Fisher information, compressed sensing, and the origins of sequence memory in neuronal networks.

Surya Ganguli

Sloan-Swartz Center for
Theoretical Neurobiology

UCSF

Joint work with:
- Ben Huh (UCSD)
- Haim Sompolinsky (Harvard/Hebrew Univ)

Funding:
- Swartz Foundation
- Burroughs Wellcome

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.

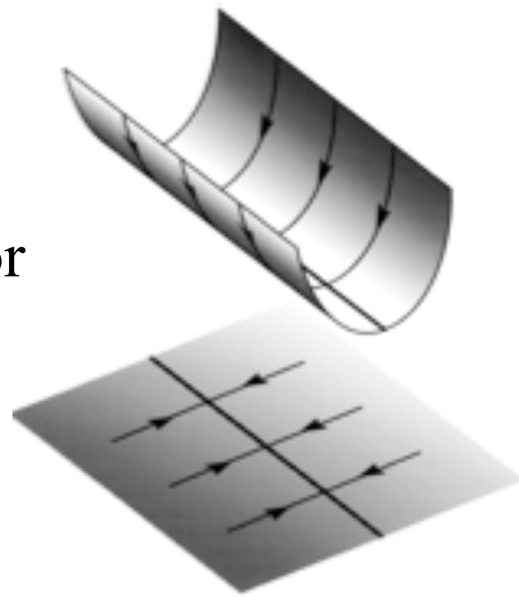
So to mediate short-term memory, networks of neurons must interact with each other to keep our memories alive.

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

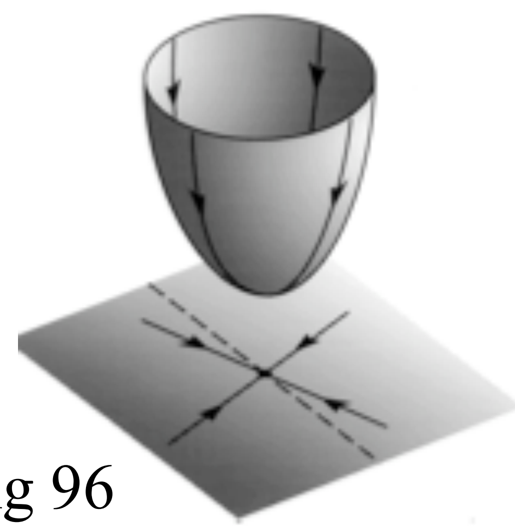
And how can networks store multiple items in a temporal sequence?

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

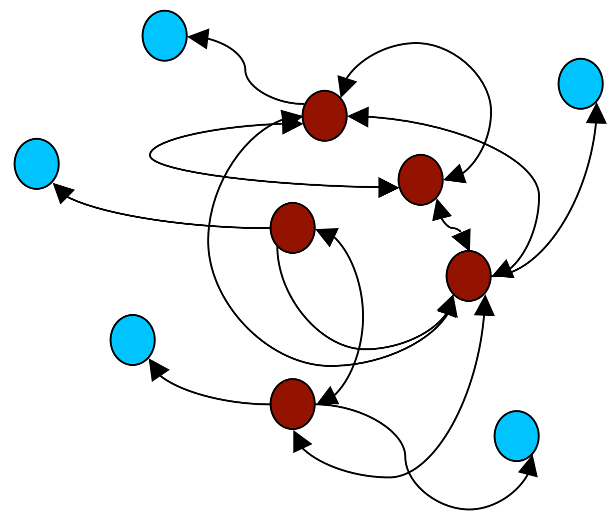
Line Attractor



Fixed Point Attractor

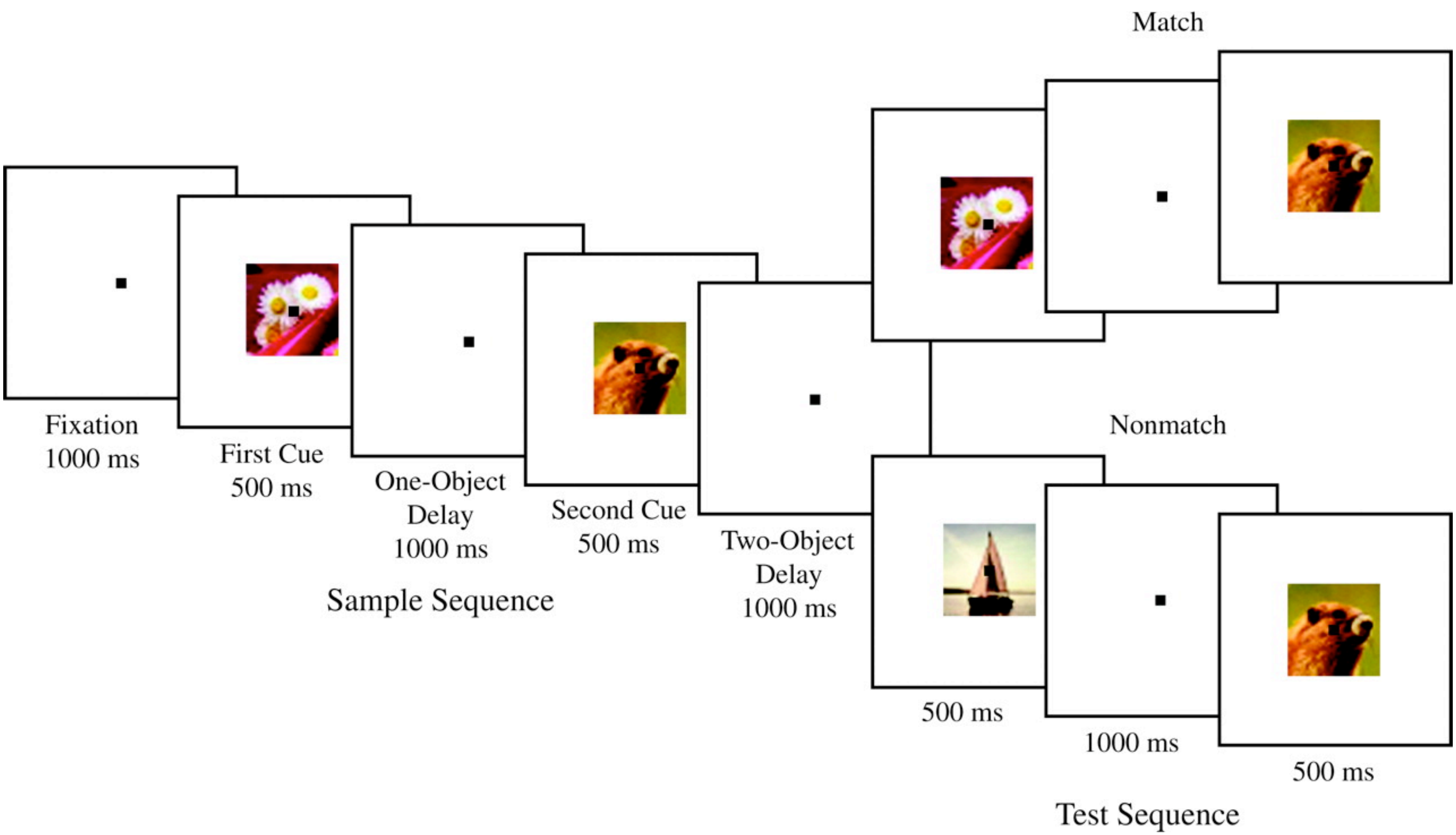


Seung 96



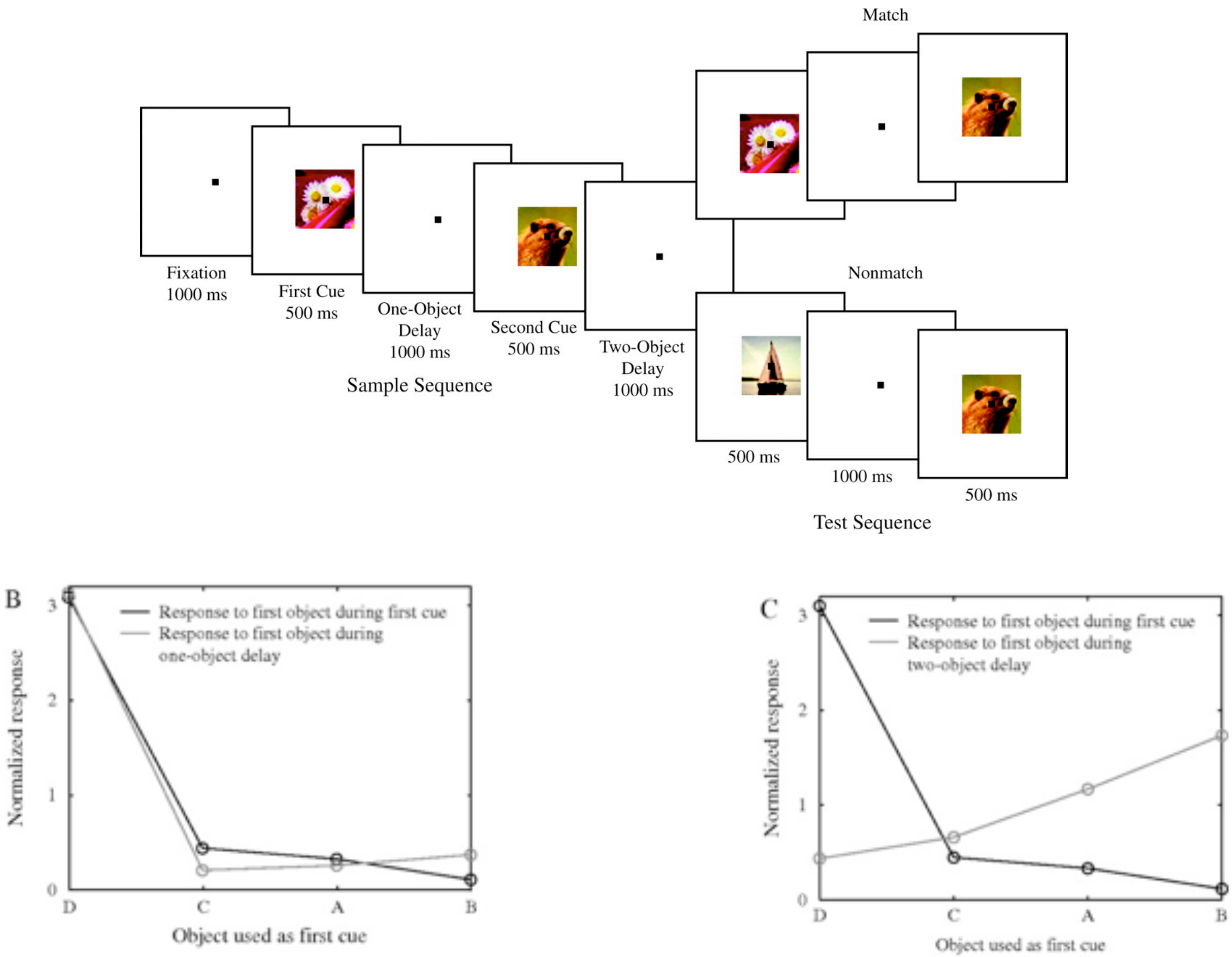
Positive Feedback

Probing sequence memory in the macaque brain.



Warden, M. R. et al. Cereb. Cortex 2007

Probing sequence memory in the macaque brain.



Warden, M. R. et al. Cereb. Cortex 2007

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 ”.

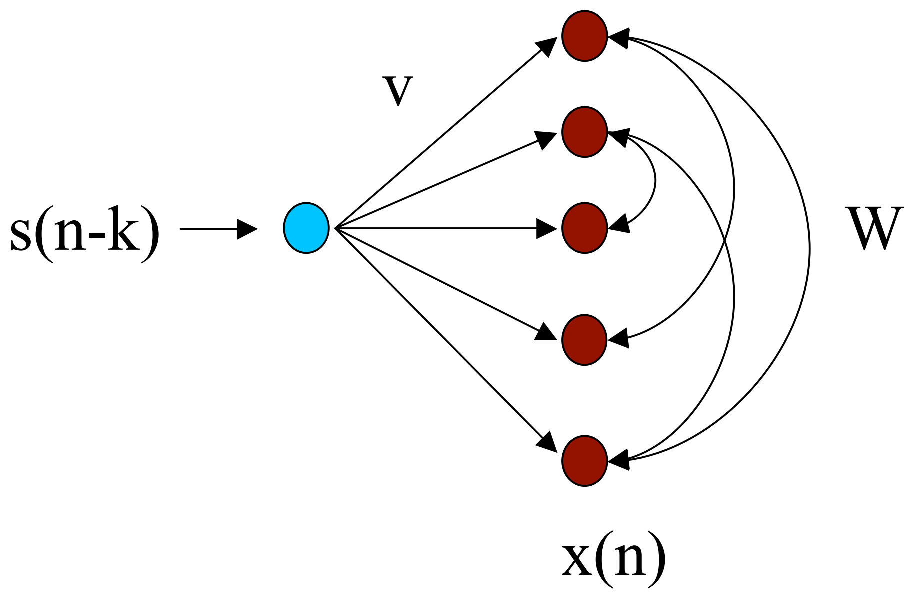
- *Markram, Natschlager, and Maas, 2001*
  also: Buonomano and Merzenich, 1995
  Mayor and Gerstner, 2003

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.

