Statistical learning algorithms for biological neural networks

New Jersey Institute of Technology Biology Colloquium

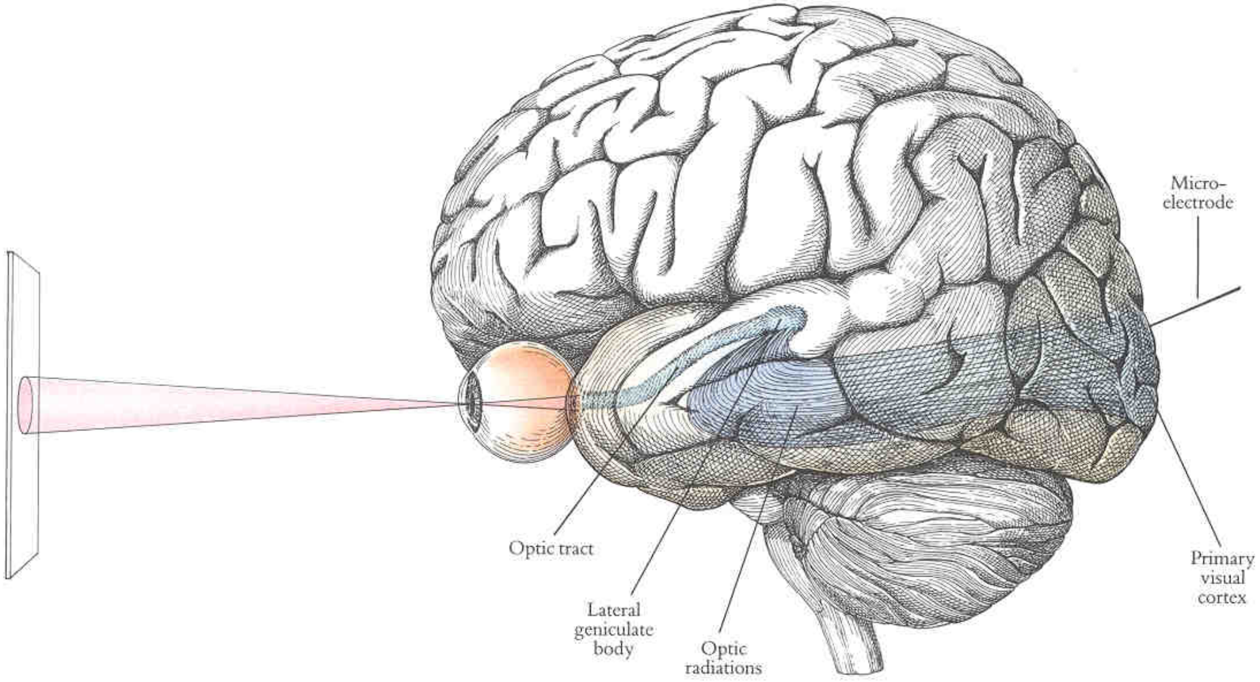
David Lipshutz

16 April 2024

FLATIRON
INSTITUTE

A theorist's view of sensory processing

Sensory systems encode natural signals as patterns of electrical activity, which are reformatted over multiple stages of processing to produce useful representations of the world.



Goal: concise mathematical descriptions of the statistical learning algorithms that support sensory processing.

Neural systems are complex. What does a concise mathematical description look like?

Neural systems are complex. What does a concise mathematical description look like?

Let’s contrast 2 models of single neurons.

J. Physiol. (1952) 117, 500–544

A QUANTITATIVE DESCRIPTION OF MEMBRANE CURRENT AND ITS APPLICATION TO CONDUCTION AND EXCITATION IN NERVE

By A. L. HODGKIN and A. F. HUXLEY

$$I = C_M \frac{\mathrm{d}V}{\mathrm{d}t} + \bar{g}_K n^4 (V - V_K) + \bar{g}_{Na} m^3 h (V - V_{Na}) + \bar{g}_l (V - V_l),$$

$$\mathrm{d}n/\mathrm{d}t = \alpha_n (1 - n) - \beta_n n,$$

$$\mathrm{d}m/\mathrm{d}t = \alpha_m (1 - m) - \beta_m m,$$

~20 parameters

$$\mathrm{d}h/\mathrm{d}t = \alpha_h (1 - h) - \beta_h h,$$

$$\alpha_n = 0{\cdot}01\ (V + 10) \Big/ \left(\exp \frac{V + 10}{10} - 1\right),$$

