Robust performance of inhomogeneous forgetful associative memory networks

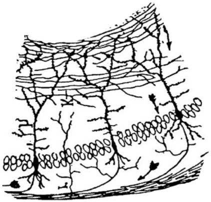
David Sterratt and David Willshaw

Institute for Adaptive & Neural Computation
School of Informatics
University of Edinburgh

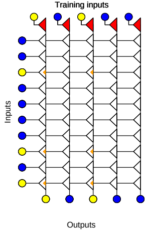
3rd October 2007

Motivation

Theoretical networks of *simple* neurons show how parts of the CNS with mutable synapses could store and retrieve memories

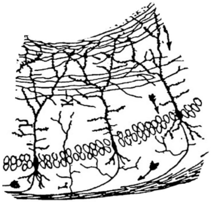


Cajal

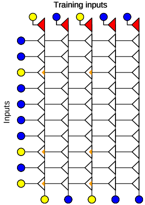


Motivation

Theoretical networks of *simple* neurons show how parts of the CNS with mutable synapses could store and retrieve memories



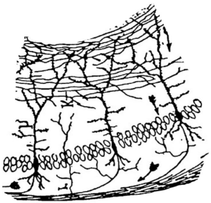
Cajal



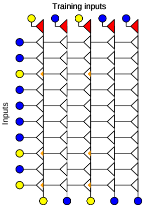
But. . .

Motivation

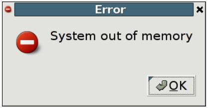
Theoretical networks of *simple* neurons show how parts of the CNS with mutable synapses could store and retrieve memories



Cajal

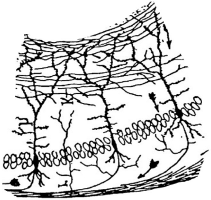


But...

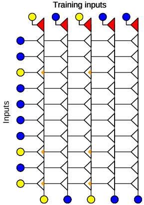


Motivation

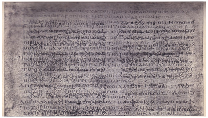
Theoretical networks of *simple* neurons show how parts of the CNS with mutable synapses could store and retrieve memories



Cajal



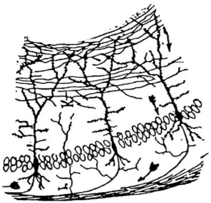
But...



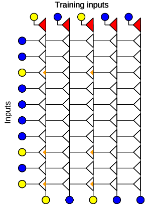
Memories need to be forgotten to make space for new ones: *Palimpsests*

Motivation

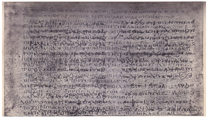
Theoretical networks of *simple* neurons show how parts of the CNS with mutable synapses could store and retrieve memories



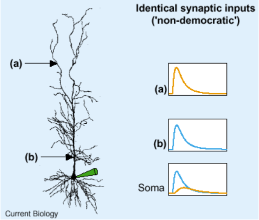
Cajal



But...



Memories need to be forgotten to make space for new ones: *Palimpsests*

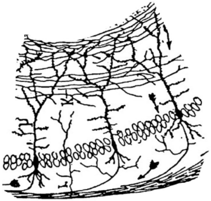


[Häusser, 2001]

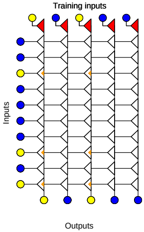
Real neurons are not homogeneous: *differential attenuation*

Motivation

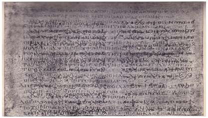
Theoretical networks of *simple* neurons show how parts of the CNS with mutable synapses could store and retrieve memories



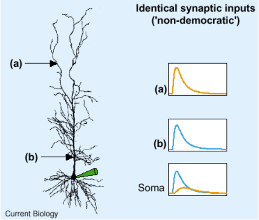
Cajal



But...

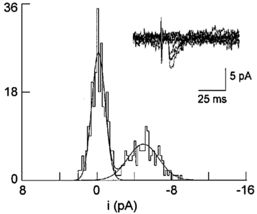


Memories need to be forgotten to make space for new ones: *Palimpsests*



[Häusser, 2001]

Real neurons are not homogeneous: *differential attenuation*



[Bolshakov et al., 1997]

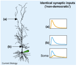
Real synapses are not deterministic: *stochastic transmission*

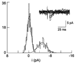
Overview

- We incorporate three “inhomogeneities” into Dayan & Willshaw’s theory of associative nets with general learning rules

Overview

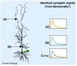
- We incorporate three “inhomogeneities” into Dayan & Willshaw’s theory of associative nets with general learning rules
    1. Forgetting
    2. Differential attenuation
    3. Stochastic transmission

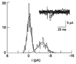




Overview

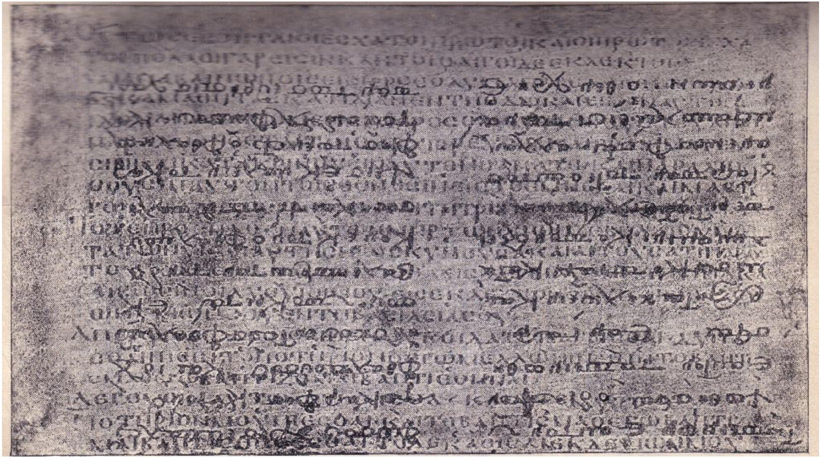
- We incorporate three “inhomogeneities” into Dayan & Willshaw’s theory of associative nets with general learning rules
  1. Forgetting
  2. Differential attenuation
  3. Stochastic transmission





- Highlights
  - There is an optimal rate at which to forget
  - Optimal network capacity scales with size
  - Differential attenuation predicted to affect performance less than stochastic transmission

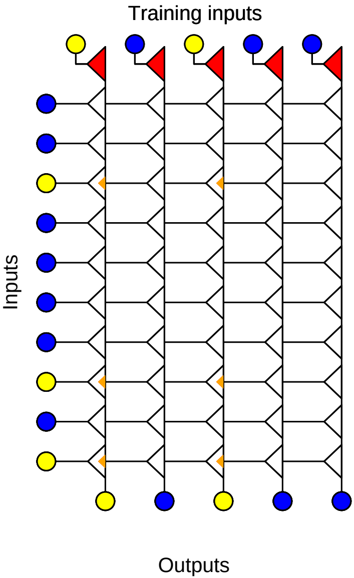
Inhomogeneity 1: Forgetting



Codex Ephraemi Rescriptus (5th century and 12th century)

