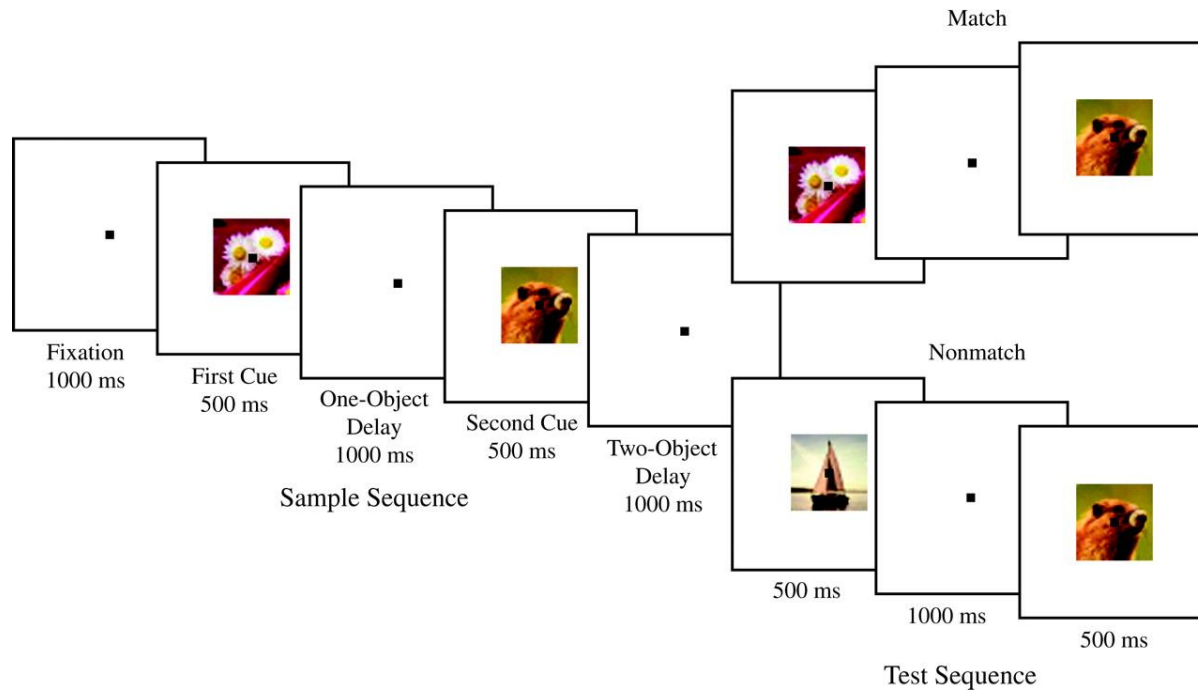


Positive Feedback

Probing sequence memory in the macaque brain.



Probing sequence memory in the macaque brain.



An Alternate Paradigm: The liquid brain / echo state hypothesis

“ If the recurrent circuit is sufficiently complex, its inherent dynamics automatically absorbs and stores information from the incoming input stream ”.

- *Markram, Natschlager, and Maas, 2001*
also: Buonomano and Merzenich, 1995
Mayor and Gerstner, 2003

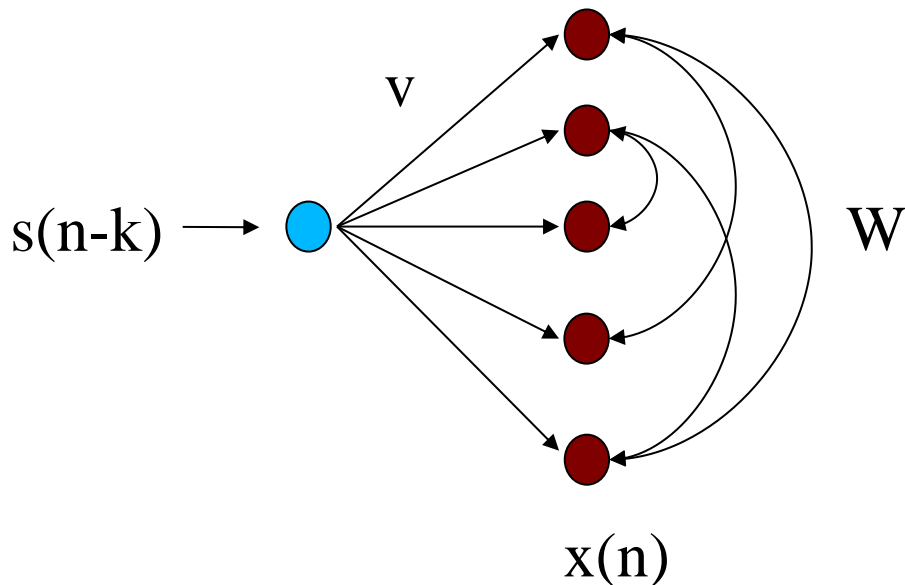
The basic idea of echo state network is to use a large reservoir RNN as a supplier of interesting dynamics from which the desired output is combined.”

- *Herbert Jaeger 2001*

An Alternate Paradigm: The liquid brain / echo state hypothesis



An Alternate Paradigm: The liquid brain / echo state hypothesis



Maass, Natschlager Markram, 2002:

$N = 135$ neurons

Membrane Time Const: 20ms

Synaptic Time Constants: 1 sec

Memory: ~ 1 sec.

Goals

Generate a theoretical framework within which one can define the memory capacity of such networks.

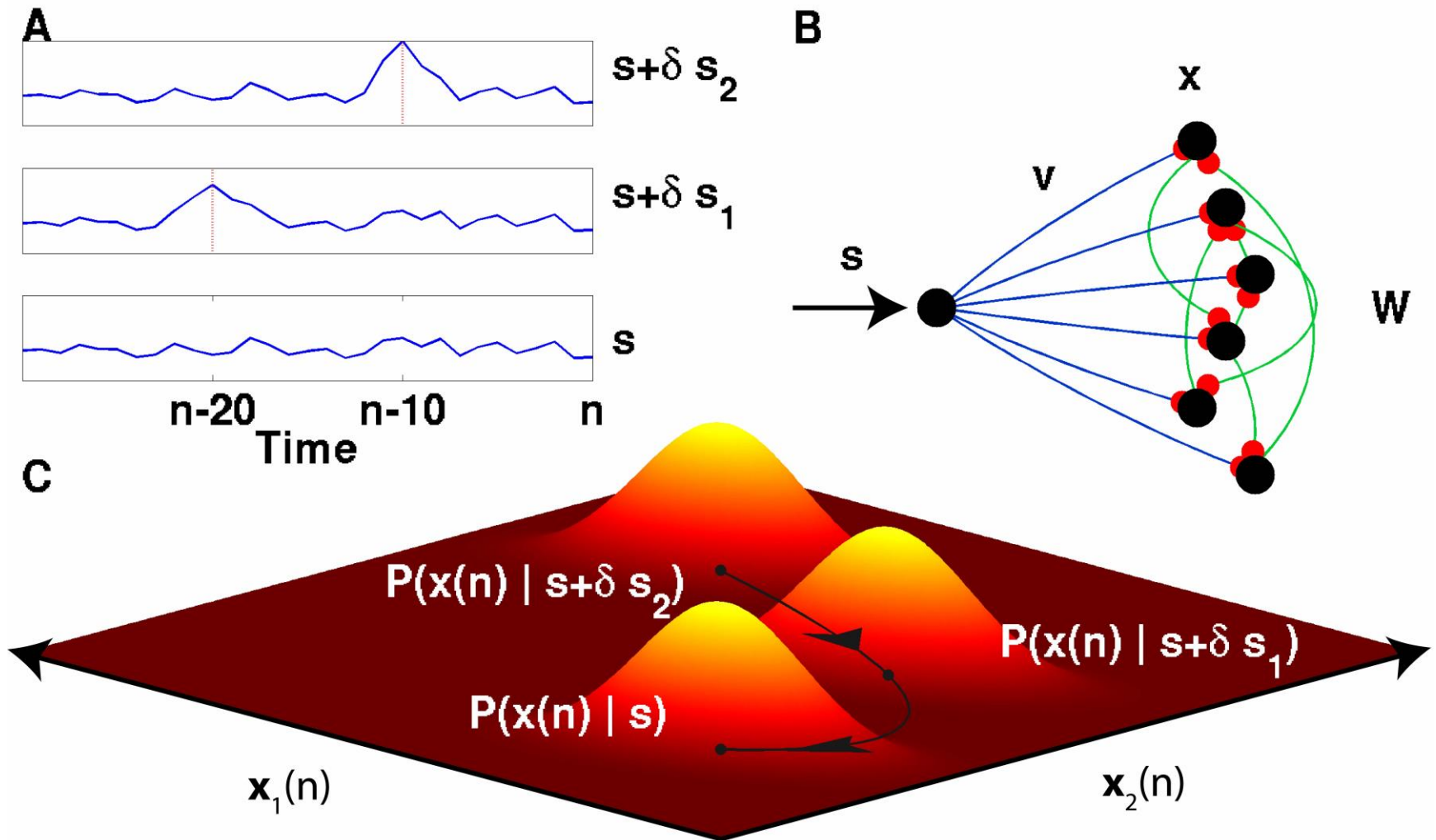
Compute this capacity analytically.

Understand its dependence on circuit connectivity and noise in the system.

Extract fundamental performance limits or tradeoffs.

Find and understand optimal networks which achieve these performance limits: **What are the design principles?**

Storing temporal information in a spatial network state.



Signal Power = 1
Noise Power = ε

$$\mathbf{x}(n) = \mathbf{W}\mathbf{x}(n-1) + \mathbf{v}s(n) + \mathbf{z}(n).$$

Memory Traces through Fisher Information

Two Dual Viewpoints on Memory:

1) Memory = Ability to use the present to reconstruct the past.

White, Lee, Sompolinsky, PRL., 2004.

2) Memory = The ability of the past to change the present.

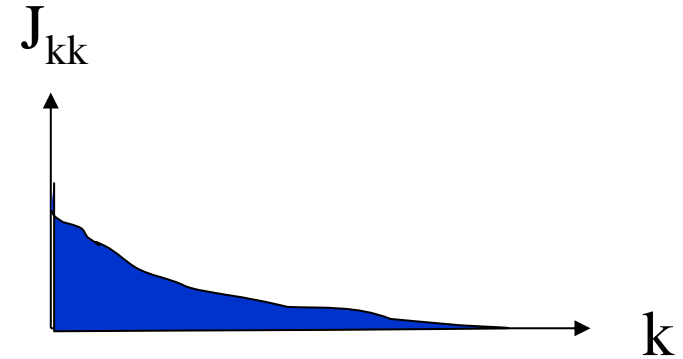
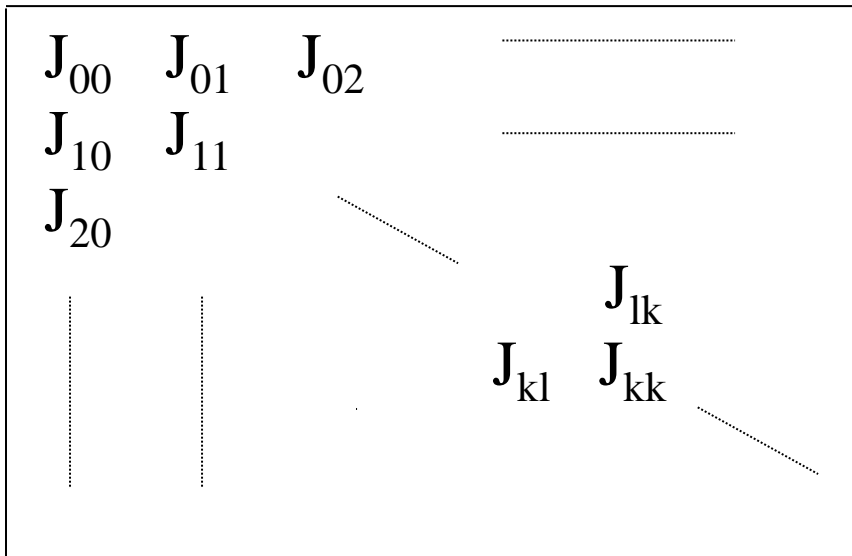
i.e. How much does $P(\mathbf{x} | \mathbf{s})$ change when you change \mathbf{s} . This is captured by the Fisher Information Matrix:

$$J_{k_1 k_2} = - \left\langle \frac{\partial^2}{\partial s(n-k_1) \partial s(n-k_2)} \log P(\mathbf{x}(n) | s(n), s(n-1), \dots) \right\rangle_{P(\mathbf{x}(n) | \vec{s})}$$

Consider a change in the signal: $\mathbf{s} \rightarrow \mathbf{s} + d\mathbf{s}$.