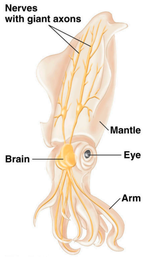
$$\beta_n = 0{\cdot}125 \exp (V/80),$$

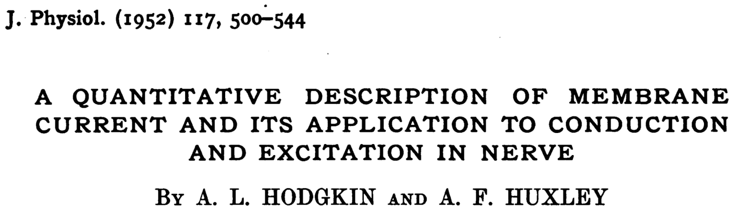
$$\alpha_m = 0{\cdot}1\ (V + 25) \Big/ \left(\exp \frac{V + 25}{10} - 1\right),$$

$$\beta_m = 4 \exp (V/18),$$

$$\alpha_h = 0{\cdot}07 \exp (V/20),$$

$$\beta_h = 1 \Big/ \left(\exp \frac{V + 30}{10} + 1\right).$$

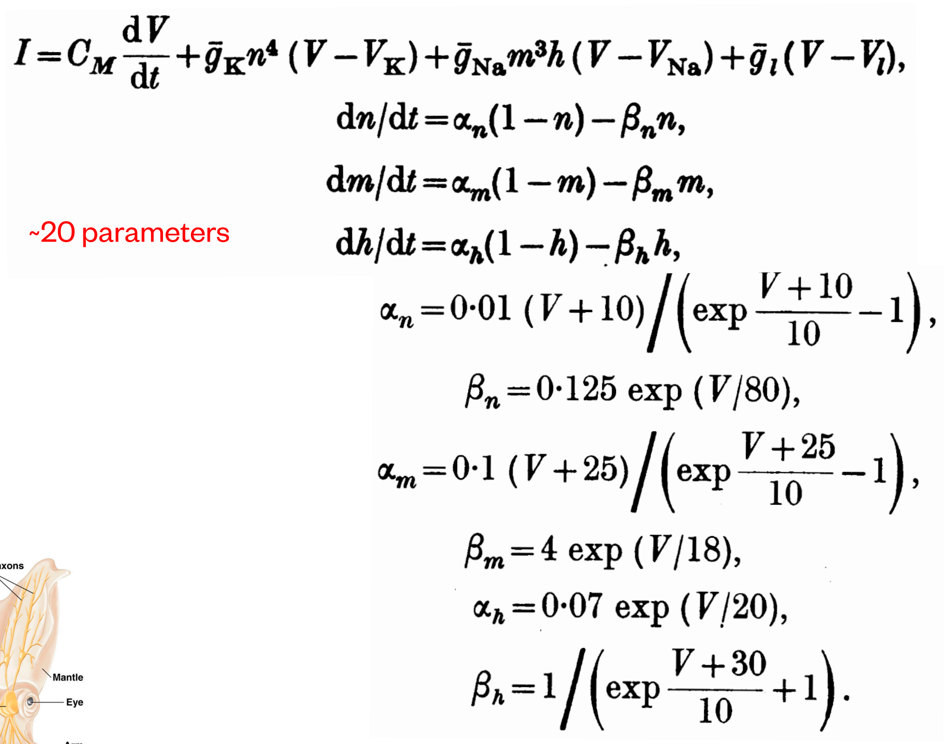




J. Physiol. (1952) 117, 500–544

A QUANTITATIVE DESCRIPTION OF MEMBRANE CURRENT AND ITS APPLICATION TO CONDUCTION AND EXCITATION IN NERVE

By A. L. HODGKIN and A. F. HUXLEY



$$I = C_M \frac{\mathrm{d}V}{\mathrm{d}t} + \bar{g}_{\mathrm{K}} n^4 (V - V_{\mathrm{K}}) + \bar{g}_{\mathrm{Na}} m^3 h (V - V_{\mathrm{Na}}) + \bar{g}_l (V - V_l),$$

