Canonic Neural Network Models

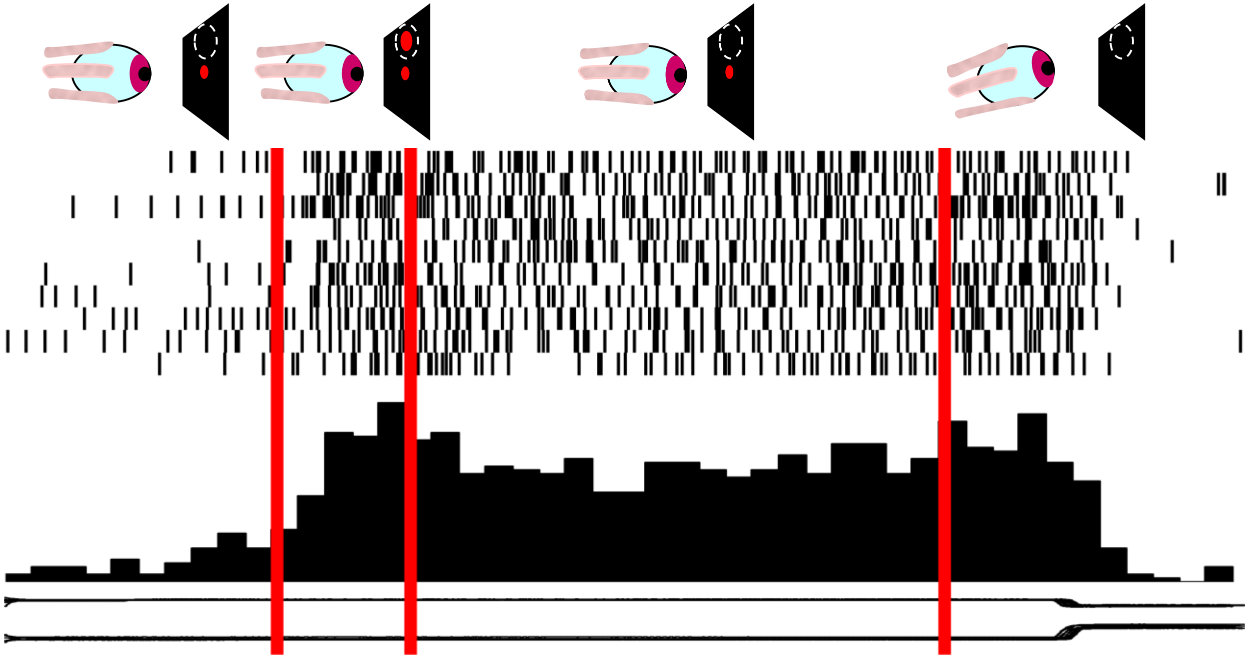
吴思
心理与认知科学学院
IDG/McGovern 脑科学研究所
北大-清华联合生命科学中心
北京大学

Attractor Neural Networks

- Networks of various types/structures, formed by large numbers of neurons, are the substrate of brain functions.
- The brain carries out computation by updating network states in response to external inputs.
- The stationary states, i.e., attractors, of networks encode the stimulus information.

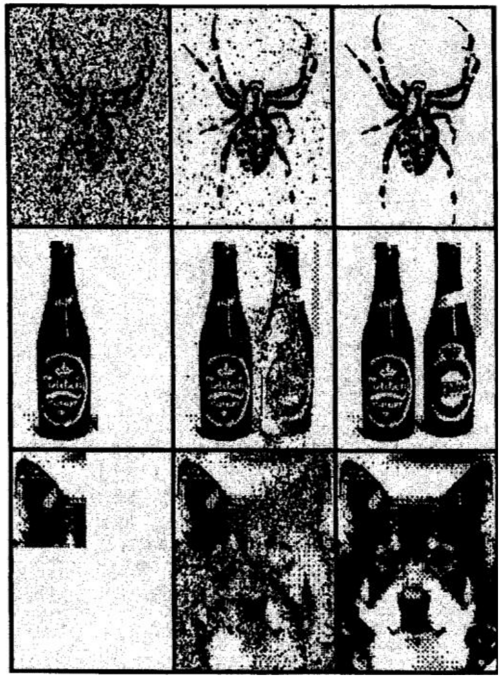
Part I. Hopfield Model

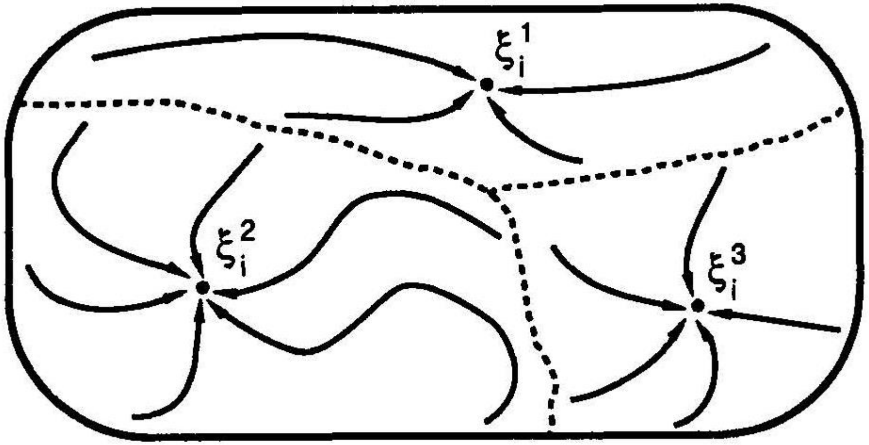
Persistent Activity in Working Memory



The Hopfield Model

- An attractor model
- The simplest model captures the computation of a network
- A model for associative memory—content-addressable memory
- Should be the Amari-Hopfield model





The mathematical formulation

$S_i = \pm 1$: the neuronal state

w_{ij} : the neuronal connection

The network dynamics:

$$S_i = \mathrm{sign}\left(\sum_j w_{ij} S_j - \theta \right), \quad \mathrm{sign}(x) = 1, \text{for } x > 0; -1, \text{ otherwise}$$

Updating rule: synchronous or asycchronous

The Energy Function

Energy function: $E = -\frac{1}{2}\sum_{i,j} w_{ij}S_iS_j + \theta\sum_i S_i$

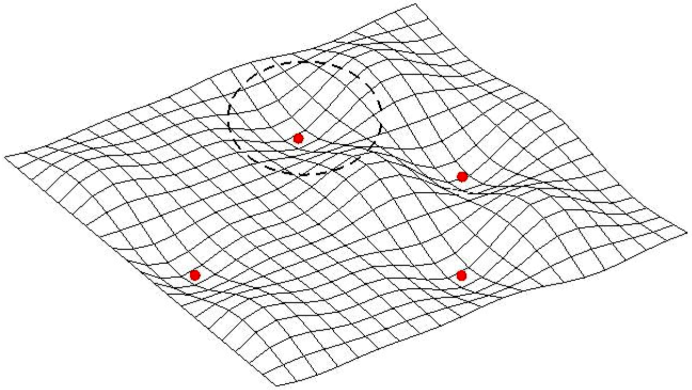
Consider S_i is updated, $S_i(t+1) = sign[\sum_j w_{ij}S_j(t) - \theta]$

$$\Delta E = E(t+1) - E(t)$$

$$= -[S_i(t+1) - S_i(t)]\sum_j w_{ij}S_j(t) + \theta\,[S_i(t+1) - S_i(t)]$$

$$= -[S_i(t+1) - S_i(t)][\sum_j w_{ij}S_j(t) - \theta]$$

$$\leq 0$$



The case of one pattern

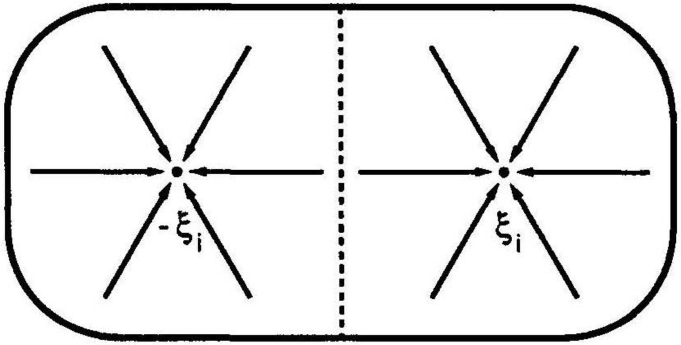
Consider the network stores only one pattern, ξ_i, for $i = 1, \ldots N$

Setting $w_{ij} = \dfrac{1}{N}\xi_i \xi_j$: analogy to the Hebb rule

The memory pattern is always stable:

$$\operatorname{sign}\left(\sum_j w_{ij}\xi_j\right) = \operatorname{sign}(\xi_i) = \xi_i$$

The attracting basin



The case of many patterns

Consider the network stores p pattern, ξ_i^μ, for μ=1,...p; $i = 1, \ldots N$

Setting $w_{ij} = \frac{1}{N}\sum_{\mu=1}^{p} \xi_i^\mu \xi_j^\mu$

The stability condition of a particular memory pattern:

sign$(h_i^v) = \xi_i^v$, for all i,

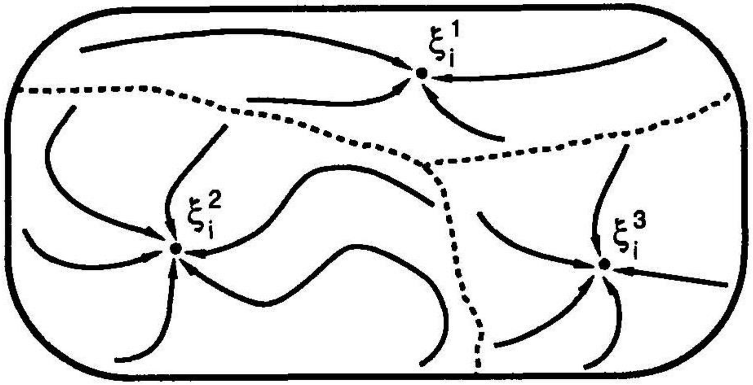
$$h_i^v = \sum_j w_{ij}\xi_j^v = \frac{1}{N}\sum_j \sum_\mu \xi_i^\mu \xi_j^\mu \xi_j^v$$

$$= \xi_i^v + \frac{1}{N}\sum_j \sum_{\mu \neq v} \xi_i^\mu \xi_j^\mu \xi_j^v$$

The error comes from the cross−talk term,

$$\frac{1}{N}\sum_j \sum_{\mu \neq v} \xi_i^\mu \xi_j^\mu \xi_j^v,$$

which is due to pattern correlation.



The capacity for storing random patterns

Consider the network stores p random pattern, ξ_i^μ, for μ=1,...p; $i = 1, \ldots N$

The stable condition of a particular memory pattern:

$\text{sign}(h_i^\nu) = \xi_i^\nu$, for all i,

$$h_i^\nu = \xi_i^\nu + \frac{1}{N}\sum_j \sum_{\mu \neq \nu} \xi_i^\mu \xi_j^\mu \xi_j^\nu$$

$$h_i^\nu \xi_i^\nu = 1 - C_i^\nu$$

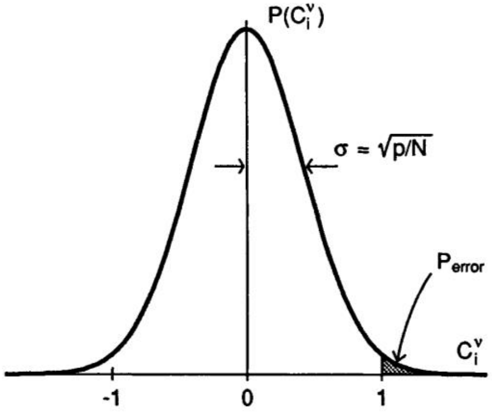
$$C_i^\nu = -\xi_i^\nu \frac{1}{N}\sum_j \sum_{\mu \neq \nu} \xi_i^\mu \xi_j^\mu \xi_j^\nu$$

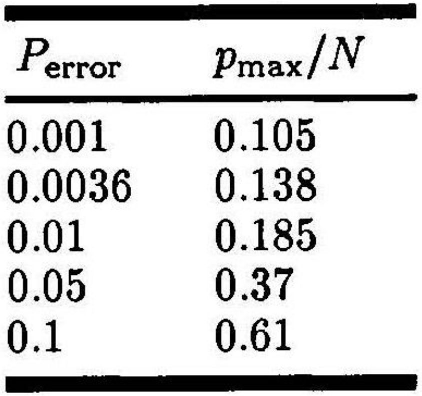
The error occurs when $C_i^\nu > 1$

$P_{error} = \text{Prob}(C_i^\nu > 1)$

In the limit of large N & p, $\text{Prob}(C_i^\nu)$ satisfies a Gassian distribution with zero mean and variance $\sigma^2 = p/N$. Thus,