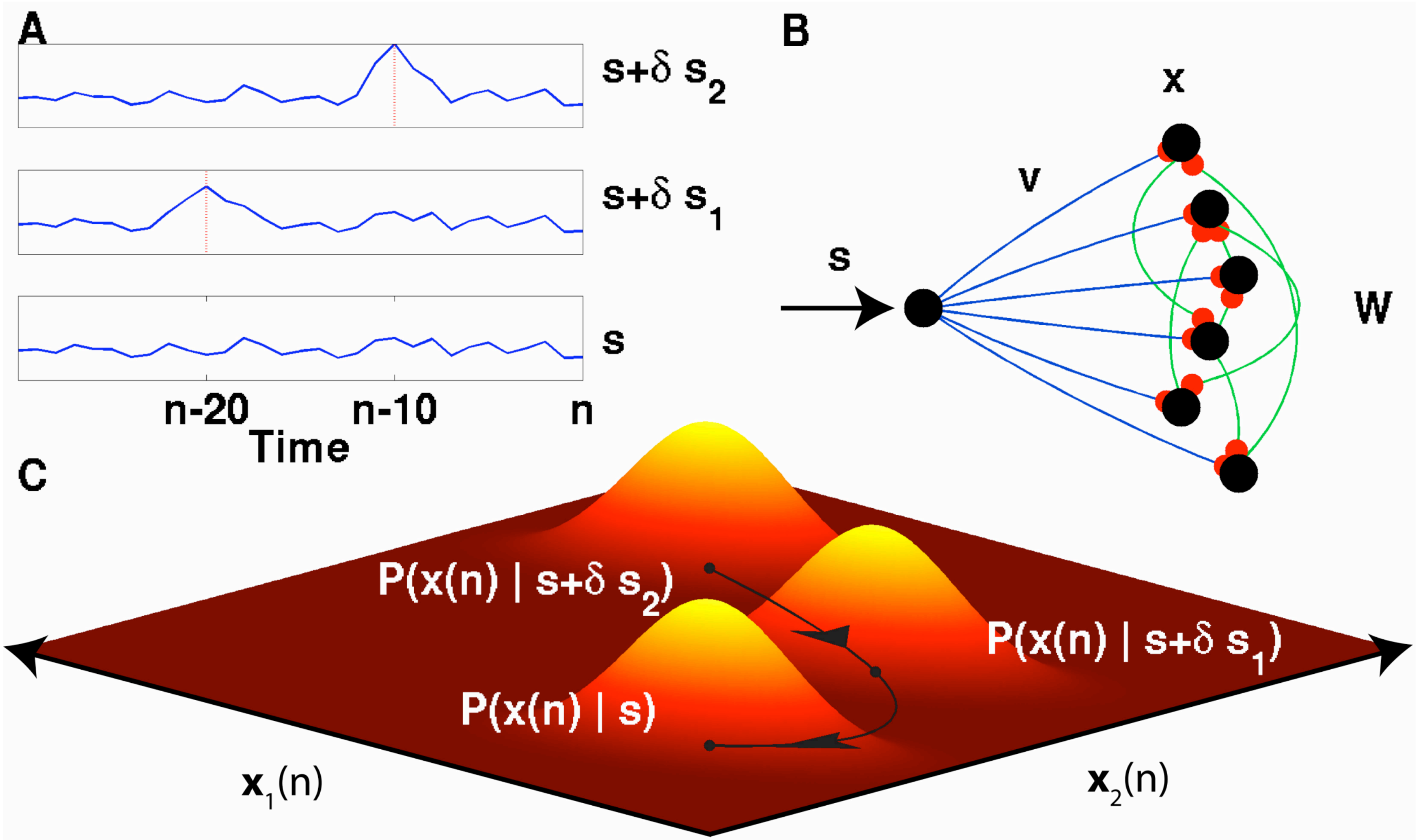
Compute this capacity analytically.

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(x | s) change when you change s. This is captured by the Fisher Information Matrix:

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

Consider a change in the signal: s -> s + ds.

Then the distribution of x(n) will change by an amount ~ $ds^T J ds$.

The Matrix Nature of Memory

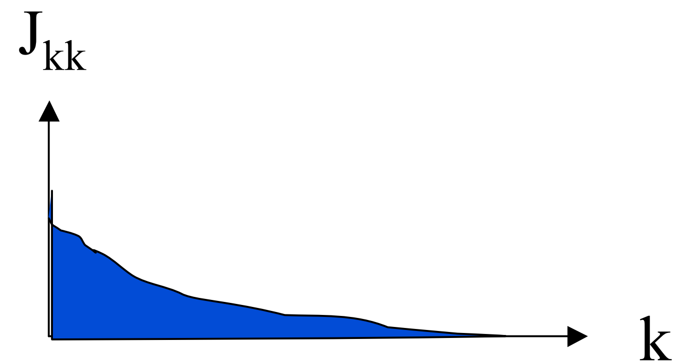
J_{00} J_{01} J_{02}

J_{10} J_{11}

J_{20}

J_{lk}

J_{kl} J_{kk}

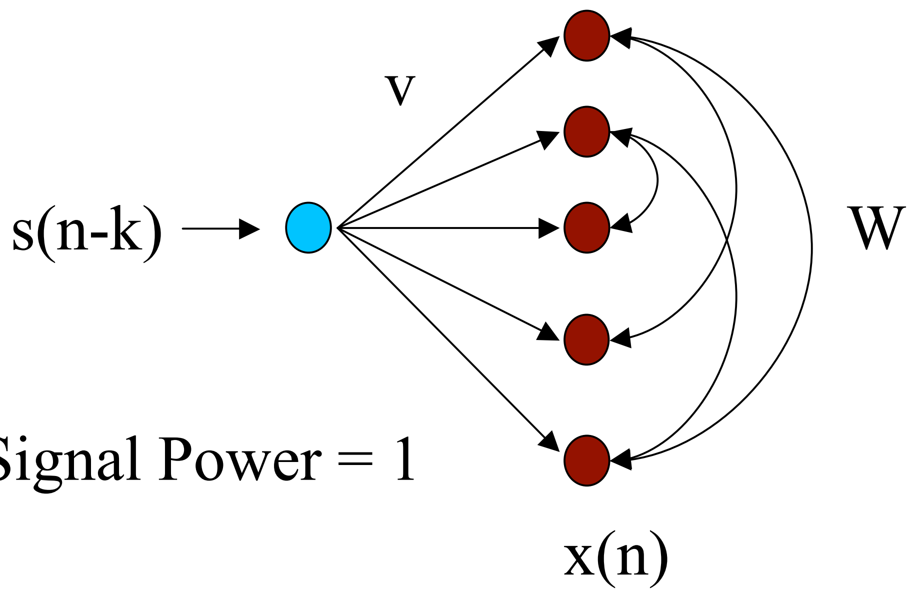


$$J_{\text{Tot}} = \sum_{k=0}^{\infty} J_{kk}$$

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

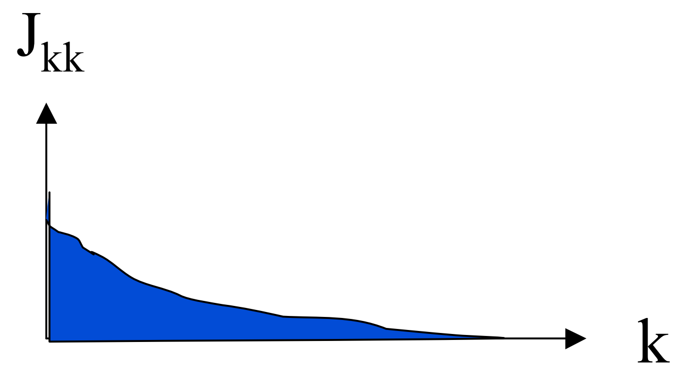
J_{kl} = Amount of interference between the memory traces of two pulses entering at different times k and l in the past.

A Fundamental Performance Limit on Memory



Signal Power = 1

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



Noise Power = ε

Instantaneous SNR at input: $1/\varepsilon$

$$\mathbf{J}_{\mathrm{Tot}} \equiv \sum_{\mathbf{k}=0}^{\infty} \mathbf{J}_{\mathbf{kk}} \leq \frac{\mathbf{N}}{\epsilon}$$

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:

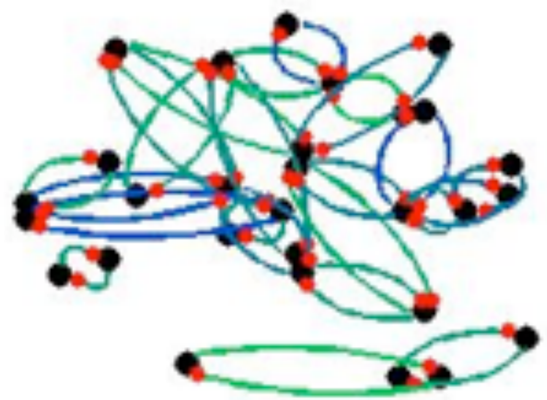
$$\mathbf{J}_{\mathrm{Tot}} \equiv \sum_{\mathbf{k}=0}^{\infty} \mathbf{J}_{\mathbf{kk}} = \frac{\mathbf{1}}{\epsilon}$$

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 it redistribute this SNR across time.

Examples of "Normal" Network Connectivities

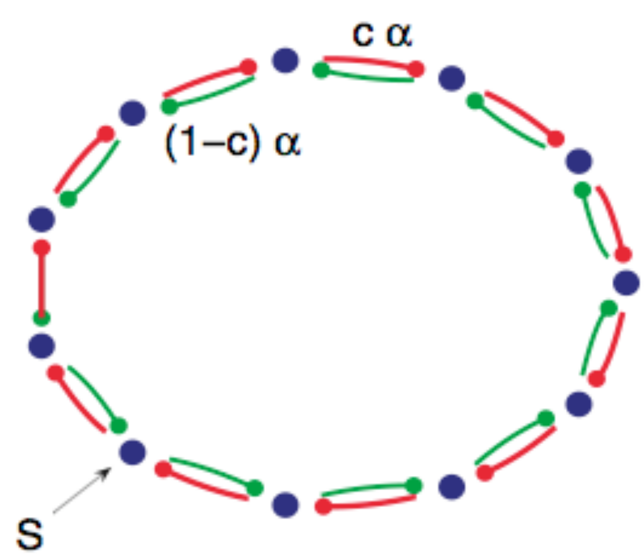
Any symmetric network.

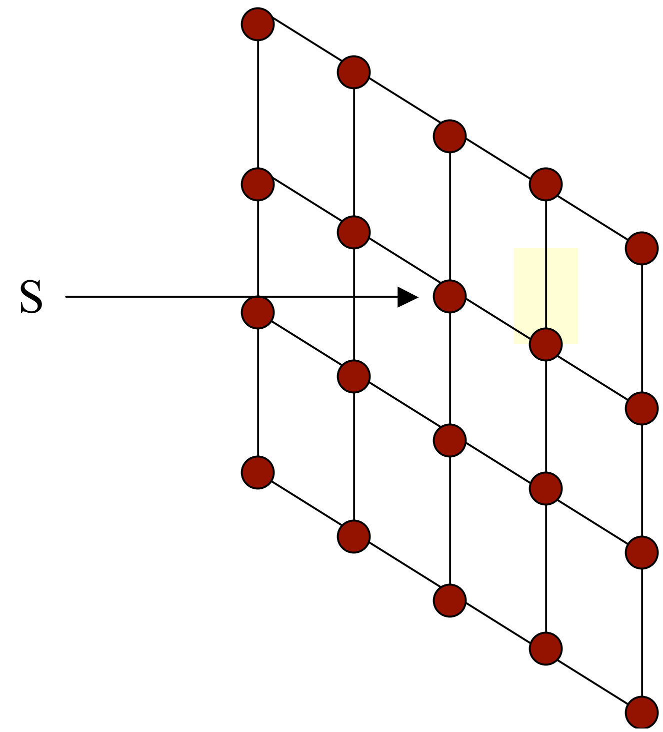


Any antisymmetric network.

Any orthogonal network.

Translation invariant lattices.



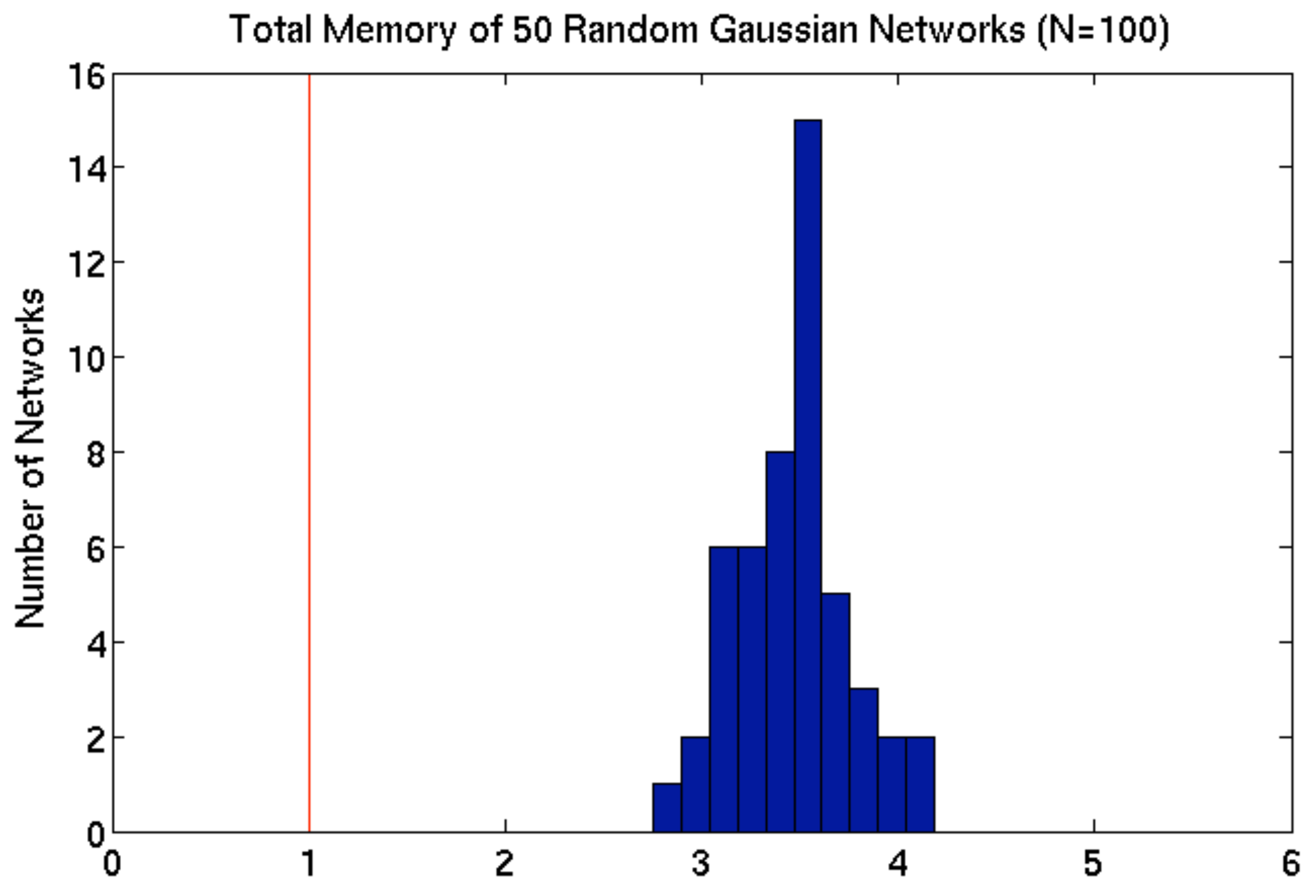


Memory Beyond the Normal Limit: Perturb Normality

Random Symmetric W ⟶ Random Asymmetric W

Now J_{Tot} depends on W and v.