$$\mathrm{d}n/\mathrm{d}t = \alpha_n (1 - n) - \beta_n n,$$

$$\mathrm{d}m/\mathrm{d}t = \alpha_m (1 - m) - \beta_m m,$$

$$\mathrm{d}h/\mathrm{d}t = \alpha_h (1 - h) - \beta_h h,$$

$$\alpha_n = 0{\cdot}01 \; (V + 10) \Big/ \left(\exp \frac{V + 10}{10} - 1 \right),$$

$$\beta_n = 0{\cdot}125 \; \exp \; (V/80),$$

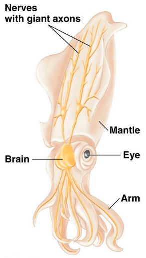
$$\alpha_m = 0{\cdot}1 \; (V + 25) \Big/ \left(\exp \frac{V + 25}{10} - 1 \right),$$

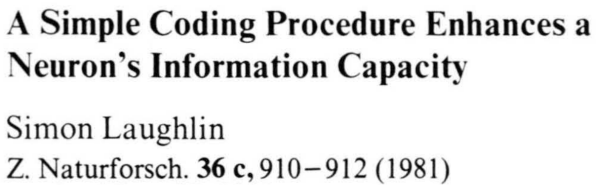
$$\beta_m = 4 \; \exp \; (V/18),$$

$$\alpha_h = 0{\cdot}07 \; \exp \; (V/20),$$

$$\beta_h = 1 \Big/ \left(\exp \frac{V + 30}{10} + 1 \right).$$

~20 parameters

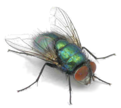


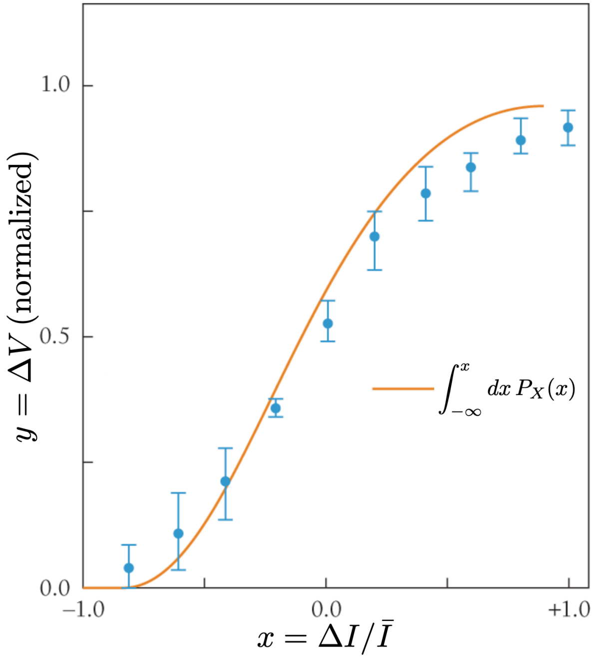


A Simple Coding Procedure Enhances a Neuron's Information Capacity

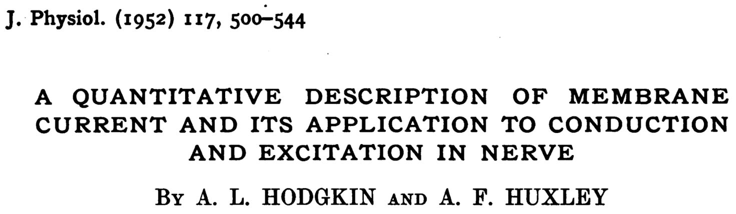
Simon Laughlin

Z. Naturforsch. **36 c**, 910–912 (1981)





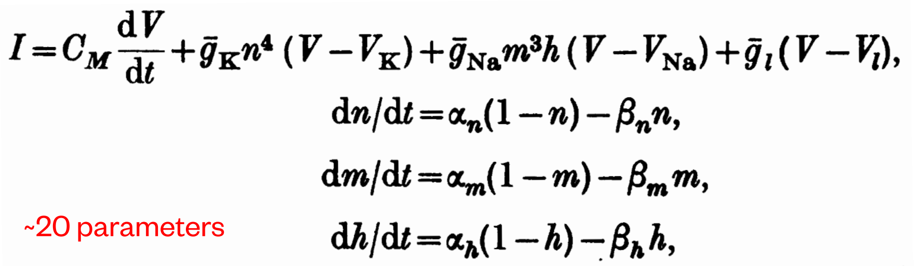