$$P_{error} = \frac{1}{\sqrt{2\pi}\sigma}\int_1^\infty e^{-x^2/2\sigma^2}\,dx$$

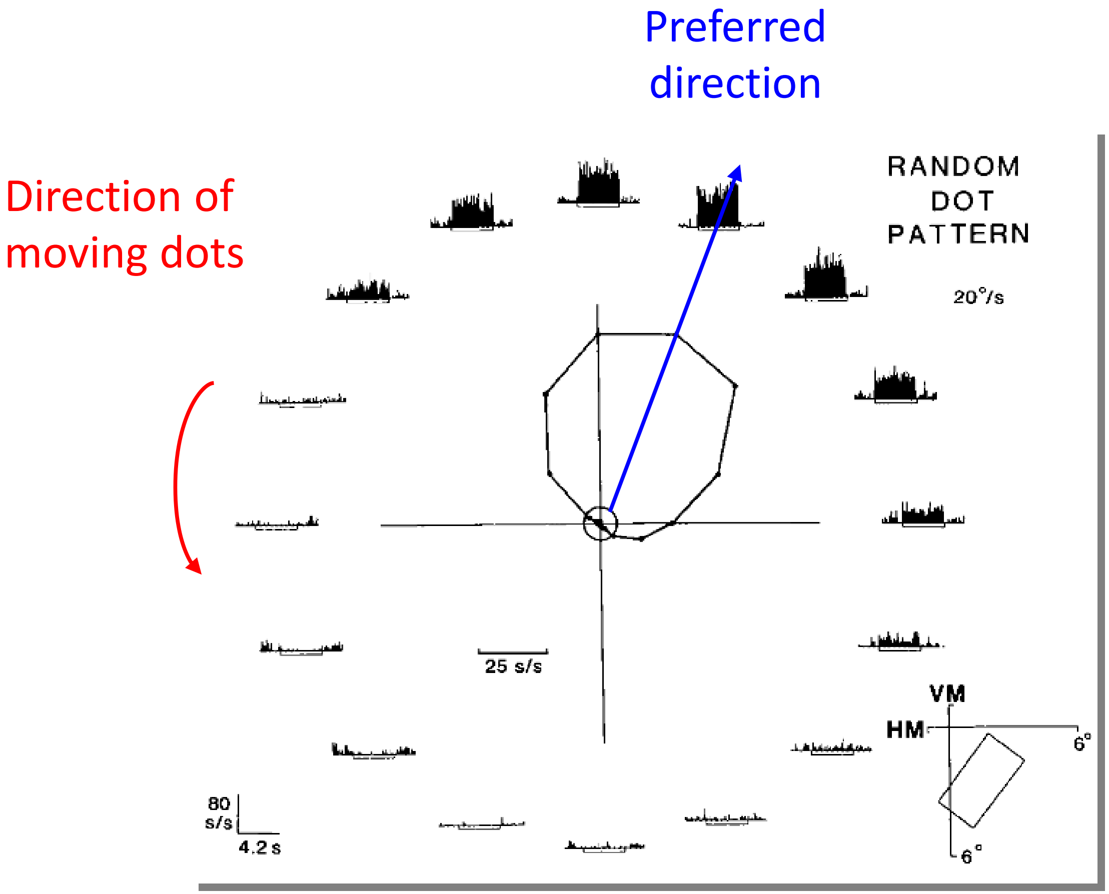




P_{error}	p_{max}/N
0.001	0.105
0.0036	0.138
0.01	0.185
0.05	0.37
0.1	0.61

Part II. Continuous Attractor Neural Networks

Neural Encoding of Motion Direction



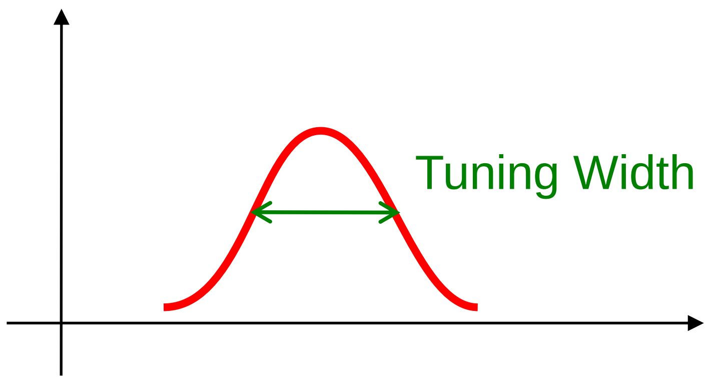
Activities of macaque Middle Temporal (MT) neurons (TD Albright 1984)

Neural Population Code

Individual neurons:

- Preferred feature value
- Bell-shape tuning function

Firing rate

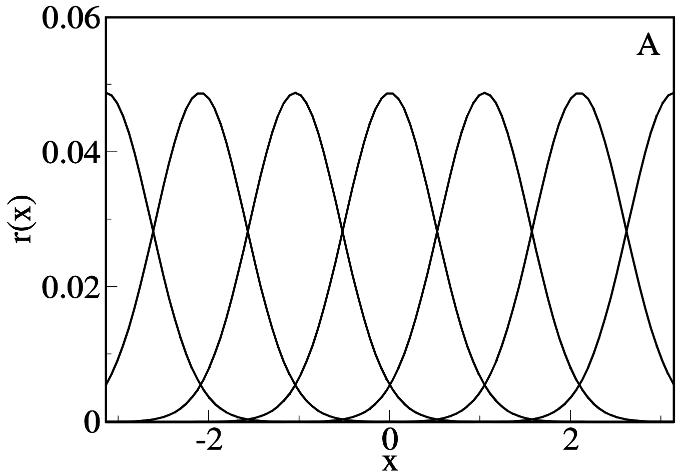


Tuning Width

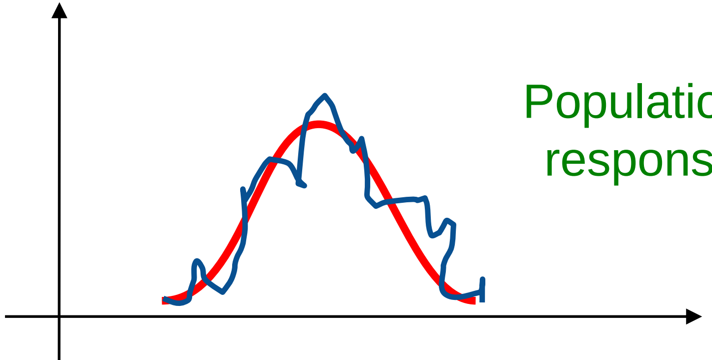
Feature value

A neural population:

- Overlapped tuning functions covering the whole space
- Largely independent responses



Feature space



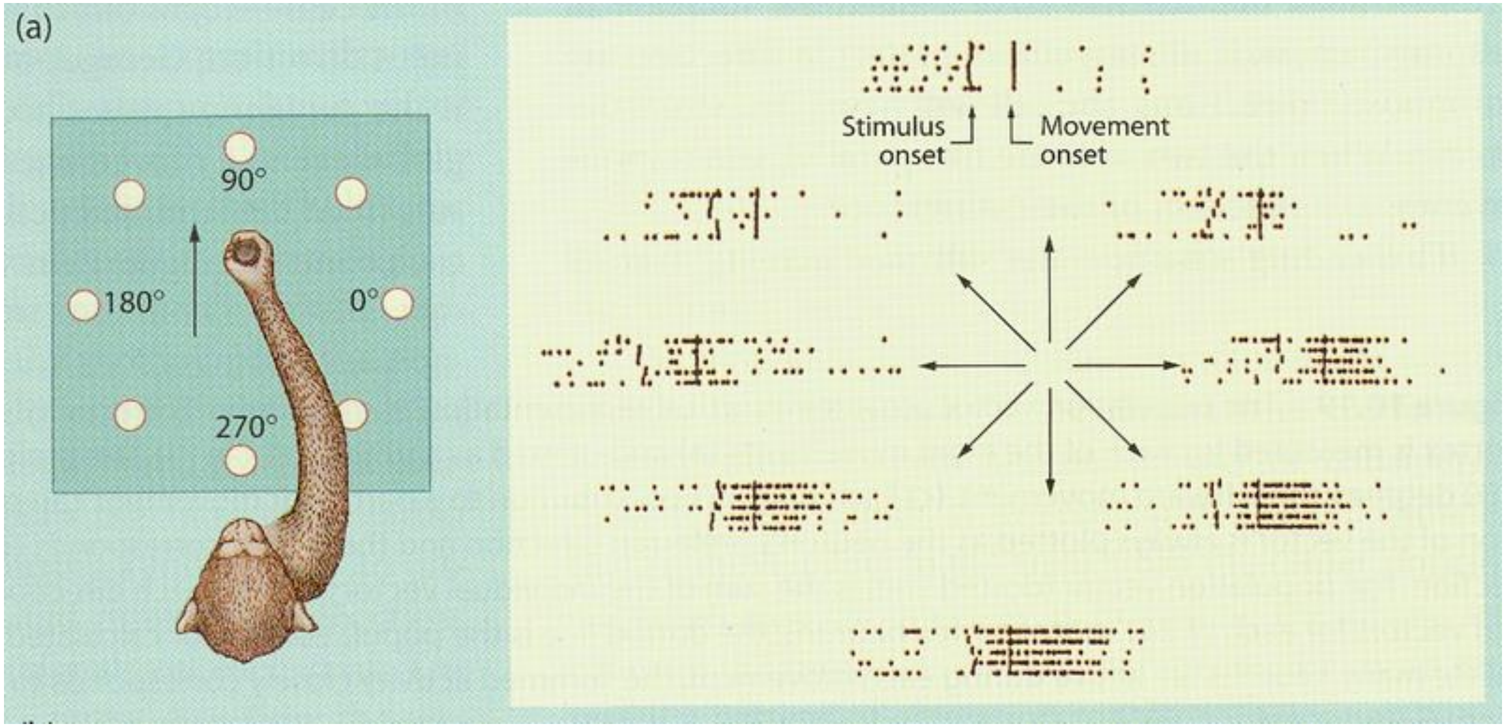
Population response

Neuron Index

Population Code for Moving Direction

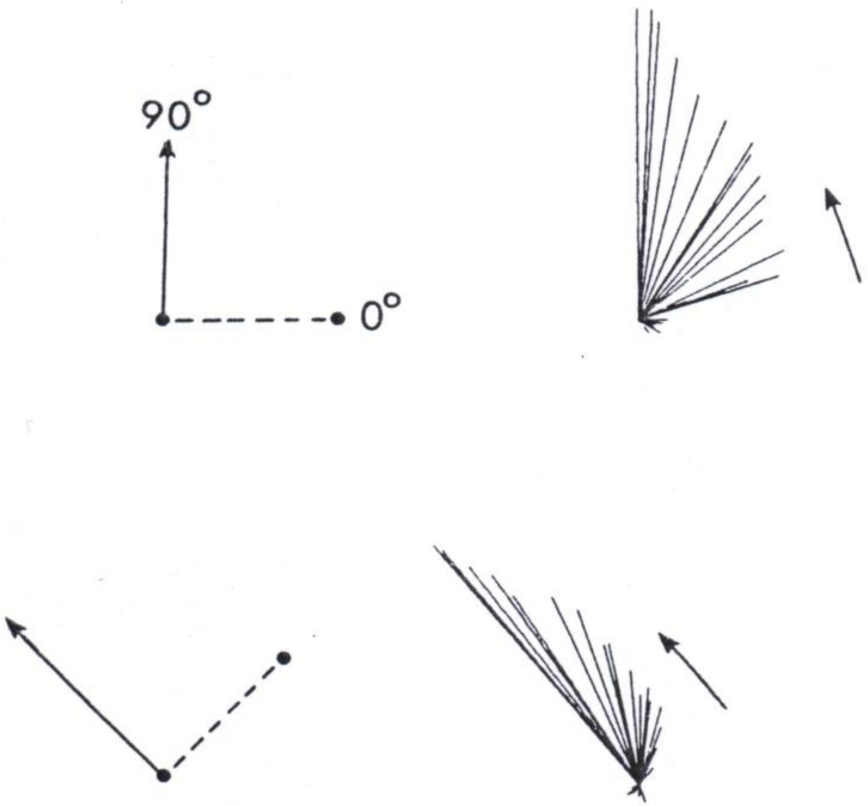
- In the experiment, the monkey was guided to move the lever in the center of apparatus to one of eight peripheral locations.

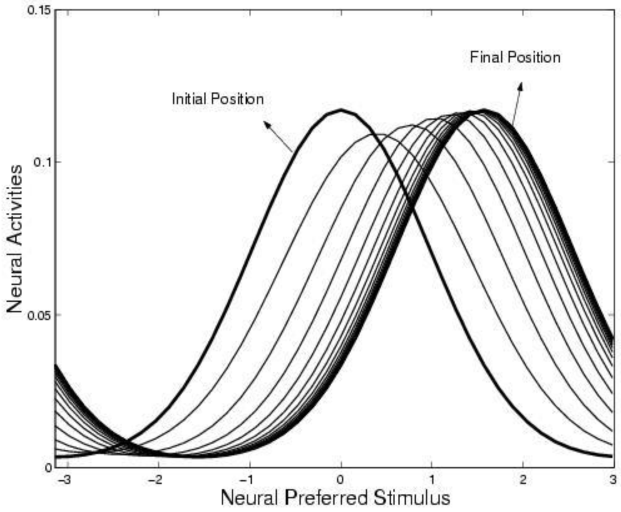
- Neural activities in the motor area were recorded.



Georgepolous et al, Science 1987

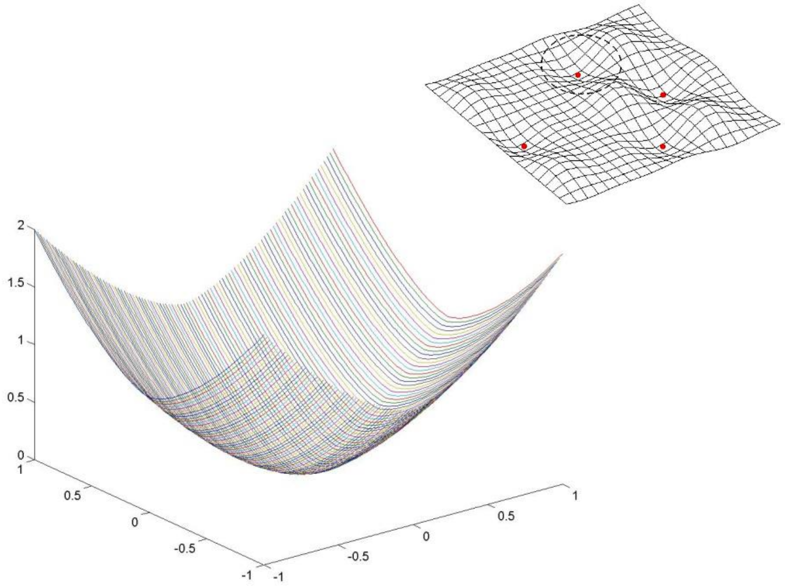
Mental rotation in the premotor cortex



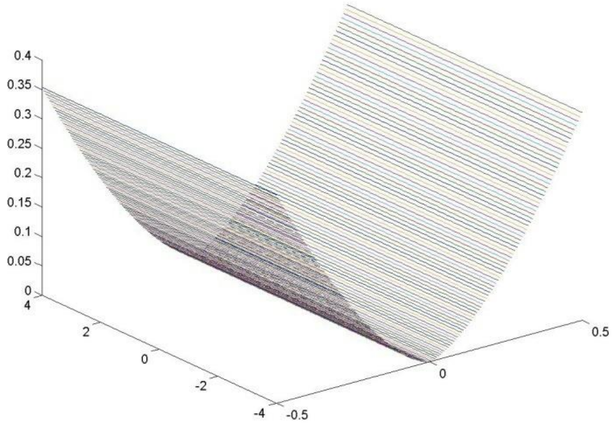


A. Georgopoulos et al., science, 1993

Discrete vs. Continuous Attractors



Discrete attractor



1D continuous attractors

For line attractor

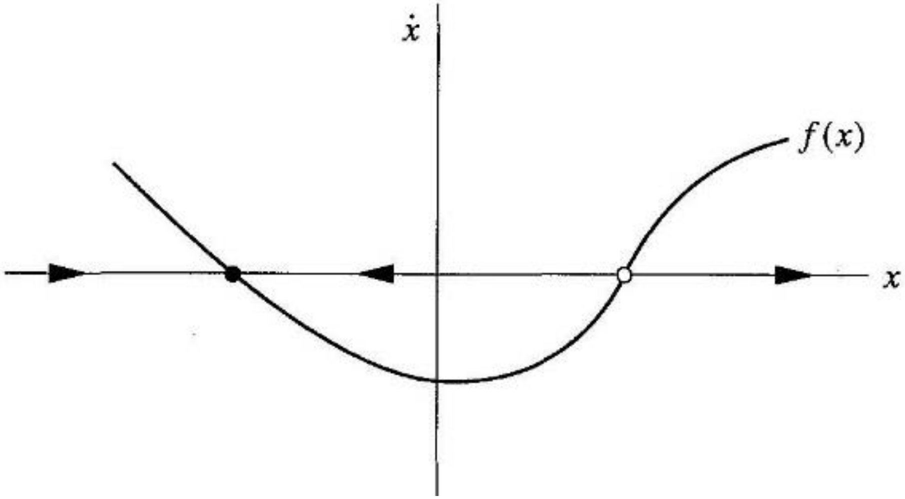
- The steady states of the system form the valley, a one-dimensional parameter space.
- The system is neutrally stable along the attractor space, i.e., no resistance when the system moving along the valley.