Then the distribution of $\mathbf{x}(n)$ will change by an amount $\sim d\mathbf{s}^T \mathbf{J} d\mathbf{s}$.

The Matrix Nature of Memory

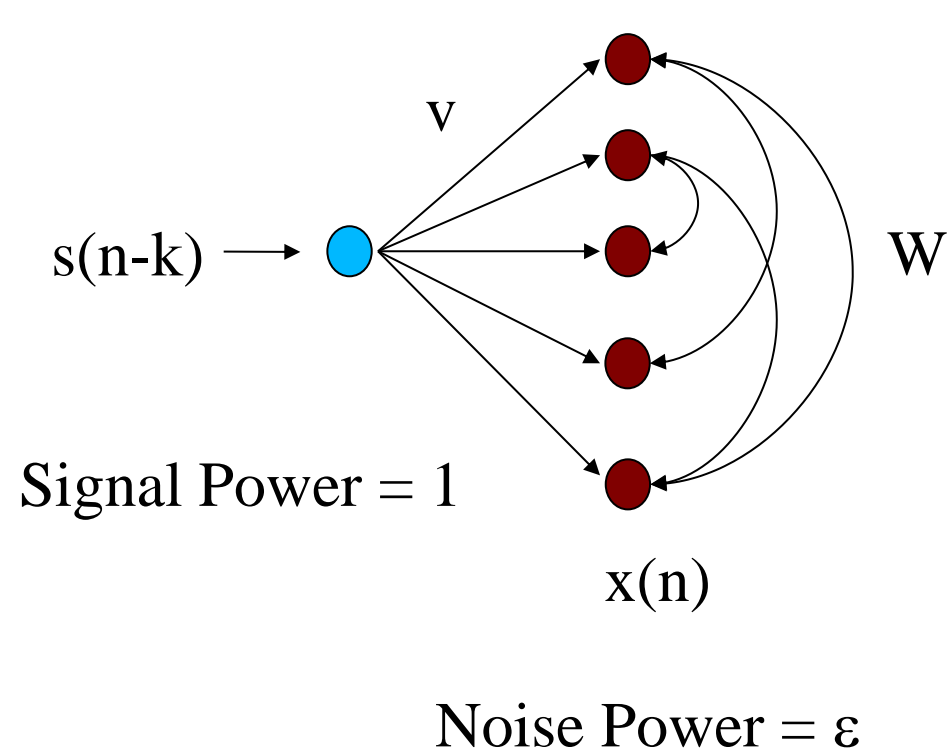


$$J_{\text{Tot}} = \sum_{k=0}^{\infty} J_{kk}$$

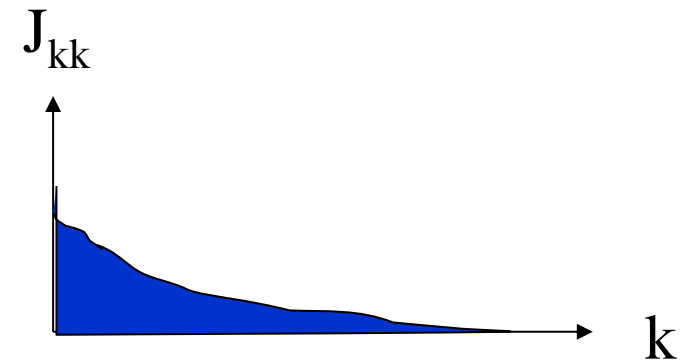
J_{kk} = Amount of information $x(n)$ retains about a single pulse that enters the network k time steps in the past.

J_{kl} = Amount of interference between the memory traces of two pulses entering at different times k and l in the past.

A Fundamental Performance Limit on Memory



$$J_{k_1 k_2} = \frac{1}{\epsilon} v^T W^T k_1 \left[\sum_{k=0}^{\infty} W^k W^T k \right]^{-1} W^T k_2 v$$



Instantaneous SNR at input:

$$1/\epsilon$$

$$J_{\text{Tot}} \equiv \sum_{k=0}^{\infty} J_{kk} \leq \frac{N}{\epsilon}$$

For *any* choice of W and v !

A simple (but large) class of networks: normality.

Assume W is normal: i.e. W has an orthogonal basis of eigenvectors.

Then the memory performance only depends on the eigenvalues of W , or the spectrum of network time constants present in the system.

Fundamental memory constraint for normal networks:

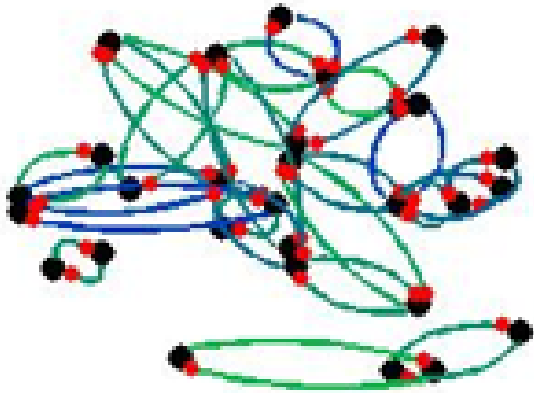
$$J_{\text{Tot}} \equiv \sum_{k=0}^{\infty} J_{kk} = \frac{1}{\epsilon}$$

Independent of W and v !

Normal networks cannot retain in their network state more SNR about the past signal history, than the instantaneous SNR at the input. They can only redistribute this SNR across time.

Examples of “Normal” Network Connectivities

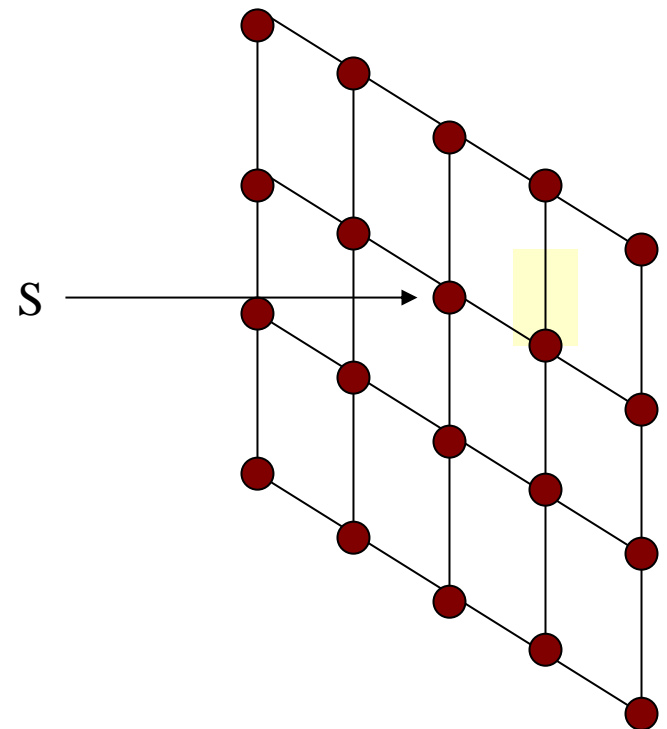
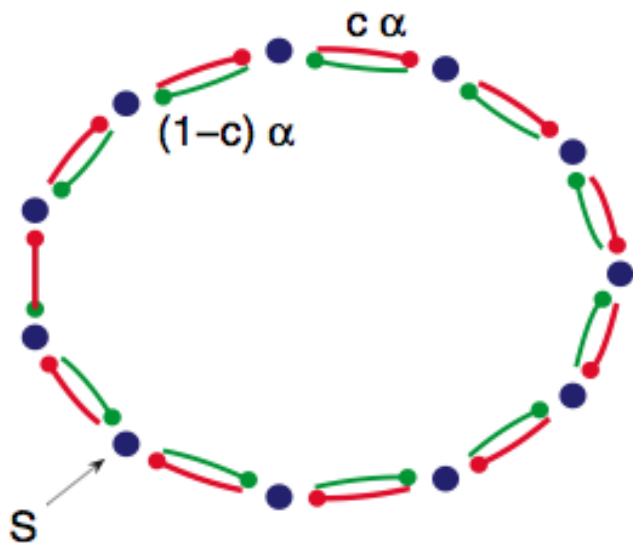
Any symmetric network.



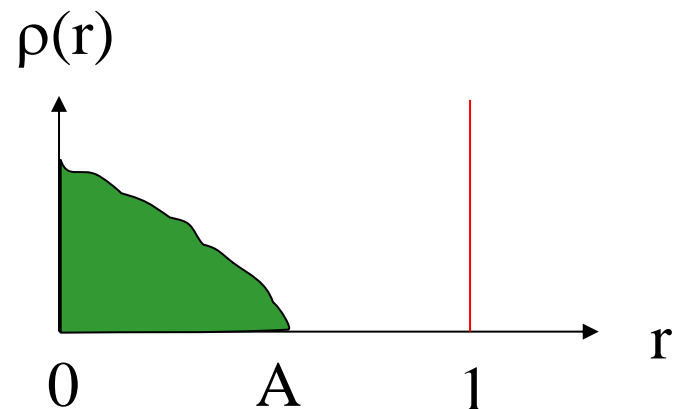
Any antisymmetric network.

Any orthogonal network.

Translation invariant lattices.



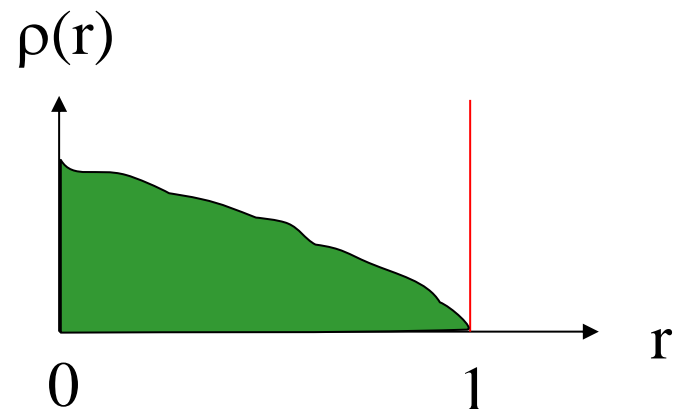
Power Law Tails in the Fisher Memory Matrix



Largest eigenvalue $A < 1$

$$J_{kk} \sim A^{-k}$$

Exponential Decay
of Memory



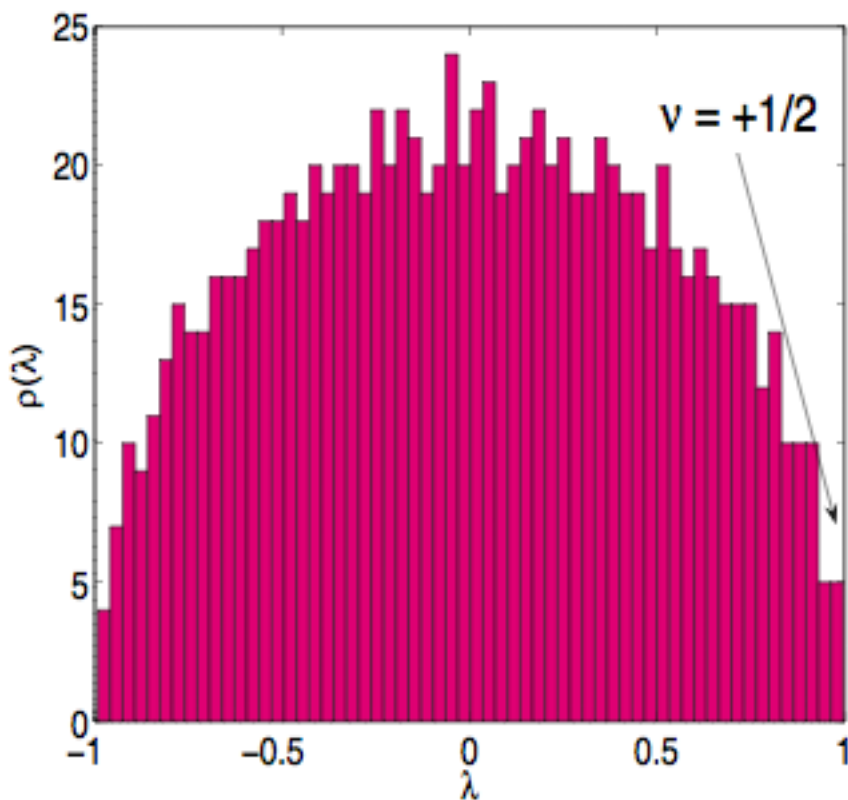
$$\lim_{r \rightarrow 1} \rho(r) = (1 - r)^\nu$$