Based on theories of **Efficient Coding** [Barlow 1961, Attneave 1954] and **Communication** [Shannon 1948]. Other successful examples [Dan et al. 1996; Bell & Sejnowski 1995, 1997; Olshausen & Field 1996; Brenner et al. 2000; Pitkow & Meister 2012; Palmer et al. 2015; ...]



J. Physiol. (1952) 117, 500–544

A QUANTITATIVE DESCRIPTION OF MEMBRANE CURRENT AND ITS APPLICATION TO CONDUCTION AND EXCITATION IN NERVE

By A. L. HODGKIN and A. F. HUXLEY



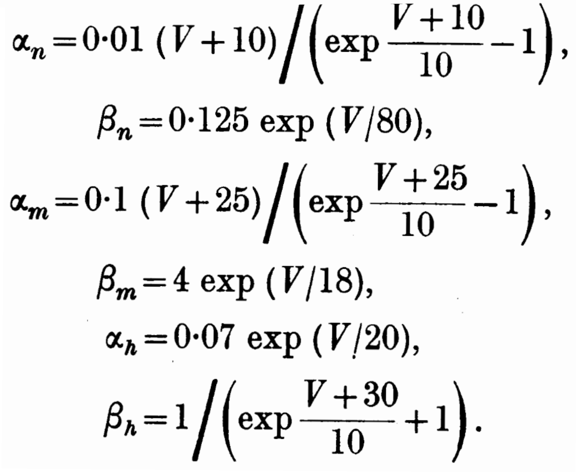
$$I = C_M \frac{dV}{dt} + \bar{g}_K n^4 (V - V_K) + \bar{g}_{Na} m^3 h (V - V_{Na}) + \bar{g}_l (V - V_l),$$

$$dn/dt = \alpha_n (1 - n) - \beta_n n,$$

$$dm/dt = \alpha_m (1 - m) - \beta_m m,$$

$$dh/dt = \alpha_h (1 - h) - \beta_h h,$$

~20 parameters



$$\alpha_n = 0{\cdot}01\ (V + 10) \Big/ \left(\exp \frac{V + 10}{10} - 1\right),$$

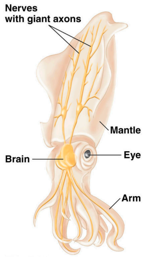
$$\beta_n = 0{\cdot}125 \exp\ (V/80),$$

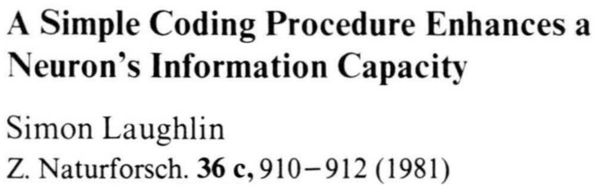
$$\alpha_m = 0{\cdot}1\ (V + 25) \Big/ \left(\exp \frac{V + 25}{10} - 1\right),$$