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



J_{Tot} optimized over v

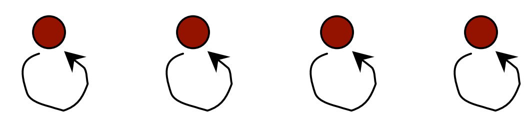
The nature of normal dynamics: independent decaying modes

Eigenvector = Preferred Pattern or Mode of Activity across Neurons
Eigenvalue = Decay time constant of that pattern (larger value -> slower deca

$$R(0) = c_1(0)\,V_1 + c_2(0)\,V_2 + \ldots\; c_N(0)\,V_N$$

$$R(k) = c_1(k)\,V_1 + c_2(k)\,V_2 + \ldots\; c_N(k)\,V_N$$

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



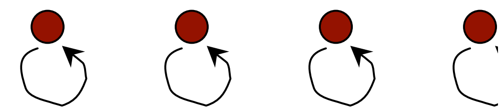
Total network activity $R^2 = c_1{}^2 + c_2{}^2 + \ldots + c_N{}^2$

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

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

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

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



$$w/N \begin{pmatrix} 1\ 1 \ldots 1\ 1 \\ 1\ 1 \ldots 1\ 1 \\ . \qquad . \\ . \qquad . \\ 1\ 1 \ldots 1\ 1 \end{pmatrix}$$

Slow mode:
(large eigenvalue)

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

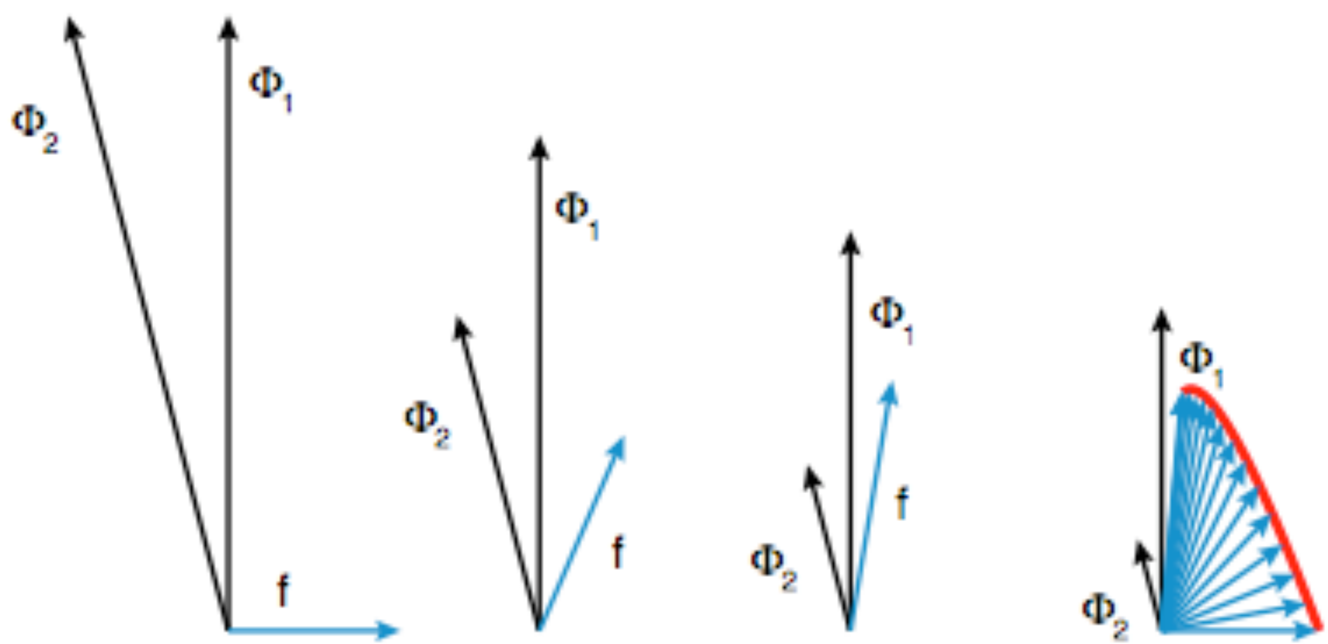
All other modes fast:
(small eigenvalues)

The nature of nonnormal dynamics: transient amplification from nonorthogonal eigenvectors

$$R(0) = c_1(0)\,V_1 + c_2(0)\,V_2 + \ldots\; c_N(0)\,V_N$$

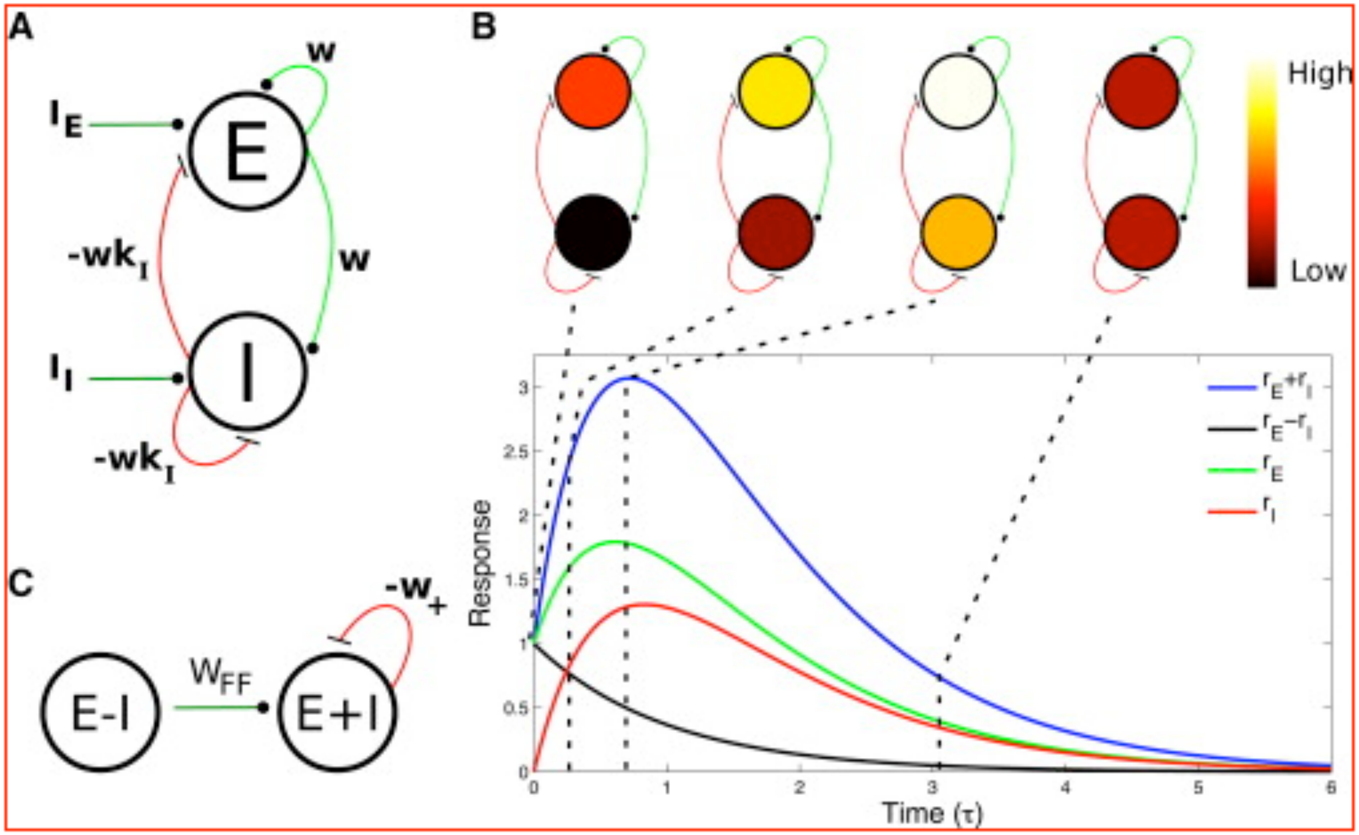
$$R(k) = c_1(k)\,V_1 + c_2(k)\,V_2 + \ldots\; c_N(k)\,V_N$$

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



Schmid Ann Rev of Fluid Mech, 20

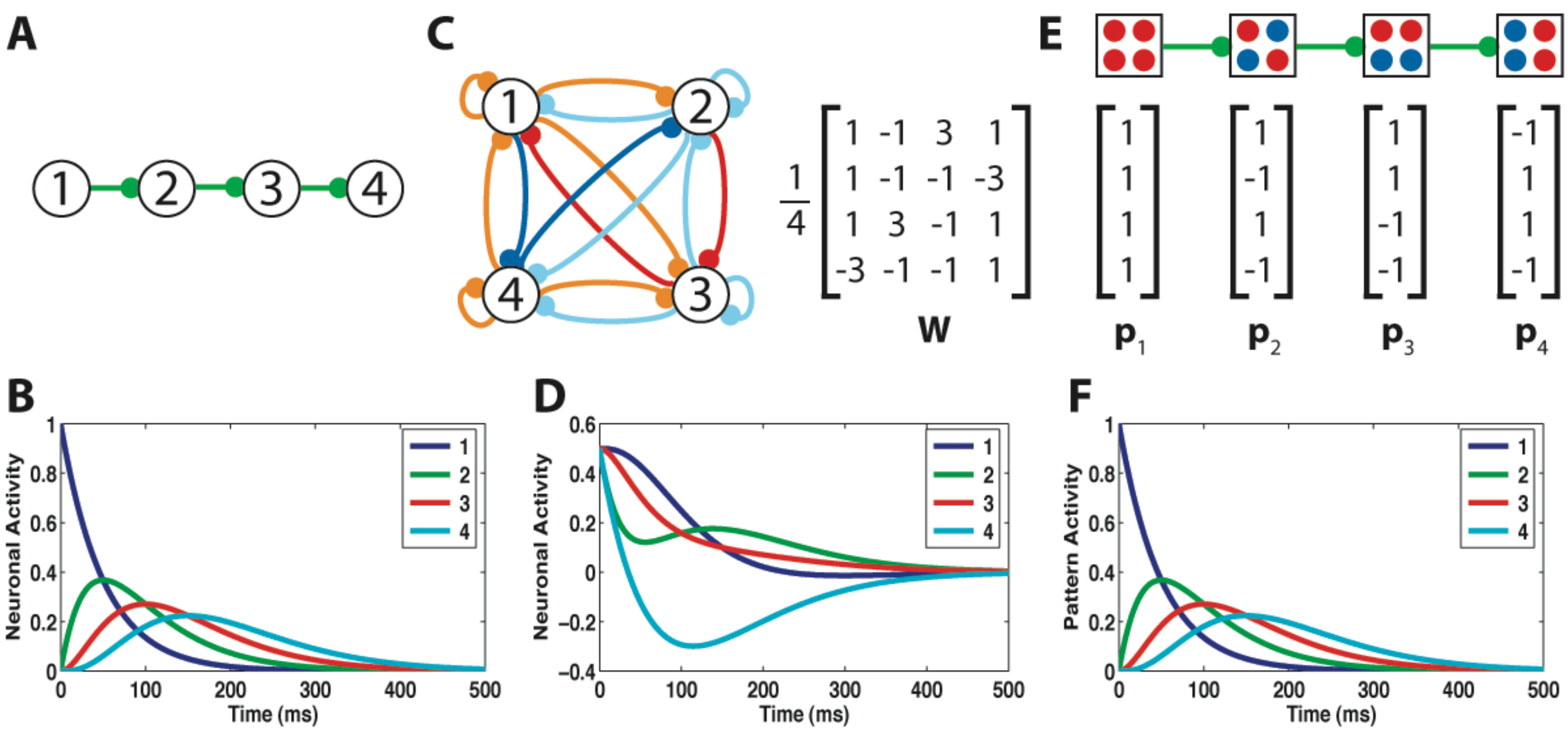
An simple two neuron example of transient amplification



Murphy Miller Neuron 2009

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

Normal Non-normal



Ganguli and Latham, Neuron 20

Related work

Normal Non-normal

Related “Nonnormal” Work:

Trefethen and Embree (2005)

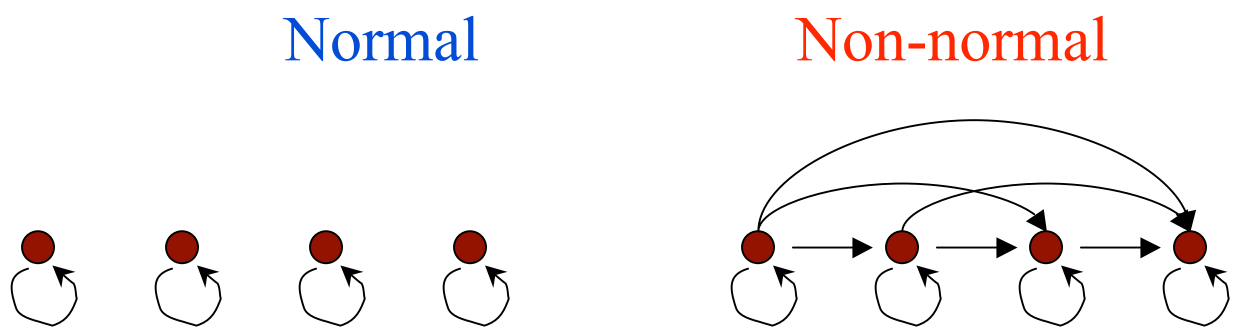
Ganguli, Huh, Sompolinsky PNAS (2008)

Murphy and Miller Neuron (2009).

Goldman, Neuron (2009).

Ganguli and Latham, Neuron (2009).

Related work



Related “Nonnormal” Work:

Trefethen and Embree (2005)

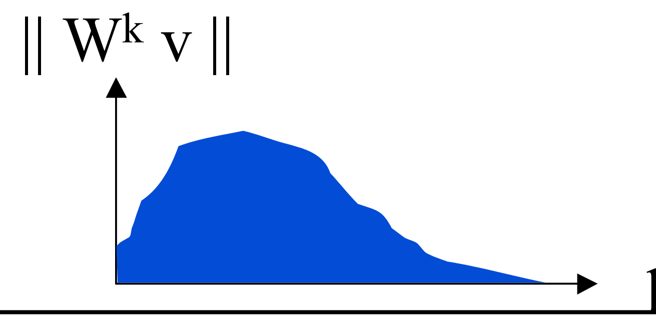
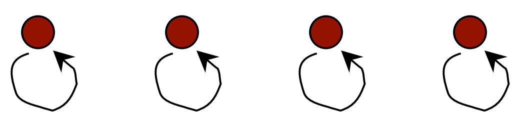
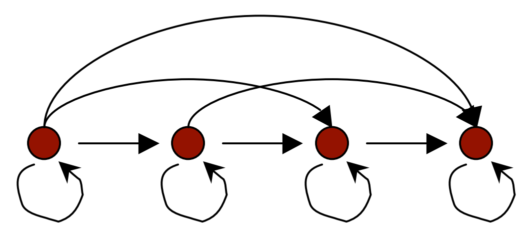
Ganguli, Huh, Sompolinsky PNAS (2008)

Murphy and Miller Neuron (2009).