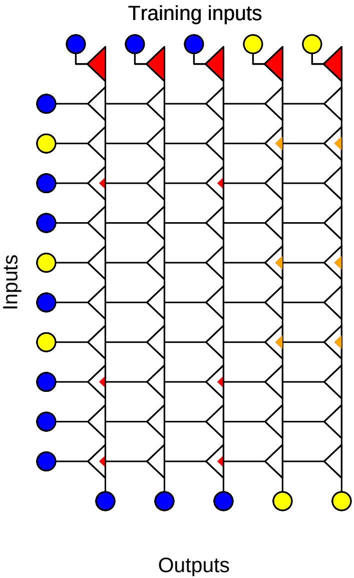
- Newer memories have a stronger trace in the network

Learning and forgetting memories



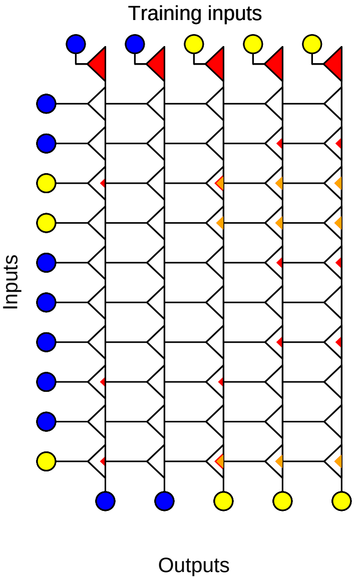
- Random binary-valued patterns
  - Input sparsity p
  - Output sparsity r
- $w_{ij}(t) = e^{-1/\tau} w_{ij}(t-1) + \Delta_{ij}(t)$
- Learning rule, e.g. Hebbian
- τ is *forgetting time constant*
  - $\tau \to \infty$: “rememberful” memory
  - Here $\tau = 3.5$

Learning and forgetting memories



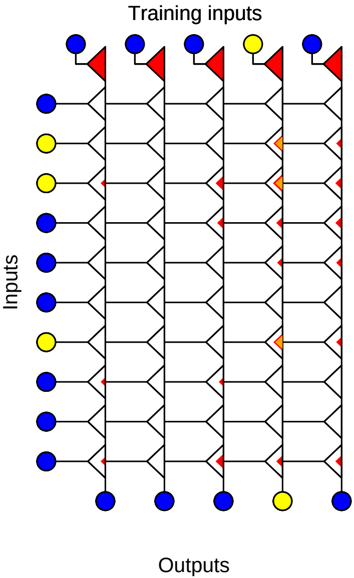
- Random binary-valued patterns
  - Input sparsity p
  - Output sparsity r
- $w_{ij}(t) = e^{-1/\tau} w_{ij}(t-1) + \Delta_{ij}(t)$
- Learning rule, e.g. Hebbian
- τ is *forgetting time constant*
  - $\tau \to \infty$: “rememberful” memory
  - Here $\tau = 3.5$

Learning and forgetting memories



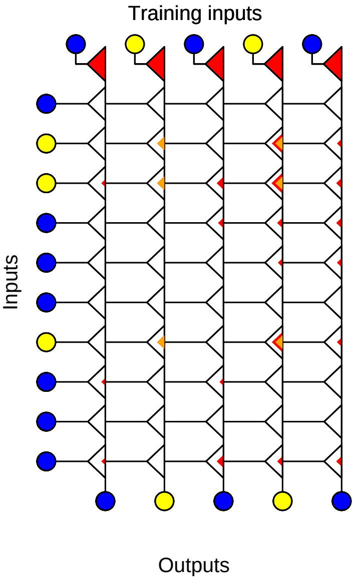
- Random binary-valued patterns
  - Input sparsity p
  - Output sparsity r
- $w_{ij}(t) = e^{-1/\tau} w_{ij}(t-1) + \Delta_{ij}(t)$
- Learning rule, e.g. Hebbian
- τ is *forgetting time constant*
  - $\tau \to \infty$: "rememberful" memory
  - Here $\tau = 3.5$

Learning and forgetting memories



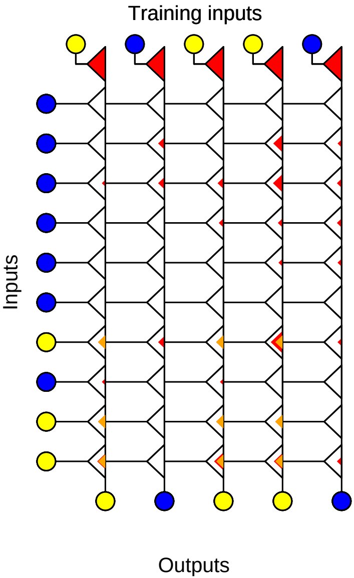
- Random binary-valued patterns
  - Input sparsity p
  - Output sparsity r
- $w_{ij}(t) = e^{-1/\tau} w_{ij}(t-1) + \Delta_{ij}(t)$
- Learning rule, e.g. Hebbian
- τ is *forgetting time constant*
  - $\tau \to \infty$: "rememberful" memory
  - Here $\tau = 3.5$

Learning and forgetting memories



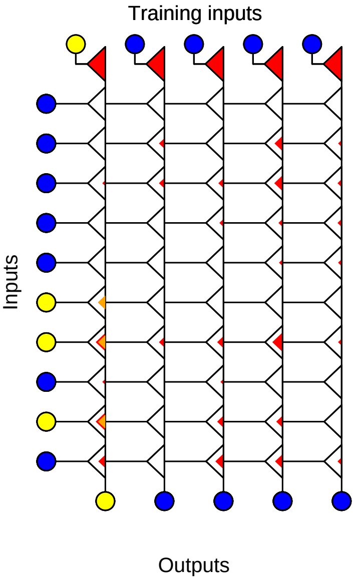
- Random binary-valued patterns
  - Input sparsity p
  - Output sparsity r
- $w_{ij}(t) = e^{-1/\tau} w_{ij}(t-1) + \Delta_{ij}(t)$
- Learning rule, e.g. Hebbian
- τ is *forgetting time constant*
  - $\tau \to \infty$: "rememberful" memory
  - Here $\tau = 3.5$

Learning and forgetting memories



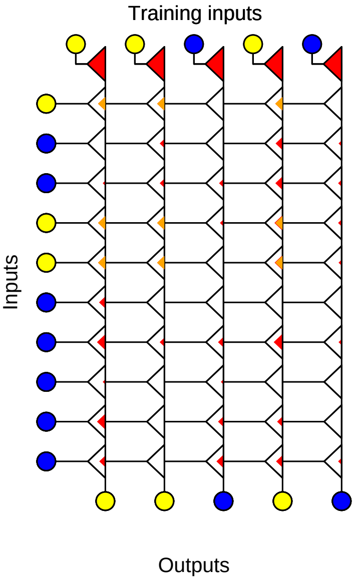
- Random binary-valued patterns
  - Input sparsity p
  - Output sparsity r
- $w_{ij}(t) = e^{-1/\tau} w_{ij}(t-1) + \Delta_{ij}(t)$
- Learning rule, e.g. Hebbian
- τ is *forgetting time constant*
  - $\tau \to \infty$: “rememberful” memory
  - Here $\tau = 3.5$