$$\beta_m = 4 \exp\ (V/18),$$

$$\alpha_h = 0{\cdot}07 \exp\ (V/20),$$

$$\beta_h = 1 \Big/ \left(\exp \frac{V + 30}{10} + 1\right).$$

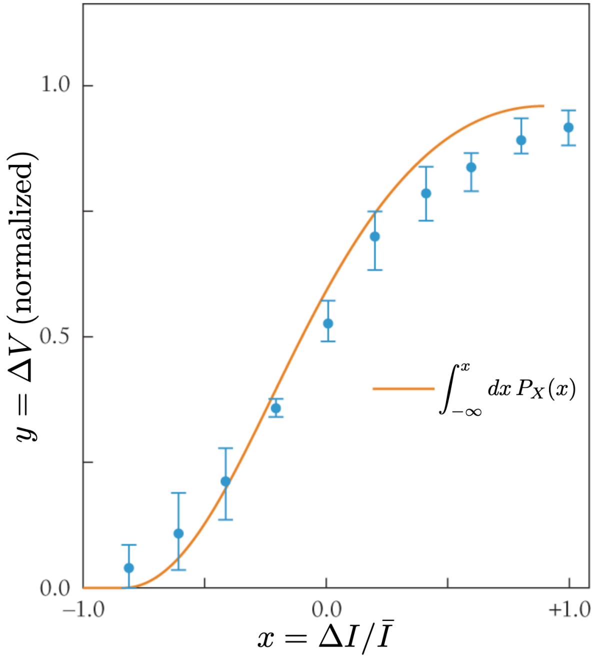




A Simple Coding Procedure Enhances a Neuron's Information Capacity

Simon Laughlin

Z. Naturforsch. **36 c**, 910–912 (1981)



normative:
- max info
- min resources

no free parameters

matches data

Based on theories of **Efficient Coding** [Barlow 1961, Attneave 1954] and **Communication** [Shannon 1948]. Other successful examples [Dan et al. 1996; Bell & Sejnowski 1995, 1997; Olshausen & Field 1996; Brenner et al. 2000; Pitkow & Meister 2012; Palmer et al. 2015; ...]

How do we apply this approach to learning in neural systems?

Wish list for learning algorithms:

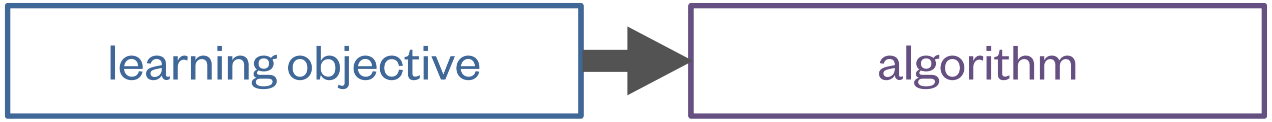
1. normative (principled)
2. sample efficient (good implicit bias, low sample complexity)
3. resource efficient (local, online, sparse, spiking, etc.)
4. no free parameters (parameters matched to input stats)
5. matches neural data (anatomical, physiological)

Normative framework for deriving biological learning algorithms

learning objective

computational goal
resource constraints

Normative framework for deriving biological learning algorithms



learning objective

computational goal
resource constraints

algorithm

algorithm (e.g., stochastic gradient-based algorithm)