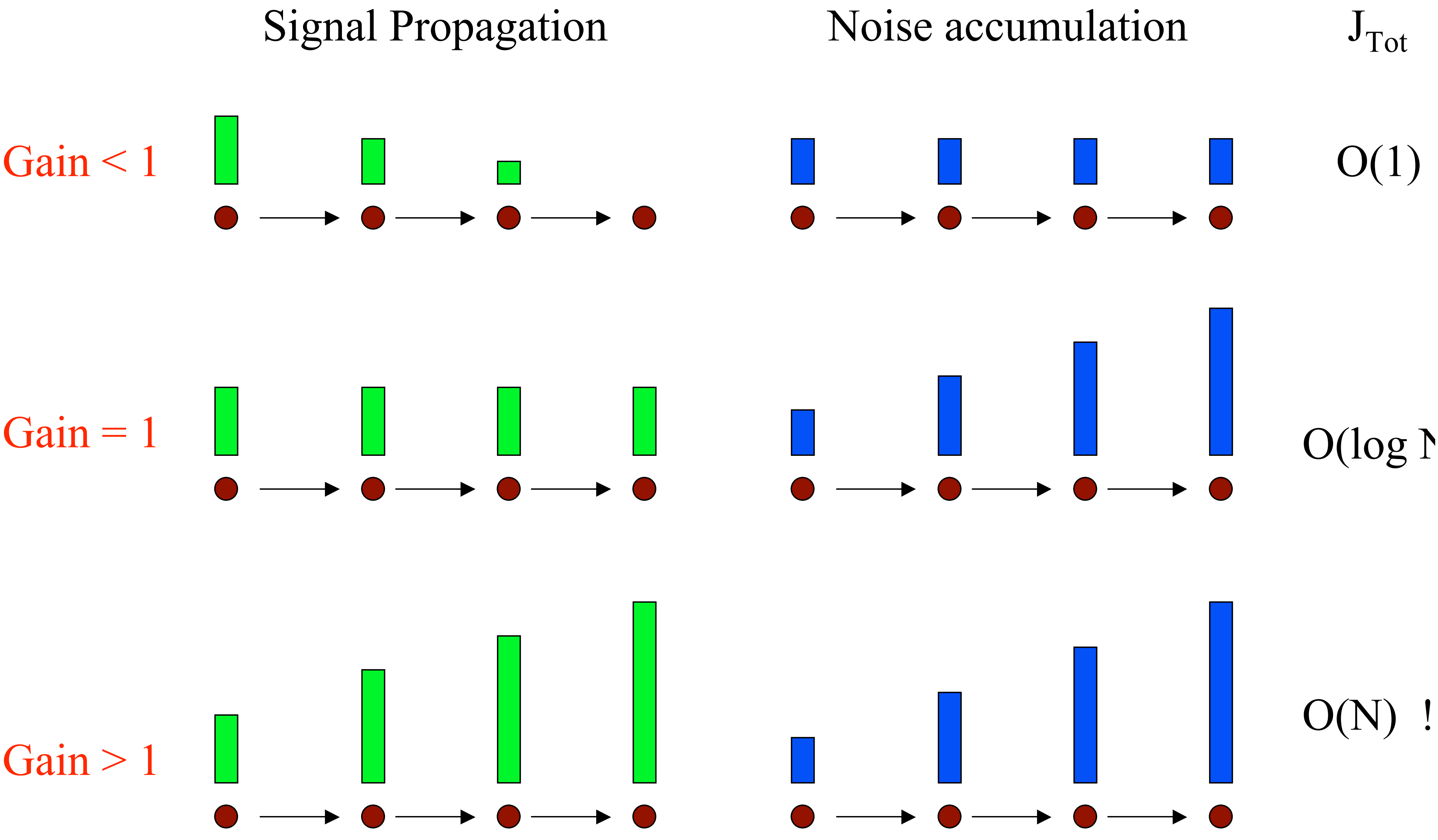
Goldman, Neuron (2009).

Ganguli and Latham, Neuron (2009).

The story so far

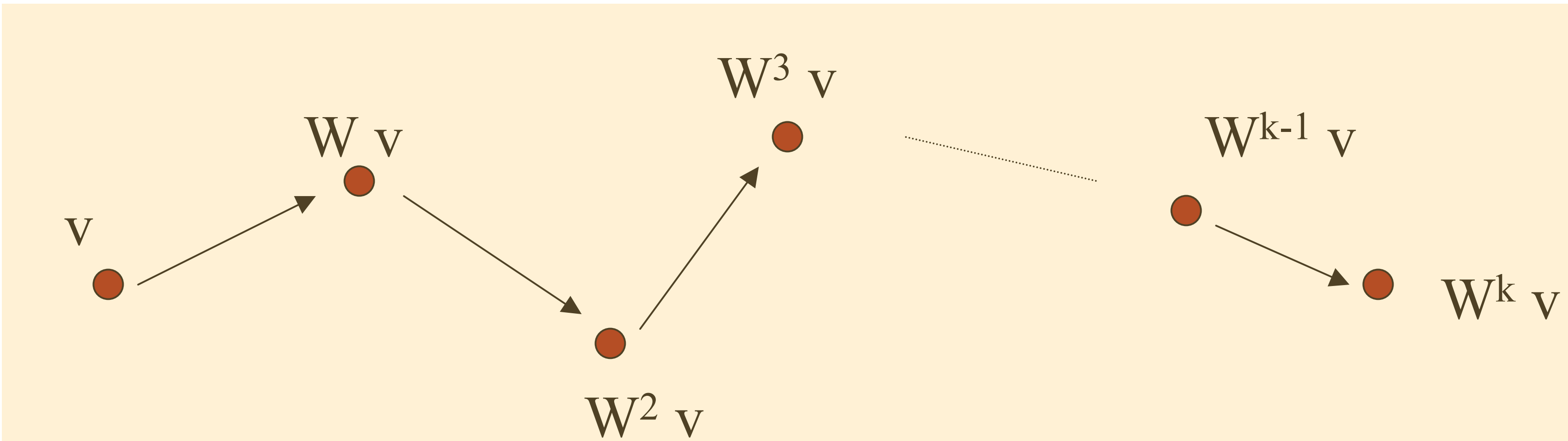
	Normal Networks	Nonnormal Networks
Information Theory	$\mathbf{J}_{\mathrm{Tot}} \equiv \sum_{\mathbf{k}=0}^{\infty} \mathbf{J}_{\mathbf{kk}} = \frac{1}{\epsilon}$	$\mathbf{J}_{\mathrm{Tot}} \equiv \sum_{\mathbf{k}=0}^{\infty} \mathbf{J}_{\mathbf{kk}} \leq \frac{\mathbf{N}}{\epsilon}$
Dynamics		
Hidden Structure		

Memory in the simplest feedforward network



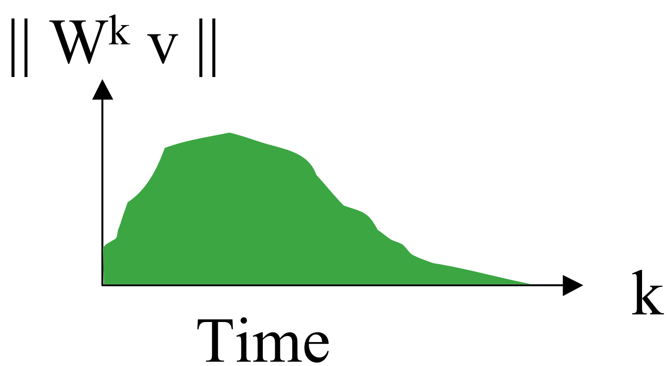
An upper bound on memory in any network

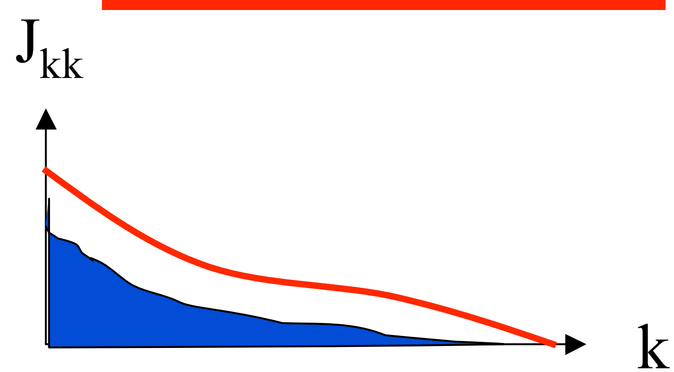
Dynamical propagation of a signal through network space.



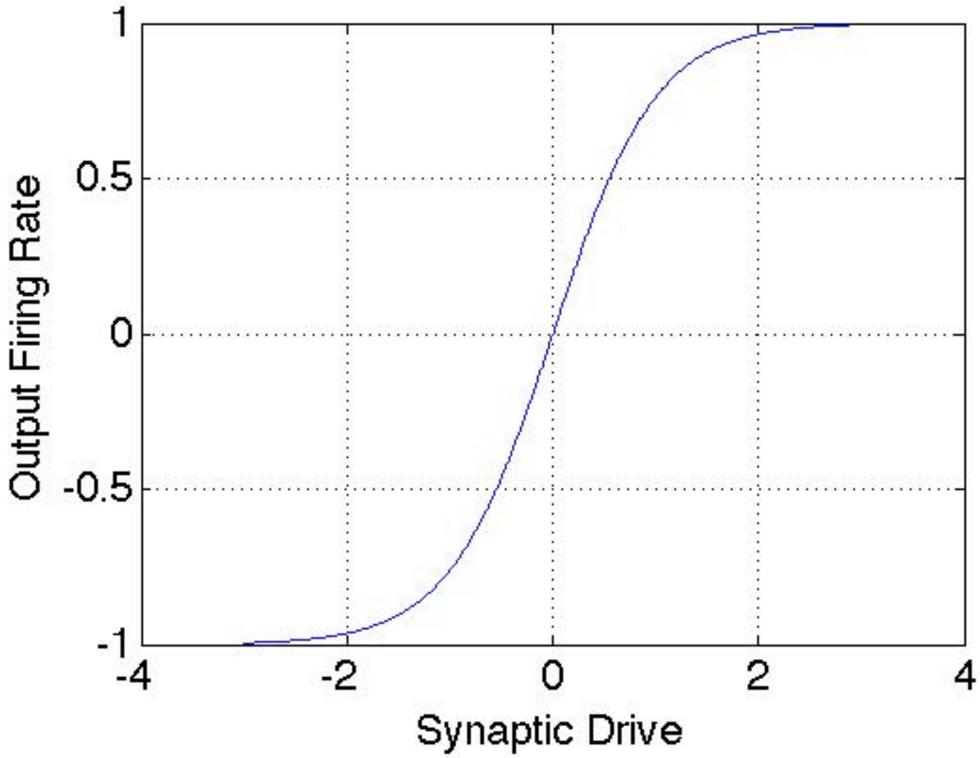
$$J_{kk} <= J_{kk}^{Delay}$$

Signal amplification profile





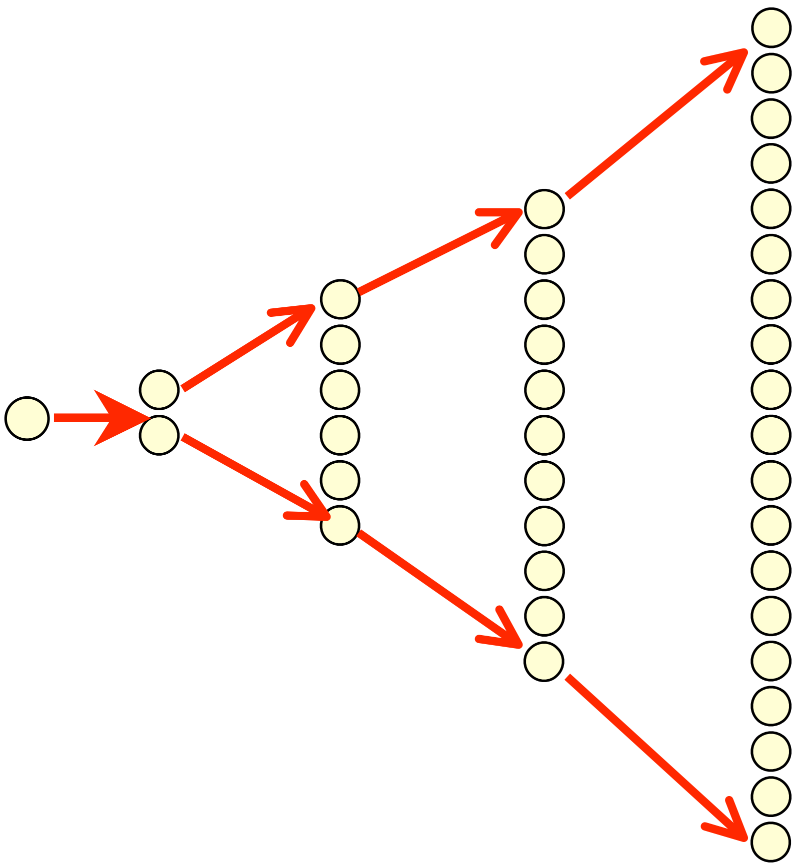
What about saturating nonlinearities?



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

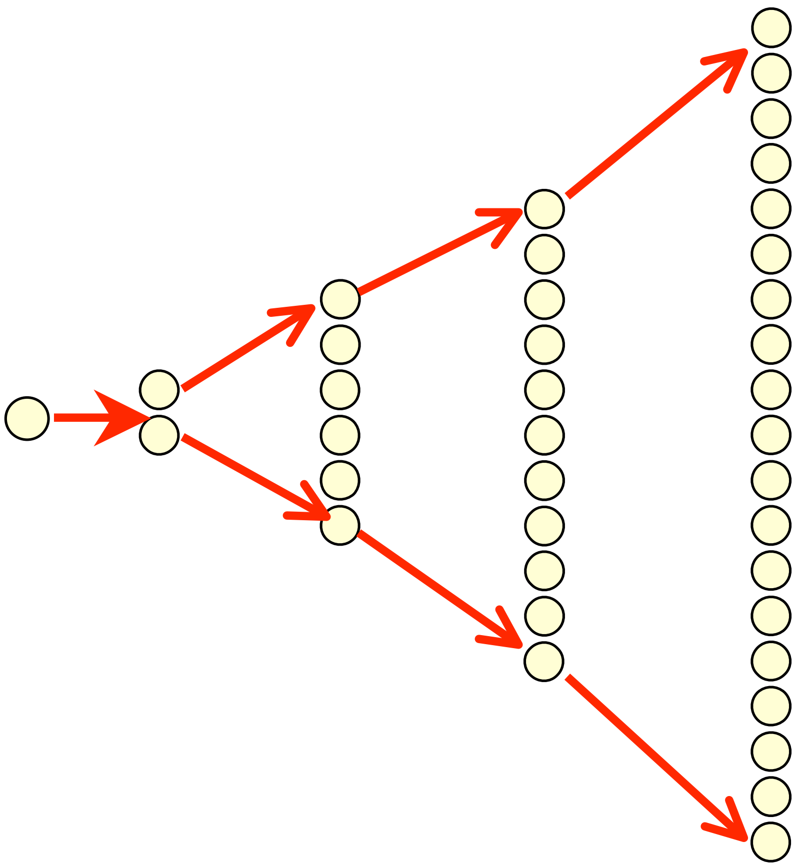
$$N_k \sim k$$

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:

$$\sim 1/k$$

Jtot = L ~ square root of N

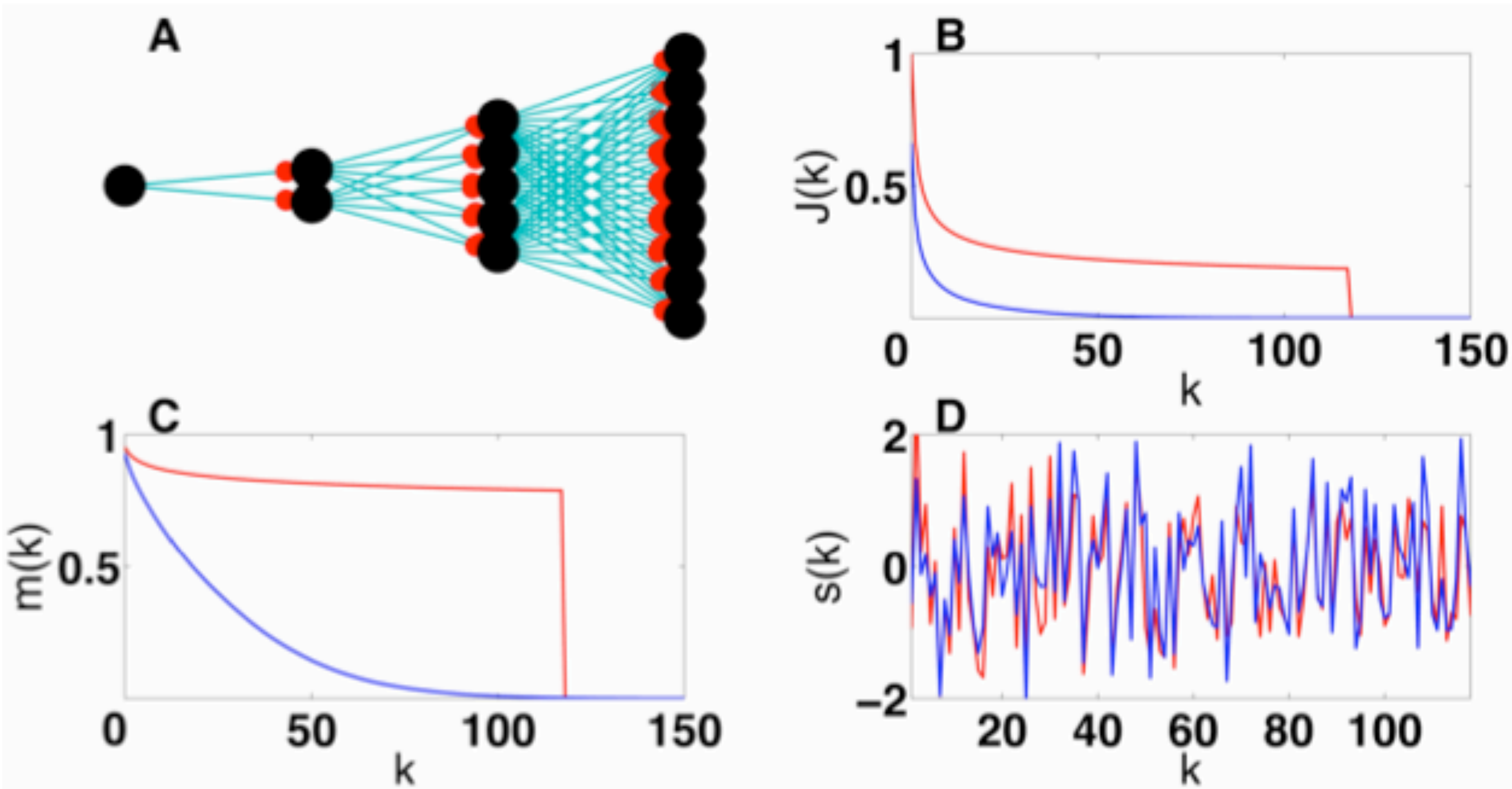
Consequences of finite dynamic range

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

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 => 1.35 seconds of full sequence memory

The Liquid State Machine??

Too Normal! :(

The Liquid State Machine??

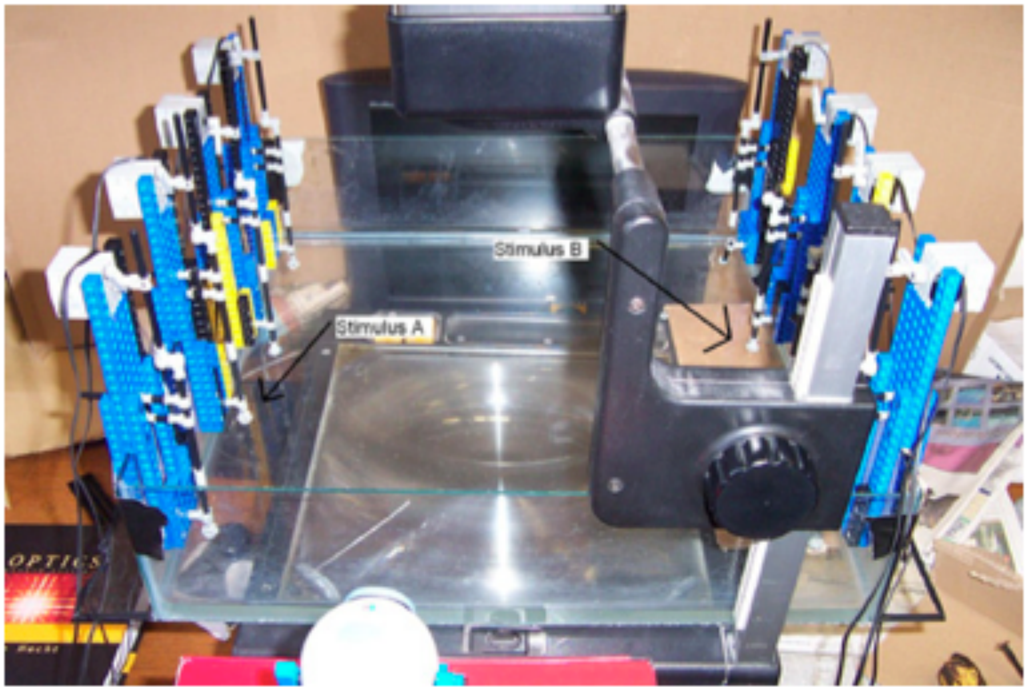
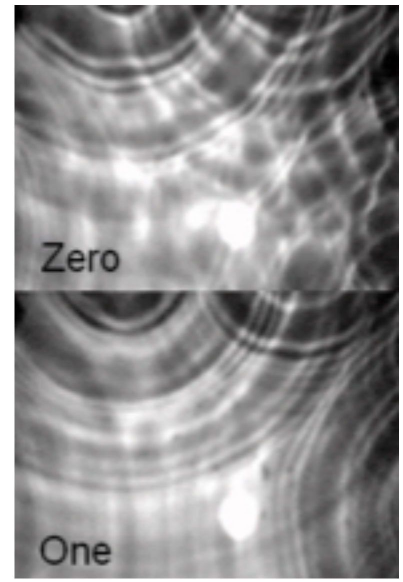


Fig. 1. The Liquid Brain.



Chrisantha Fernando and Sampsa Sojakka, ECAL 2003

Better liquid states.

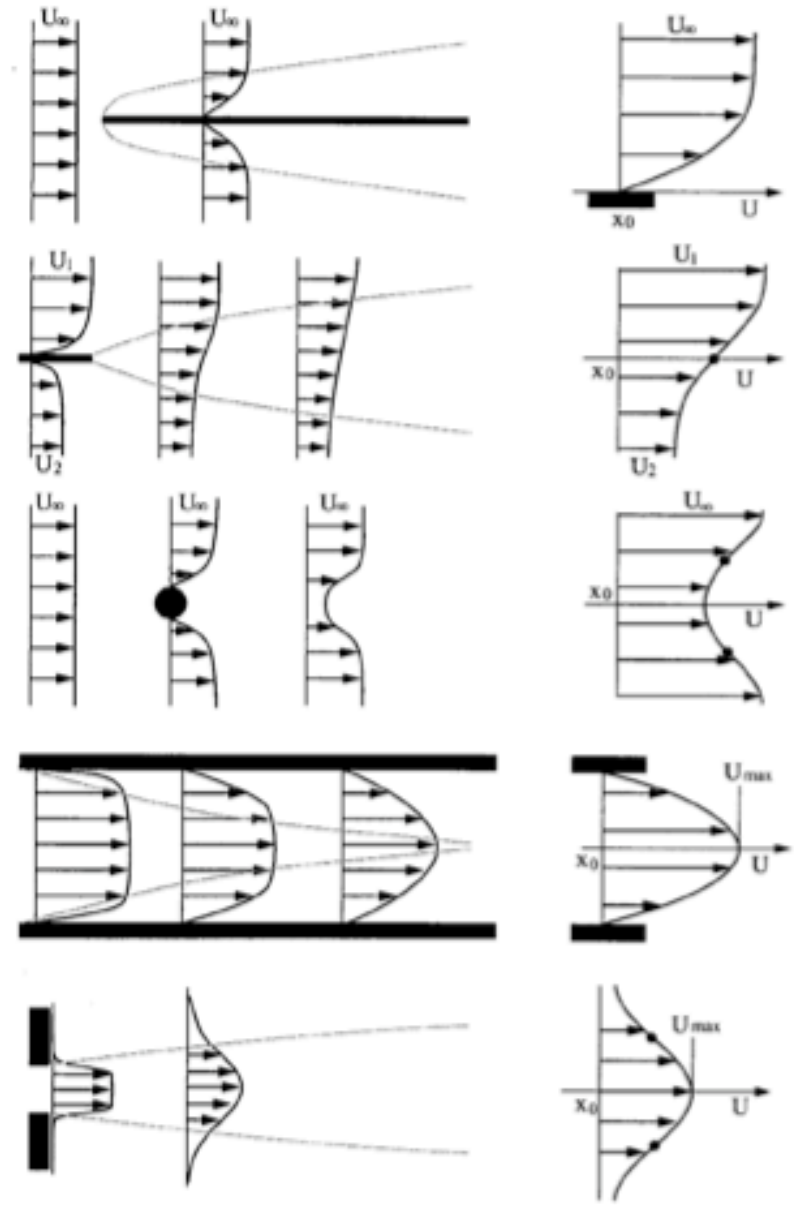
Flat plate boundary layer

Mixing layer

Cylinder wake

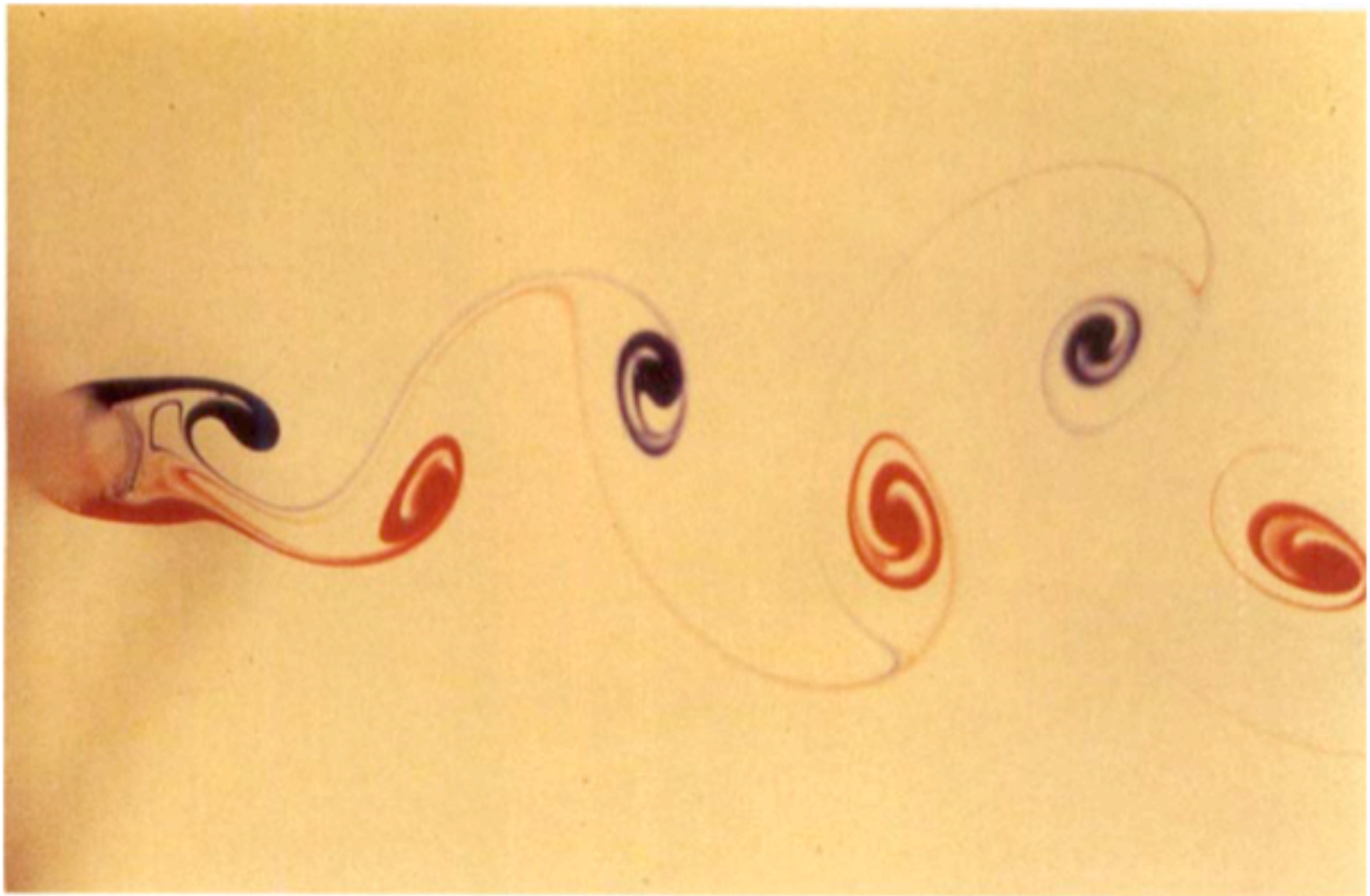
Plane channel flow

2D jet



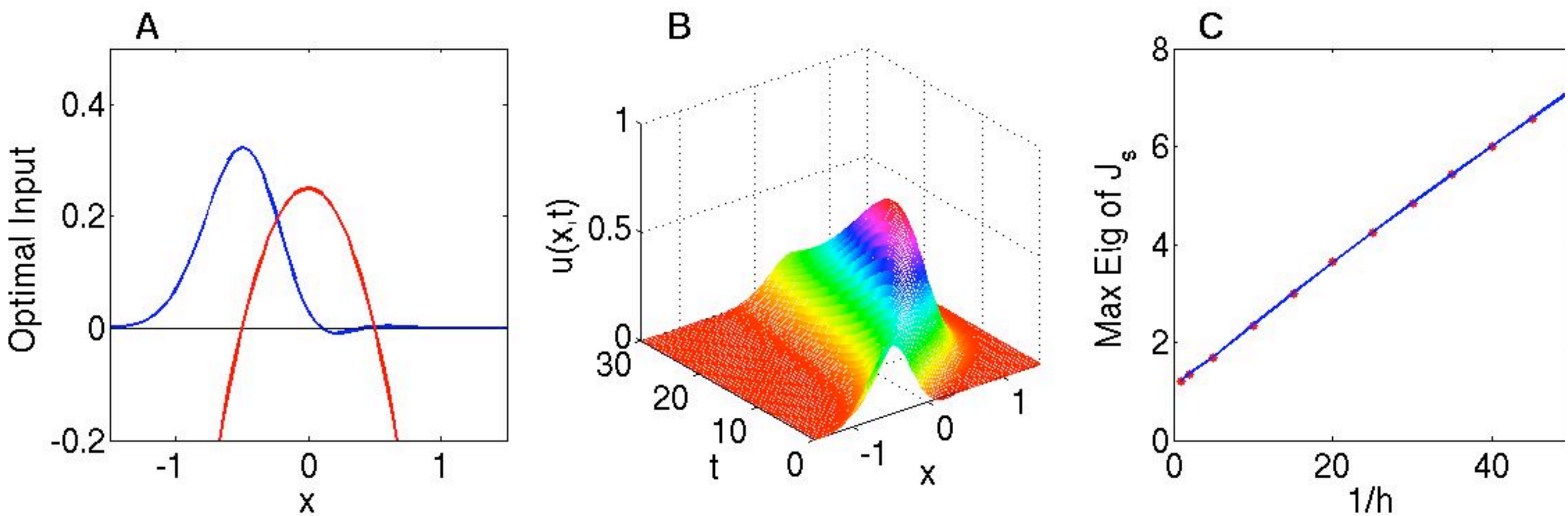
Patrick Huer

Cylinder Wake Beyond the Instability.