Learning and forgetting memories



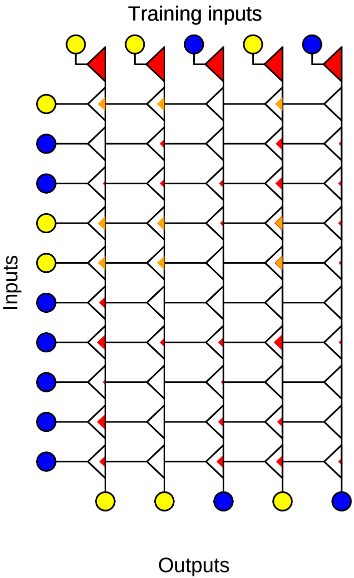
- Random binary-valued patterns
  - Input sparsity p
  - Output sparsity r
- $w_{ij}(t) = e^{-1/\tau} w_{ij}(t-1) + \Delta_{ij}(t)$
- Learning rule, e.g. Hebbian
- τ is *forgetting time constant*
  - $\tau \to \infty$: "rememberful" memory
  - Here $\tau = 3.5$

Learning and forgetting memories



- Random binary-valued patterns
  - Input sparsity p
  - Output sparsity r
- $w_{ij}(t) = e^{-1/\tau} w_{ij}(t-1) + \Delta_{ij}(t)$
- Learning rule, e.g. Hebbian
- τ is *forgetting time constant*
  - $\tau \to \infty$: "rememberful" memory
  - Here $\tau = 3.5$

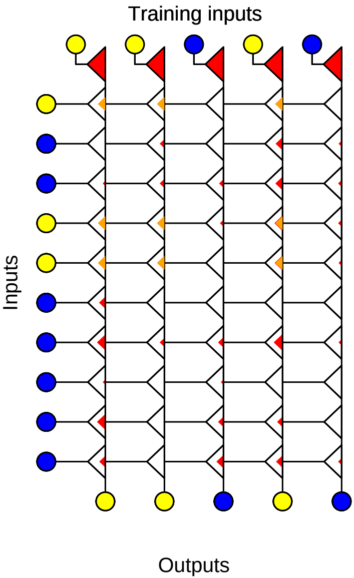
General learning rules



- General learning rule

		Post	
	$\Delta_{ij}(t)$	0	1
Pre	0	α	β
	1	γ	δ

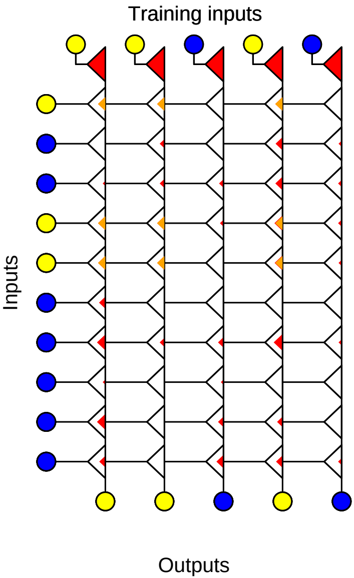
General learning rules



- e.g. Hebbian

Pre	$\Delta_{ij}(t)$	Post 0	Post 1
	0	0	0
	1	0	1

General learning rules



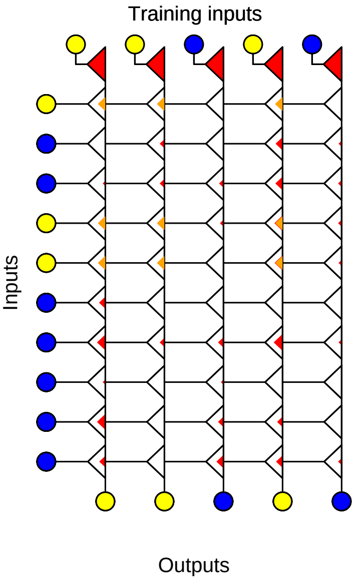
- e.g. Hebbian

$\Delta_{ij}(t)$		Post 0	Post 1
Pre	0	0	0
Pre	1	0	1

- e.g. Heterosynaptic

$\Delta_{ij}(t)$		Post 0	Post 1
Pre	0	0	-p
Pre	1	0	1-p

General learning rules



- e.g. Hebbian

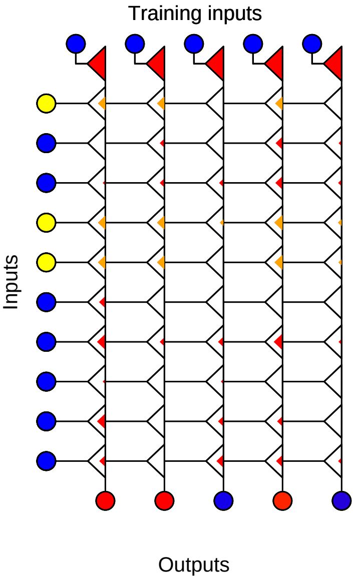
	$\Delta_{ij}(t)$	Post 0	Post 1
Pre	0	0	0
Pre	1	0	1

- e.g. Heterosynaptic

	$\Delta_{ij}(t)$	Post 0	Post 1
Pre	0	0	-p
Pre	1	0	1-p

- A rule is *balanced* if the expected weight is zero
  - e.g. the Heterosynaptic rule

Recalling memories

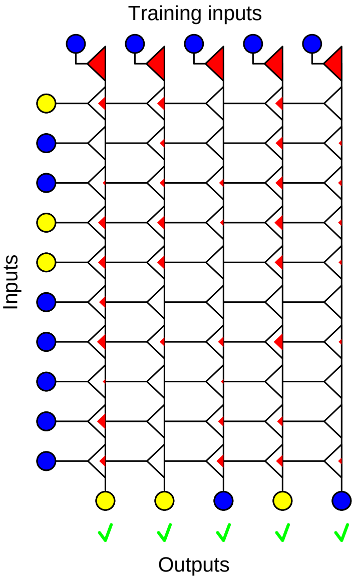


- Dendritic sum (like membrane potential at soma)

$$d_j = \sum_i w_{ij} a_i$$

Most recent memory

Recalling memories



Most recent memory

- Dendritic sum (like membrane potential at soma)

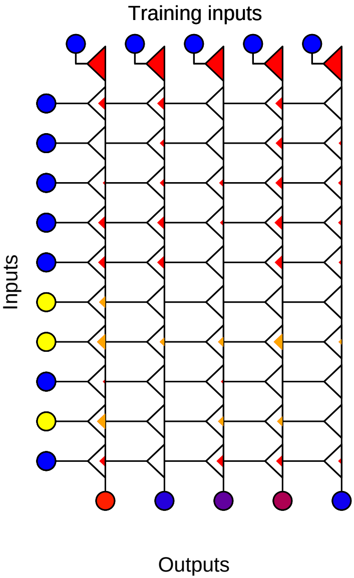
$$d_j = \sum_i w_{ij} a_i$$

- Output (like spike)

$$o_j = \begin{cases} 0 & \text{if } d_j < \theta_j \\ 1 & \text{if } d_j > \theta_j \end{cases}$$

where θ_j is the *per-unit* optimal threshold [Dayan and Willshaw, 1991]

Recalling memories



Older memory

- Dendritic sum (like membrane potential at soma)