Normative framework for deriving biological learning algorithms

learning objective

computational goal
resource constraints

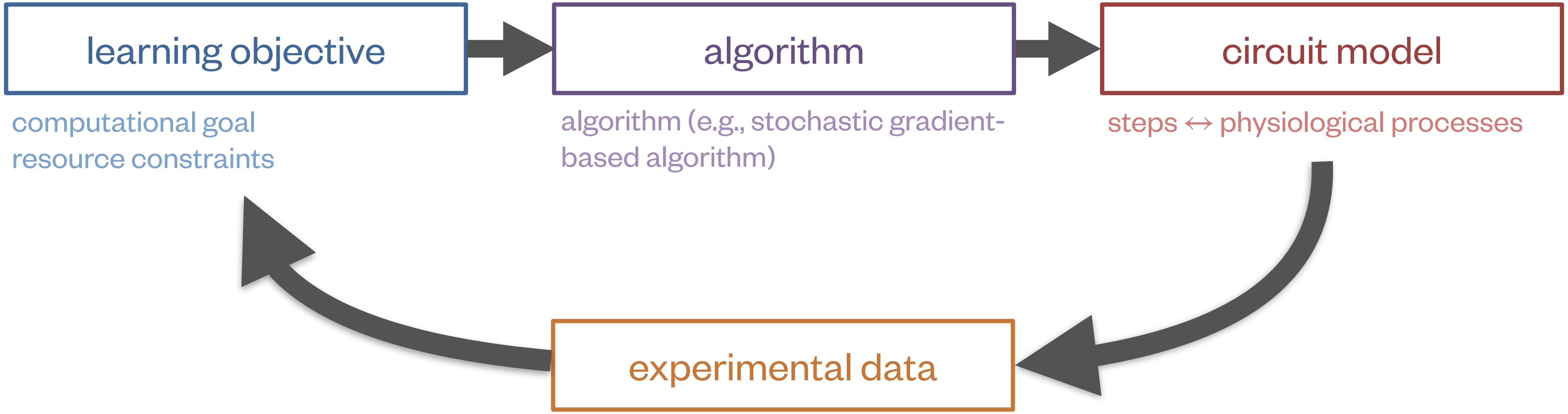
algorithm

algorithm (e.g., stochastic gradient-based algorithm)

circuit model

steps ↔ physiological processes

Normative framework for deriving biological learning algorithms



learning objective

computational goal
resource constraints

algorithm

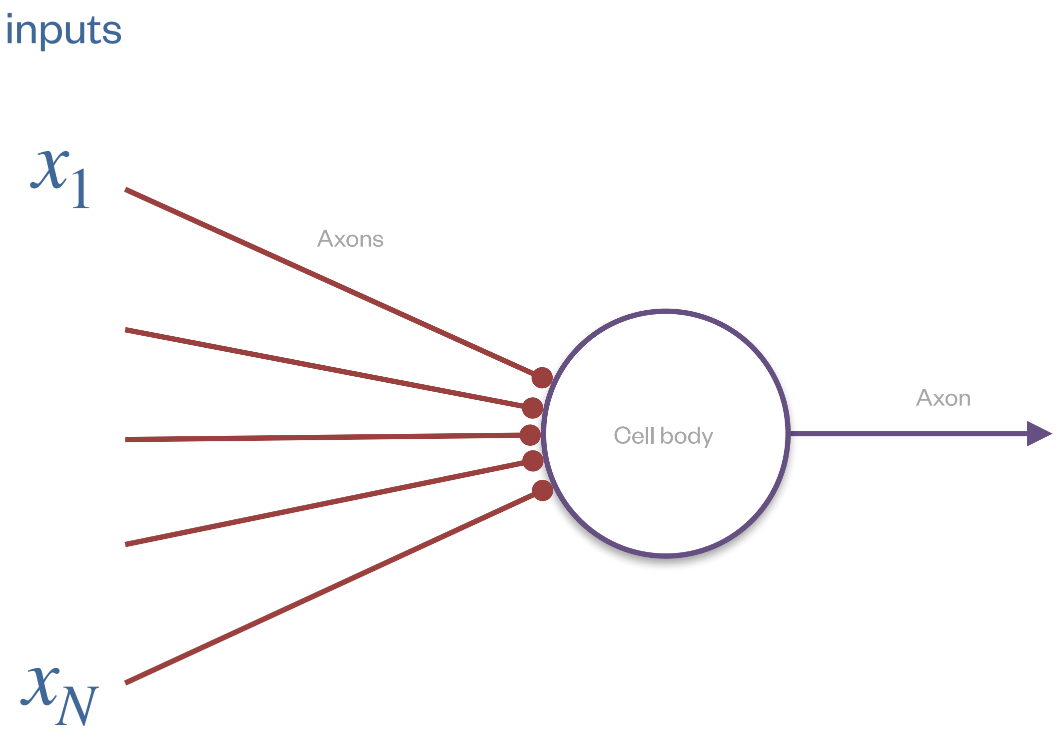
algorithm (e.g., stochastic gradient-based algorithm)