$$J_{kk} \sim 1/k^{(\nu+2)}$$

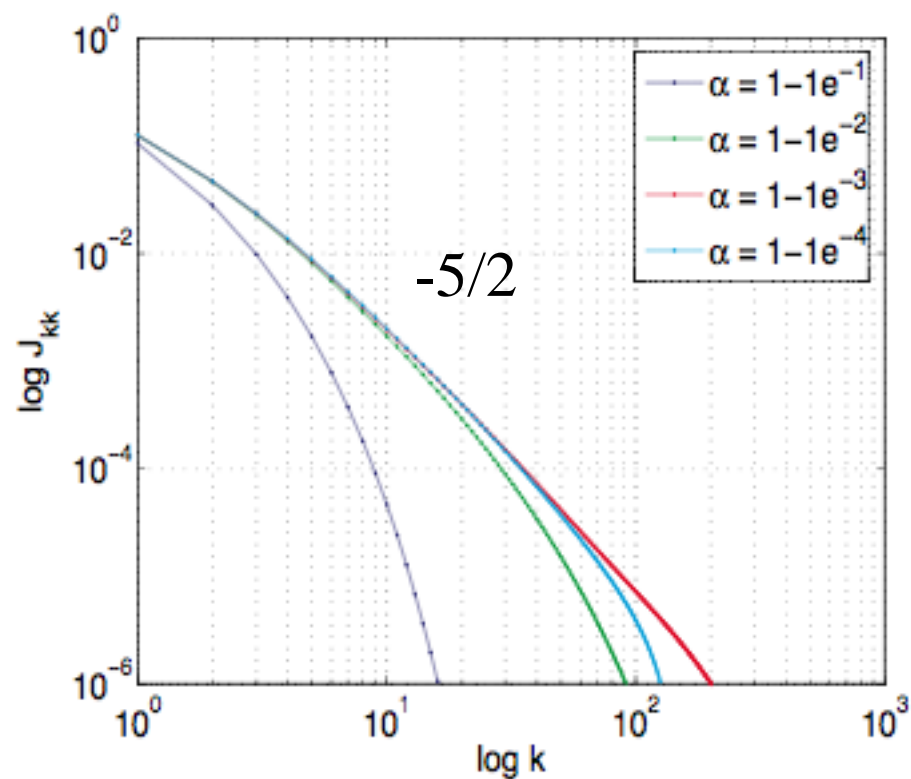
Power Law Decay
of Memory

An Example: Random Symmetric Matrices

Eigenvalue Spectrum



Power Law Tail of Memory

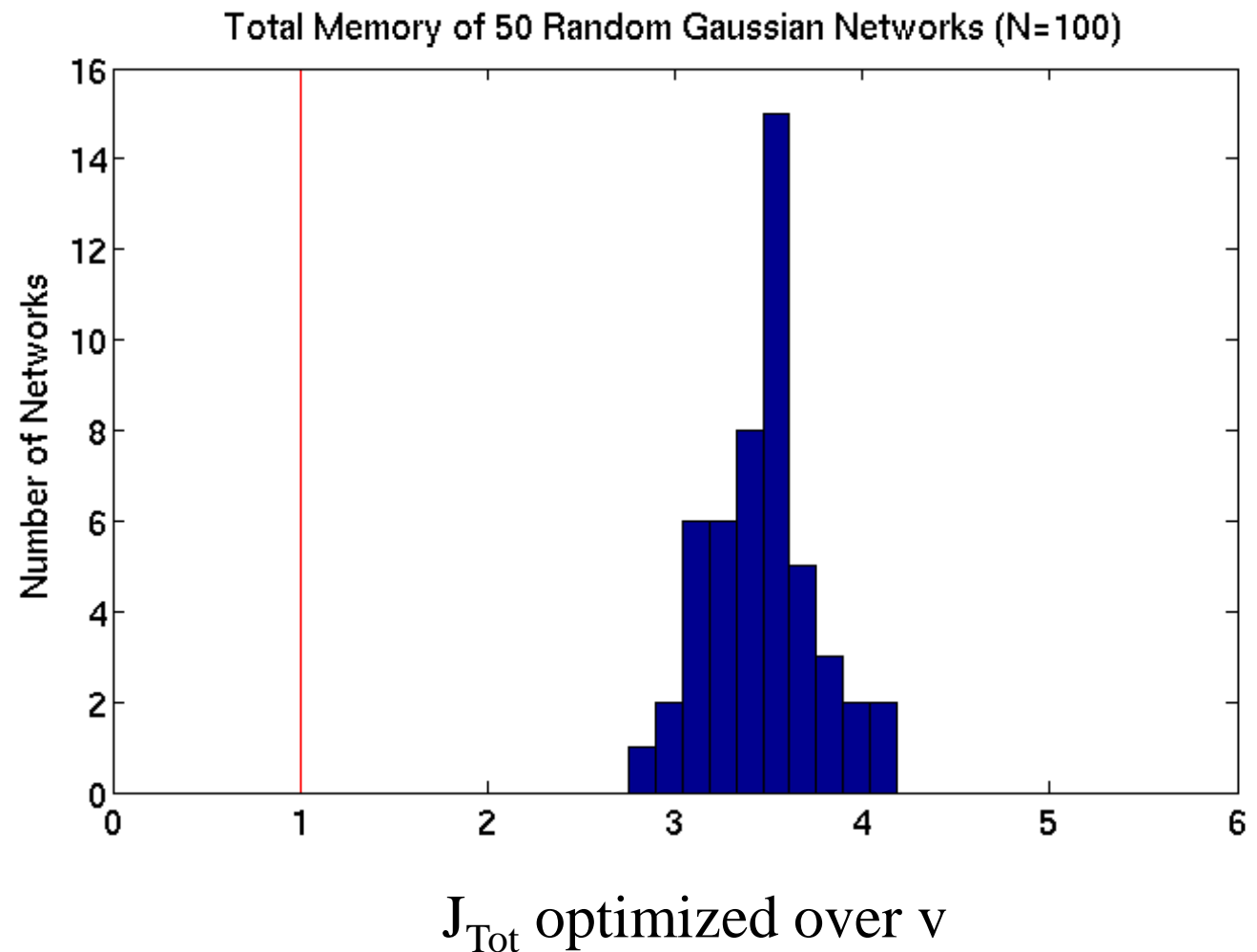


Memory Beyond the Normal Limit: Perturb Normality

Random Symmetric W \longrightarrow Random Asymmetric W

Now J_{Tot} depends
on W and v .

For a given W ,
optimize J_{Tot}
over v .



The nature of normal dynamics: independent decaying modes

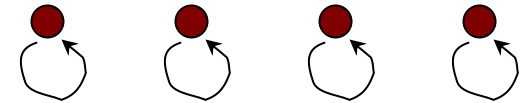
Eigenvector = Preferred Pattern or Mode of Activity across Neurons

Eigenvalue = Decay time constant of that pattern (larger value -> slower decay)

$$R(0) = c_1(0) V_1 + c_2(0) V_2 + \dots c_N(0) V_N$$

$$R(k) = c_1(k) V_1 + c_2(k) V_2 + \dots c_N(k) V_N$$

$$c_i(k) = a_i^k \quad |a_i| < 1$$



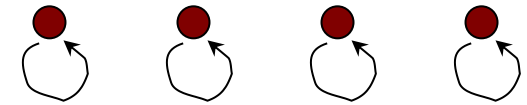
$$\text{Total network activity } R^2 = c_1^2 + c_2^2 + \dots + c_N^2$$

The nature of normal dynamics: independent decaying modes - the line attractor example

$$R(0) = c_1(0) V_1 + c_2(0) V_2 + \dots c_N(0) V_N$$

$$R(k) = c_1(k) V_1 + c_2(k) V_2 + \dots c_N(k) V_N$$

$$c_i(k) = a_i^k \quad |a_i| < 1$$



$$w/N \begin{pmatrix} 1 & 1 & \dots & 1 & 1 \\ 1 & 1 & \dots & 1 & 1 \\ \cdot & & & & \cdot \\ \cdot & & & & \cdot \\ 1 & 1 & \dots & 1 & 1 \end{pmatrix}$$

Slow mode:
(large eigenvalue)

$$\begin{pmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ 1 \end{pmatrix}$$

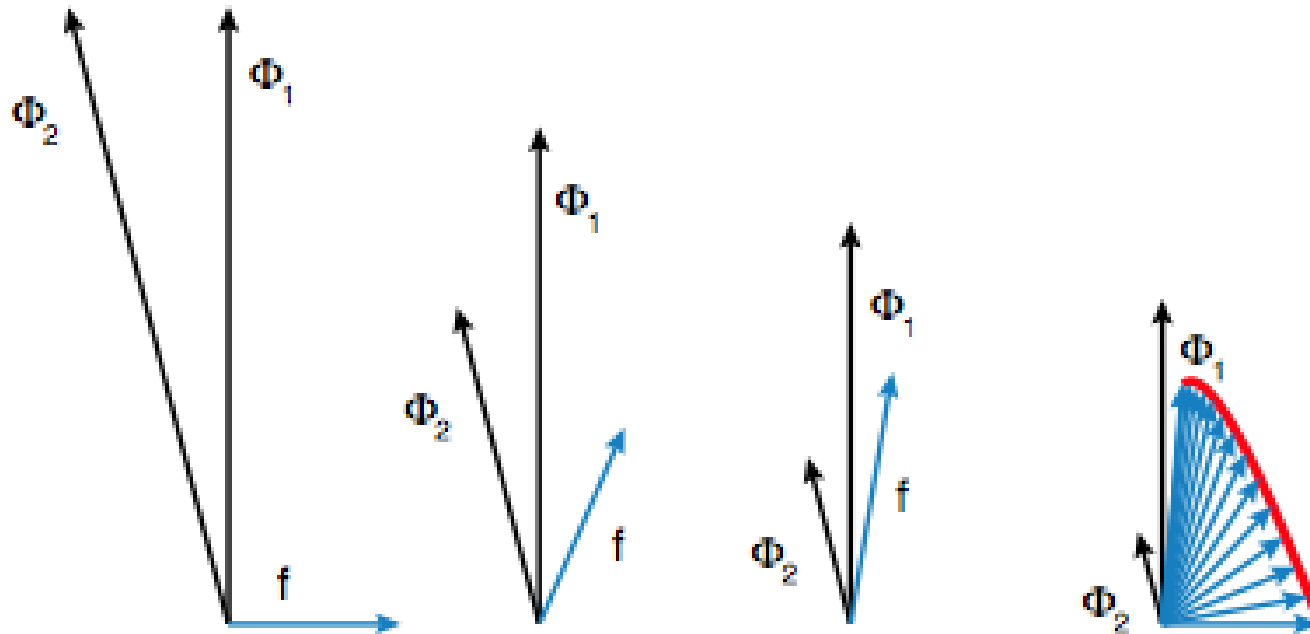
All other modes fast:
(small eigenvalues)

The nature of nonnormal dynamics: transient amplification from nonorthogonal eigenvectors