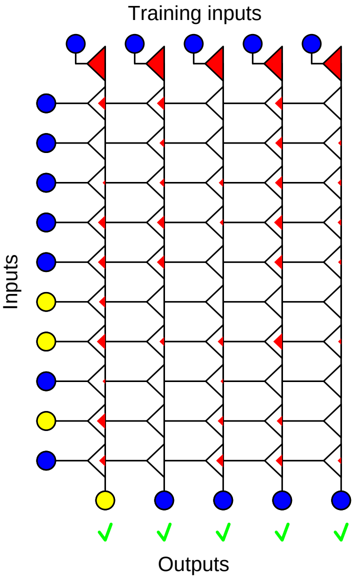
$$d_j = \sum_i w_{ij} a_i$$

- Output (like spike)

$$o_j = \begin{cases} 0 & \text{if } d_j < \theta_j \\ 1 & \text{if } d_j > \theta_j \end{cases}$$

where θ_j is the *per-unit* optimal threshold [Dayan and Willshaw, 1991]

Recalling memories



Older memory

- Dendritic sum (like membrane potential at soma)

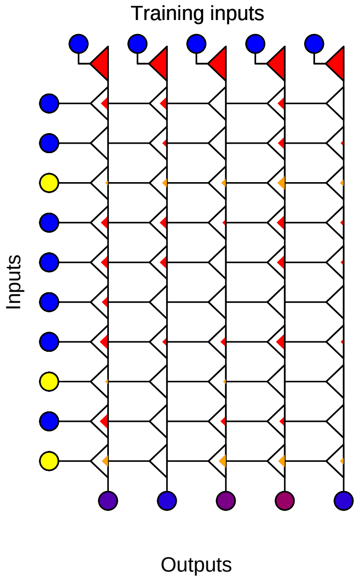
$$d_j = \sum_i w_{ij} a_i$$

- Output (like spike)

$$o_j = \begin{cases} 0 & \text{if } d_j < \theta_j \\ 1 & \text{if } d_j > \theta_j \end{cases}$$

where θ_j is the *per-unit* optimal threshold [Dayan and Willshaw, 1991]

Recalling memories



Oldest memory

- Dendritic sum (like membrane potential at soma)

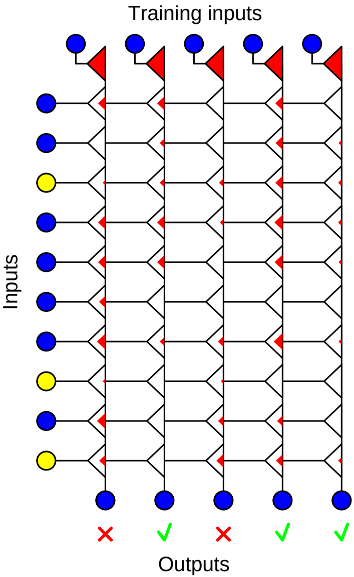
$$d_j = \sum_i w_{ij} a_i$$

- Output (like spike)

$$o_j = \begin{cases} 0 & \text{if } d_j < \theta_j \\ 1 & \text{if } d_j > \theta_j \end{cases}$$

where θ_j is the *per-unit* optimal threshold [Dayan and Willshaw, 1991]

Recalling memories



Oldest memory

- Dendritic sum (like membrane potential at soma)

$$d_j = \sum_i w_{ij} a_i$$

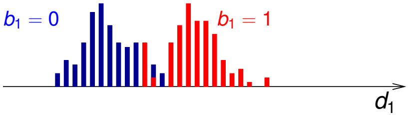
- Output (like spike)

$$o_j = \begin{cases} 0 & \text{if } d_j < \theta_j \\ 1 & \text{if } d_j > \theta_j \end{cases}$$

where θ_j is the *per-unit* optimal threshold [Dayan and Willshaw, 1991]

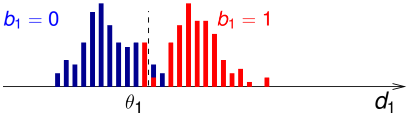
Overview of analysis

1. Compute histogram of dendritic sums of first output unit for patterns
   - where target is 1
   - and where target is 0



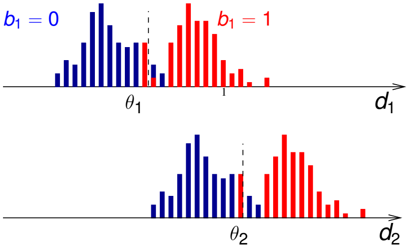
Overview of analysis

1. Compute histogram of dendritic sums of first output unit for patterns
   - where target is 1
   - and where target is 0
2. Find threshold that discriminates optimally between the two distributions



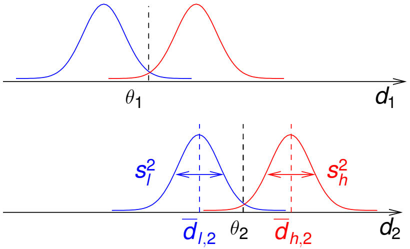
Overview of analysis

1. Compute histogram of dendritic sums of first output unit for patterns
   - where target is 1
   - and where target is 0
2. Find threshold that discriminates optimally between the two distributions
3. Repeat for units $2, 3, \ldots$
   - Different optimal thresholds



Overview of analysis

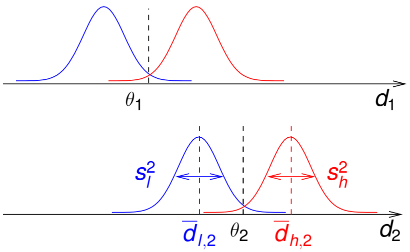
1. Compute histogram of dendritic sums of first output unit for patterns
   - where target is 1
   - and where target is 0
2. Find threshold that discriminates optimally between the two distributions
3. Repeat for units $2, 3, \ldots$
   - Different optimal thresholds
4. Evaluate the *expected* Signal to Noise Ratio (SNR)



$$\mathrm{SNR} = \frac{\left(\overline{d}_h - \overline{d}_l\right)^2}{\frac{1}{2}\left(s_h^2 + s_l^2\right)}$$

Overview of analysis

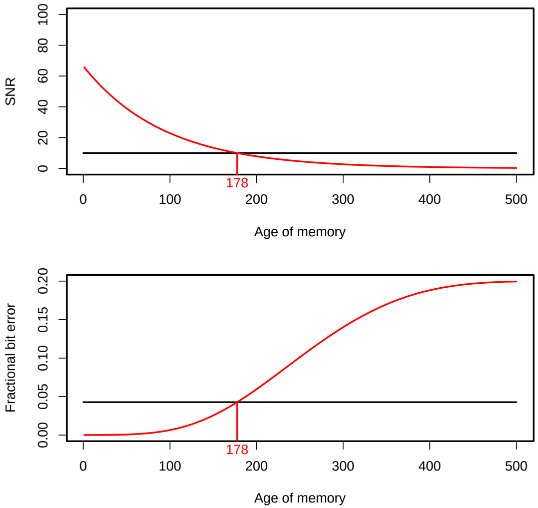
1. Compute histogram of dendritic sums of first output unit for patterns
    - where target is 1
    - and where target is 0
2. Find threshold that discriminates optimally between the two distributions
3. Repeat for units $2, 3, \ldots$
    - Different optimal thresholds
4. Evaluate the *expected* Signal to Noise Ratio (SNR)



$$\text{SNR} = \frac{\left(\overline{d}_h - \overline{d}_l\right)^2}{\frac{1}{2}(s_h^2 + s_l^2)}$$
$$= f(p, r, N_{\text{units}}, \tau)$$