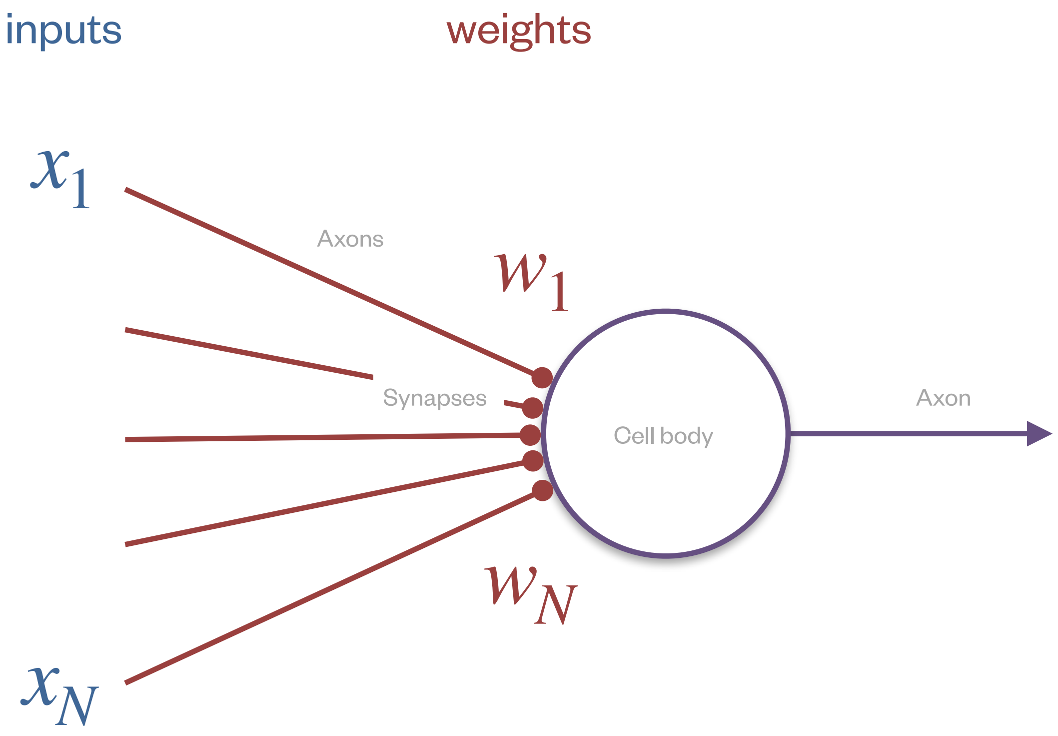
circuit model

steps ↔ physiological processes

experimental data

compare w. anatomical / physiological data

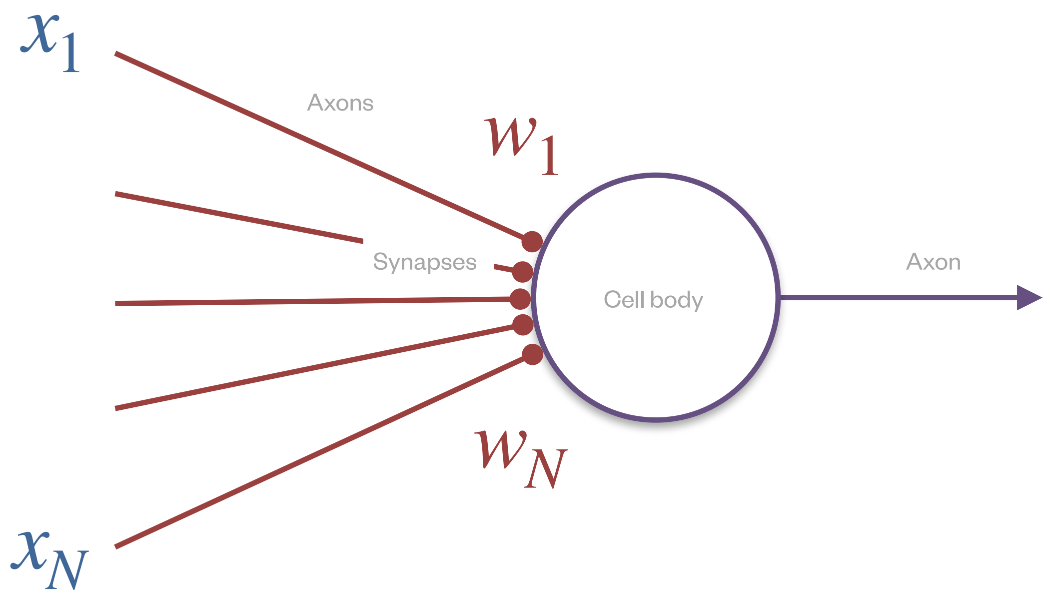