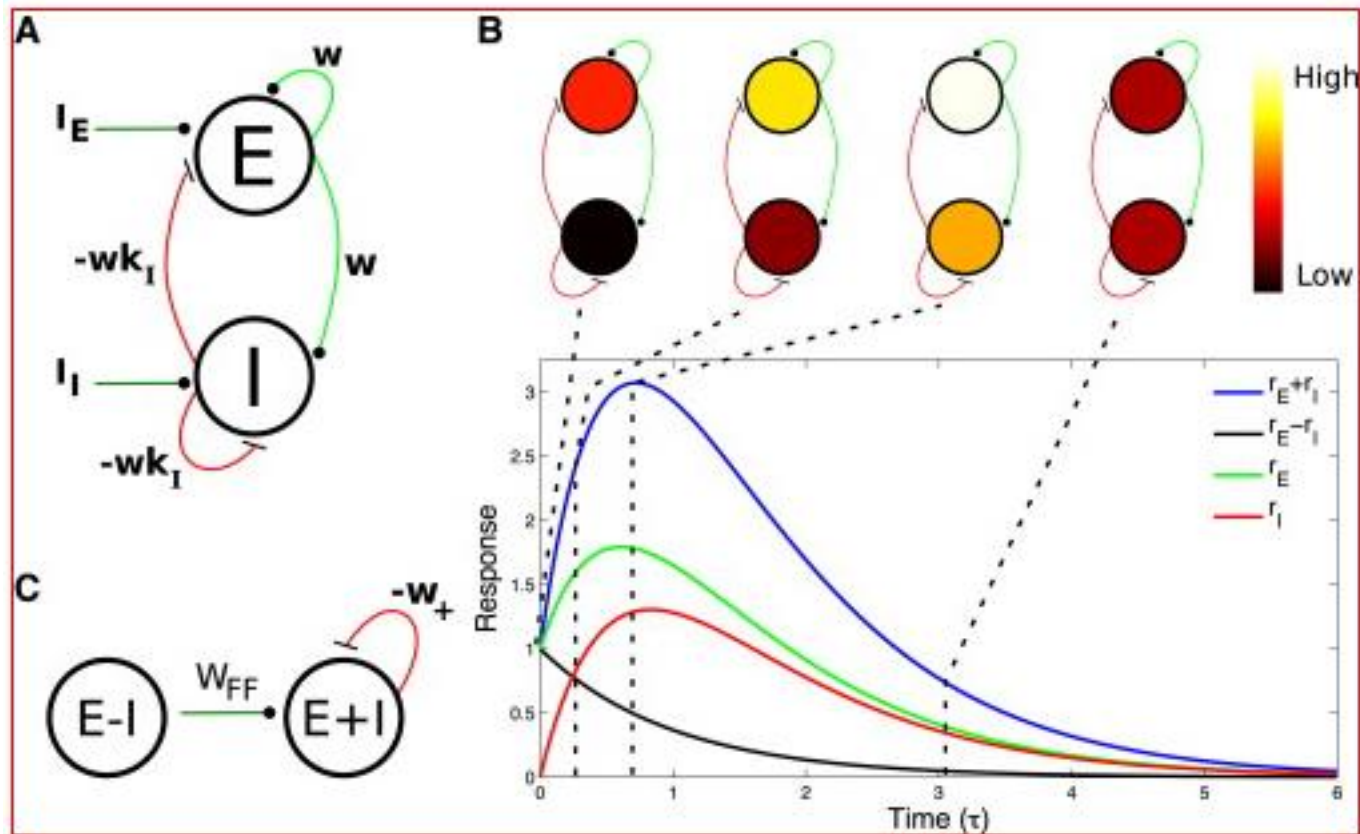
$$\mathbf{R}(0) = c_1(0) \mathbf{V}_1 + c_2(0) \mathbf{V}_2 + \dots c_N(0) \mathbf{V}_N$$

$$\mathbf{R}(k) = c_1(k) \mathbf{V}_1 + c_2(k) \mathbf{V}_2 + \dots c_N(k) \mathbf{V}_N$$

$$c_i(k) = a_i^k \quad |a_i| < 1$$

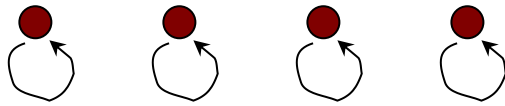


An simple two neuron example of transient amplification

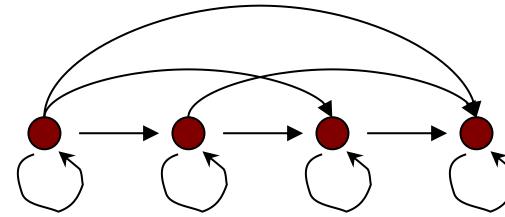


A Key Property of Nonnormal Networks: (Hidden) Feedforward Structure

Normal



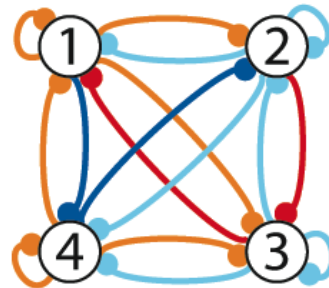
Non-normal



A



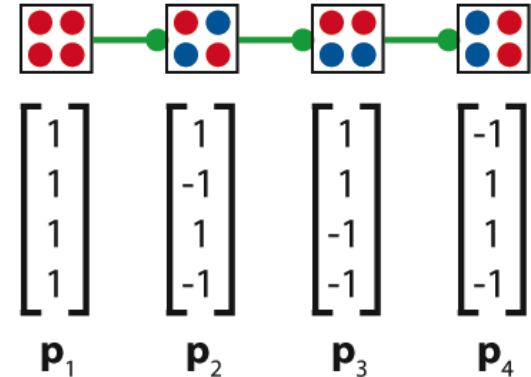
C



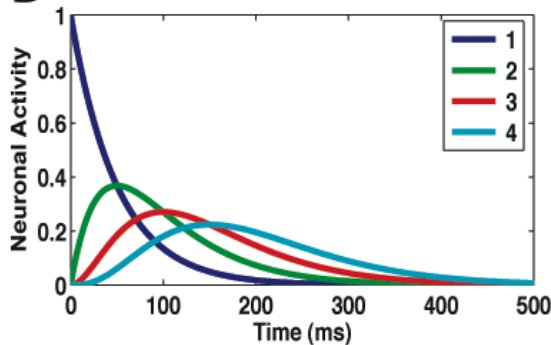
$$\frac{1}{4} \begin{bmatrix} 1 & -1 & 3 & 1 \\ 1 & -1 & -1 & -3 \\ 1 & 3 & -1 & 1 \\ -3 & -1 & -1 & 1 \end{bmatrix}$$

W

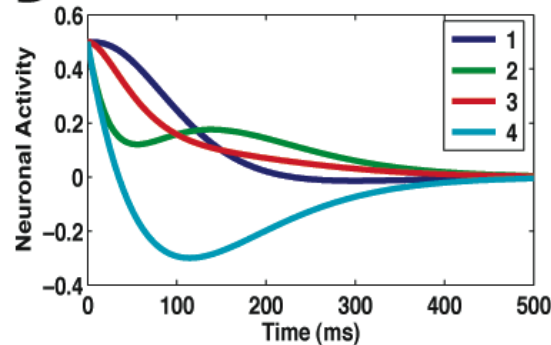
E



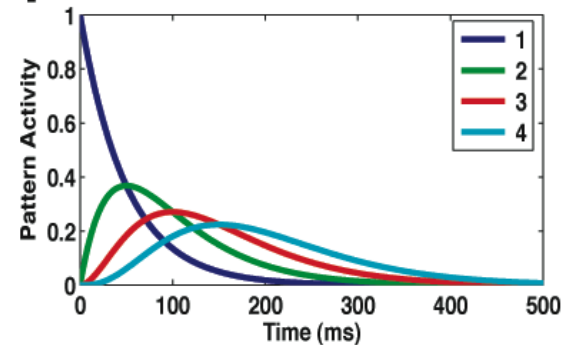
B



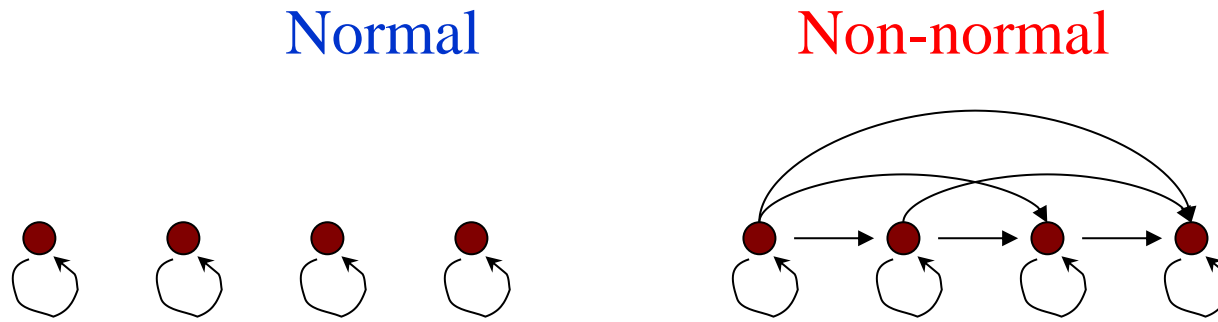
D



F



Related work



Related “Nonnormal” Work:

Trefethen and Embree (2005)


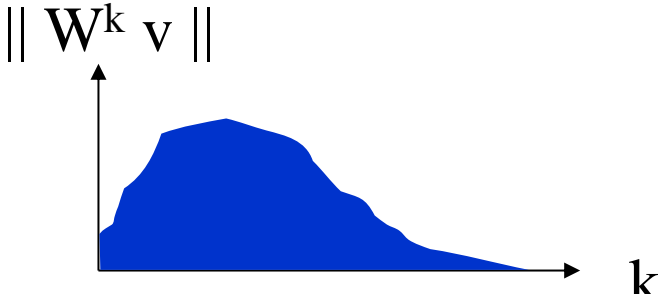
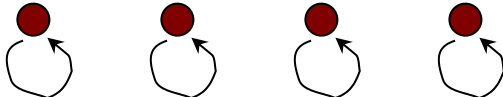
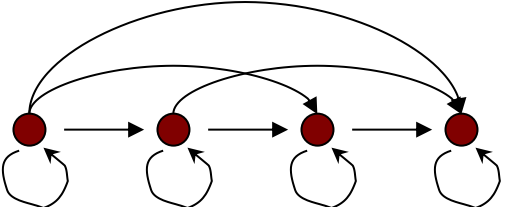
Ganguli, Huh, Sompolinsky PNAS (2008)

Murphy and Miller Neuron (2009).

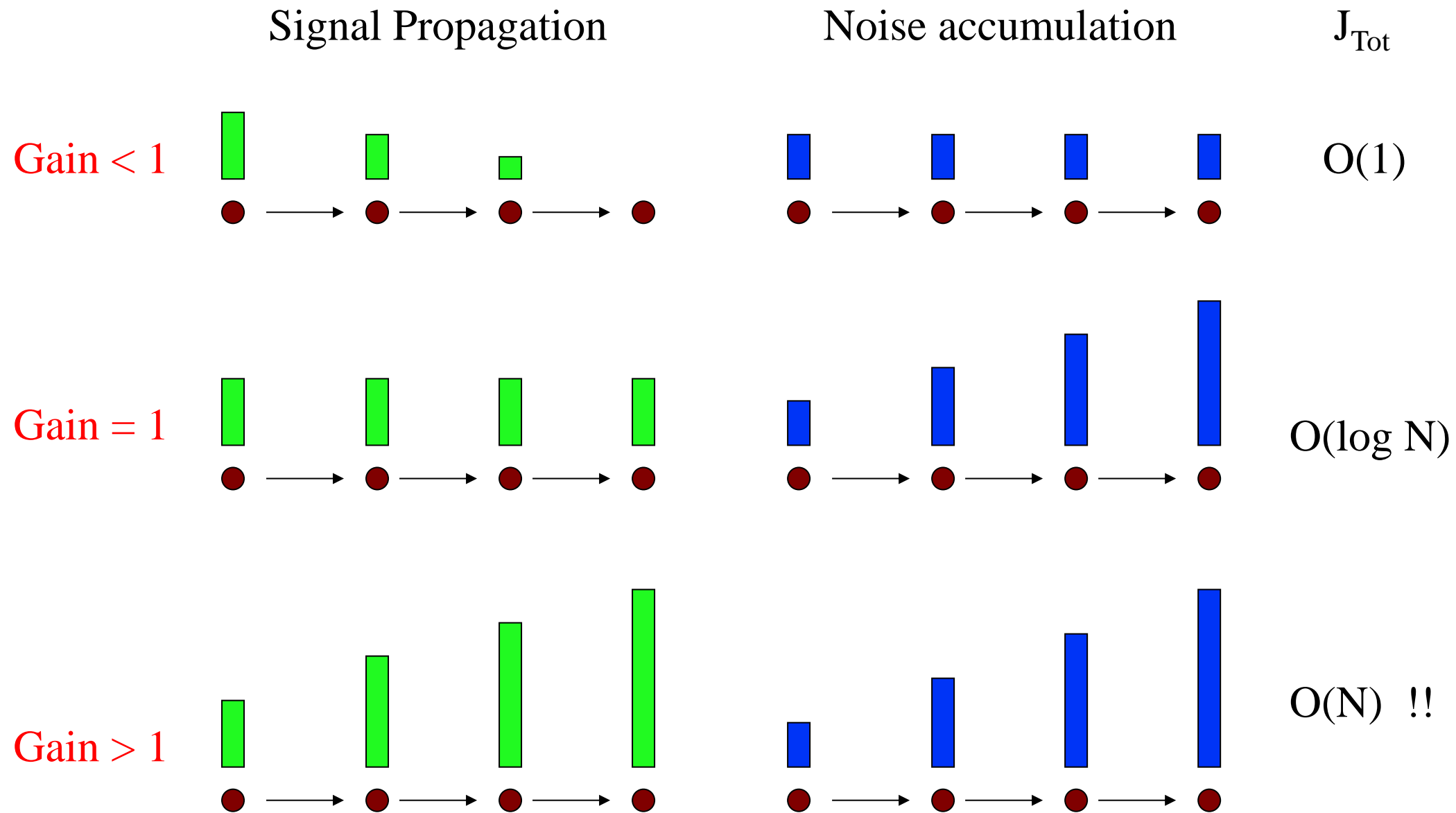
Goldman, Neuron (2009).