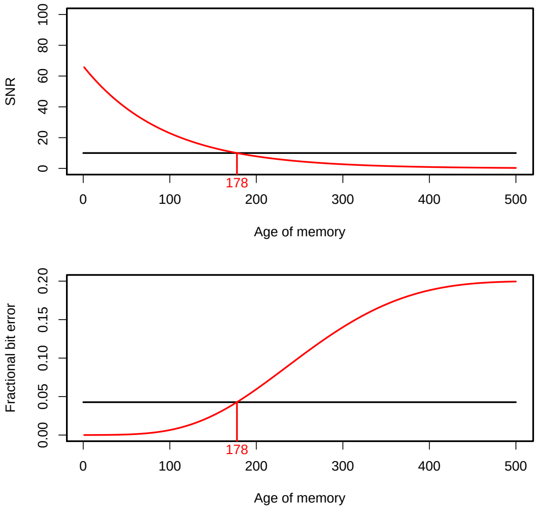
where $f(\cdot)$ is complicated

Palimpsest properties



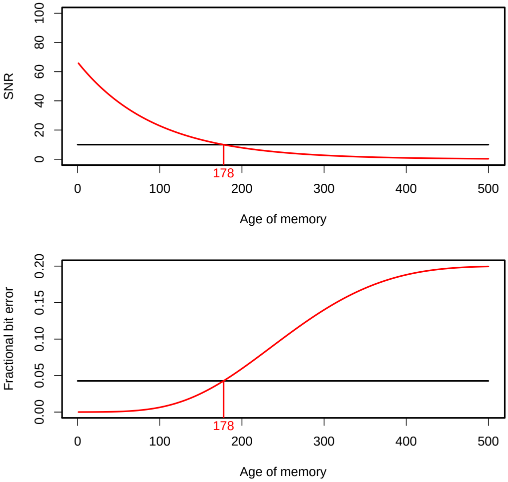
- **Associative nets with general learning rules act as palimpsests**

Palimpsest properties



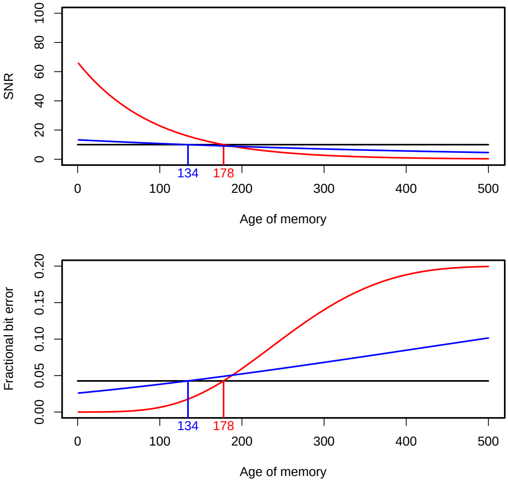
- Associative nets with general learning rules act as palimpsests
- Define *capacity* of network as age of memory whose error (or equivalently SNR) reaches an criterion value

How best to forget?



Quickly

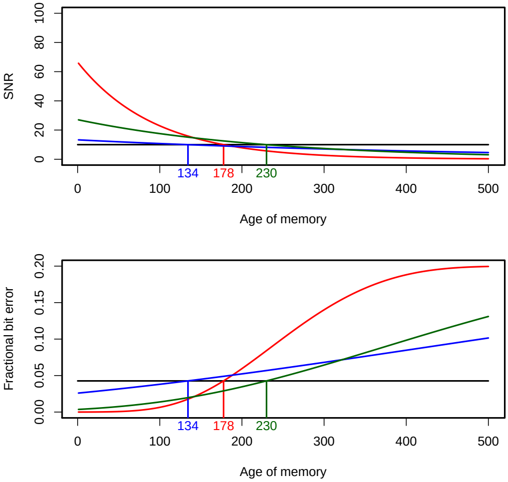
How best to forget?



Quickly

Slowly

How best to forget?



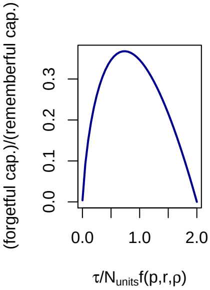
Quickly
Slowly
Optimally

How best to learn?

- Hebbian Learning Rule
  - Large number of input neurons (N_{units}):

$$\text{capacity} \approx \frac{\tau}{2}\left(\ln N_{\text{units}} - 2\ln\tau + \ln\frac{1-p}{pr^2\rho_c}\right)$$

where ρ_c is the SNR criterion.

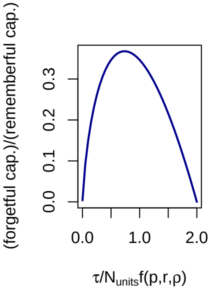


How best to learn?

- Hebbian Learning Rule
  - Large number of input neurons (N_{units}):

$$\text{capacity} \approx \frac{\tau}{2}\left(\ln N_{\text{units}} - 2\ln\tau + \ln\frac{1-p}{pr^2\rho_c}\right)$$

    where ρ_c is the SNR criterion.
  - If $\tau \propto \sqrt{N_{\text{units}}}$, capacity $\propto \sqrt{N_{\text{units}}}$.



How best to learn?

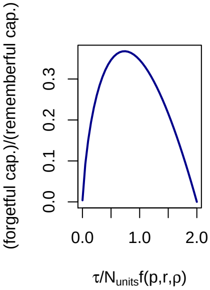
- Hebbian Learning Rule
  - Large number of input neurons (N_{units}):

    $$\text{capacity} \approx \frac{\tau}{2}\left(\ln N_{\text{units}} - 2\ln\tau + \ln\frac{1-p}{pr^2\rho_c}\right)$$

    where ρ_c is the SNR criterion.
  - If $\tau \propto \sqrt{N_{\text{units}}}$, capacity $\propto \sqrt{N_{\text{units}}}$.
- Balanced learning rules
  - For large N_{units}

    $$\text{capacity} \approx \frac{\tau}{2}\left(\ln N_{\text{units}} - \ln\tau + \ln f(p, r, \rho_c)\right)$$

  - If $\tau \propto N_{\text{units}}$, capacity $\propto N_{\text{units}}$.

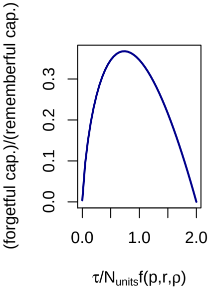


How best to learn?

- Hebbian Learning Rule
  - Large number of input neurons (N_{units}):

    $$\text{capacity} \approx \frac{\tau}{2}\left(\ln N_{\text{units}} - 2\ln\tau + \ln\frac{1-p}{pr^2\rho_c}\right)$$

    where ρ_c is the SNR criterion.
  - If $\tau \propto \sqrt{N_{\text{units}}}$, capacity $\propto \sqrt{N_{\text{units}}}$.