A little bit of dynamical system

- **Fixed point and stability**

$$\frac{dx}{dt} = f(x)$$

x^* : fixed point, if $f(x^*) = 0$



A little bit of dynamical system

$\eta = x - x^*$: a small perturbation away from the fixed point

Linearizing the dynamics around the fixed point

$$\frac{d\eta}{dt} = \frac{dx}{dt} = f(x^* + \eta)$$
$$= f(x^*) + \eta \nabla f(x^*) + O(\eta^2)$$
$$\approx \eta \nabla f(x^*)$$

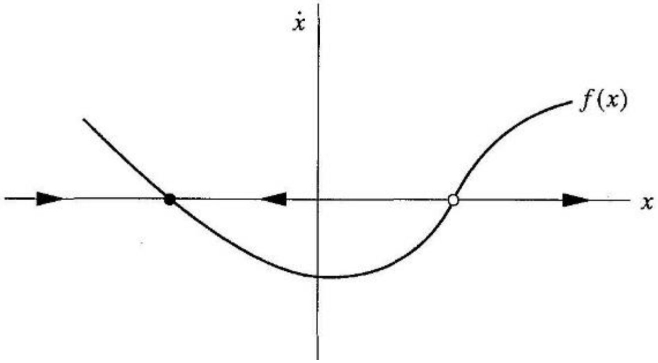
The fixed point is unstable, if $\nabla f(x^*) > 0$

The fixed point is stable, if $\nabla f(x^*) < 0$

Around the fixed point,

$$\eta(t) \approx \eta(t=0) e^{\nabla f(x^*)t} = \eta(t=0) \exp\left[\frac{sign(\nabla f(x^*))t}{|1/\nabla f(x^*)|}\right]$$

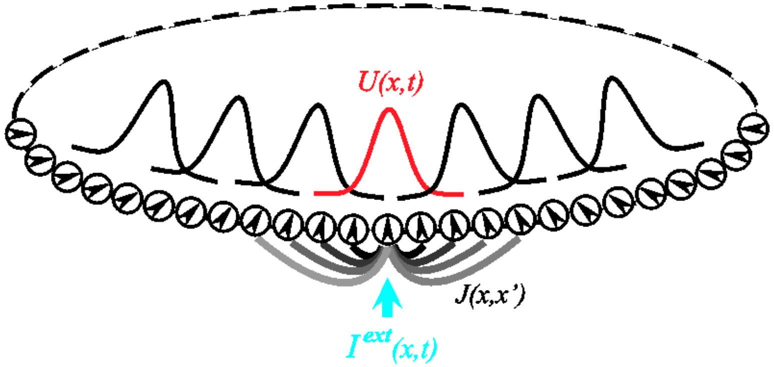
$|1/\nabla f(x^*)|$ is the time constant



Continuous Attractor Neural Network (CANN)

$$\tau \frac{\partial U(x,t)}{\partial t} = -U(x,t) + \rho \int dx' J(x-x') r(x',t) + I^{ext}(x,t)$$

$$r(x,t) = \frac{U(x,t)^2}{1 + k\rho \int dx' U(x',t)^2}; \quad J(x-x') = \frac{J}{\sqrt{2\pi a}} \exp\left[-\frac{(x-x')^2}{2a^2}\right]$$



Key Structure:

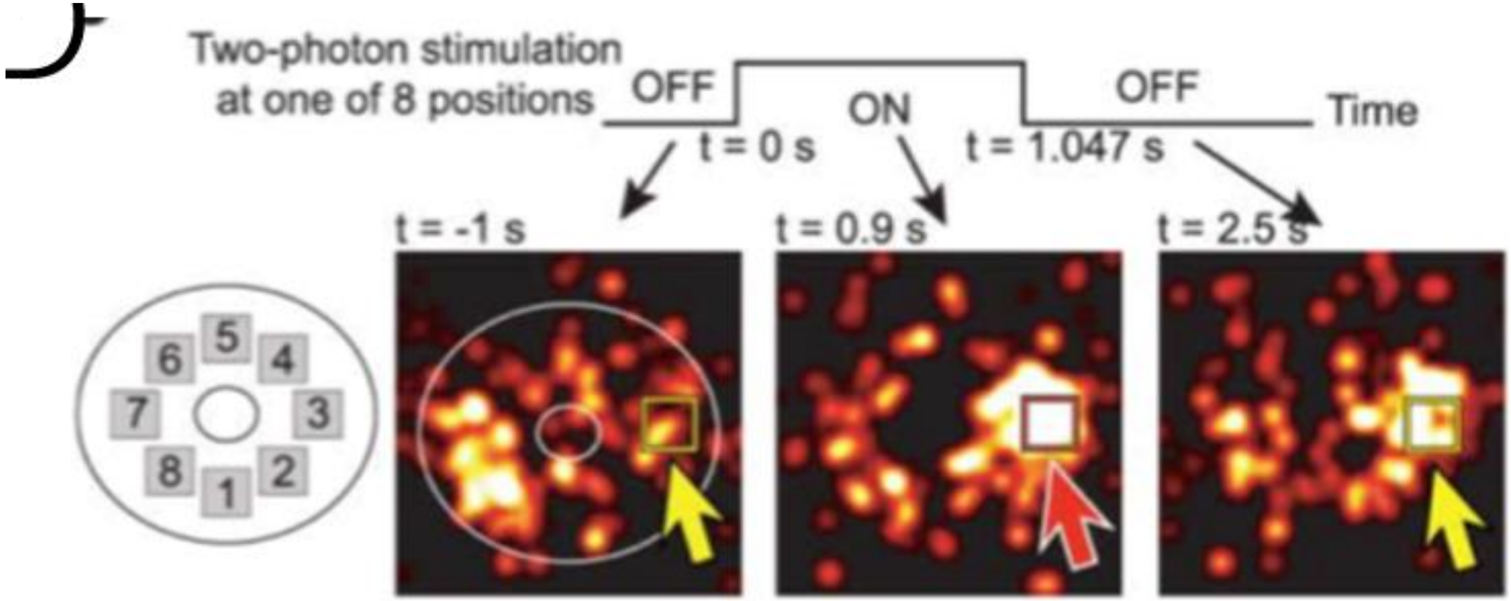
- Bell-shaped recurrent connection strength
- Translation-invariant connection pattern
- Global divisive normalization

Key Mathematic Properties:

- Recurrent positive-feedback generates attractor, retaining input information
- Divisive normalization avoids exploration
- Translation-invariance ensures many attractors

References: 1. Amari, 1977, 2. Ben-Yishai et al., 1995, 3. Zhang, 1996, 4. Seung, 1996,
5. Deneve et al, 1999, 6. Wu et al, 2002, 2005, 2008, 2010, 2012

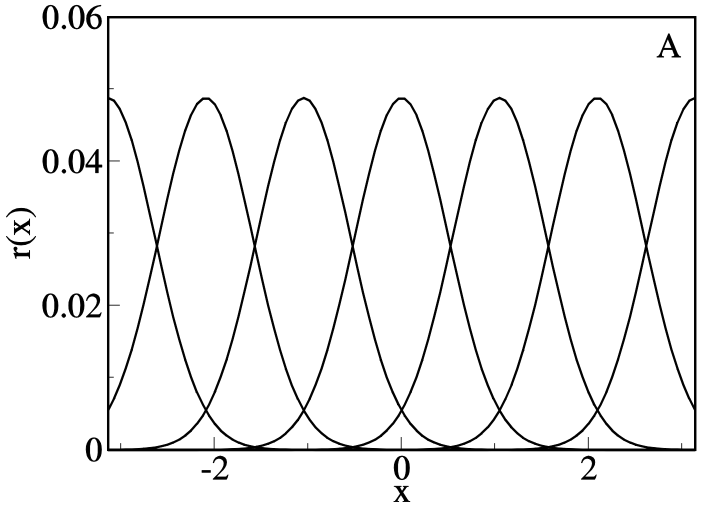
1D CANN for Head-direction in Fruit fly



A Continuous Family of Stationary States

$$\overline{U}(x\,|\,z) = \frac{A\rho J}{\sqrt{2}}\exp\left[-\frac{(x-z)^2}{4a^2}\right]$$

$$\bar{r}(x\,|\,z) = A\exp\left[-\frac{(x-z)^2}{4a^2}\right]$$



Stability Analysis （1）

Consider small fluctuations around a stationary state at z:

$$\delta U(x|z) = U(x|z) - \overline{U}(x|z)$$

$$\tau \frac{\partial \delta U(x|z)}{\partial t} = -\delta U(x|z) + \rho \int dx' J(x,x')\delta r(x'|z)$$

$$= -\delta U(x|z) + \int dx' F(x,x')\delta U(x')$$

Where

$$F(x,x') = \int dx'' \rho J(x,x'') \frac{\partial \overline{r}(x''|z)}{\partial \overline{U}(x'|z)}$$

Stability Analysis （2）

$$\tau \frac{\partial \boldsymbol{\delta}\mathbf{U}}{\partial t} = -(\mathbf{I} - \mathbf{F})\boldsymbol{\delta}\mathbf{U}, \qquad \boldsymbol{\delta}\mathbf{U} = \{\delta U(x|z)\},\ \text{for all } x$$

Projecting $\boldsymbol{\delta}\mathbf{U}$ on the *i*th right eigenvector of F

$$(\boldsymbol{\delta}\mathbf{U})_i(t) = (\boldsymbol{\delta}\mathbf{U})_i(0)e^{-(1-\lambda_i)t/\tau}$$

Two cases:

1. If $\lambda_i < 1,$ the projection decays exponentially;

2. If $\lambda_i = 1,$ the projection is sustained.

Spectra of the Kernel F

$$F(x,x'|z)=\frac{AJ^2\rho^2}{B\sqrt{\pi a}}e^{-(x-x')^2/2a^2}-\frac{kA^3\rho^5J^4}{\sqrt{3}B^2}e^{-(x-z)^2/4a^2}e^{-(x'-z)^2/4a^2}$$

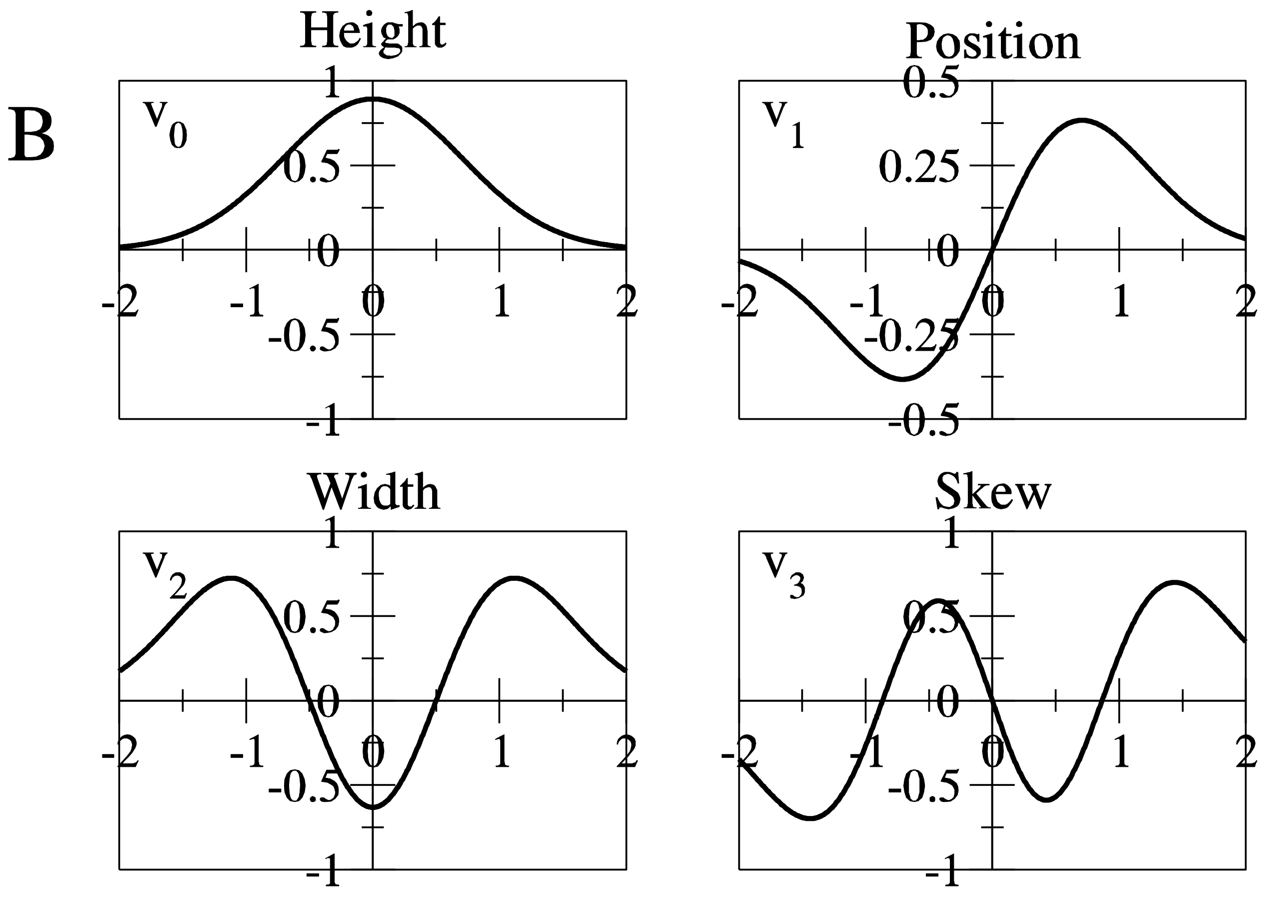
- $\lambda_0=1-2k\rho A\sqrt{2\pi a}<1,\quad \mathbf{u}_0(x|z)=\bar{\mathbf{U}}(x|z);$
- $\lambda_1=1,\quad \mathbf{u}_1(x|z)=\frac{d\bar{\mathbf{U}}(x|z)}{dz},$ the tangent of the valley
- $\lambda_n=\frac{1}{2^{n-2}},\quad \mathbf{u}_n(z)=\text{Combination of }\mathbf{v}_n(z)$

$\mathbf{v}_n(z)\sim e^{-(c-z)^2/4a^2}(\frac{d}{dc})^n e^{-(c-z)^2/2a^2},$ the wave functions of quntumn harmonic osscilator

Note the decay time constant is : $\frac{\tau}{1-\lambda_n}$

Physical meaning of basis functions

$$\mathbf{v}_n(z) \sim e^{-(c-z)^2/4a^2}\left(\frac{d}{dc}\right)^n e^{-(c-z)^2/2a^2}$$