Ganguli and Latham, Neuron (2009).

The story so far

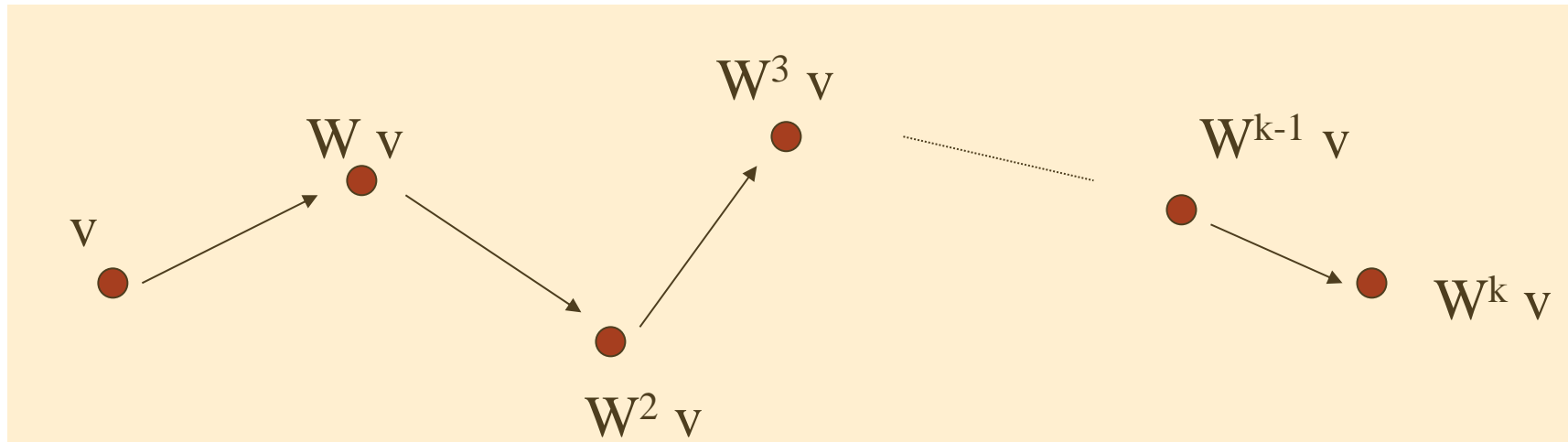
	Normal Networks	Nonnormal Networks
Information Theory	$\mathbf{J}_{\text{Tot}} \equiv \sum_{k=0}^{\infty} \mathbf{J}_{kk} = \frac{1}{\epsilon}$	$\mathbf{J}_{\text{Tot}} \equiv \sum_{k=0}^{\infty} \mathbf{J}_{kk} \leq \frac{N}{\epsilon}$
Dynamics	 <p>A plot showing the norm $\ W^k v\$ on the y-axis versus k on the x-axis. The curve is a decaying exponential-like shape, starting high and decreasing towards zero. A small yellow rectangular region highlights the area under the curve for small values of k.</p>	 <p>A plot showing the norm $\ W^k v\$ on the y-axis versus k on the x-axis. The curve is a bell-shaped distribution, starting at zero, rising to a peak, and then decaying towards zero.</p>
Hidden Structure	 <p>Four red circular nodes arranged horizontally. Each node has a self-loop arrow pointing back to itself.</p>	 <p>Four red circular nodes arranged horizontally. Each node has a self-loop arrow pointing back to itself. Additionally, there are curved arrows connecting the first node to the second, third, and fourth nodes, and the second node to the third and fourth nodes.</p>

Memory in the simplest feedforward network

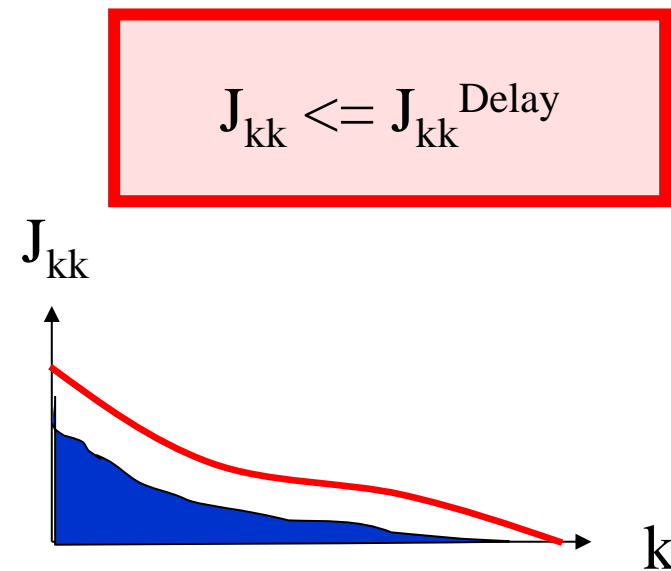
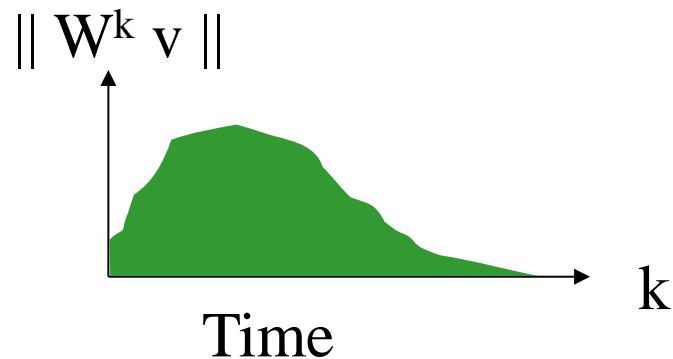


An upper bound on memory in any network

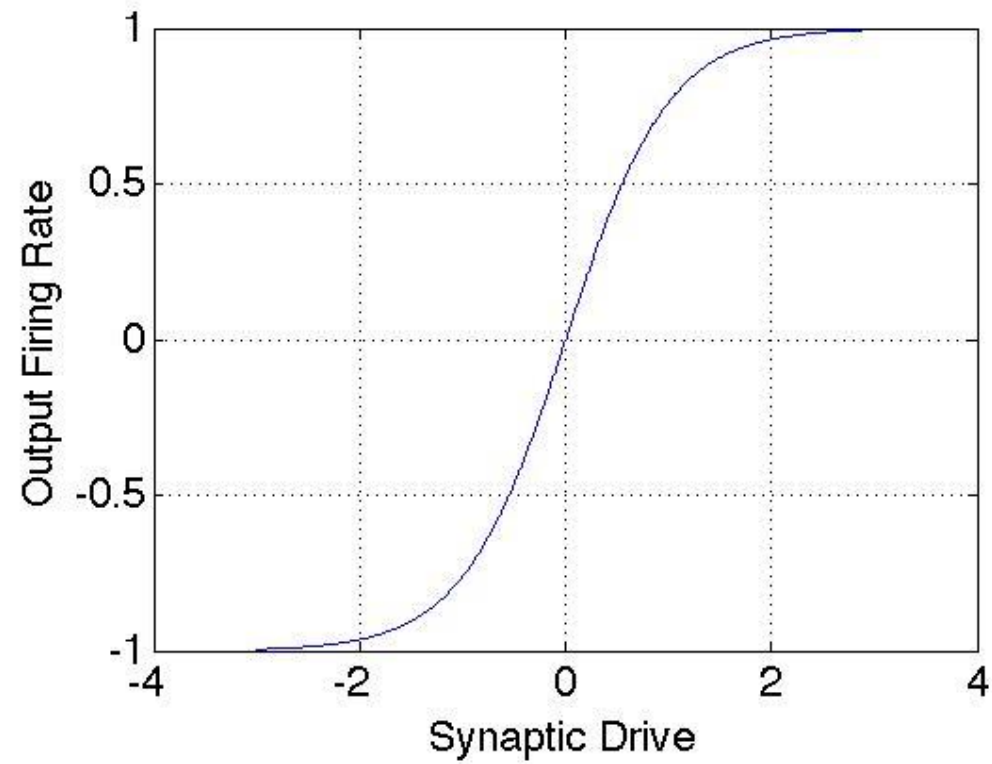
Dynamical propagation of a signal through network space.



Signal amplification profile



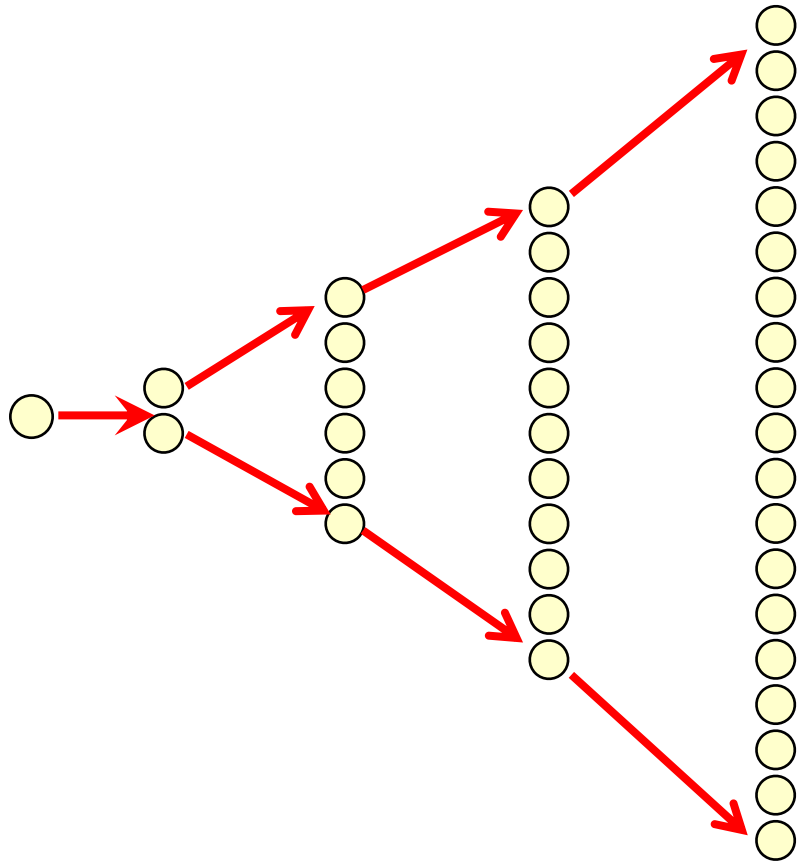
What about saturating nonlinearities?



Single neuron input output response

Signal Amplification in Nonlinear Dynamics

A Divergent Chain



Number of neurons
in a layer grows linearly in
the depth of the layer, so in
layer k

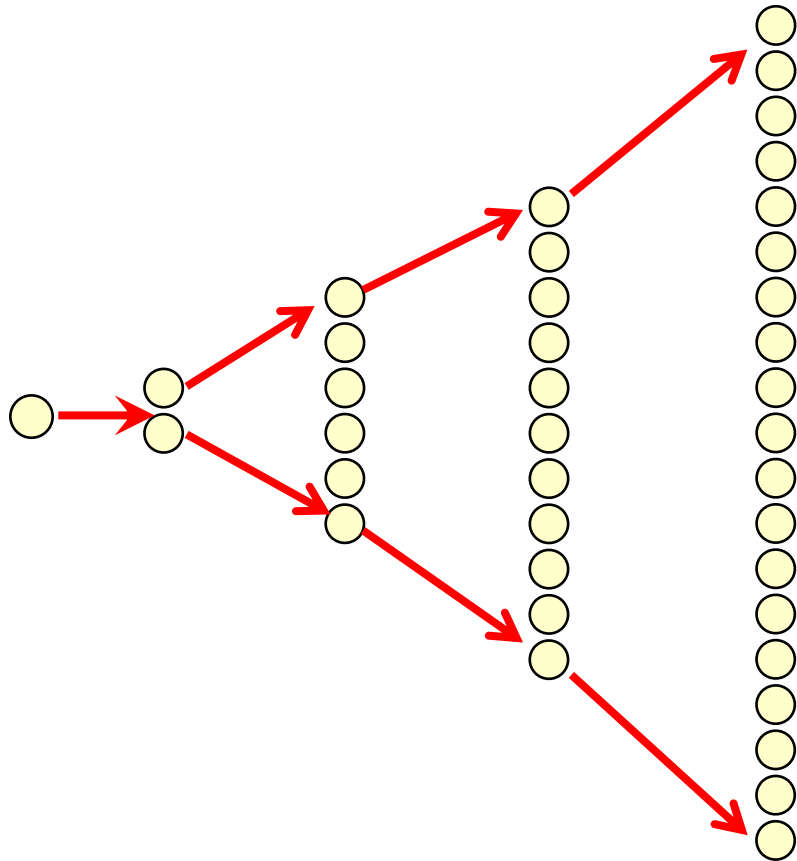
$$N_k \sim k$$

Strength of connections
between layer k and $k+1$:

$$\sim 1/k$$

Signal Amplification in Nonlinear Dynamics

A Divergent Chain with L layers



Number of neurons
in a layer grows linearly in
the depth of the layer, so in
layer k

$$N_k \sim k$$

Strength of connections
between layer k and $k+1$:

$$\sim 1/k$$

$$J_{\text{tot}} = L \sim \text{square root of } N$$

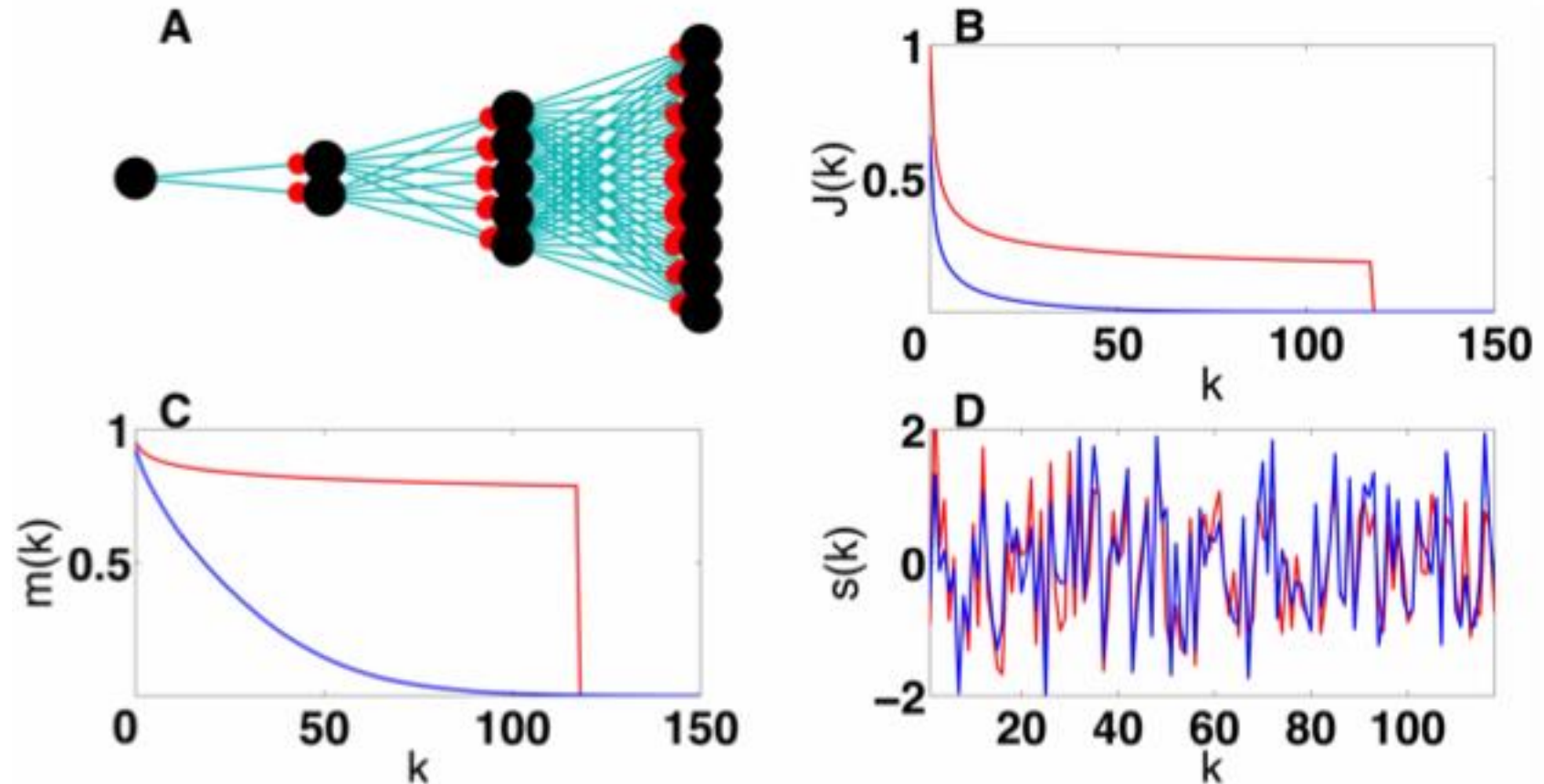
Consequences of finite dynamic range

$$\mathbf{J}_{\text{Tot}} \leq O\left(\frac{\sqrt{N}}{\epsilon}\right)$$

For any network operating in a linear regime in which neurons have a finite dynamic range.

Memory in nonlinear networks

Divergent chain: 135 layers, ~ 9000 neurons



Memory that lasts 135 times in intrinsic neuronal processing time scale!
Intrinsic scale = 10ms \Rightarrow 1.35 seconds of full sequence memory

The Liquid State Machine??

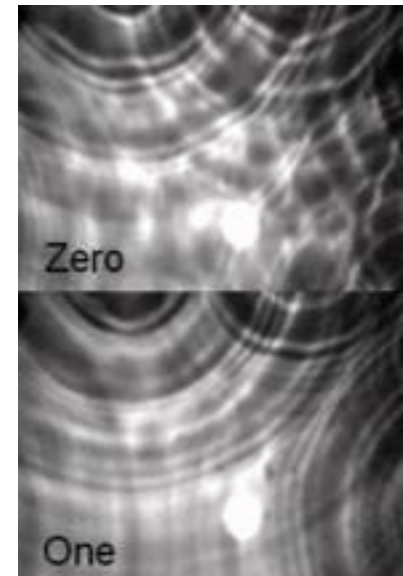


Too Normal! :(

The Liquid State Machine??



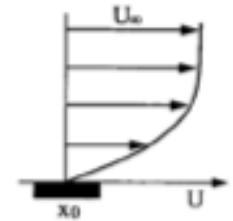
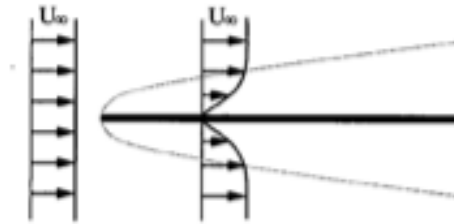
Fig. 1. The Liquid Brain.



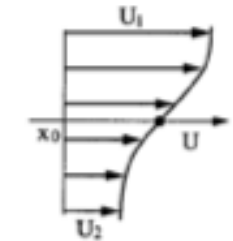
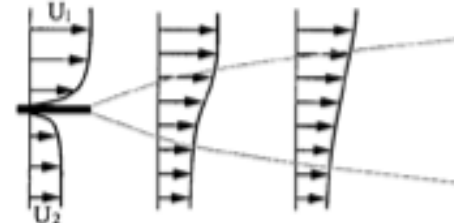
Chrisantha Fernando and Sampsa Sojakka, ECAL 2003

Better liquid states.

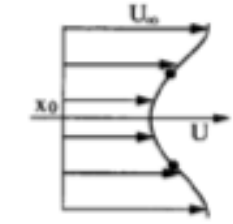
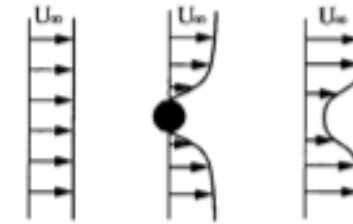
Flat plate boundary layer



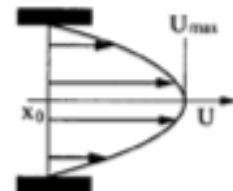
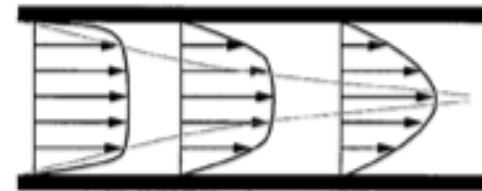
Mixing layer



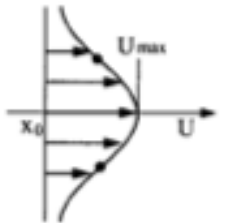
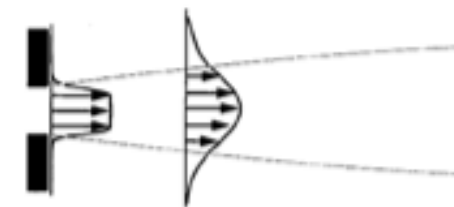
Cylinder wake