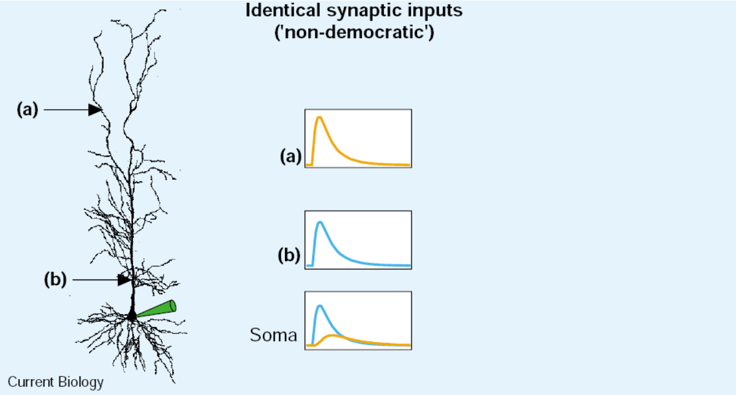


- Balanced learning rules
  - For large N_{units}

    $$\text{capacity} \approx \frac{\tau}{2}\left(\ln N_{\text{units}} - \ln\tau + \ln f(p, r, \rho_c)\right)$$

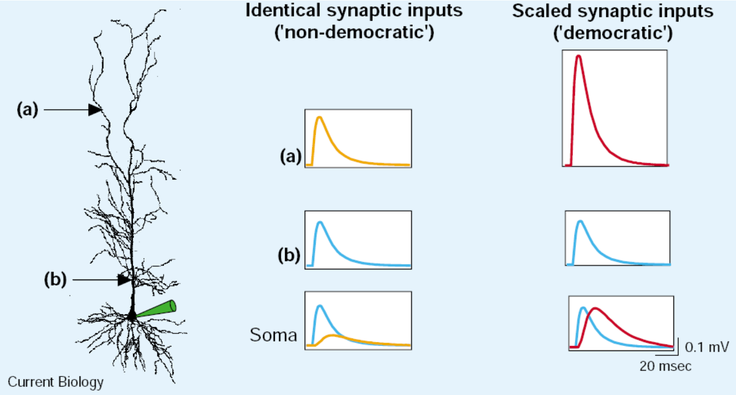
  - If $\tau \propto N_{\text{units}}$, capacity $\propto N_{\text{units}}$.
- Conclusion: **balanced rules scale best**
  - As in rememberful nets [Dayan and Willshaw, 1991]
  - Optimal capacity approximately a third ($1/e$) of "rememberful" net

Inhomogeneity 2: Differential Attenuation



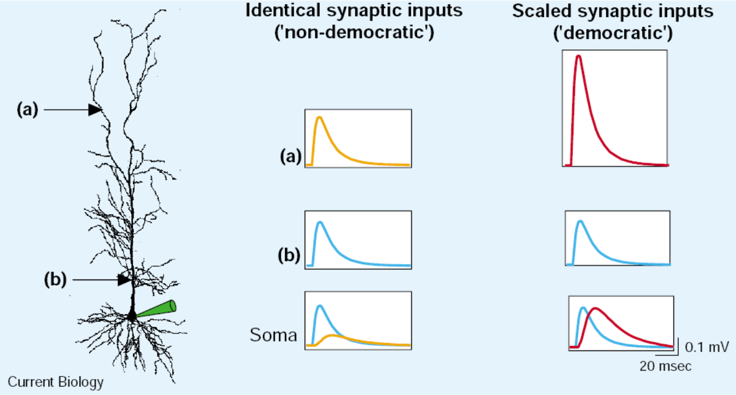
[Häusser, 2001]

Inhomogeneity 2: Differential Attenuation



[Häusser, 2001]

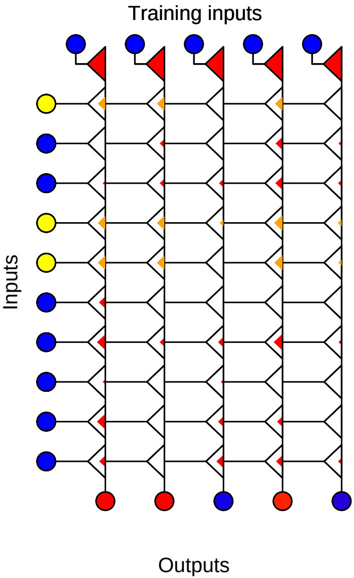
Inhomogeneity 2: Differential Attenuation



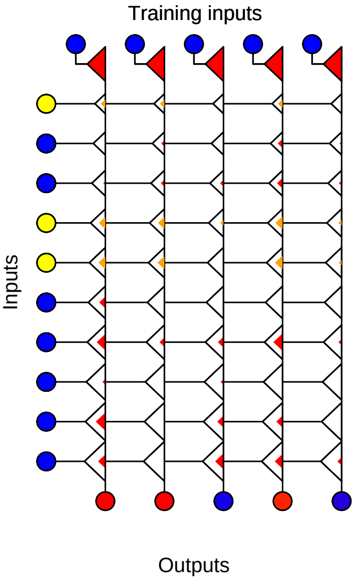
[Häusser, 2001]

- What effect does differential attenuation have on performance?