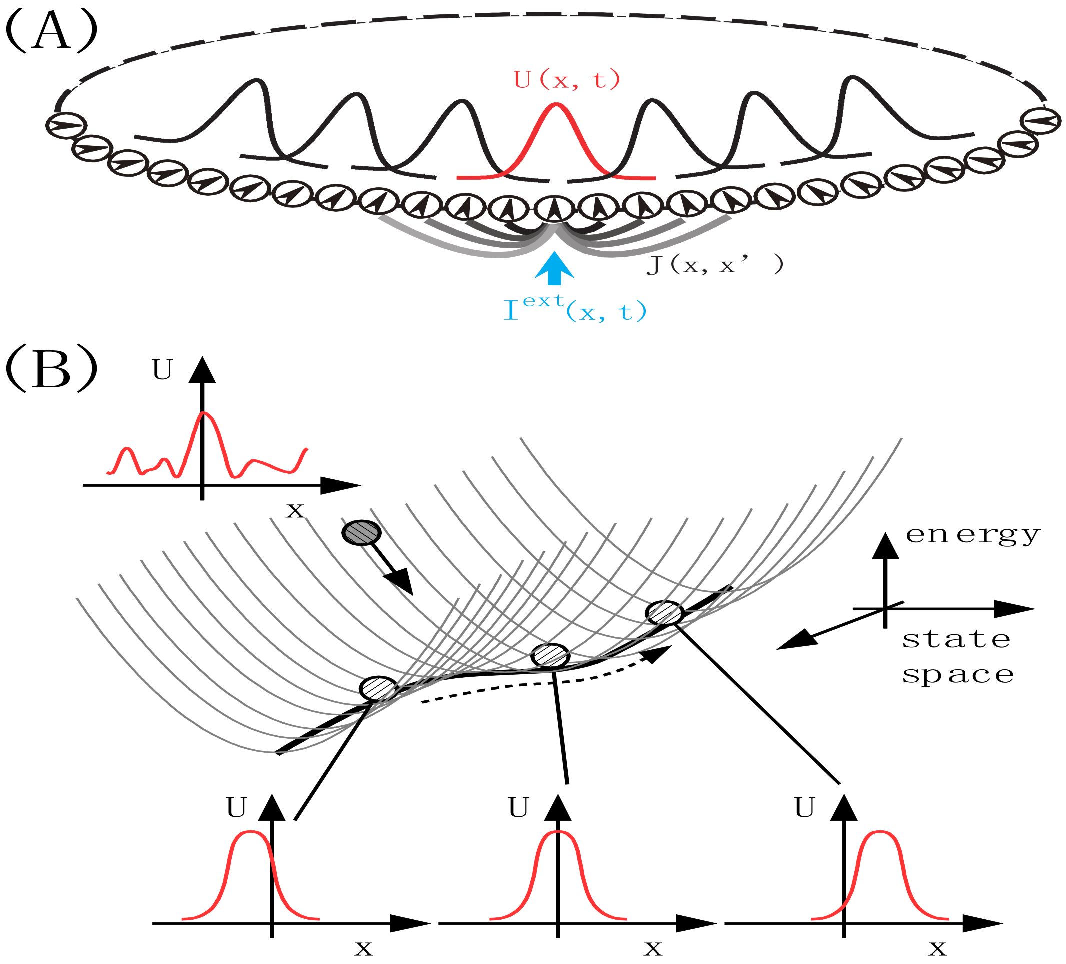
Height

Position

Width

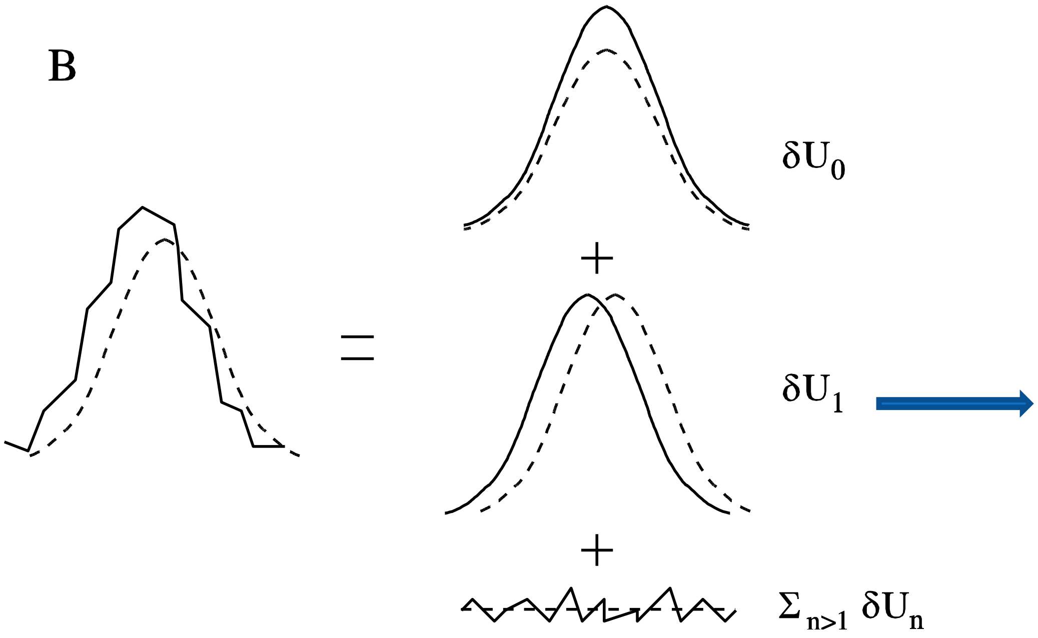
Skew

The landscape of CANN



1D CANN

Contributions of motion modes



Position shift,
the dominant one

Simplifying Network Dynamics: Projection Method

Consider

$$U(x,t)=\overline{U}(x\,|\,z(t))+\sum_{n=0}^{\infty}a_n(t)v_n(x\,|\,z(t))$$

The perturbative equation for $a_n(t)$

$$\begin{aligned}(\frac{d}{dt}+\frac{1-\lambda_n}{\tau})a_n &= \frac{I_n}{\tau}-\left[U_0\sqrt{(2\pi)^{1/2}a}\,\delta_{n1}+\sqrt{n}a_{n-1}-\sqrt{n+1}a_{n+1}\right]\frac{1}{2a}\frac{dz}{dt} \\ &\quad +\frac{1}{\tau}\sum_{r=1}^{\infty}\sqrt{\frac{(n+2r)!}{n!}}\frac{(-1)^r}{2^{n+3r-1}r!}a_{n+2r}\end{aligned} \qquad (1)$$

The peak position

$$\frac{dz}{dt}=\frac{2a}{\tau}\frac{I_1+\sum_{n=3,odd}^{\infty}\sqrt{\frac{n!!}{(n-1)!!}}I_n+a_1}{U_0\sqrt{(2\pi)^{1/2}a}+\sum_{n=0,even}^{\infty}\sqrt{\frac{(n-1)!!}{n!!}}a_n} \qquad (2)$$

Fung et al. Neural computation, 2010

1D Projection

Project the network dynamics on $\mathbf{v}_1(t)$

$$\tau \frac{\partial \mathbf{U} * \mathbf{v}_1}{\partial t} = -\mathbf{U} * \mathbf{v}_1 + (\mathbf{J} * \mathbf{r}) * \mathbf{v}_1 + \mathbf{I}^{ext} * \mathbf{v}_1$$

Consider

$$I^{ext}(t) = \alpha \bar{U}(x \mid z_0) + \sigma \xi_c(t)$$

$$\mathbf{U} * \mathbf{v}_1 \equiv \int_x dx U(x \mid z) v_1(x \mid z)$$

1D dynamics for position movement

$$\tau \frac{dz}{dt} = -\alpha(z - z_0)e^{-(z-z_0)^2/8a^2} + \beta\xi(t)$$

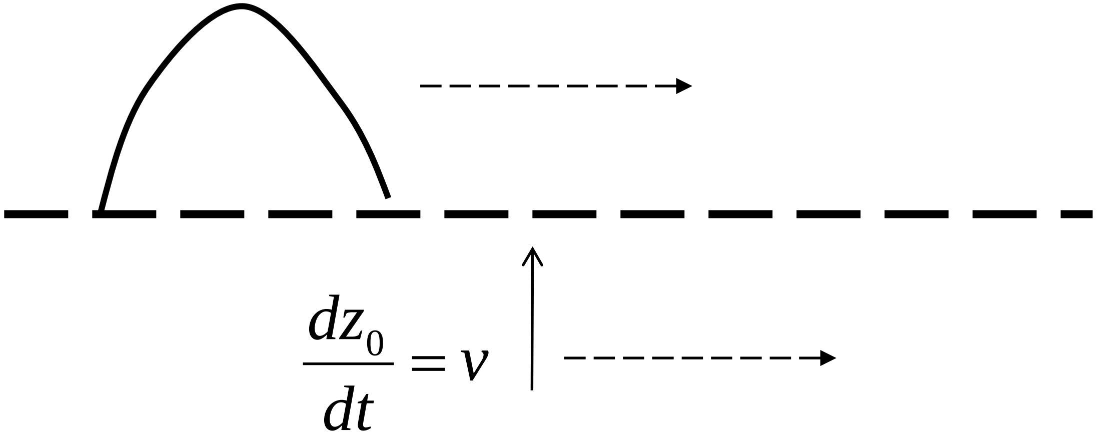
$1st$ term: the force of the signal that pulls the bump back to the stimulus position

$2nd$ term : random shift

Wu et al, Neural Computation, 2008

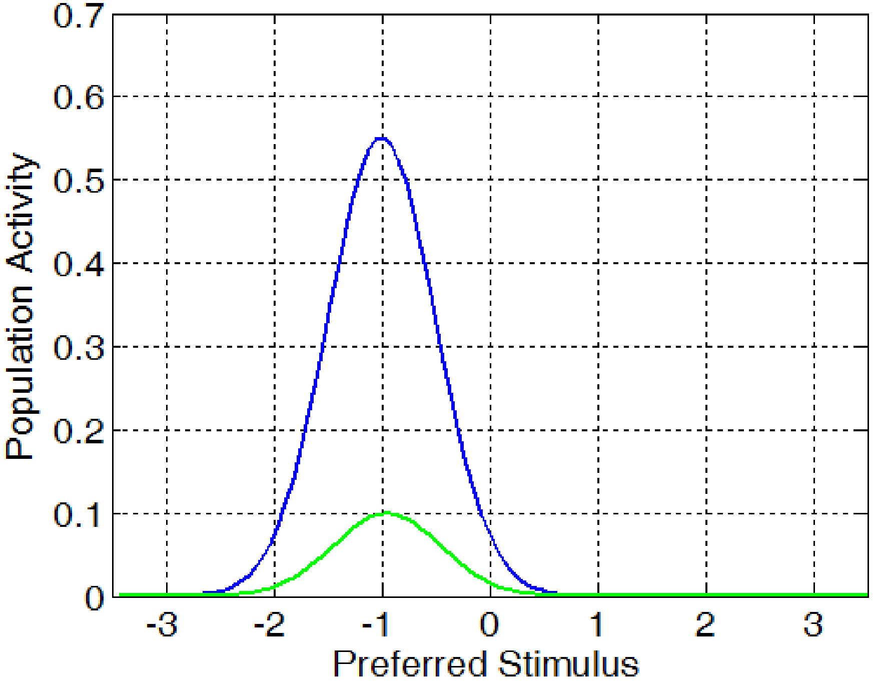
Tracking Behaviours (1)

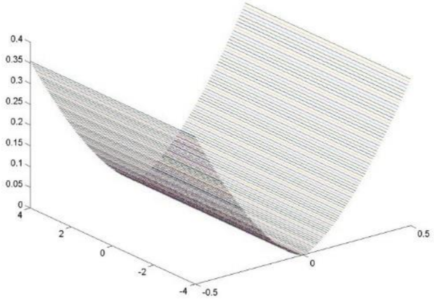
- Tracking a moving stimulus



$$\frac{dz_0}{dt} = v$$

Smooth Tracking by a CANN





Wu et al., Neural Computation 2005,2010

The tracking property

Define $s = z_0 - z$

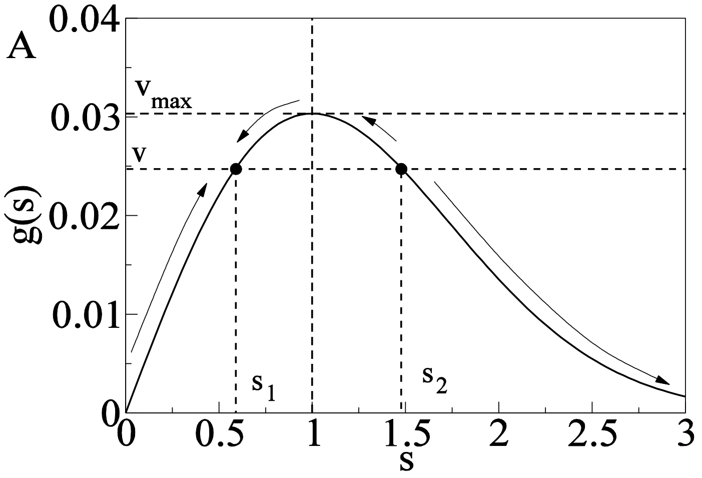
$$\frac{ds}{dt} = \frac{dz_0}{dt} - \frac{dz}{dt}$$

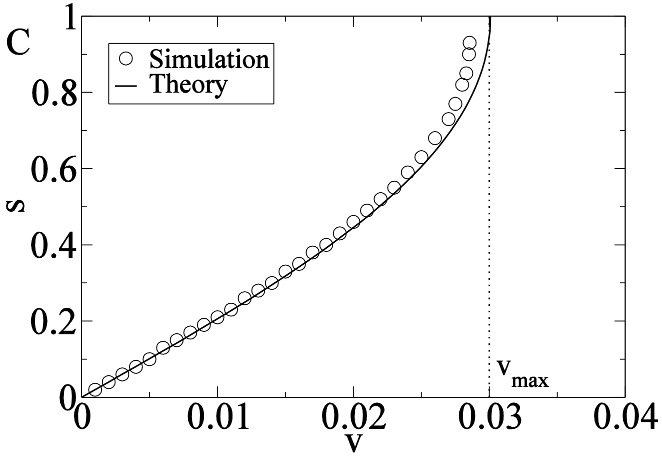
$$= v - \frac{\alpha s}{\tau} e^{-s^2/8a^2}$$

$$= v - g(s)$$

Condition for successful tracking :

$v = g(s)$

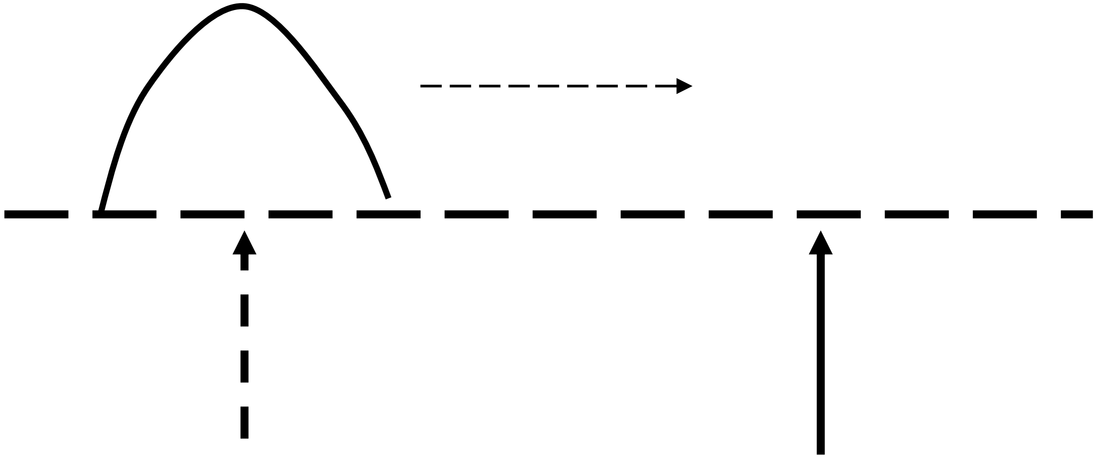




Tracking Behaviours (2)

- **Reaction time:**

How long it takes for the network to catch up an abrupt stimulus change.



