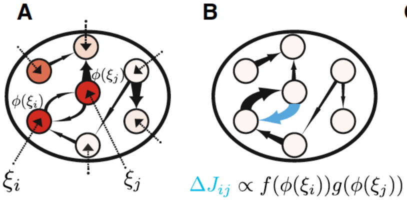
Information bounds and attractor dynamics of a Hebbian associative memory

Clayton Seitz

May 26, 2021

RNNs trained with a Hebbian learning rule

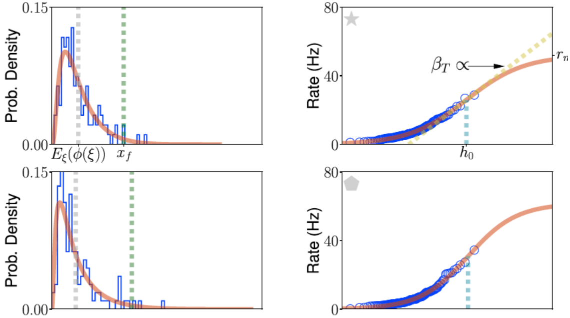
It is difficult to infer learning rules *in vivo* solely from spikes



Learning rules can be inferred (with assumptions) from firing rates [1]

[1] [Peirera and Brunel, Neuron. 2018]

Inferring the transfer function from ITC data



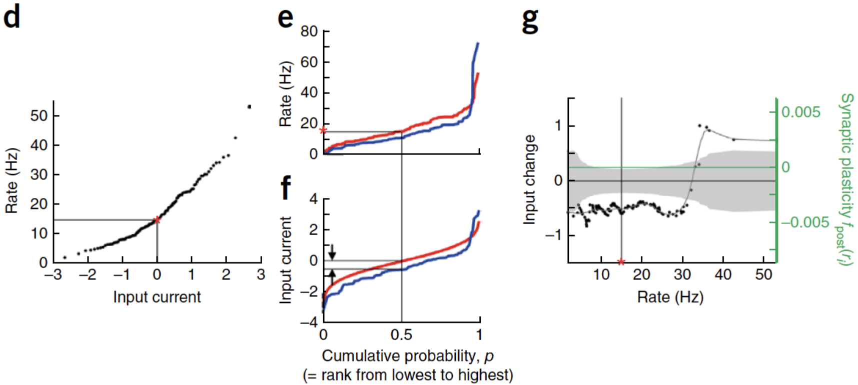
Measuring the *static* transfer function from novel images assuming that input currents are Gaussian variables

$$\phi(\xi) = \frac{r_{max}}{1 + \exp \beta(\xi - \xi_0)}$$

2

[2] [Peirera and Brunel, Neuron. 2018]

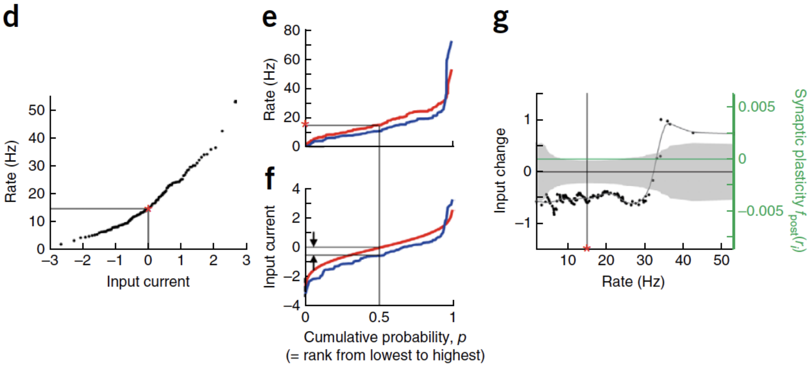
Inferring the learning rule from ITC data



Inferring the change in input current ξ_{in} from the change in firing rate in novel relative to familiar stimuli [3]

[3] [Lim et al., Nature Neuroscience. 2015]

Inferring the learning rule from ITC data



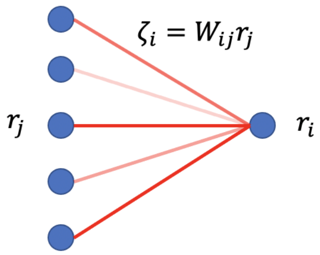
The change in input current to a neuron can then be read from the firing rate of that neuron when presented a novel stimulus

$$\Delta\xi_i(r) \propto (2q + 1 - \tanh(\beta(r - x)))$$

4

[4] [Lim et al., Nature Neuroscience. 2015]