Modelling differential attenuation



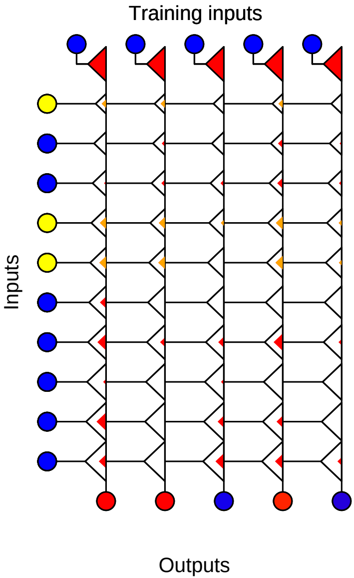
Modelling differential attenuation



- Input *attenuation factors* f_i

$$d_j = \sum_i f_i w_{ij} a_i$$

Modelling differential attenuation



- Input *attenuation factors* f_i

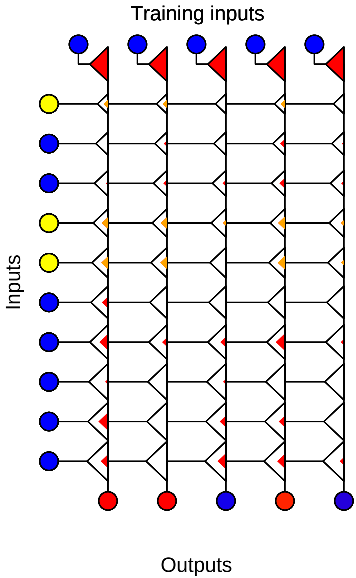
$$d_j = \sum_i f_i w_{ij} a_i$$

- Balanced capacity reduction

$$1 + (\mathrm{CV}(f))^2$$

$$\mathrm{CV}(f) = \frac{\mathrm{SD}(f)}{\mathrm{MEAN}(f)}$$

Modelling differential attenuation



- Input *attenuation factors* f_i

$$d_j = \sum_i f_i w_{ij} a_i$$

- Balanced capacity reduction

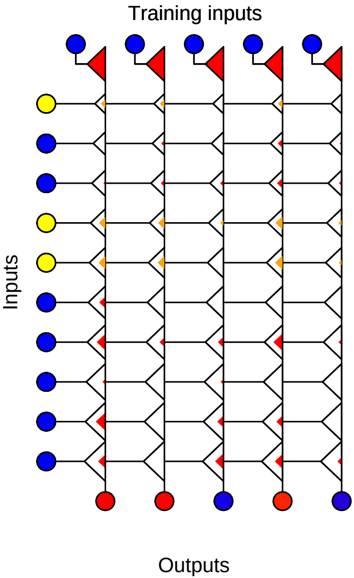
$$1 + (\mathrm{CV}(f))^2$$

$$\mathrm{CV}(f) = \frac{\mathrm{SD}(f)}{\mathrm{MEAN}(f)}$$

- Hebbian capacity reduction

$$\sqrt{1 + (\mathrm{CV}(f))^2}$$

Modelling differential attenuation



- Input *attenuation factors* f_i

$$d_j = \sum_i f_i w_{ij} a_i$$

- Balanced capacity reduction

$$1 + (\mathrm{CV}(f))^2$$

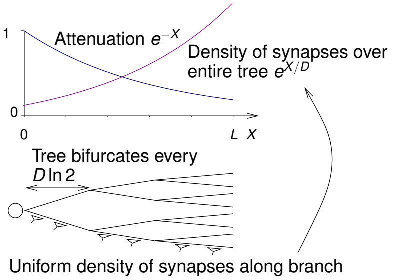
$$\mathrm{CV}(f) = \frac{\mathrm{SD}(f)}{\mathrm{MEAN}(f)}$$

- Hebbian capacity reduction

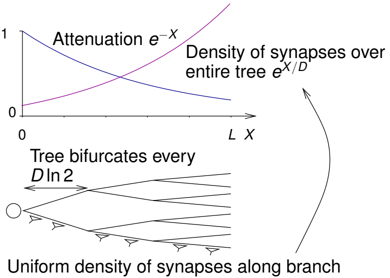
$$\sqrt{1 + (\mathrm{CV}(f))^2}$$

- $\sim$independent of forgetting

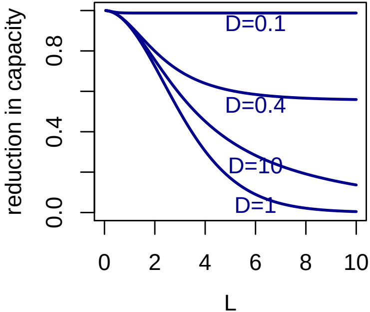
Example: a branching dendritic tree

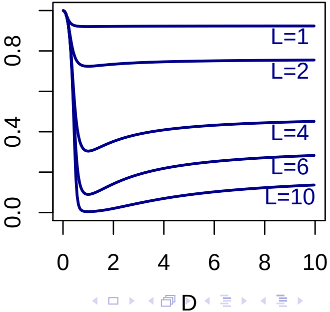


Example: a branching dendritic tree

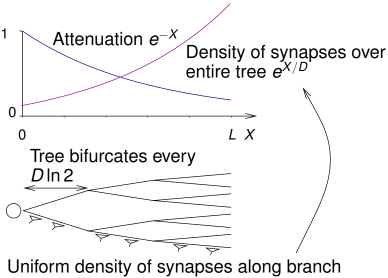


$$\frac{(1-2D)\left(e^{L(\frac{1}{D}-1)}-1\right)^2}{(D-1)^2(e^{\frac{L}{D}}-1)(e^{L(\frac{1}{D}-2)}-1)}$$



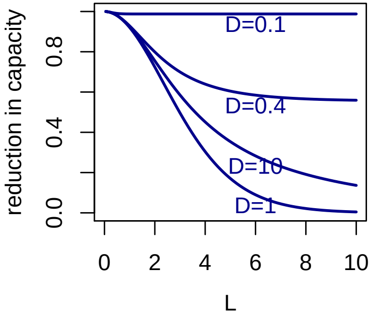


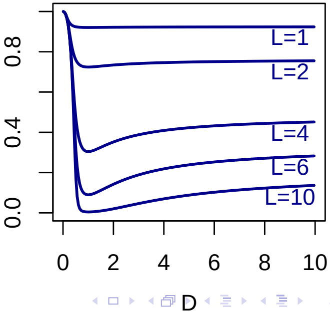
Example: a branching dendritic tree



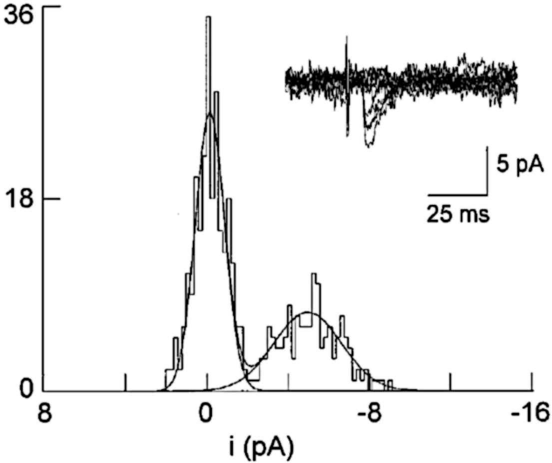
- Capacity is reduced at worst by factor of 1.4 for Schaffer Collaterals ($L = 2$)

$$\frac{(1-2D)\left(e^{L(\frac{1}{D}-1)}-1\right)^2}{(D-1)^2(e^{\frac{L}{D}}-1)(e^{L(\frac{1}{D}-2)}-1)}$$





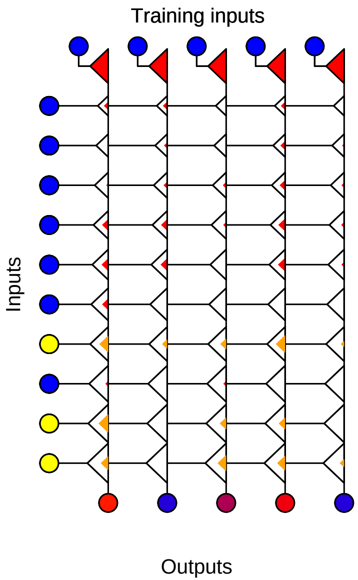
Inhomogeneity 3: Stochastic transmission



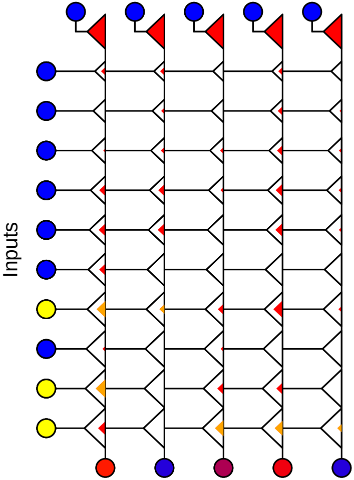
[Bolshakov et al., 1997]