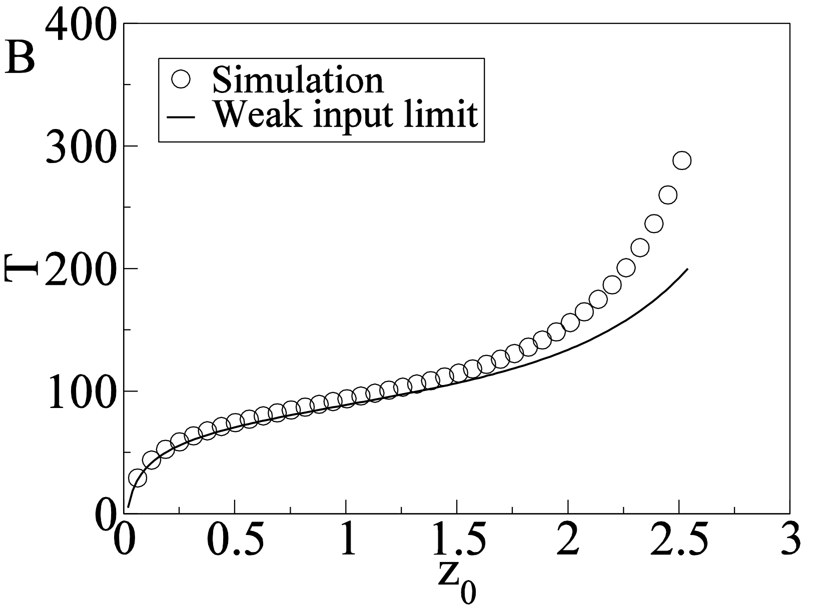
Logarithm Reaction Time (small rotation angle)

$$\tau \frac{dz}{dt} = -\alpha(z - z_0)e^{-(z-z_0)^2/8a^2} + \beta\xi(t)$$

When $|z - z_0| << a$, and noise small

$$\tau \frac{dz}{dt} = -\alpha(z - z_0)$$

$$T = \frac{\tau}{\alpha}\ln|z_0| + const$$



Logarithm reaction time of mental rotation?

- Backward Alignment

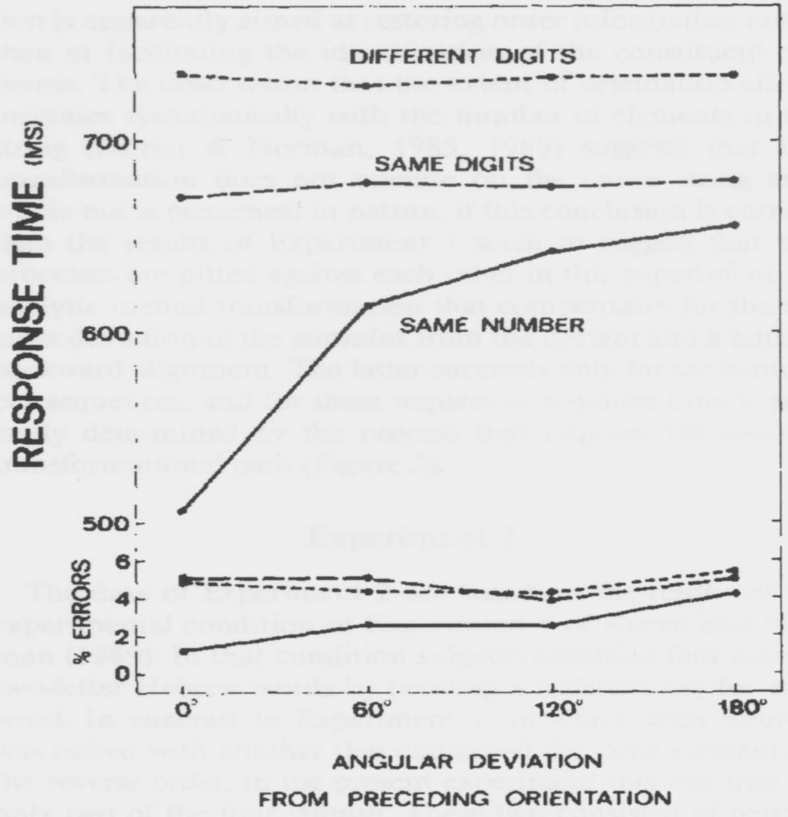
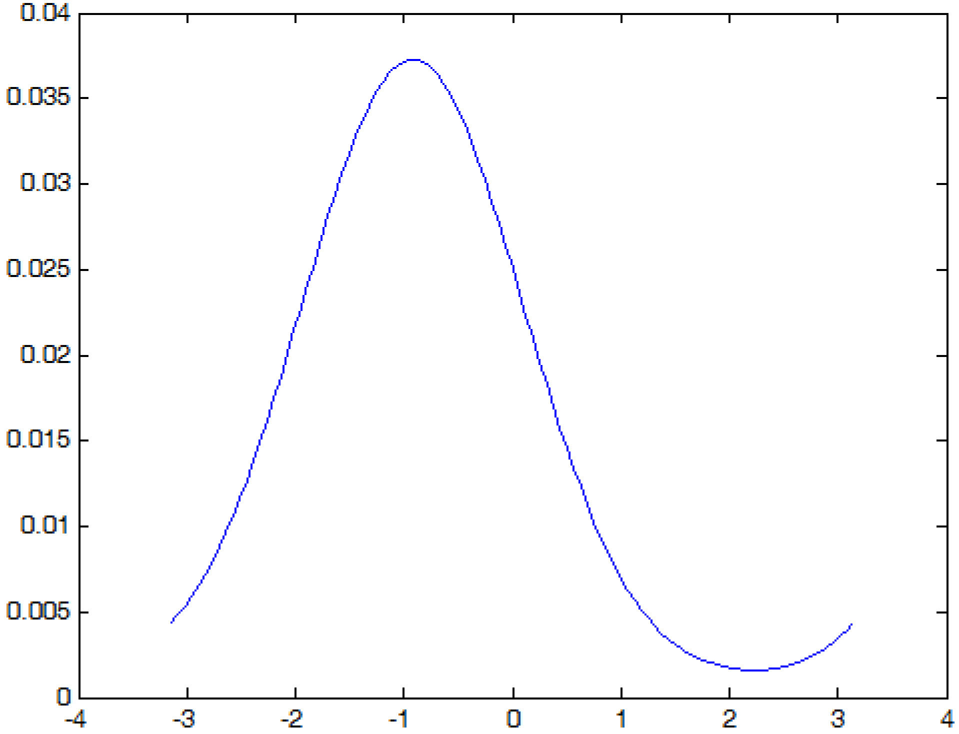
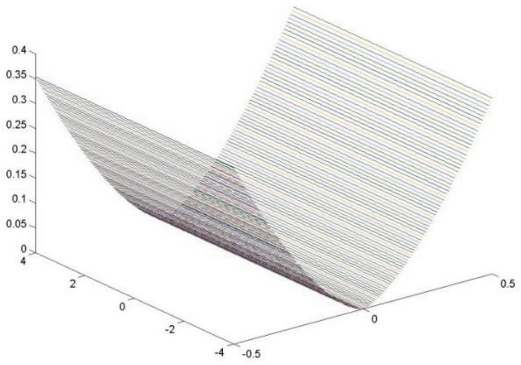


Figure 1. Response time and percentage of errors as a function of angular deviation from preceding orientation for same number, same-digits, and different-digits sequences (Experiment 1).

A. Koriat & J. Norman (1989)
J. Experimental Psychology

Efficient population decoding via a CANN





A CANN achieves template-matching, a statistical efficient decoding strategy

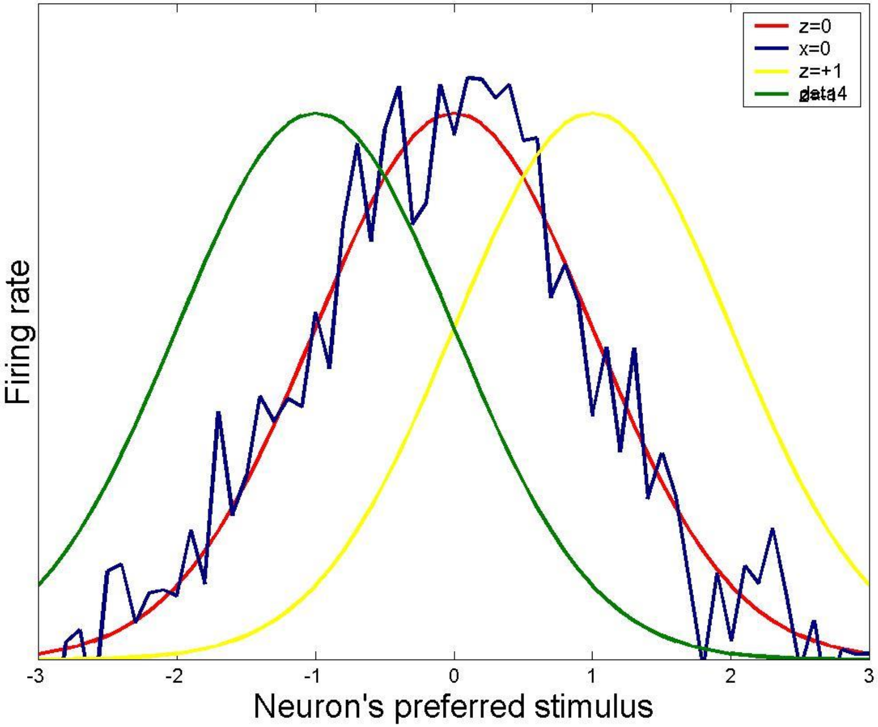
Deneve et al. 1999
Wu at al. 2002

Template Matching

$$\tau_s \frac{\partial U(x,t)}{\partial t} = -U(x,t) + \rho \int dx' J(x,x') r(x',t) + \varepsilon I^{ext}(x,t)$$

For small inputs, the final position

$$\hat{z} = \max_z \int dx \bar{U}(x \mid z) I^{ext}(x)$$



Performance of template-matching

- MLI for independent Gaussian noise

For independent Gaussian noises,

$$p(\mathbf{r} \mid z) = \prod_i p(r_i \mid z) \propto \prod_i \exp[-(r_i - f_i(z))^2 / 2\sigma^2]$$

Thus

$$\begin{aligned} \hat{x} &= \max_z \log p(\mathbf{r} \mid z) \\ &= \max_z \sum_i -[r_i - f_i(z)]^2 \\ &= \max_z \sum_i r_i f_i(z) \xrightarrow{\qquad\qquad} \text{Template-matching} \end{aligned}$$

To get the last equality, we have used the condition

$\sum_i f_i(z) \approx$ constant, which is true when the number of neuron is large

Sequential Bayesian Decoding

The Decoding Procedure

$$\textit{Step}\ 1:\quad \hat{x}_t \quad (\textit{When}\ t=1,\ \textit{Maximum Likelihood})$$

$$\textit{Step}\ 2:\quad \text{Gaussian Prior}:\ P(x) = e^{-(x-\hat{x}_t)^2/2\tau_t^2} \Big/ (\sqrt{2\pi}\ \tau_t)$$

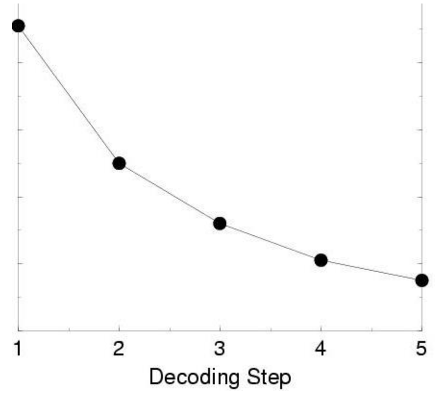
$$\textit{Step}\ 3:\quad \hat{x}_{t+1} \quad (\textit{Maximum a Posterior})$$

$$\textit{Step}\ 4:\quad \text{Repeat}\ \ \textit{Step}\ 2$$

The optimal result

$$\boldsymbol{\Omega_t^2 = \frac{1}{t}\Omega_1^2}$$

$$\boldsymbol{@\ \tau_t^2 = \frac{1}{t}\cdot\frac{1}{-\nabla\nabla \ln P(\mathbf{r}\mid x)}}$$