Perry, Chong and Lim 1982

A Phenomenological Convective Instability.

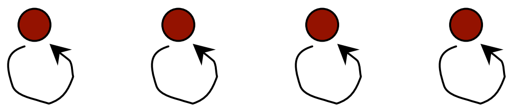


$$\partial_t u = h^2\,\partial_x^2 u - h\,\partial_x u + (\frac{1}{4} - x^2)u + v(x)s(t) + \eta.$$

Cossou and Chomaz PRL 97

Summary so far:

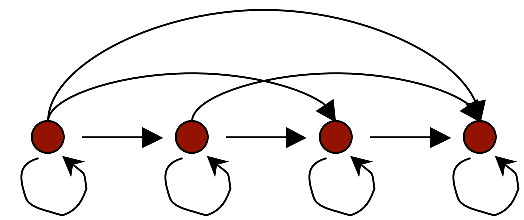
Normal



Homogenous
Feedback Loops

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

Non-normal

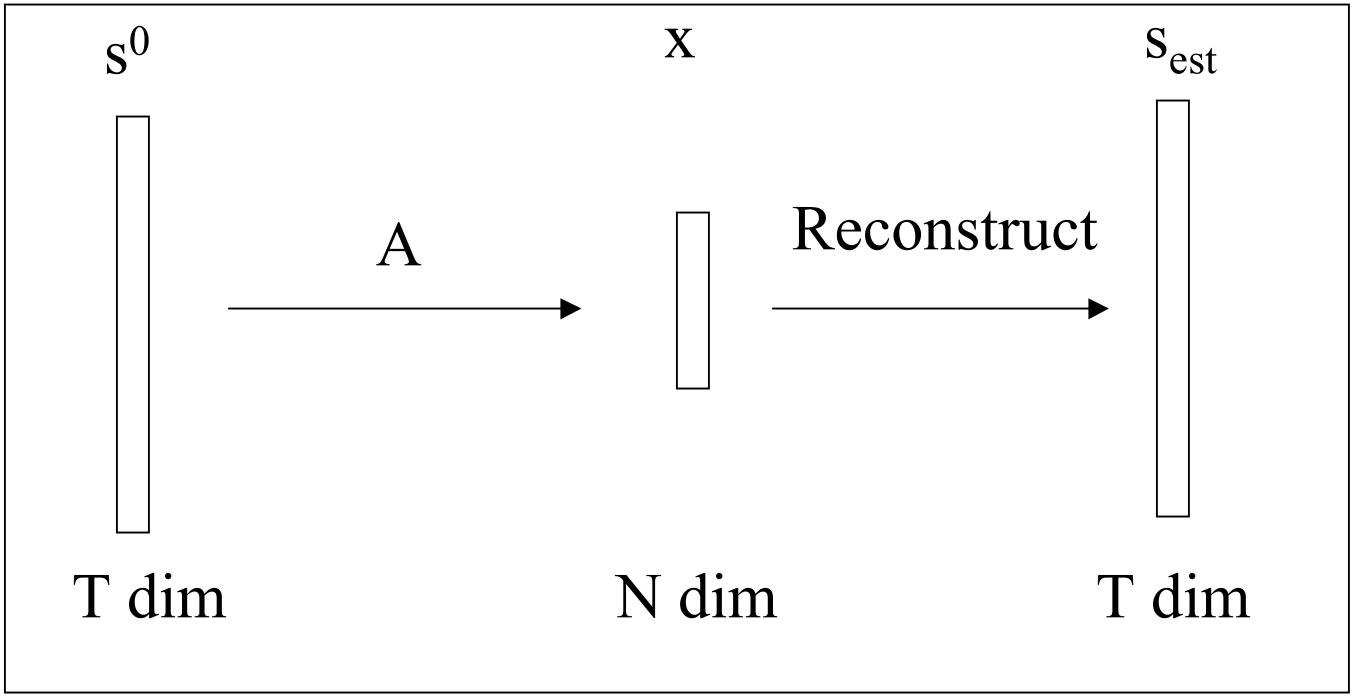


Hidden feedforward amplification cascades

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 numbe of measurements.

Compressed Sensing



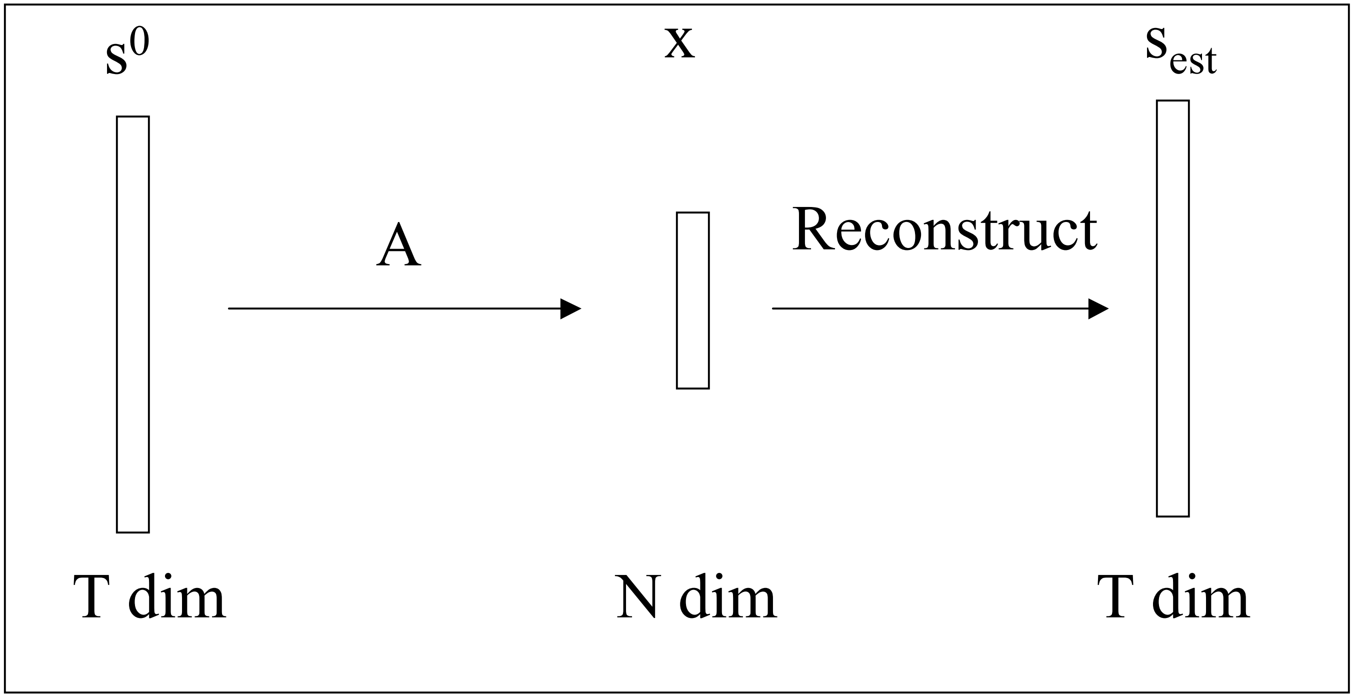
$\mathbf{s}_0$: T dimensional signal with a fraction f elements nonzero

$\mathbf{x} = \mathbf{A}\mathbf{s}_0$: 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 constrain

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 a = N/T < 1

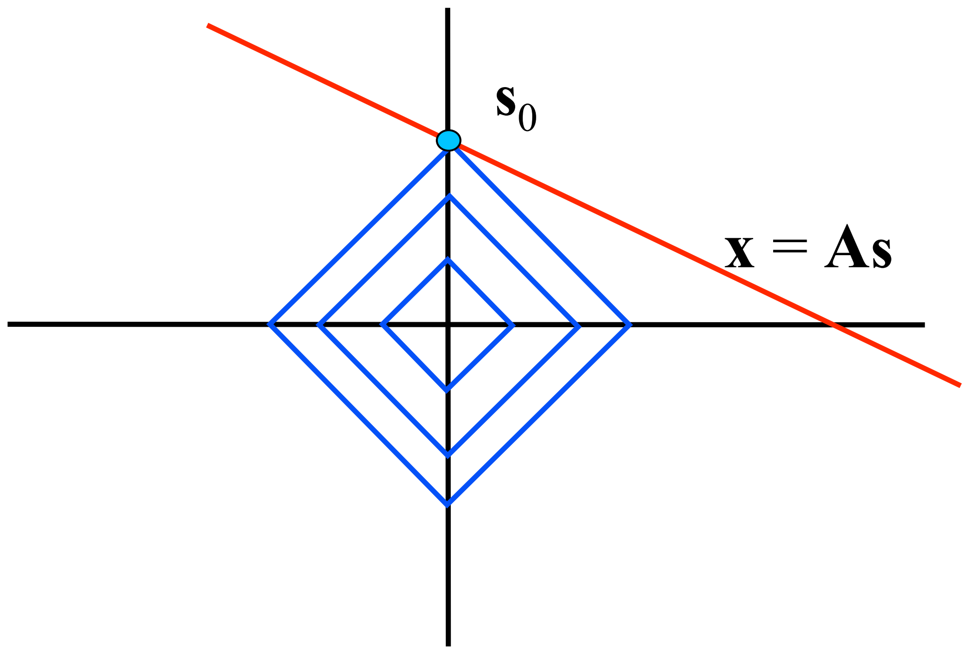
Approaches to constructing an estimate $\mathbf{s}_{est}$ of $\mathbf{s}_0$ from $\mathbf{x}$ when $\mathbf{s}_0$ is sparse:

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

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

Why L1? Geometry behind compressed sensing

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



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: $\mathbf{s}_{est} = \arg\min_{\mathbf{s}} \Sigma_i |s_i|^1$ 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 **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: $\mathbf{s}_{est} = \arg\min_{\mathbf{s}} \Sigma_i |s_i|^1$ subject to $\mathbf{x} = \mathbf{A}\mathbf{s}$

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

$$E(\mathbf{s}) = \lambda/2 \,||\, \mathbf{A}\mathbf{s} - \mathbf{A}\mathbf{s}_0 \,||^2 + \Sigma_i |s_i|$$

later will take λ -> infinity

- This yields a Gibbs distribution

$$P_G(\mathbf{s}) = \exp(-\beta\, E(\mathbf{s}))$$

later will take β -> infinity

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

$$\left\langle\!\left\langle \int D\mathbf{s} \,||\, \mathbf{s} - \mathbf{s}_0 \,||^2 P_G(\mathbf{s}) \right\rangle\!\right\rangle_{\mathbf{A},\mathbf{s}0}$$

Mean field theory of compressed sensing

- The full theory:

$$E(\mathbf{u}) = \lambda/2 \, \| \mathbf{A u} \|^2 + \Sigma_i \, | u_i + s^0_i| \qquad \mathbf{u} = \mathbf{s}\text{-}\mathbf{s}_0$$

A_{nk} is zero mean unit variance gaussian

- Mean field effective theory:

$\Delta Q = Q_1 - Q_0$

z = zero mean unit variance gaussian

$$H(u) = \frac{\alpha}{2\Delta Q}(u - z\sqrt{Q_0/\alpha})^2 + \beta \, | u + s_0 |$$

- Self consistent equations for order parameters:

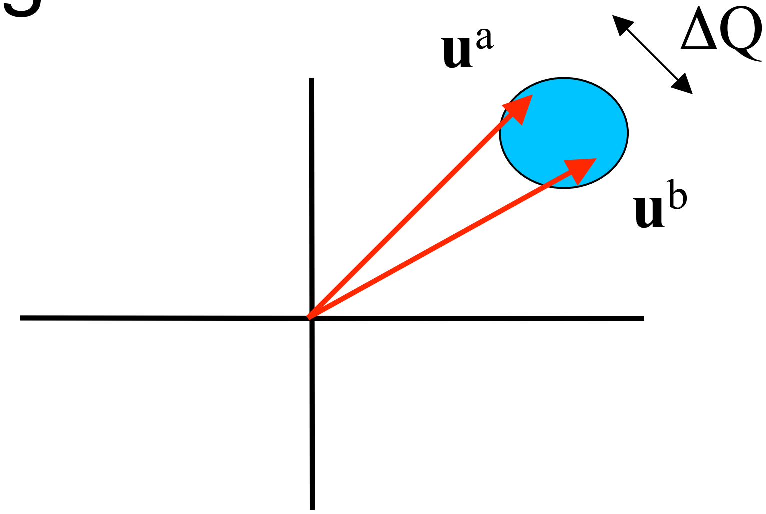
$$Q_0 = \langle\langle \; \langle u \rangle^2_H \; \rangle\rangle_{z,s0}$$
$$\Delta Q = \langle\langle \; \langle (\delta u)^2 \rangle_H \; \rangle\rangle_{z,s0}$$

- Interpretation of order parameters in terms of original problem:
  Let $\mathbf{s}^a$ and $\mathbf{s}^b$ be two candidate signals drawn from the Gibbs distribution P_G

Q_1 = typical value of $1/T \langle \mathbf{u}^a \, \mathbf{u}^a \rangle_{PG}$
Q_0 = typical value of $1/T \langle \mathbf{u}^a \, \mathbf{u}^b \rangle_{PG}$

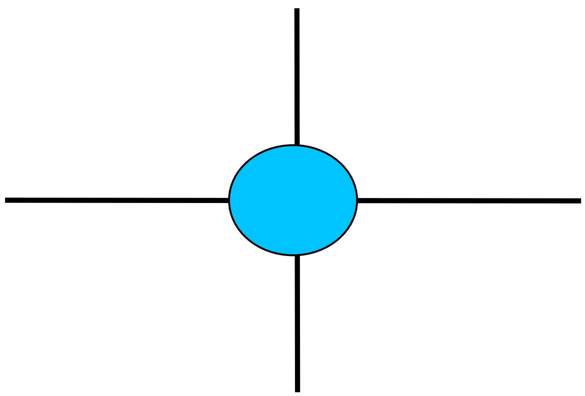
Order parameters and the geometry of low energy configurations