- Quantal failure
- Fluctuations in quantal amplitude

Modelling stochastic transmission



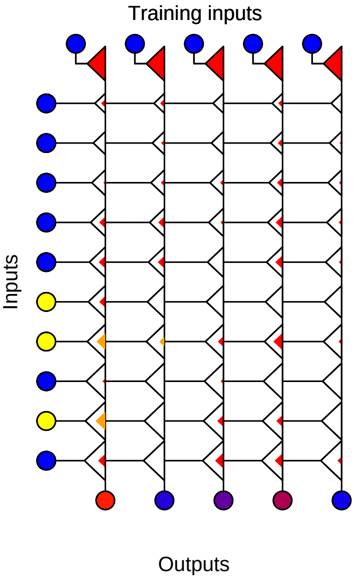
Modelling stochastic transmission



- *Stochastic transmission factors $g_{ij}(t)$*

$$d_j = \sum_i f_i g_{ij}(t) w_{ij} a_i$$

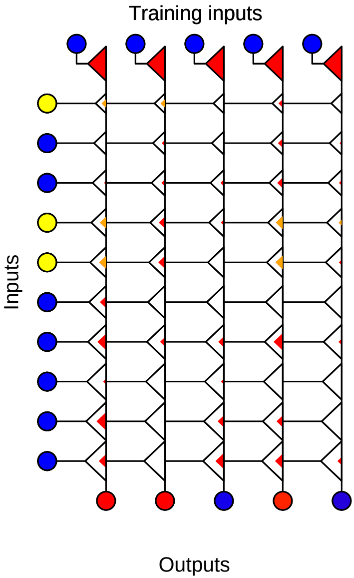
Modelling stochastic transmission



- *Stochastic transmission factors $g_{ij}(t)$*

$$d_j = \sum_i f_i g_{ij}(t) w_{ij} a_i$$

Modelling stochastic transmission



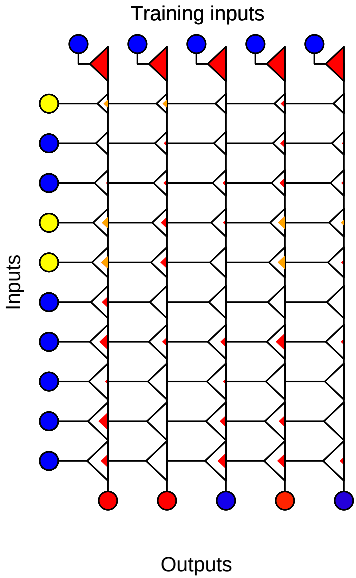
- *Stochastic transmission factors $g_{ij}(t)$*

$$d_j = \sum_i f_i g_{ij}(t) w_{ij} a_i$$

- Balanced capacity reduction

$$1 + (\mathrm{CV}(g))^2 / (1 - p)$$

Modelling stochastic transmission



- *Stochastic transmission factors* $g_{ij}(t)$

$$d_j = \sum_i f_i g_{ij}(t) w_{ij} a_i$$

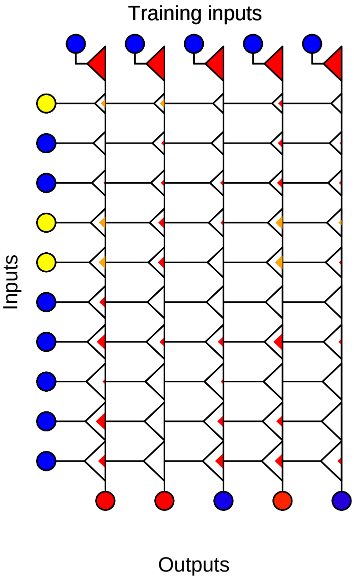
- Balanced capacity reduction

$$1 + (\mathrm{CV}(g))^2 / (1 - p)$$

- Hebbian capacity reduction

$$\sqrt{1 + (\mathrm{CV}(g))^2 / (1 - p)}$$

Modelling stochastic transmission



- *Stochastic transmission factors* $g_{ij}(t)$

$$d_j = \sum_i f_i g_{ij}(t) w_{ij} a_i$$

- Balanced capacity reduction

$$1 + (\mathrm{CV}(g))^2 / (1 - p)$$

- Hebbian capacity reduction

$$\sqrt{1 + (\mathrm{CV}(g))^2 / (1 - p)}$$

- $\sim$independent of forgetting and differential attenuation

Effect of stochastic transmission in the hippocampus

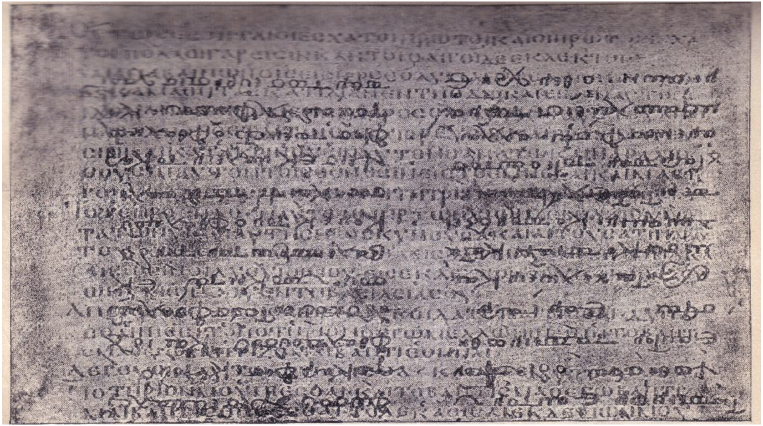
- **Factorial reduction of capacity is in range**
  - **2–15 (balanced)**
  - **1.4–4 (Hebbian)**
- CV(g) Estimated from quantal analysis:
  - Maximum value of $(\mathrm{CV}(g))^2 \approx 10$ [Stricker et al., 1996]
  - $(\mathrm{CV}(g))^2 \approx 0.5$ in a potentiated state [Bolshakov et al., 1997]
  - $(\mathrm{CV}(g))^2 \approx 2$ for unpotentiated synapses [Bolshakov et al., 1997]
- p in range 0.2–0.3 (*in vivo* recordings from CA3 [Leutgeb et al., 2004, Barnes et al., 1990]

Conclusions

- First unified analysis of the effects of differential attenuation and stochastic transmission on the performance of forgetful associative memories
- Different types of inhomogeneity are approximately independent from one another
- There is an optimal rate at which to forget
- Optimal network capacity scales with size
- Stochastic transmission is likely to have a more pronounced effect on the performance of memory networks in the hippocampus than differential attenuation
  - Surprising in the context of the findings that synaptic conductances are scaled according to distance from the soma
- Performance is robust

Palimpsests

- Forgetful (Hopfield) neural network [Nadal et al., 1986, Parisi, 1986] ...
- ... a document written on vellum over the incompletely-erased trace of an earlier document



Codex Ephraemi Rescriptus (5th century and 12th century)

Barnes, C. A., McNaughton, B. L., Mizumori, S. J. Y., Leonard, B. W., and Lin, L.-H. (1990).
Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing.
Progress in Brain Research, 83:287–300.

Bolshakov, V. Y., Golan, H., Kandel, E. R., and Siegelbaum, S. A. (1997).
Recruitment of new sites of synaptic transmission during the cAMP-dependent late phase of LTP at CA3–CA1 synapses in the hippocampus.
Neuron, 19:635–651.

Dayan, P. and Willshaw, D. J. (1991).
Optimising synaptic learning rules in linear associative memories.
Biological Cybernetics, 65:253–265.

Häusser, M. (2001).
Dendritic democracy.