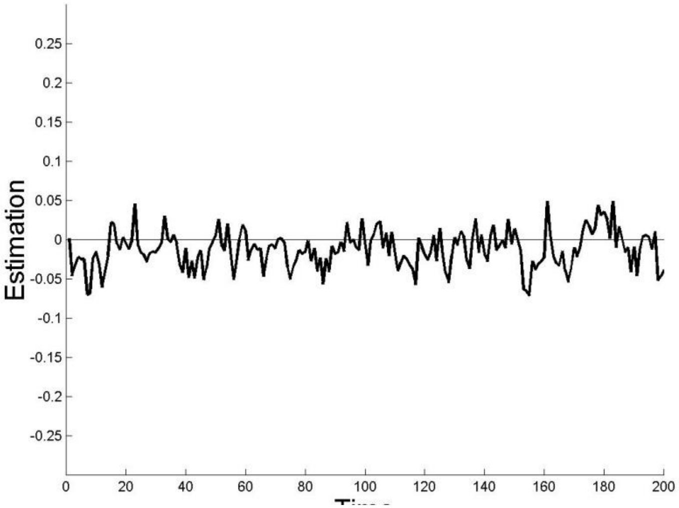


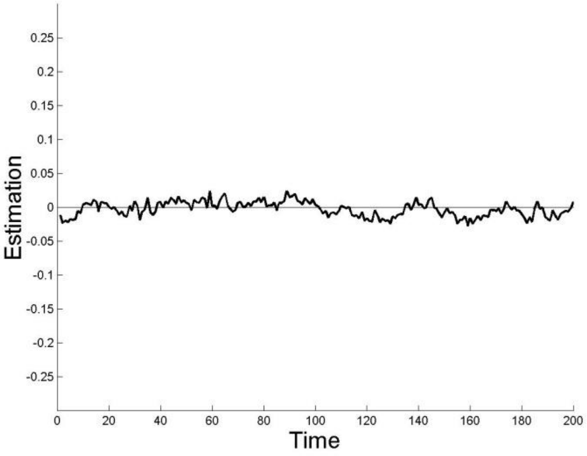
Wu et al., NIPS*2001

Fast Hebbian learning improves decoding

$$J(x,x',t) = J_0(x,x') + W(x,x',t)$$

$$\tau \frac{dW(x,x',t)}{dt} = -W(x,x',t) + \lambda r(x) r(x')$$





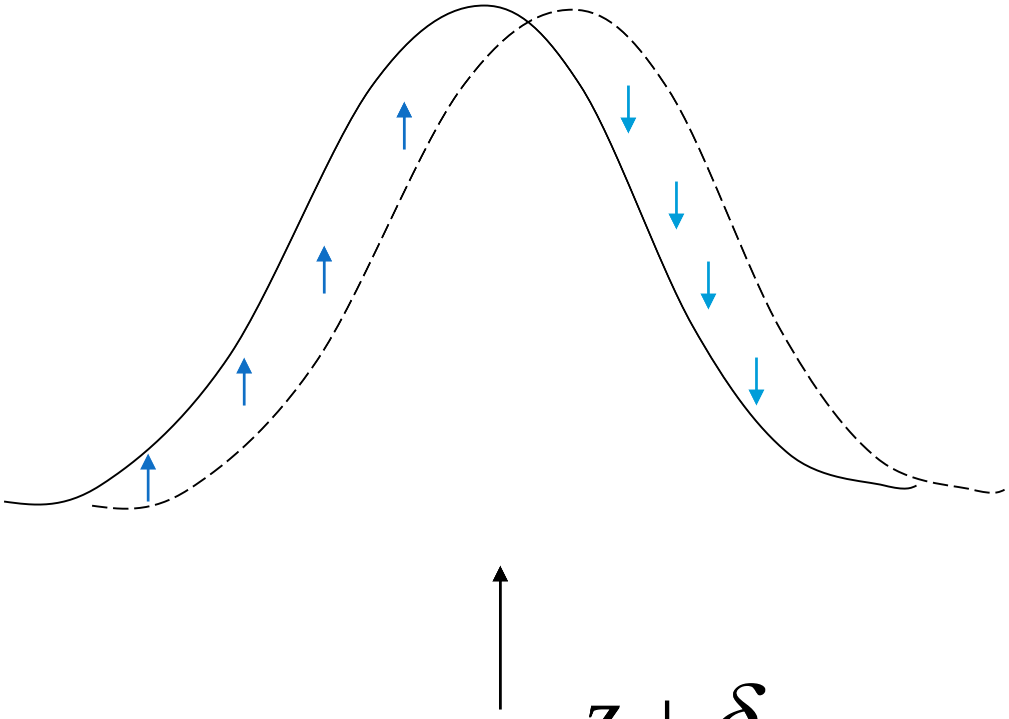
Wu et al, Neural Computation, 2005

The asymmetric correlation

The correlation between neural response variability

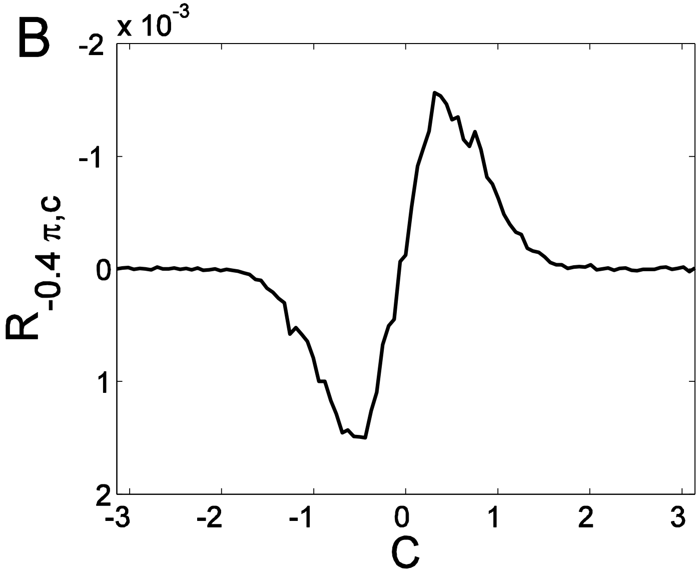
$$R(x,y) = \left\langle (r(x) - <r(x)>)(r(y) - <r(y)>) \right\rangle$$

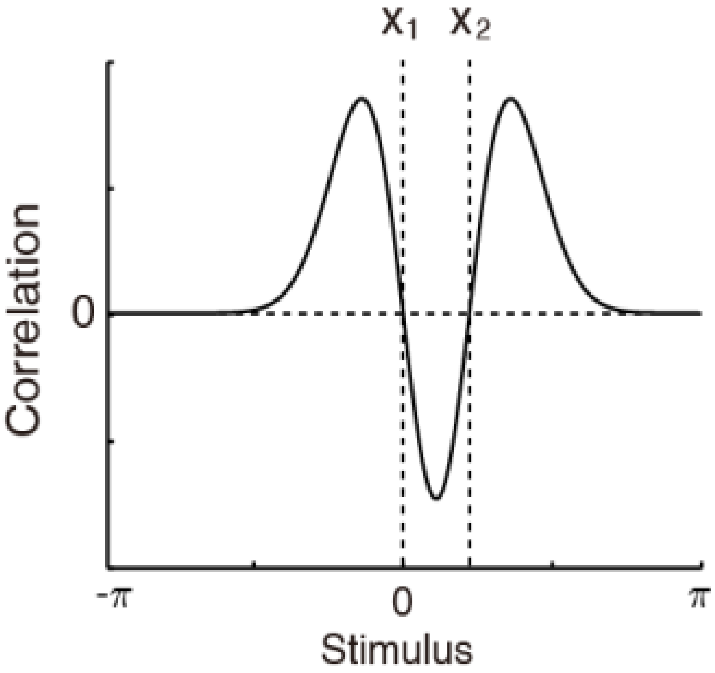
$$R(x,y) \propto (x-z)(y-z)e^{-(x-z)^2/2a^2} e^{-(y-z)^2/2a^2}$$



$z+\delta$

Neural Signature of a CANN

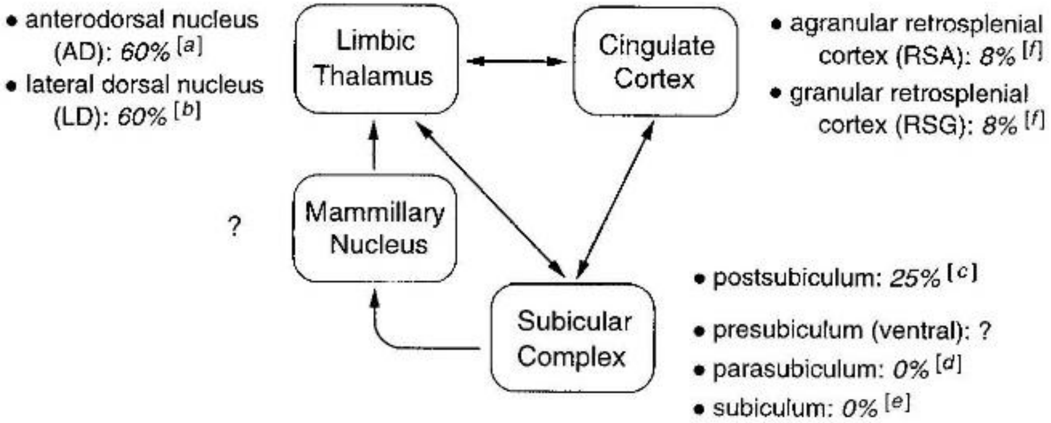




Wu et al, Neural Computation 2008
Ponce-Alvarez et al., PNAS 2013
Wimmer et al., Nature Neuroscience, 2014

Head-direction neurons

Head-direction neurons in the limbic system



- anterodorsal nucleus (AD): *60%* [a]
- lateral dorsal nucleus (LD): *60%* [b]

Limbic Thalamus

Cingulate Cortex

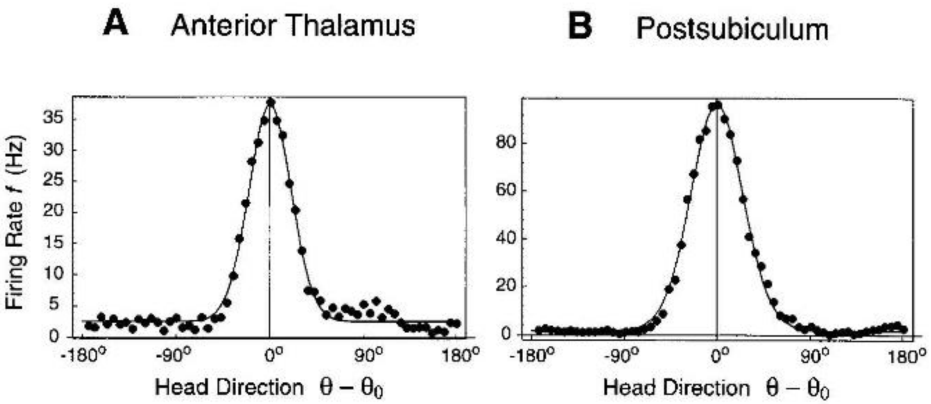
- agranular retrosplenial cortex (RSA): *8%* [f]
- granular retrosplenial cortex (RSG): *8%* [f]

?

Mammillary Nucleus

Subicular Complex

- postsubiculum: *25%* [c]
- presubiculum (ventral): ?
- parasubiculum: *0%* [d]
- subiculum: *0%* [e]



A CANN for Head-direction Representation

$$\tau \frac{\partial U(x,t)}{\partial t} = -U(x,t) + \rho \int dx' J(x-x',t) r(x',t)$$

$r(x,t) = F\left[U(x,t)\right],\ F:$ a sigmoid function

$J(x-x',t) = W(x-x') + v(t)\tau\nabla W(x-x')$

$W(x-x'):$ symmetric; $\nabla W(x-x')$: asymmetric

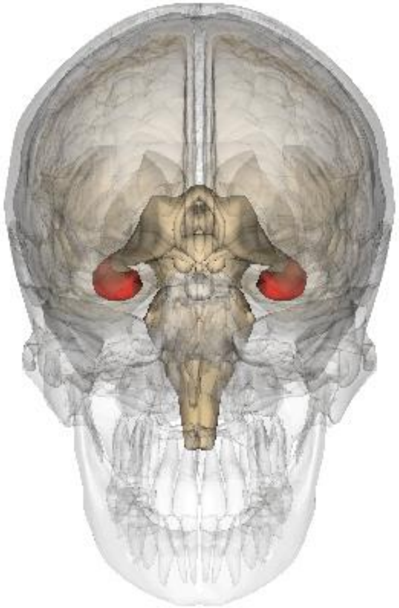
$v(t)$: the rotating speed

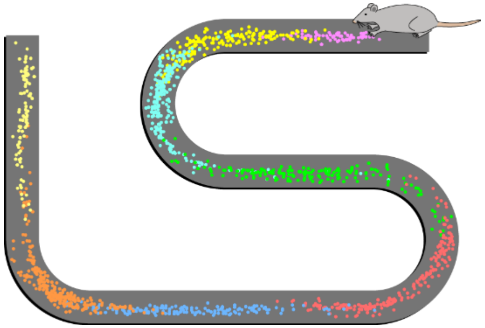
Suppose $\bar{\mathrm{U}}(x\,|\,z)$ is the static solution when $v(t)=0$,

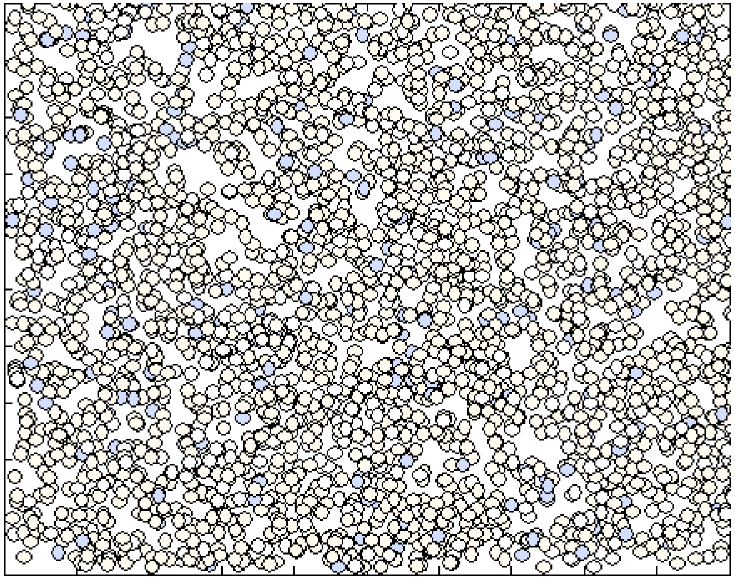
Then $\bar{\mathrm{U}}(x\,|\,z(t))$ with $z(t) = z_0 + \int_0^t v dt$ is the moving solution

Kechen Zhang, Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory. J. Neurosci. 16:2112-2126, 1996