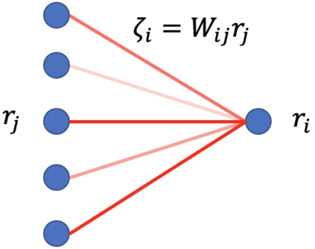
Determining the input change

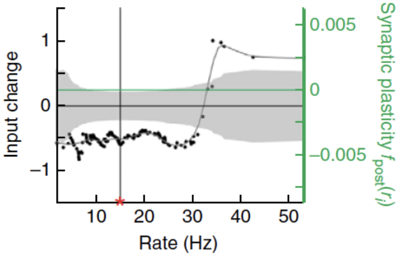


Assuming that $\Delta W_{ij} \propto f(r_i)g(r_j)$, the change in input current ξ_i is related to the learning rule by

$$\Delta \xi_i \propto f(r_i) \sum_j g(r_j) r_j$$

Determining the learning rule





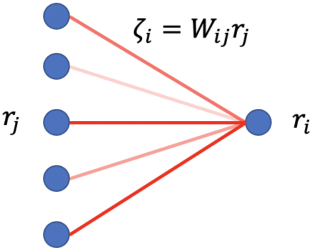
We can fit input change $\Delta\xi_i(r)$ from the data (top right)

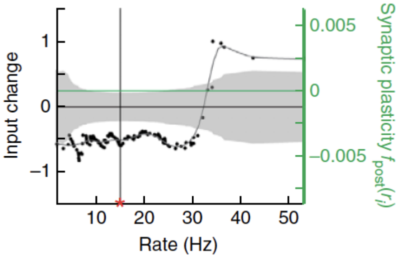
$$\Delta\xi_i(r_i) \propto (2q + 1 - \tanh(\beta(r_i - r_0)))$$

5

[5] [Lim et al., Nature Neuroscience. 2015]

Determining the learning rule





We can infer the dependence of the learning rule on the post-synaptic firing rate $f(r_i)$

$$f(r_i) = \Delta\xi_i(r_i) / \sum_j g(r_j) r_j$$

6

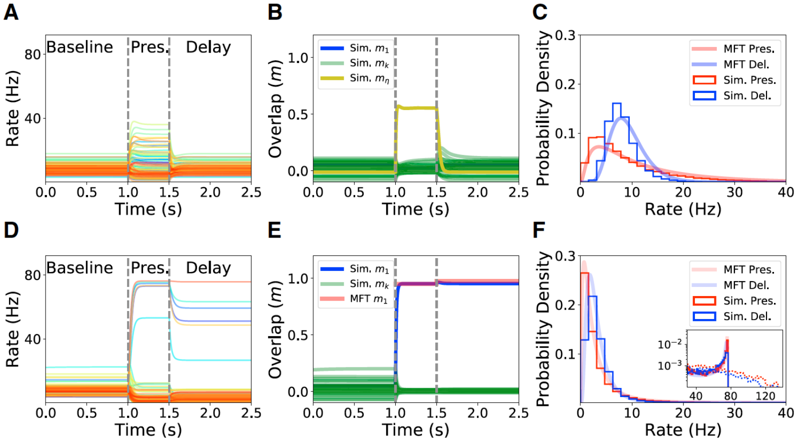
[6][Lim et al., Nature Neuroscience. 2015]

Training the model inferred from ITC data

During training, we stimulate the network with

$$\xi_{in}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp -\frac{1}{2}(\boldsymbol{r} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{r} - \boldsymbol{\mu})$$

Attractor dynamics of the trained model



Attractor states can be observed from the overlap m [7]

[7] [Peirera and Brunel, Neuron. 2018]

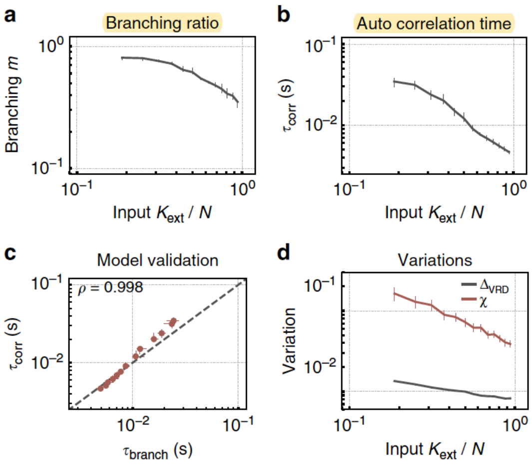
Signatures of critical points from stimuli addition

Information content of firing rates at a fixed point

The effect of noisy stimuli

Do these networks optimize information transmission?

Are these networks functioning at a critical point? What about the balance between input and recurrence? (Cramer et al. 2020)



A coding theory perspective

How much information does the response R carry about the input pattern S i.e. $I(R; S)$ on novel and familiar stimuli?

What is the fundamental coding capacity of these networks?