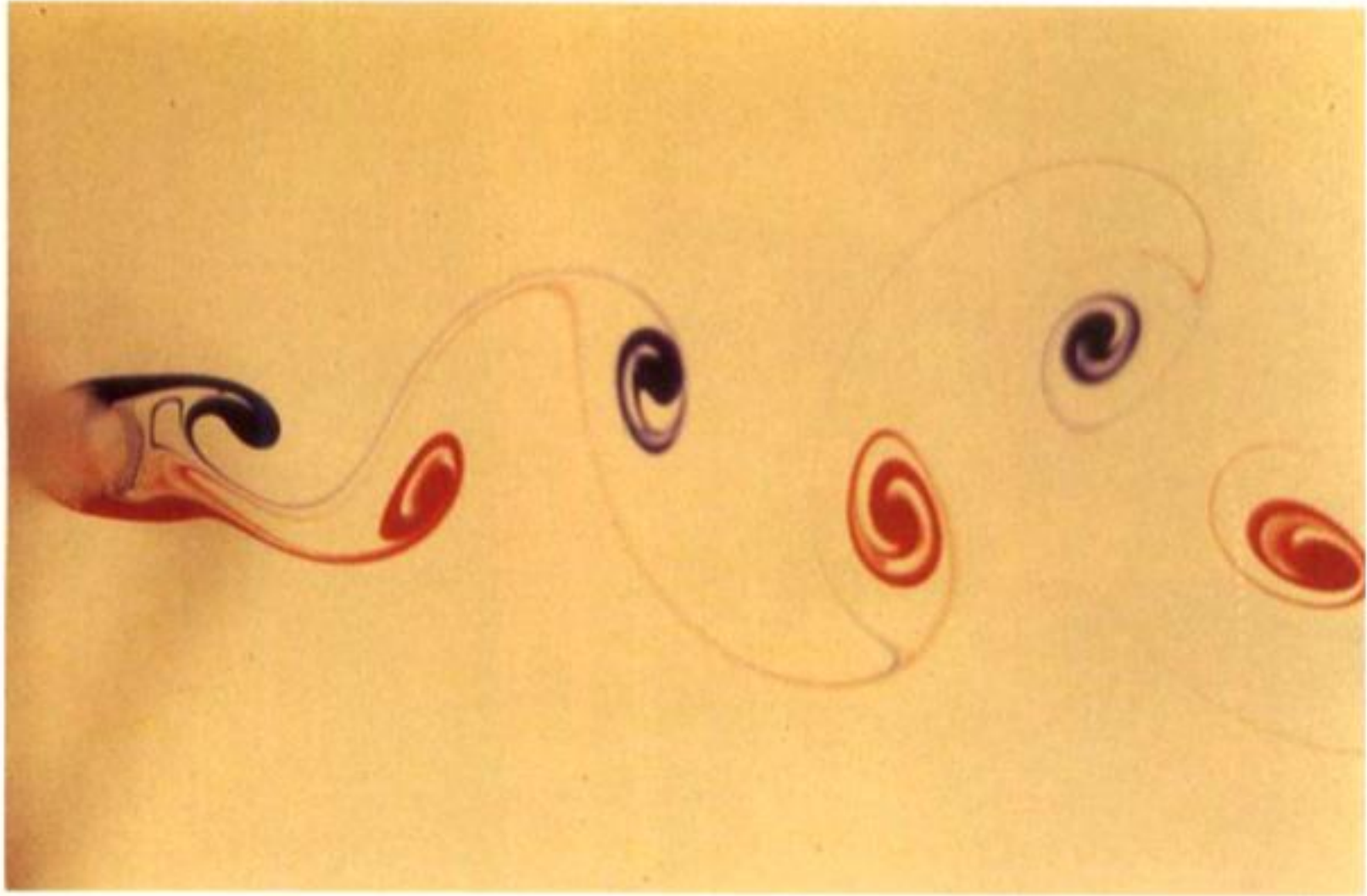
Plane channel flow



2D jet

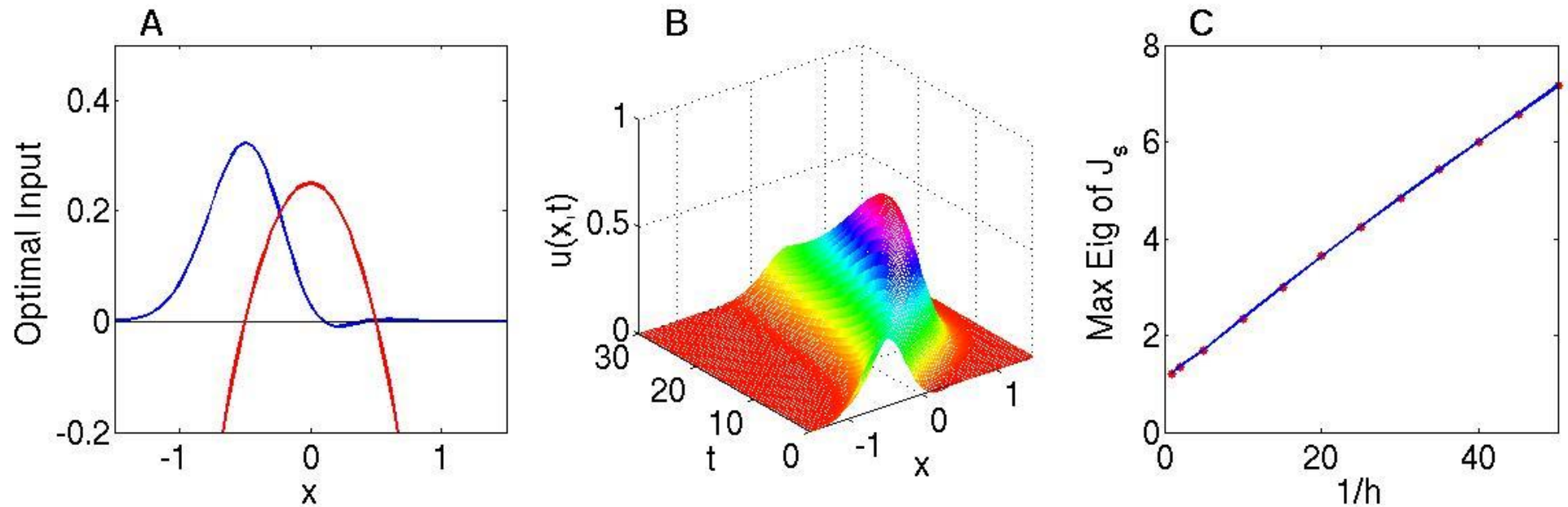


Cylinder Wake Beyond the Instability.



Perry, Chong and Lim 1982

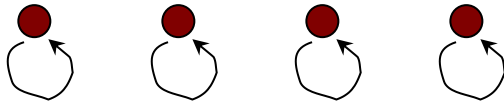
A Phenomenological Convective Instability.



$$\partial_t u = h^2 \partial_x^2 u - h \partial_x u + \left(\frac{1}{4} - x^2\right)u + v(x)s(t) + \eta.$$

Summary so far:

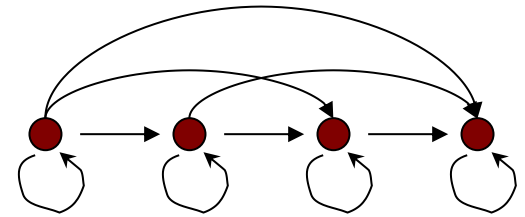
Normal



Homogenous Feedback Loops

No matter how
signal enters, cannot
amplify signal, without
amplifying noise.

Non-normal

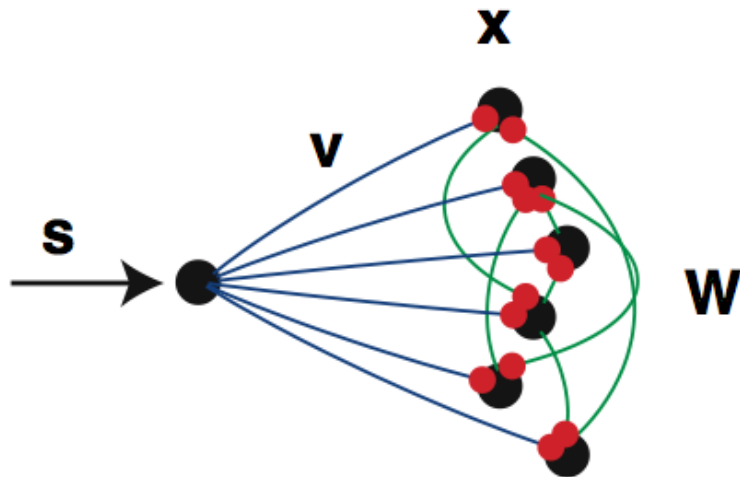


Hidden feedforward amplification cascades

Allows differential
amplification of signal
versus noise.

-
- Question: What if noise is negligible?
 - Jaeger 2001: Even with zero noise, one cannot accurately reconstruct gaussian inputs more than N time units into the past.
 - Can one do better if the input signal is temporally sparse? Idea: Use compressed sensing to recover high dim sparse signals from small numbers of measurements.

Memory as compressed sensing



Network dynamics of N neurons:

$$\mathbf{x}(i) = \mathbf{W} \mathbf{x}(i-1) + \mathbf{v} s^0(i)$$

$s^0(i-k)$ = scalar signal in the past

$\mathbf{x}(i)$ = current state of network

The network is continuously sensing a temporal stream of T inputs using N neurons via the $N \times T$ Measurement matrix: $\mathbf{A}_{nk} = (\mathbf{W}^k \mathbf{v})_n$.

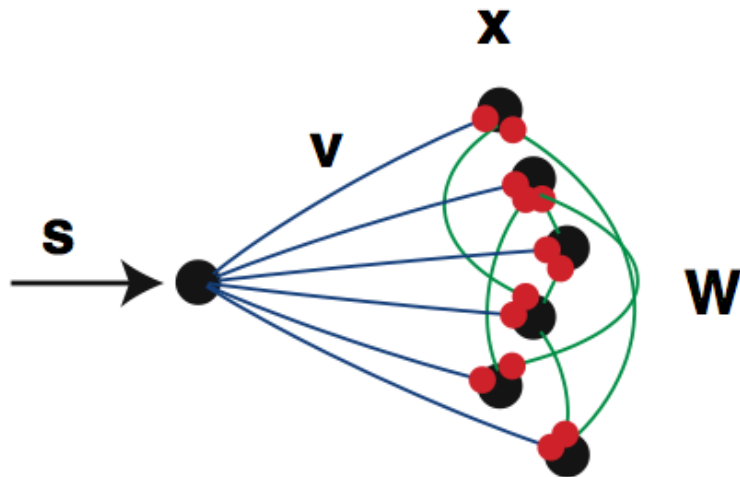
$$\mathbf{A}_{nk} = (\mathbf{W}^k \mathbf{v})_n : \begin{bmatrix} | & | & | & & | \\ \mathbf{v} & \mathbf{W}\mathbf{v} & \mathbf{W}^2\mathbf{v} & \dots & \mathbf{W}^k\mathbf{v} & \dots \\ | & | & | & & | \end{bmatrix}$$

New features:

- 1) A is N by Infinity: Deep within error regime
- 2) Columns of A decay over time (stability of dynamical system)
- 3) Columns of A are correlated (temporal correlations in response of network)

Question: What W (if any) yields good effective dynamical measurements A ?

Annealed approximation to a dynamical system



Network dynamics of N neurons:

$$\mathbf{x}(i) = \mathbf{W} \mathbf{x}(i-1) + \mathbf{v} s^0(i)$$

$s^0(i-k)$ = scalar signal in the past

$\mathbf{x}(i)$ = current state of network

The network is continuously sensing a temporal stream of T inputs using N neurons via the $N \times T$ Measurement matrix: $\mathbf{A}_{nk} = (\mathbf{W}^k \mathbf{v})_n$.

Annealed approximation (AA): \mathbf{A}_{nk} = zero mean gaussian with var ρ^{2k} where $\rho = \exp(-1/\tau N)$. Reflects decay in dynamical system, but not correlations.

$$\begin{bmatrix} | & | & | & & | \\ \mathbf{v} & \mathbf{v}1 & \mathbf{v}2 & \dots & \mathbf{v}3 & \dots \\ | & | & | & & | \end{bmatrix}$$

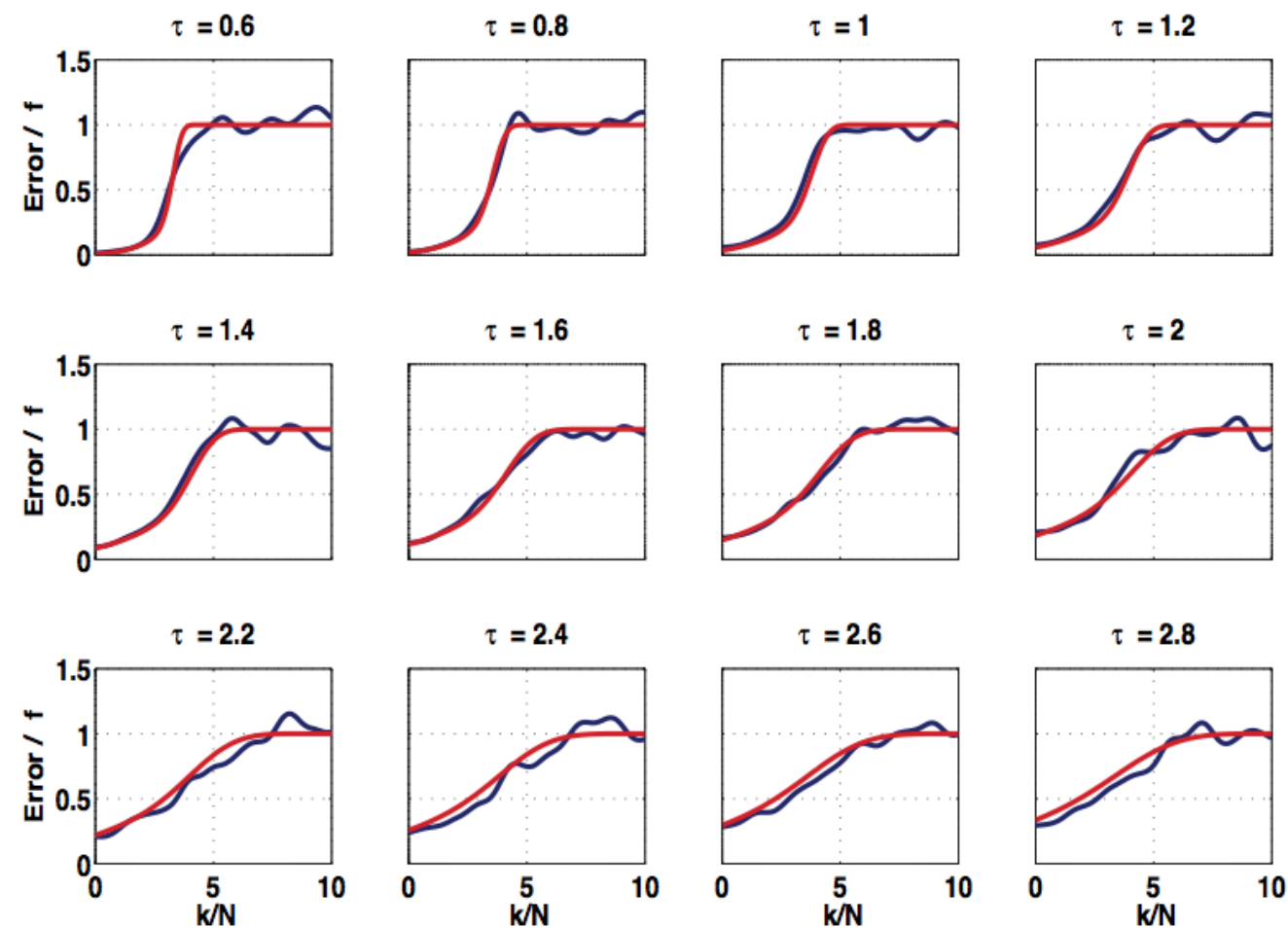
New features:

- 1) \mathbf{A} is N by Infinity: Deep within error regime
- 2) Columns of \mathbf{A} decay over time (stability of dynamical system)
- ~~3) Columns of \mathbf{A} are correlated (temporal correlations in response of network)~~