2D CANN for Spatial Navigation

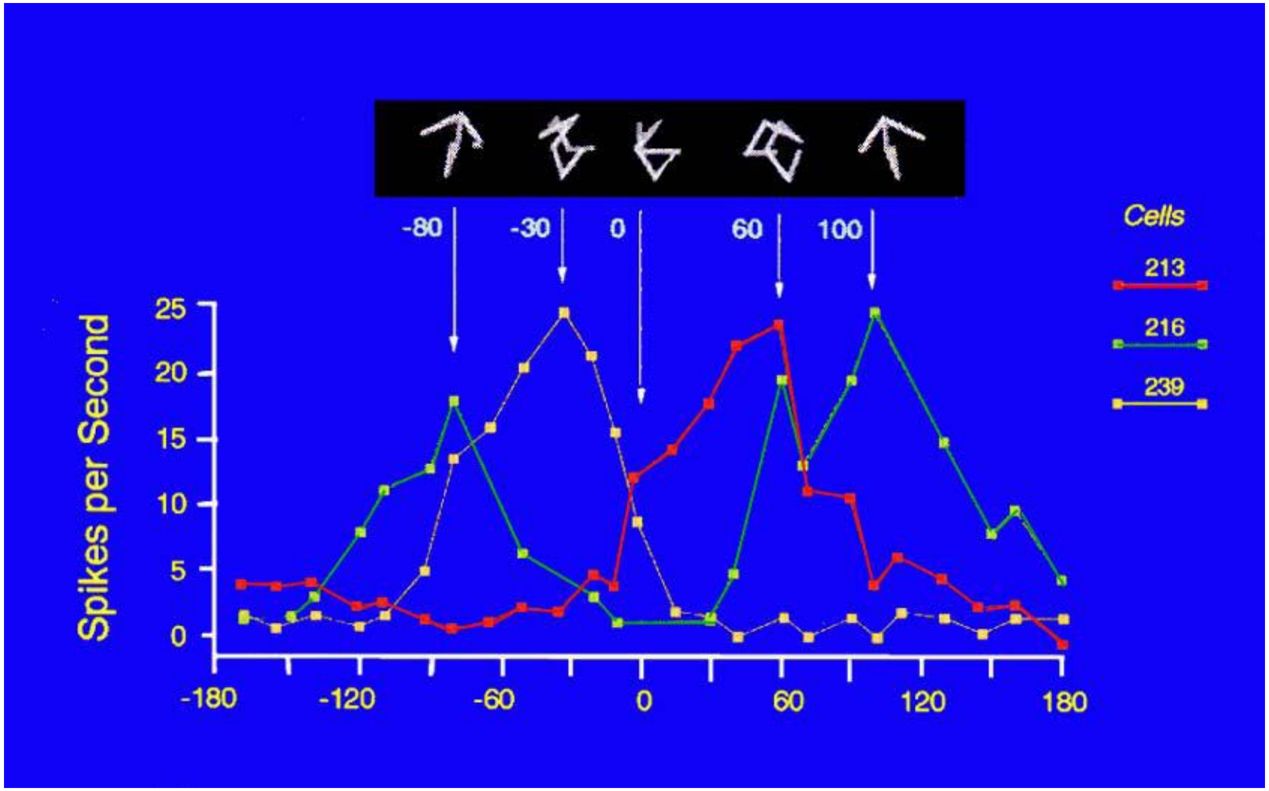






Samnovich, CANN, Scholarpedia

CANN for a General Feature

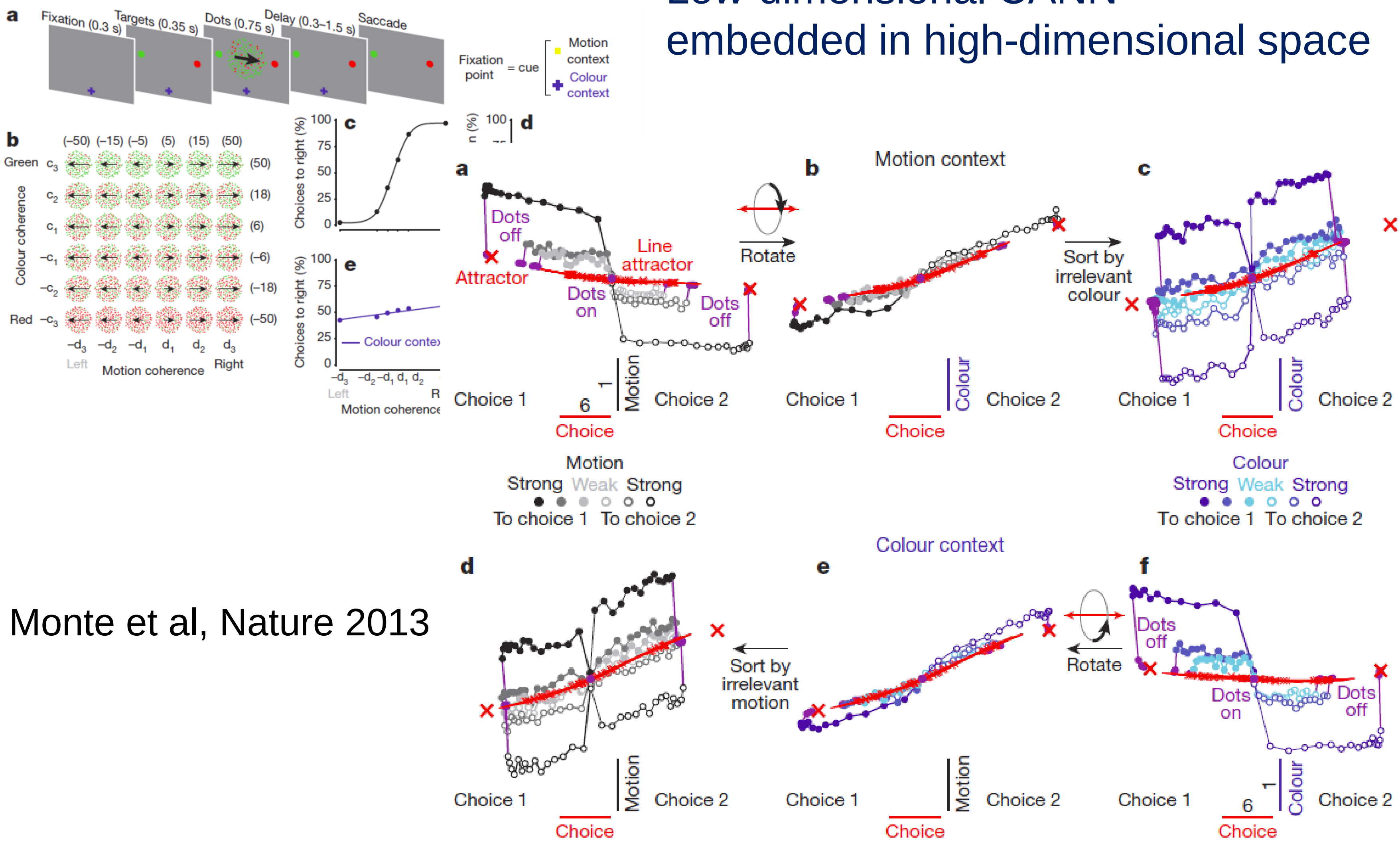


View-based object representation

Logothetis, Pauls & Poggio, 1995

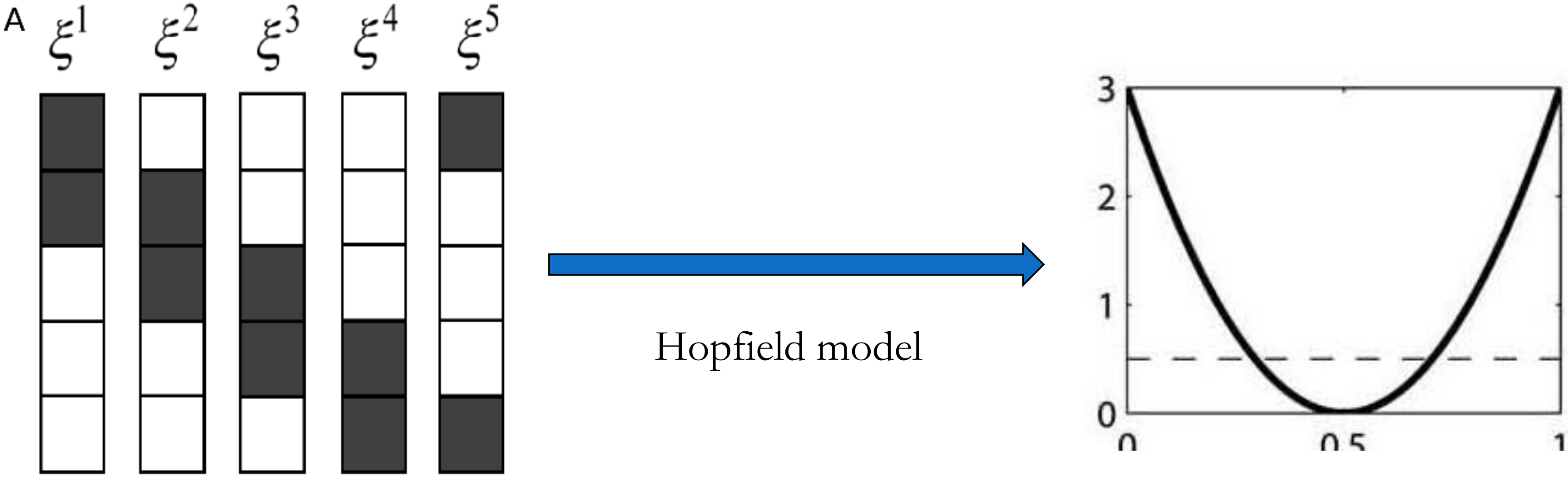
Beyond Simple Features

Low-dimensional CANN embedded in high-dimensional space



Monte et al, Nature 2013

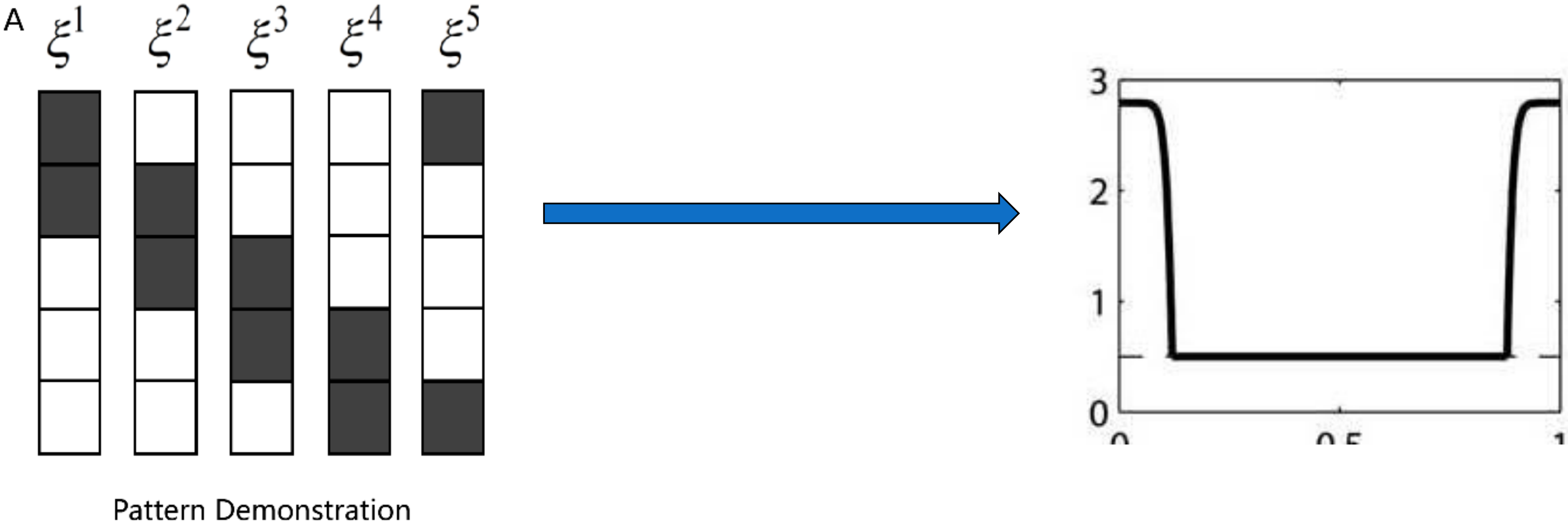
Hopfield model for correlated patterns



Pattern Demonstration

$$W_{ij} = \frac{1}{N}\sum_{\mu} \xi_i^{\mu}\xi_j^{\mu}$$

Novelty-facilitated Hebb learning



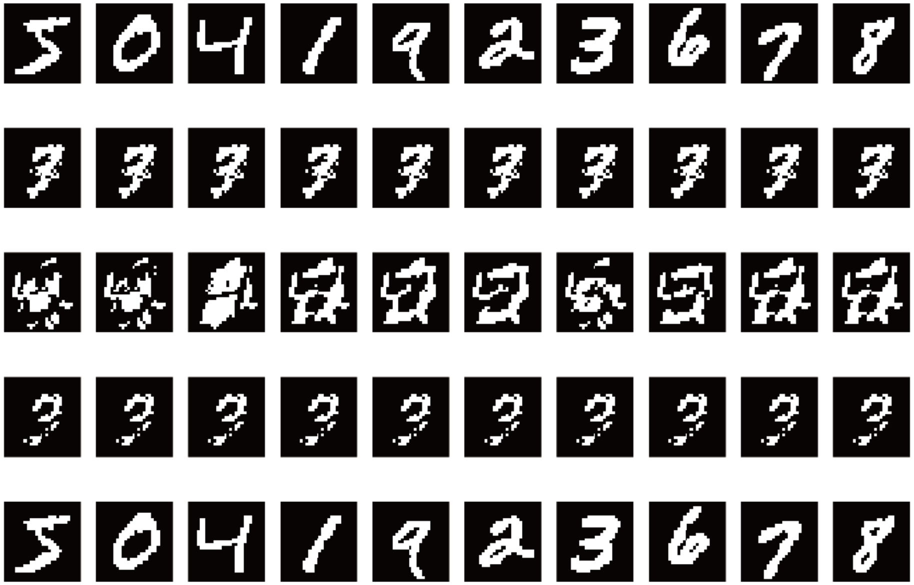
$$W_{ij} = \frac{1}{N} \sum_{\mu} w_{\mu} \xi_i^{\mu} \xi_j^{\mu}$$

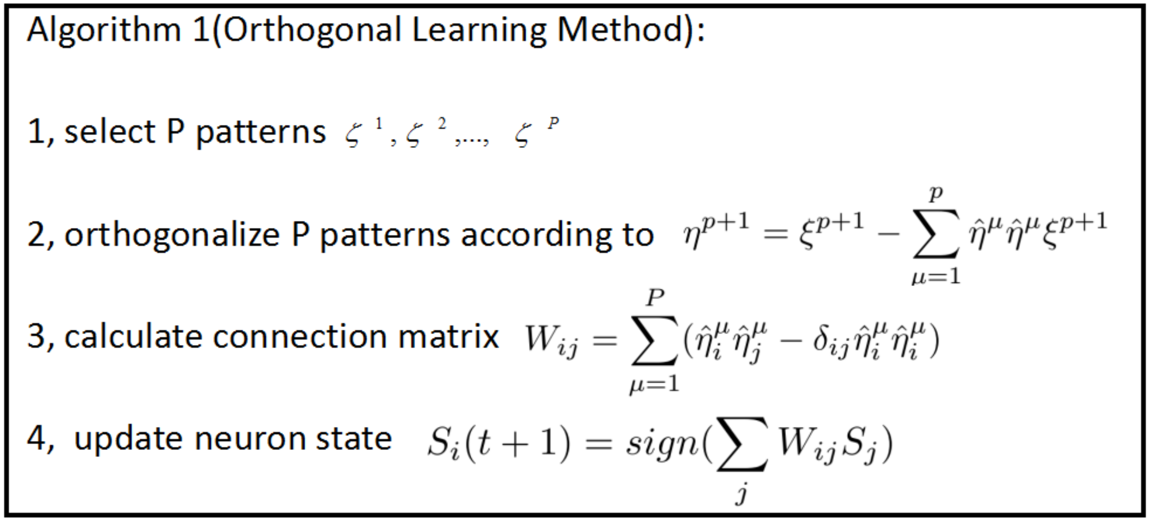
$$w_{\mu} \to w_{\mu} + \eta H$$

H: the Hamming distance between the input pattern and the memorized one

Blumenfeld et al. Dynamics of memory representations in networks with novelty-facilitated synaptic plasticity. Neuron 52: 383-394, 2006