Q_1 = typical value of $1/T < \mathbf{u}^a \, \mathbf{u}^a >_{PG}$
Q_0 = typical value of $1/T < \mathbf{u}^a \, \mathbf{u}^b >_{PG}$



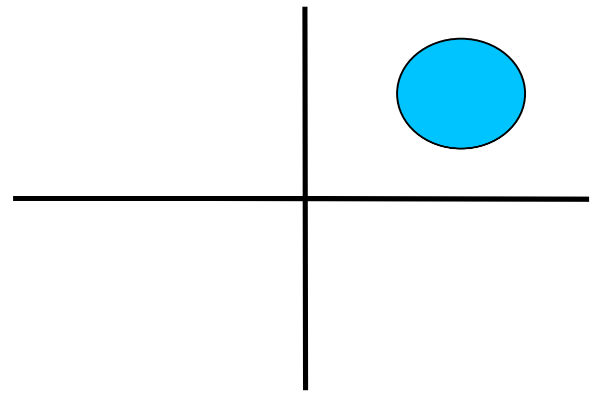
Perfect Reconstruction Solutions

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



Error Solutions

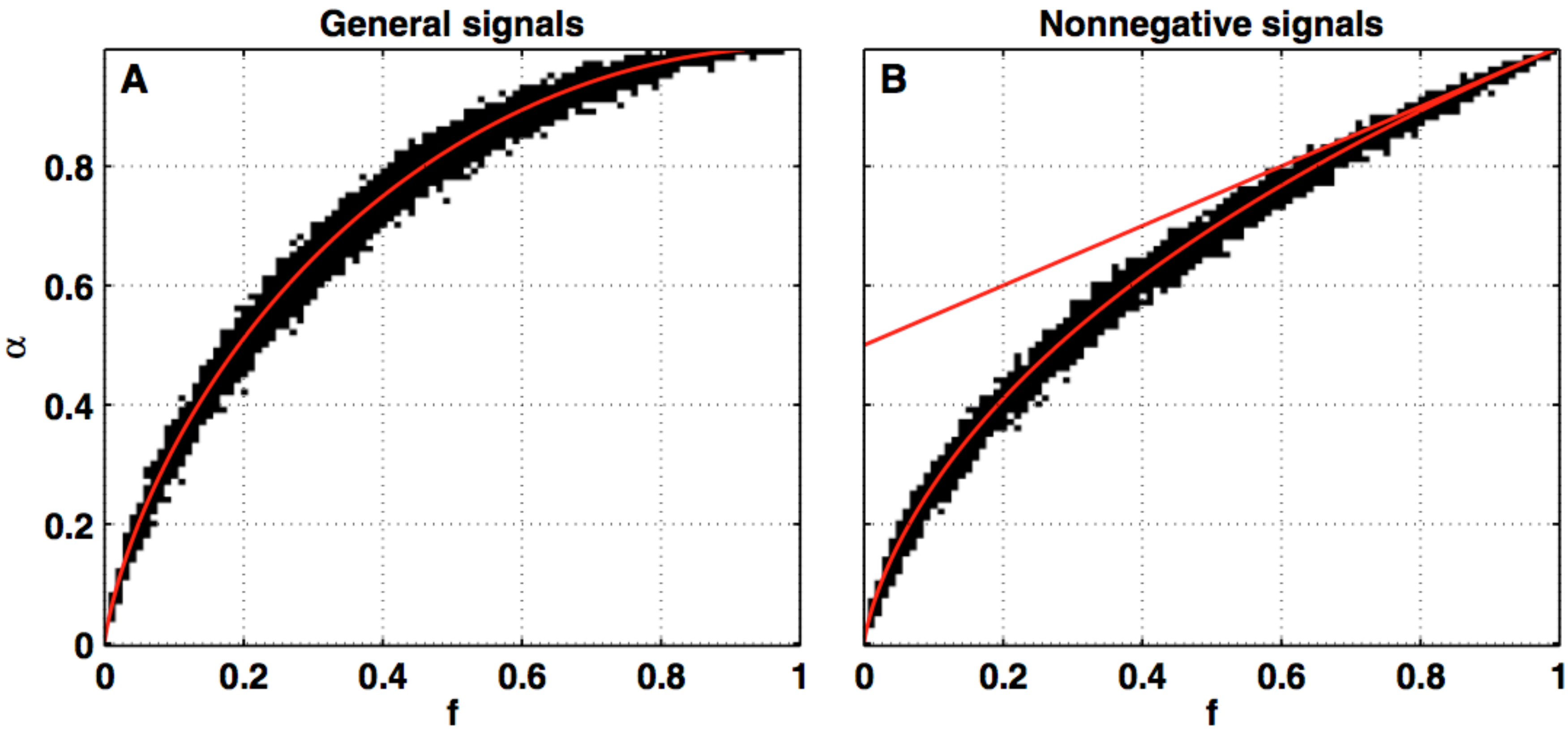
$\Delta Q \sim O(1/\beta)$
$Q_0 \sim O(1)$



Phase transitions in compressed sensing

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

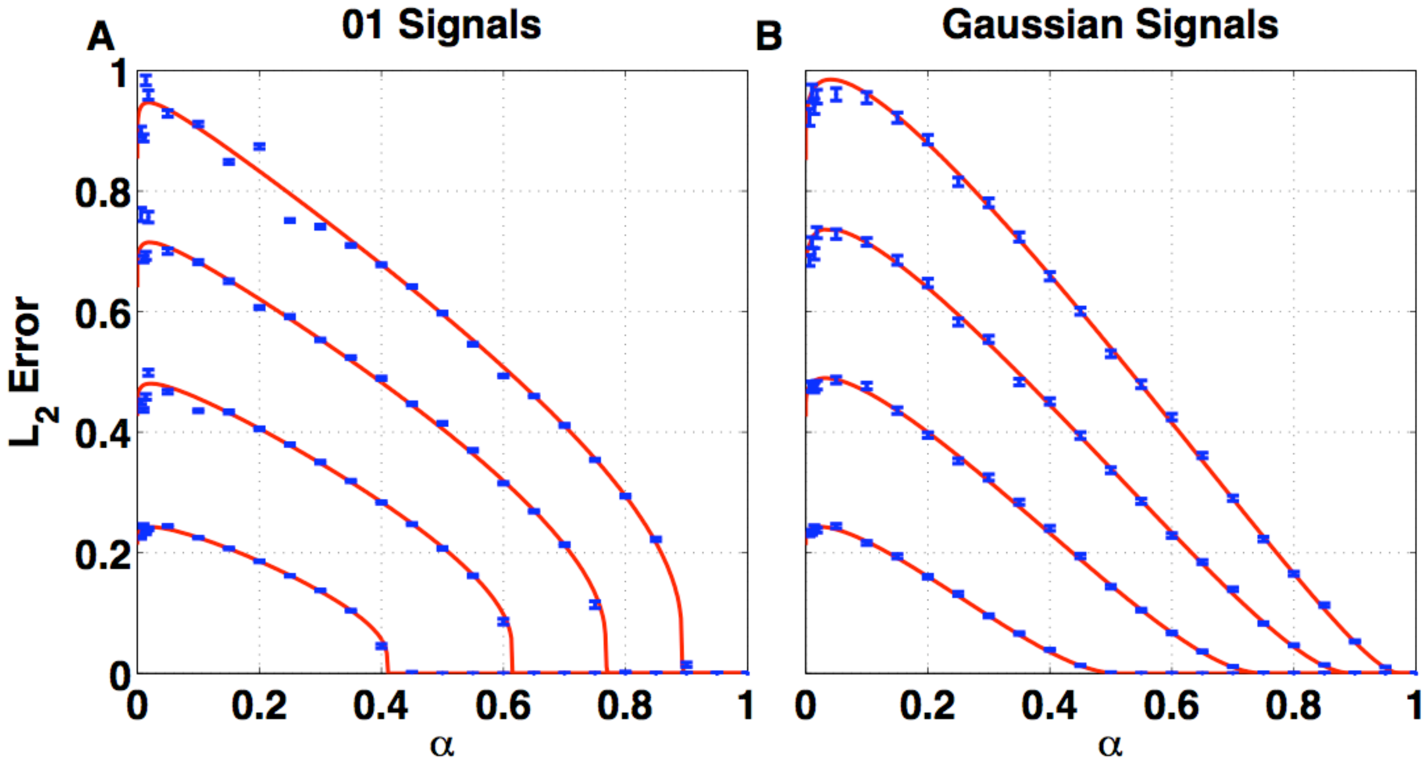
See also Donoho et.al. 2006



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

Ganguli, Sompolinsky PRL 20

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

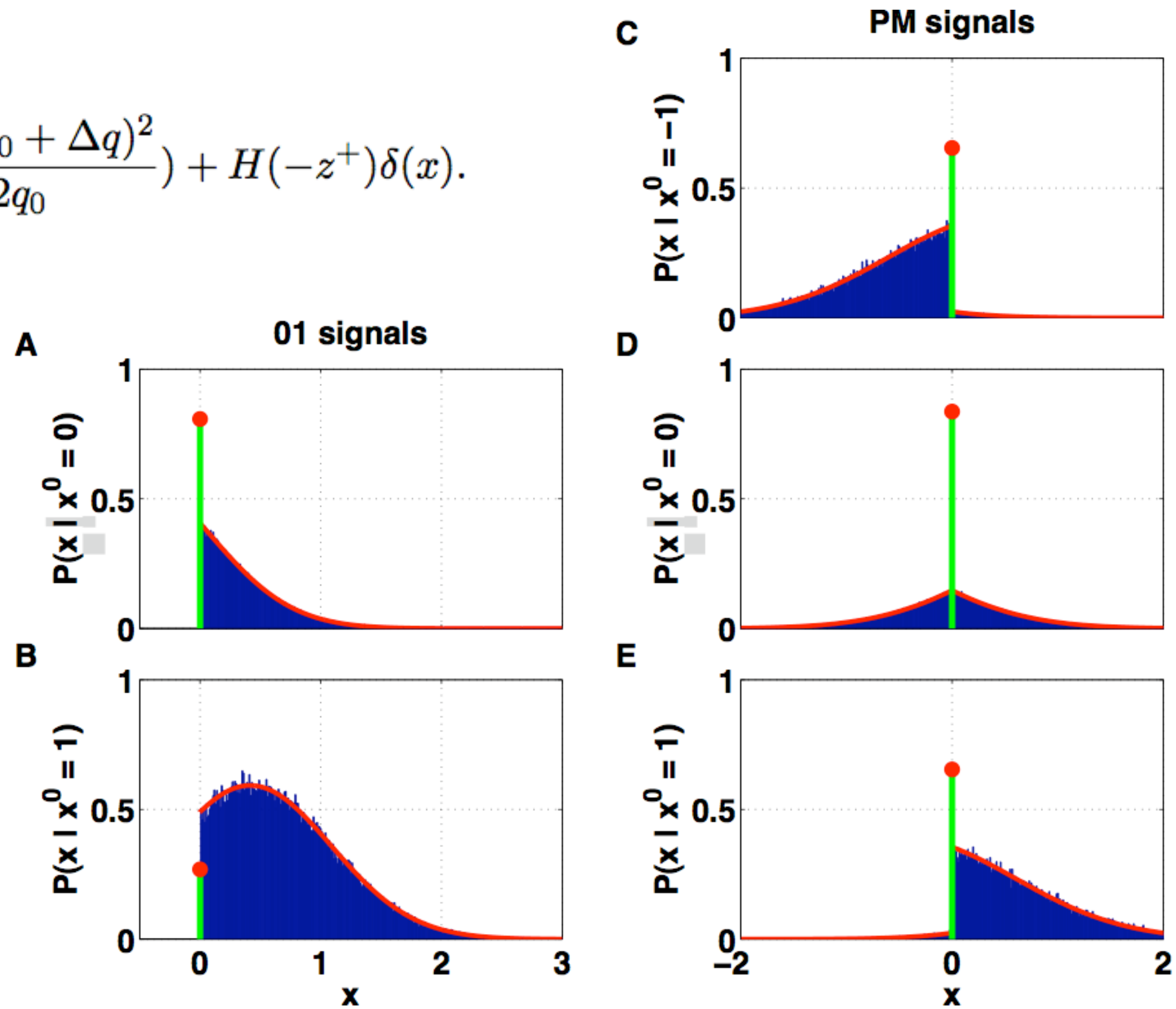
Sharper confinement of nonzeros to origin (smaller ν) => shallower rise of error

Ganguli, Sompolinsky PRL 20

The nature of errors in compressed sensing

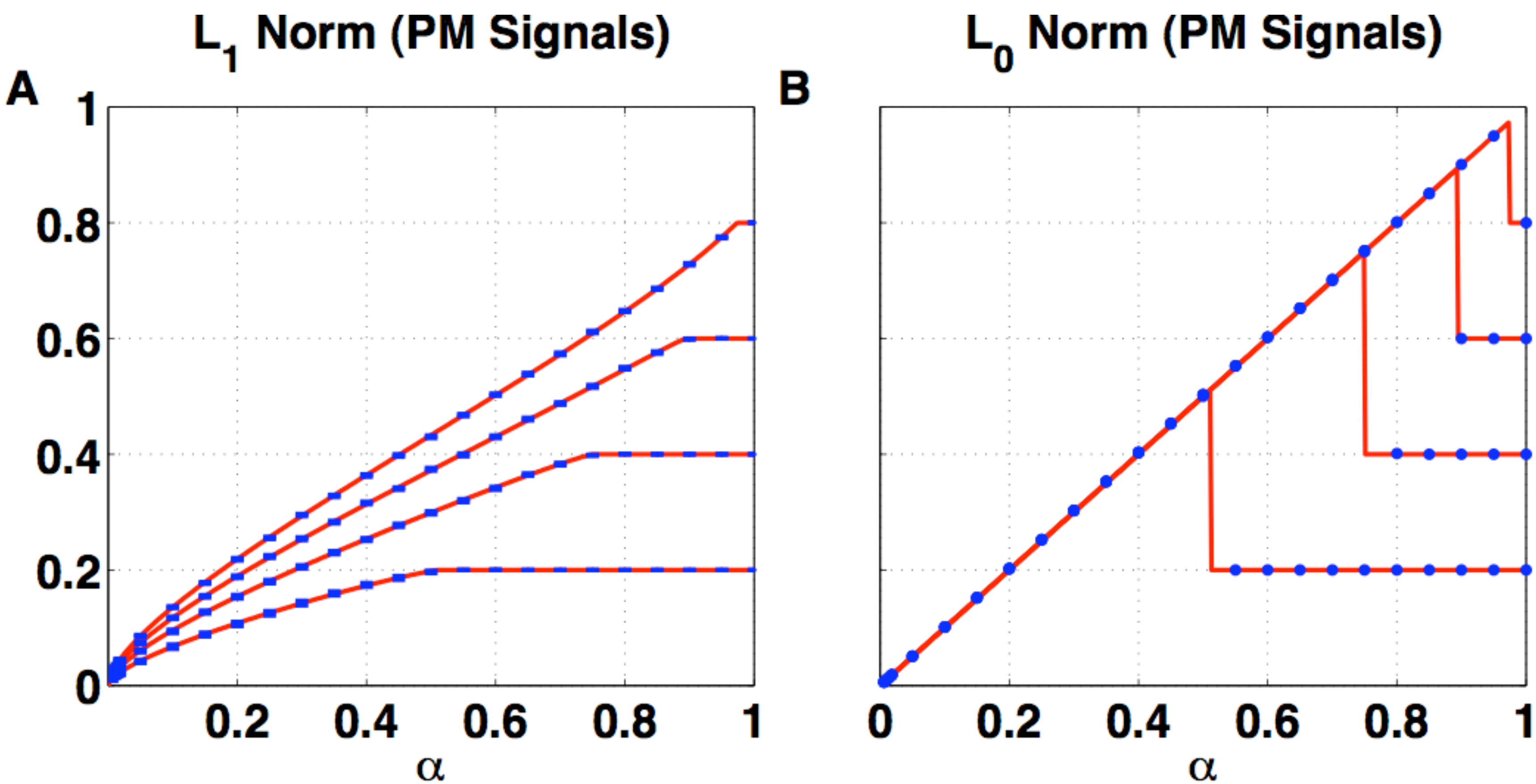
$$P(x|x^0) = \frac{1}{\sqrt{2\pi q_0}} \exp(-\frac{(x - x_0 + \Delta q)^2}{2q_0}) + H(-z^+)\delta(x).$$