Memory performance in the annealed approximation

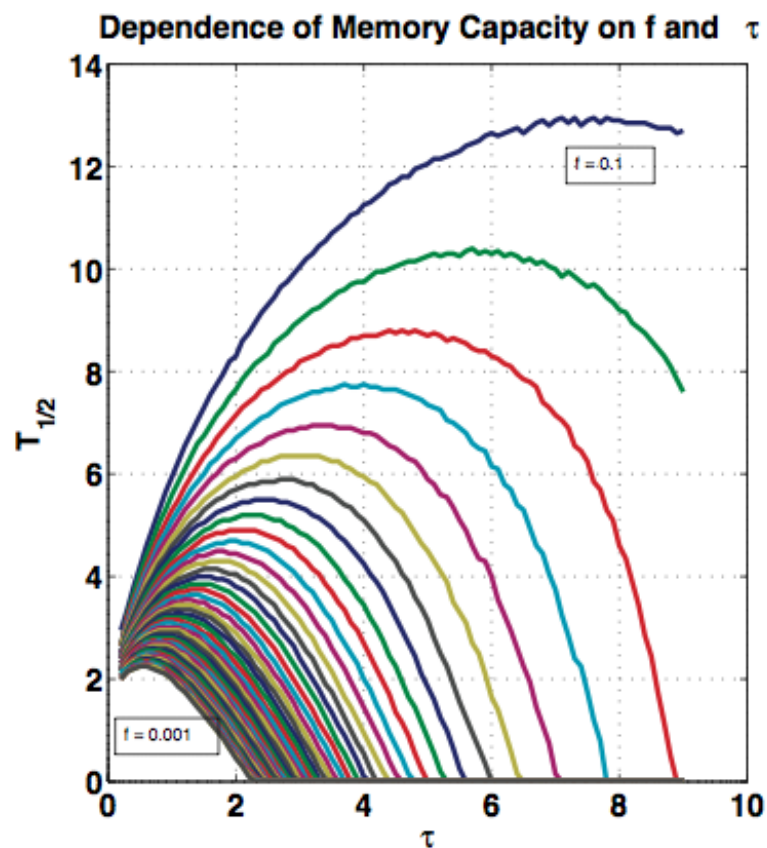
$$E(k/N) = \langle (s_{\text{est}}(n-k) - s^0(n-k))^2 \rangle_{A, s^0}$$

Memory curve =
reconstruction error as a
function of time into the past



Memory curves for $f = 0.04$
Red: theory
Blue: simulations

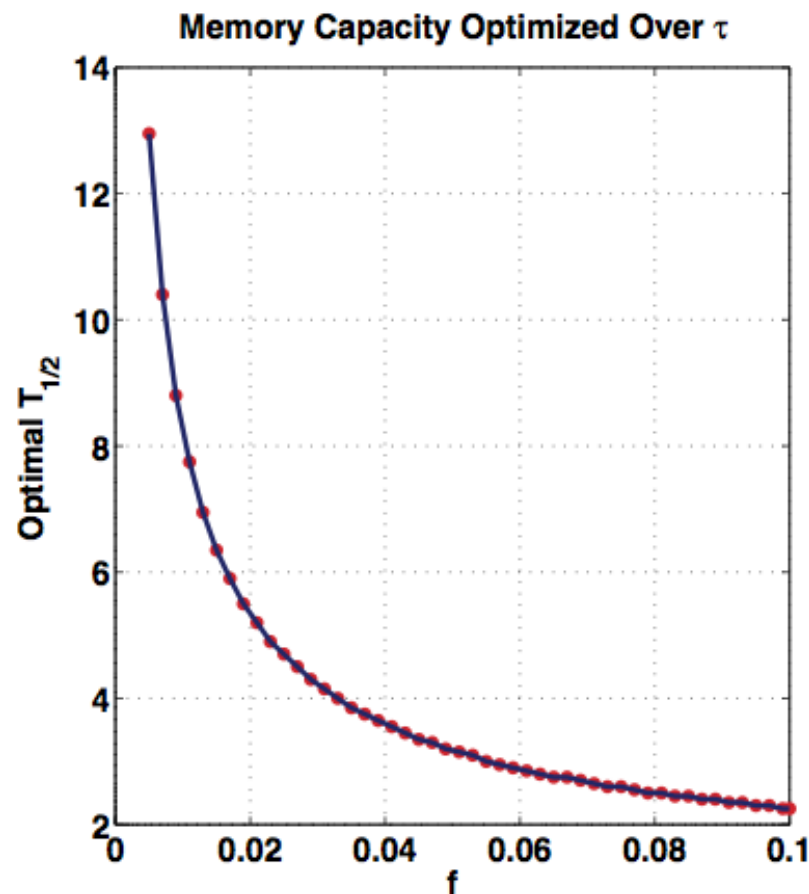
Memory performance in the annealed approximation



Tradeoff in memory capacity:

Small τ : forget quickly

Long τ : stimulus interference



Memory capacity
can exceed number of neurons:

$$\sim N / (f \log 1/f)$$

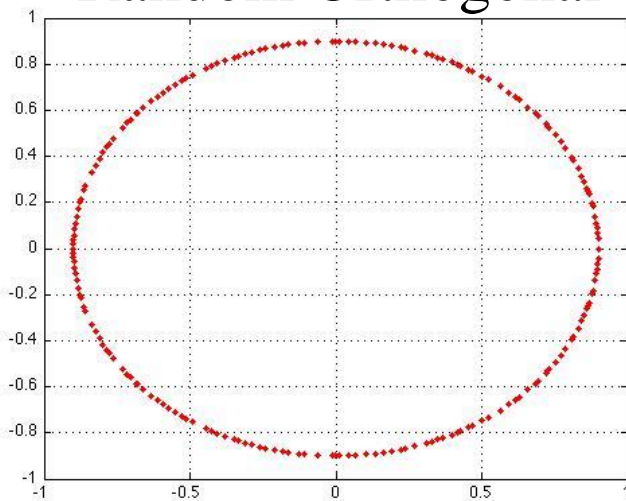
Implementing the annealed approximation

$$A_{nk} = (W^k \mathbf{v})_n :$$

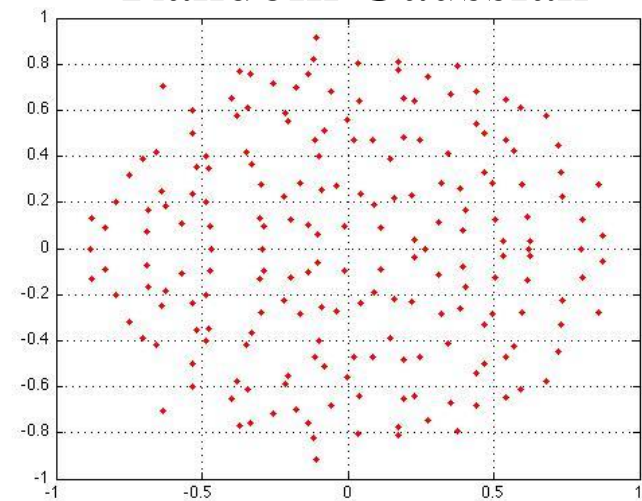
$$\begin{bmatrix} | & | & | & | & | \\ \mathbf{v} & W\mathbf{v} & W^2\mathbf{v} & \dots & W^k\mathbf{v} & \dots \\ | & | & | & | & | \end{bmatrix}$$

- $A_{nk} \sim$ Activity pattern across neurons k time steps after an input stimulus
- Want A_{nk} and A_{nl} to be as random and uncorrelated as possible.
- This can be achieved if the network connectivity is orthogonal: $W = \rho O$
- But not if W is a random gaussian matrix, or all to all connected, etc...

Random Orthogonal

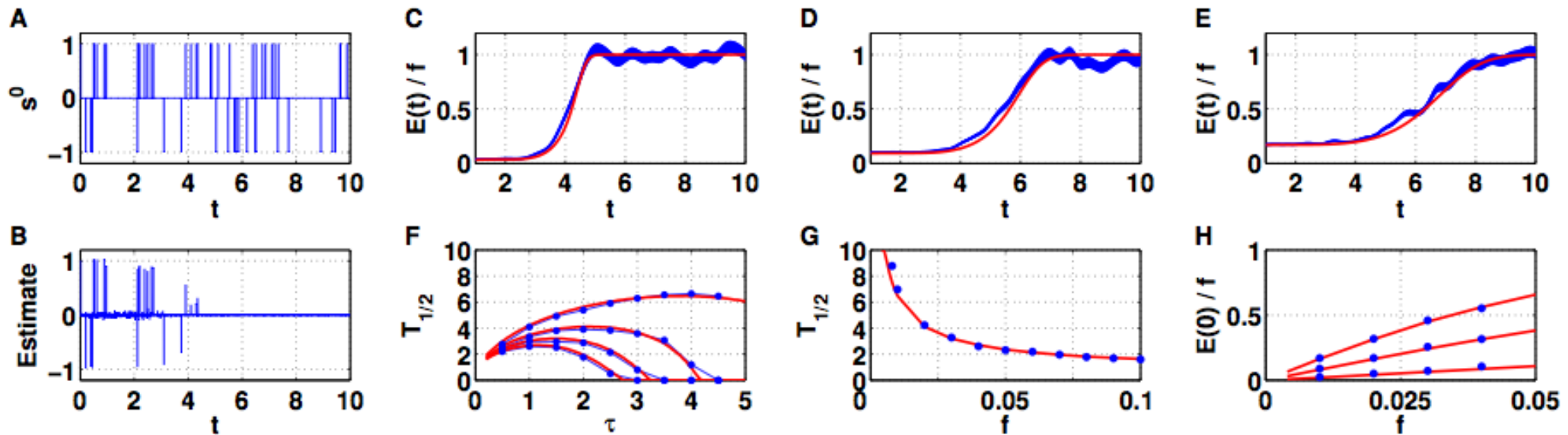


Random Gaussian

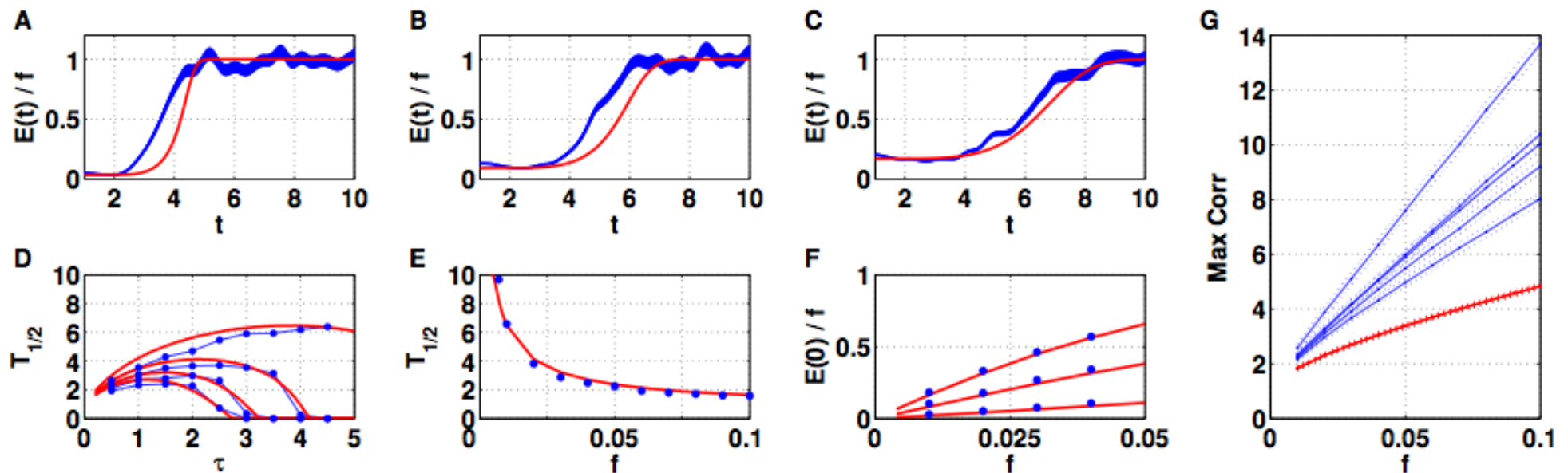


Comparison between Annealed and Orthogonal Cases

Annealed Theory versus Annealed simulations



Annealed Theory versus Orthogonal simulations



Summary

- Developed a new statistical mechanics approach to compressed sensing (CS)

Reproduces phase diagram of correct/incorrect reconstruction

Derives a new phase of CS in which optimization is not required

Yields insight into the magnitude and nature of errors made by CS

Can be extended to derive null models for sparse regression (i.e. LASSO)

- Developed a new view of memory capacity (MC) in neural networks for nongaussian, sparse sequences.

MC can exceed the number of neurons (unlike gaussian case)

Enhanced MC is attained by recurrent, orthogonal networks, not feedforward networks or random gaussian networks.

MC approaches the information theoretic limit.