An orthogonal learning method



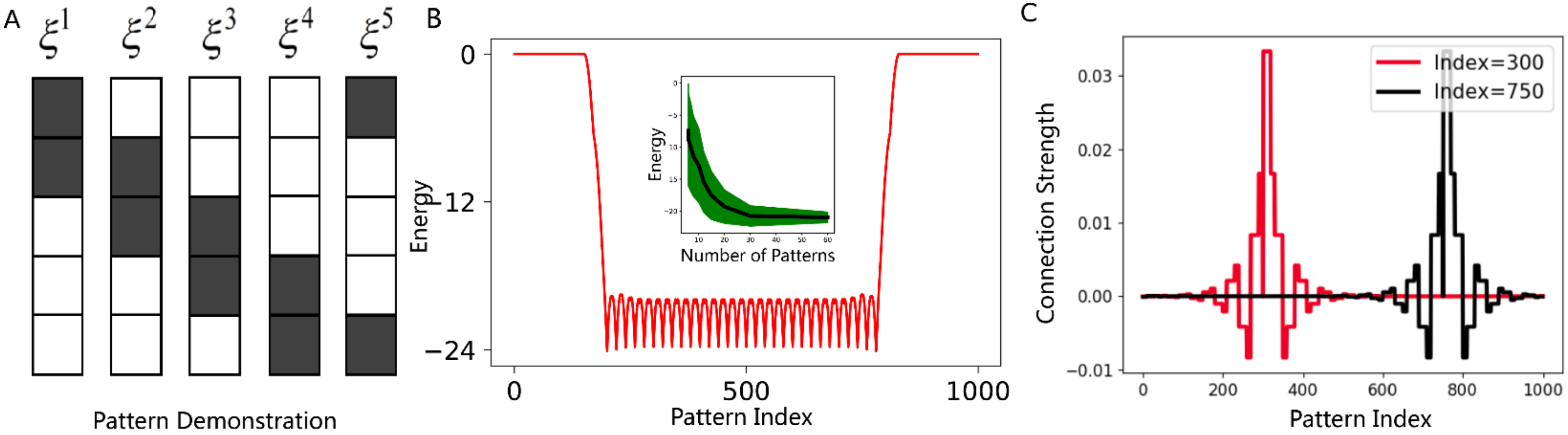


Algorithm 1(Orthogonal Learning Method):

```
1, select P patterns $\zeta^{1}, \zeta^{2}, ..., \zeta^{P}$
2, orthogonalize P patterns according to $\eta^{p+1} = \xi^{p+1} - \sum_{\mu=1}^{p} \hat{\eta}^{\mu} \hat{\eta}^{\mu} \xi^{p+1}$
3, calculate connection matrix $W_{ij} = \sum_{\mu=1}^{P} (\hat{\eta}_i^{\mu} \hat{\eta}_j^{\mu} - \delta_{ij} \hat{\eta}_i^{\mu} \hat{\eta}_i^{\mu})$
4, update neuron state $S_i(t+1) = sign(\sum_{j} W_{ij} S_j)$
```

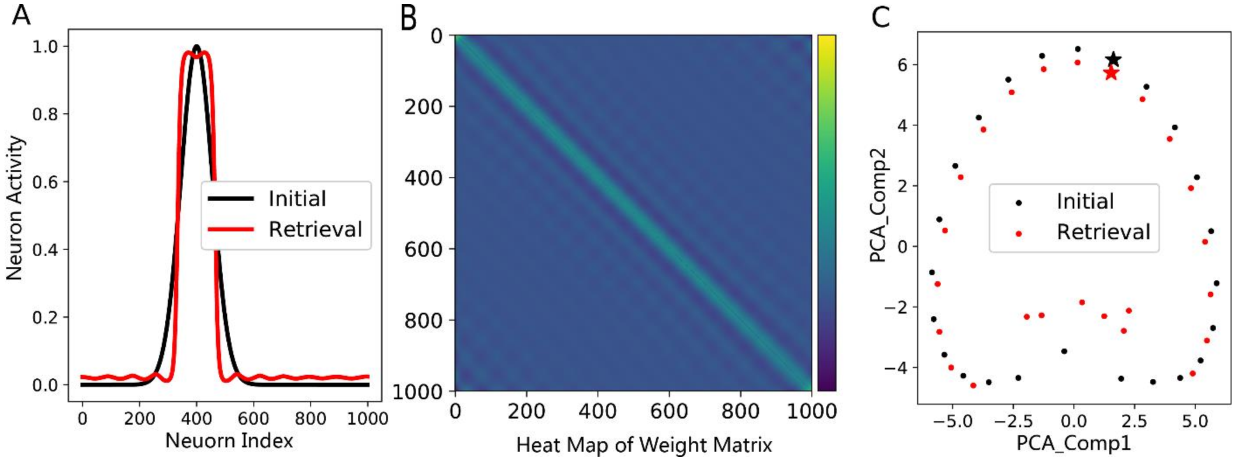
Two key operations: orthogonalizing (pattern segregation, Dental Gyrus) + novel detection (CA1).

Learning a CANN from Continuous Morphed Patterns (discrete dynamics)

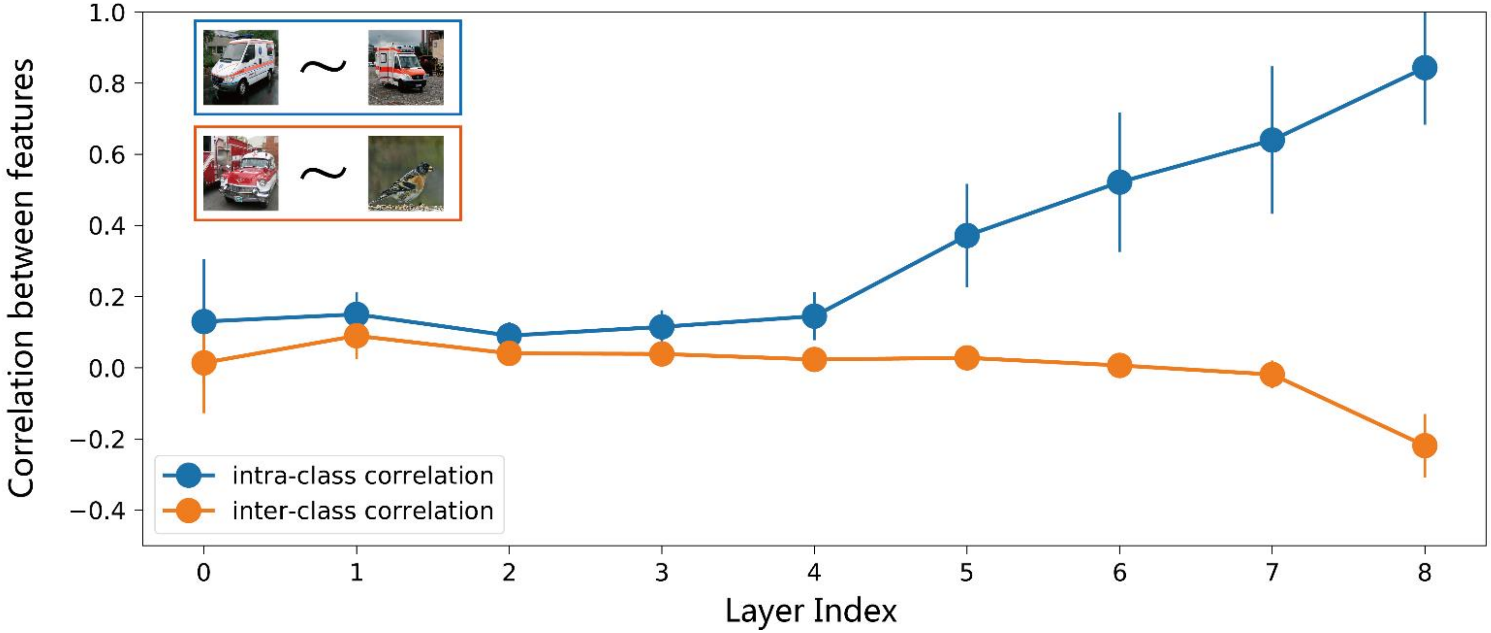


Learning a CANN from Continuous Morphed Patterns (continuous dynamics)

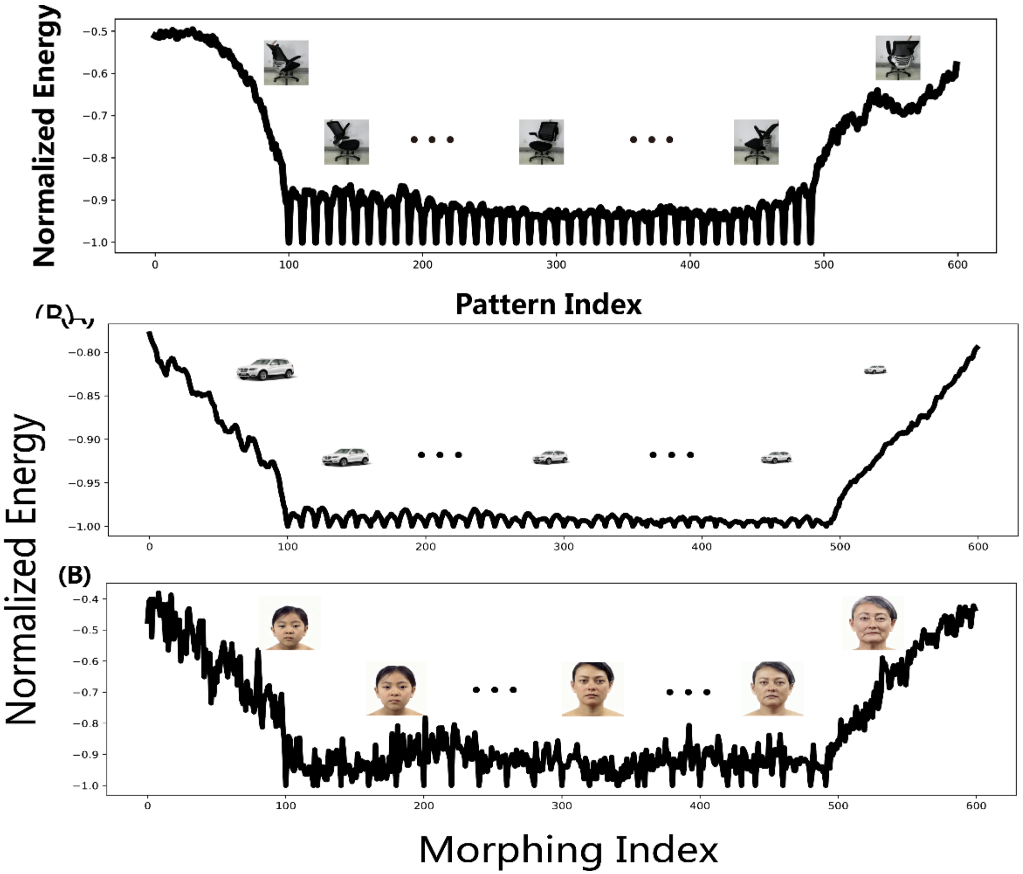
$$\tau \frac{dV_i}{dt} = -V_i + \sum_j W_{ij} g(V_i) + I_i$$



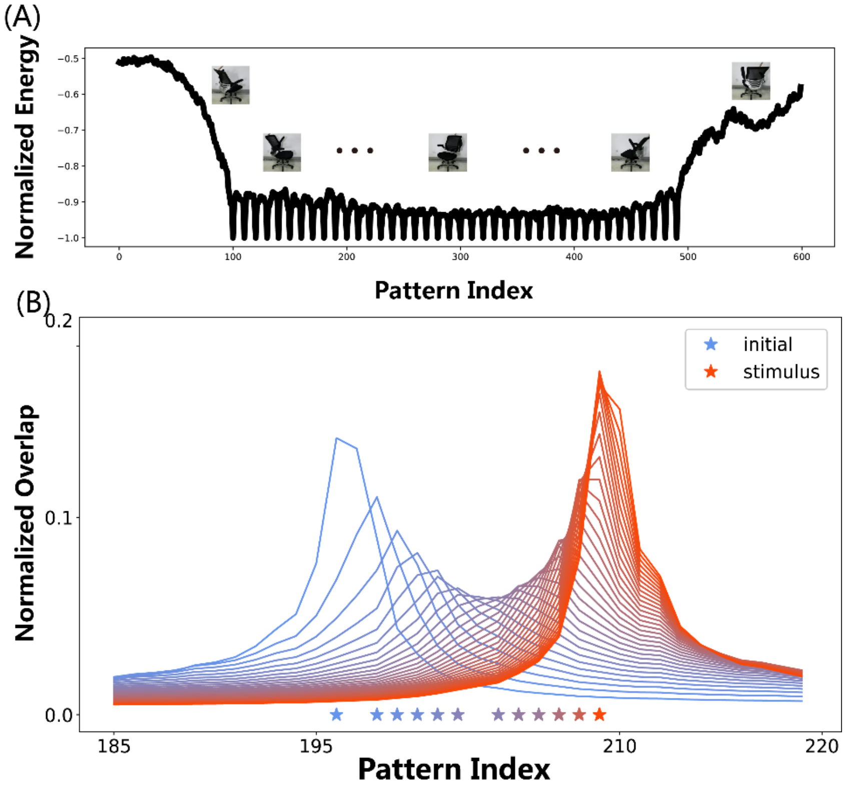
Neural correlation in the sematic sense



Learning CANNs from Images Linked by a Continuous Feature



Learning a CANN from continuously rotating chairs (Mental Rotation)



A CANN encodes similarity between objects

- Categorization of objects is based on the similarity between objects in sematic sense.

- The similarity between objects is encoded by the overlap/correlation between neural representations of objects; via synaptic plasticity, correlated neural representations are encoded in CANNs.

- Current AI mainly focuses on classification.

- The implications of the dynamics of CANNs?

References

1. **Wu, S.**, Wong, KYM., Fung, CCA., Mi, Y., and Zhang, W. (2016). Continuous attractor neural networks: candidate of a canonical model for neural information representation. **F1000 Invited Review**, 66(16), 209-226.
2. C. C.Fung, K.Y.Michael Wong and **S. Wu** (2010). A Moving Bump in a Continuous Manifold: A Comprehensive Study of the Tracking Dynamics of Continuous Attractor Neural Networks. **Neural Computation**, v.22, p.752-792.
3. S. Wu (2007). Behaviour Signatures of Continuous Attractors. International Conference on Cognitive Neurodynamics (ICCN'07).
4. **S. Wu** and S. Amari (2005). Computing with Continuous Attractors: Stability and On-Line Aspects. **Neural Computation**, v.17, 2215-2239.
5. **S. Wu**, S. Amari and H. Nakahara. (2002). Population Coding and Decoding in a Neural Field: A Computational Study. **Neural Computation,** v14, no.5, p.999-1026.