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



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

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

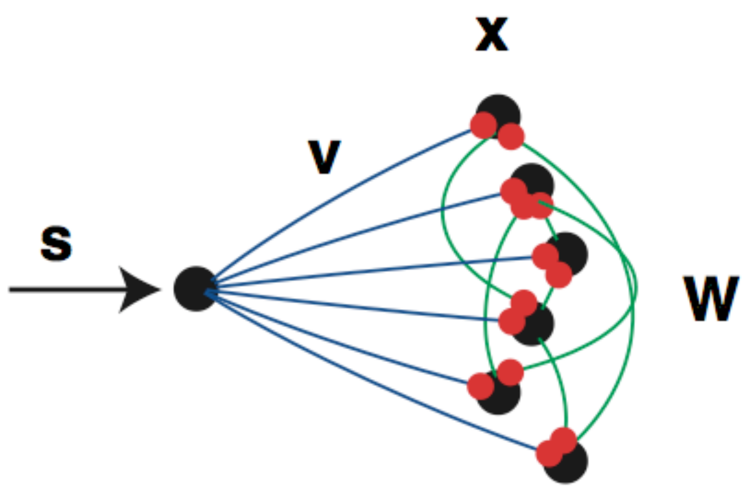
We wish to explain the relation between inputs and outputs via a sparse rule x: I.e. $y_n = \Sigma_k A_{nk} x_k$ for each n

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

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

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 NxT Measurement matrix: $\mathbf{A}_{nk} = (\mathbf{W}^k\mathbf{v})_n$.

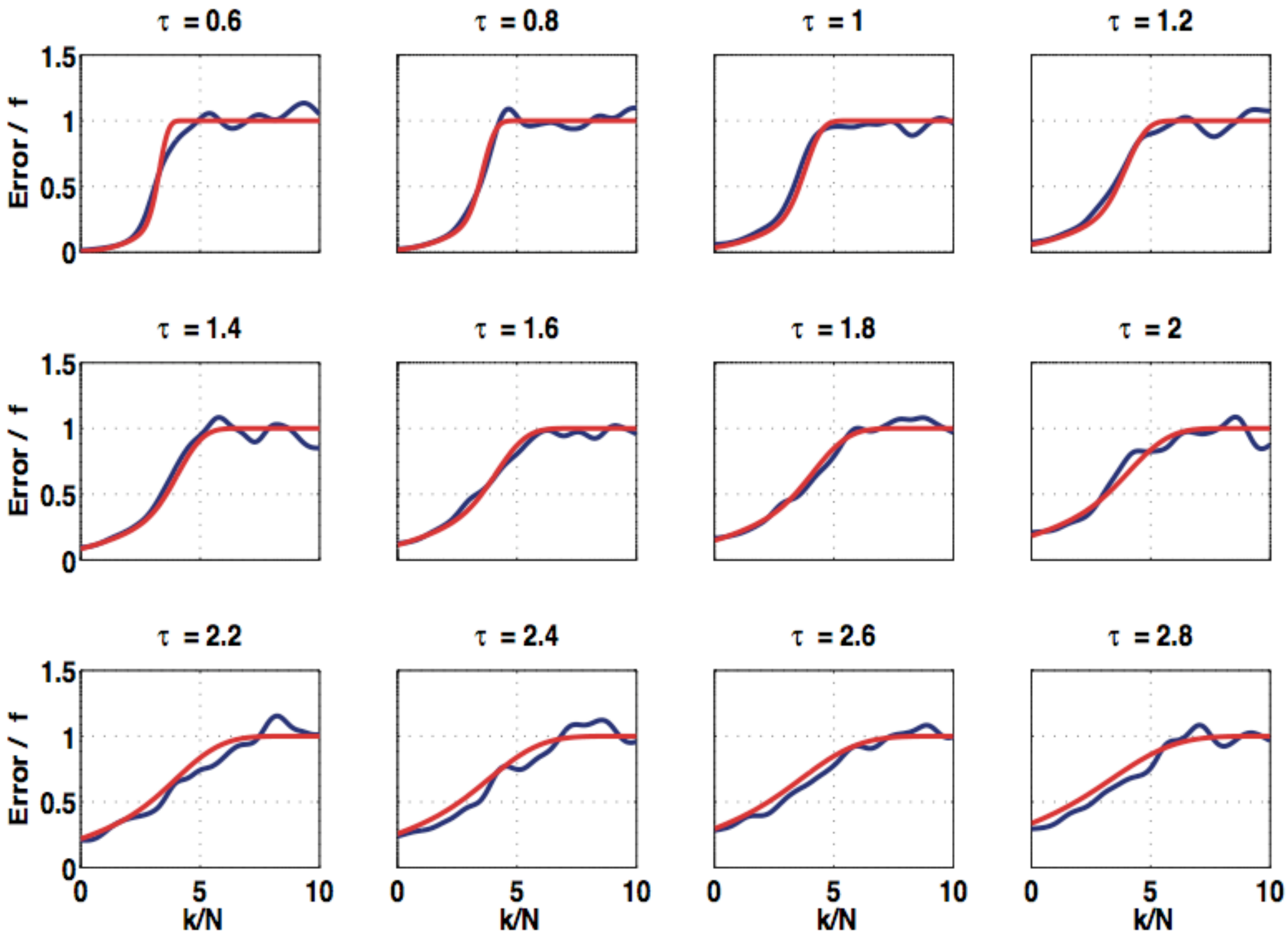
Annealed approximation (AA): $\mathbf{A}_{nk}$ = zero mean gaussian with var ρ^{2k} where $\rho = \exp(-1/\tau N)$. Reflects decay in dynamical system, but not correlations.

$$A_{nk} = (W^k v)_n \quad : \quad \begin{bmatrix} | & | & | & & | & \\ v & Wv & W^2v & \ldots & W^kv & \ldots \\ | & | & | & & | & \end{bmatrix}$$

Memory performance in the annealed approximation

$$E(k/N) = < (s_{est}(n-k) - s^0(n-k))^2 >_{A,s0}$$

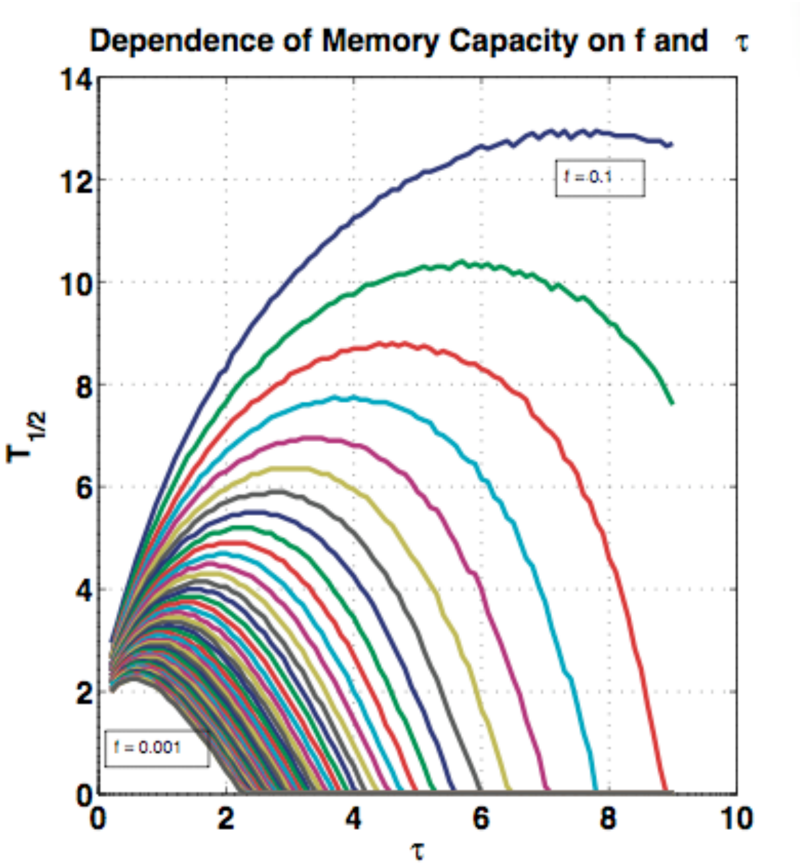
Memory curve = reconstruction error as a function of time into the past

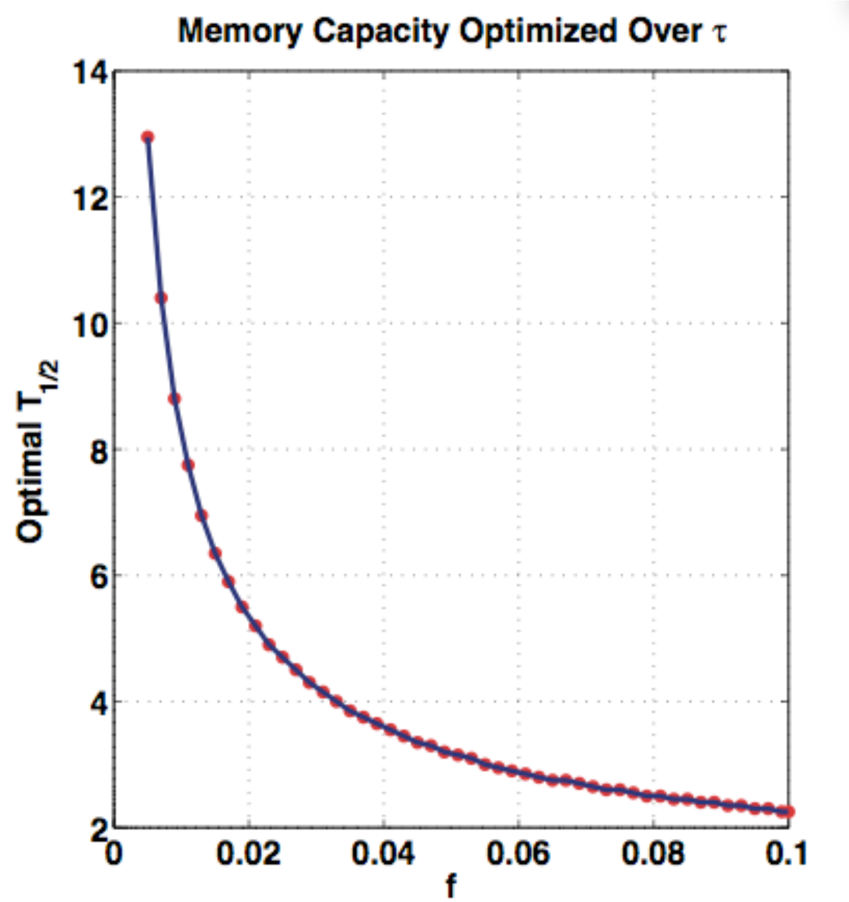


Memory curves for f = 0.(
Red: theory
Blue: simulations

Ganguli, Sompolinsky NIPS 201(

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:

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

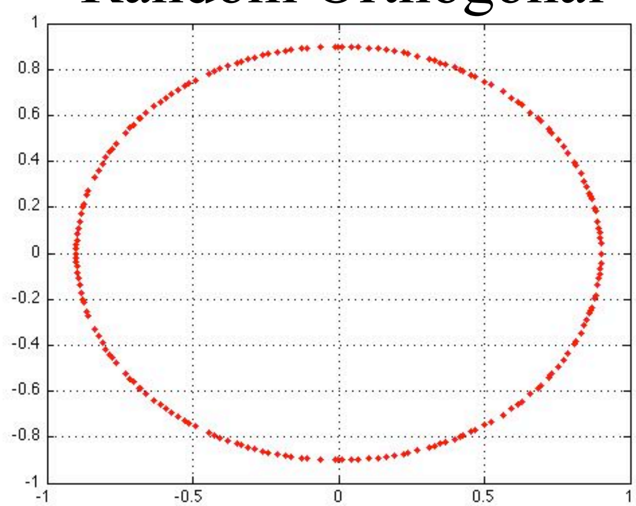
Ganguli, Sompolinsky NIPS 2010

Implementing the annealed approximation

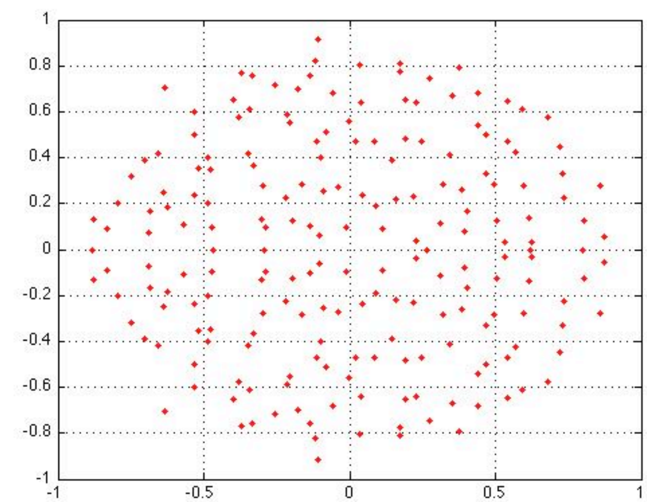
$$A_{nk} = (W^k v)_n \quad : \quad \begin{bmatrix} | & | & | & & | & \\ v & Wv & W^2 v & \dots & W^k v & \dots \\ | & | & | & & | & \end{bmatrix}$$

- $A_{nk} \sim$ 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



Network Design Principles Underlying Sequence Memory

- Multiple conflicting design constraints on sequence memory networks:
  (1) Stability of internal representations => remember distant past
  (2) Flexibility of internal representations => acquire more recent inpu
  (3) Amplification of input signals without destructive noise amplificatic

- Nonnormal networks, characterized by (possibly hidden) feedforward structure, but not feedback networks, achieve all three, and exhibit dynamical short-term memory representations

- Within the class of general networks considered, only nonnormal networks can do so.

- At high SNR, compressed sensing can lead to improved memory performance for temporally sparse inputs, but again, only with dynamical short-term memory representations.