inputs
x_1
Axons
Cell body
Axon
x_N


inputs
weights
x_1
Axons
w_1
Synapses
Cell body
Axon
w_N
x_N

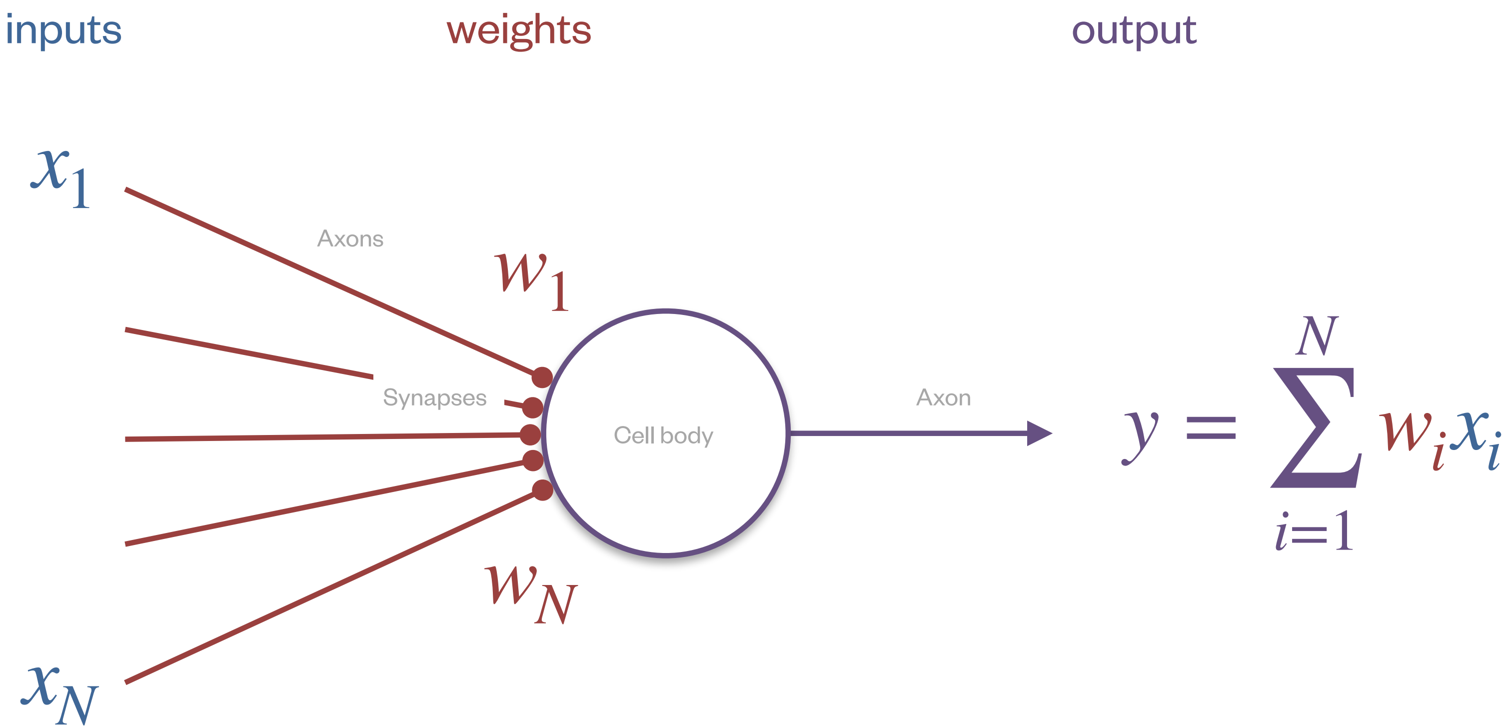
inputs

weights

output



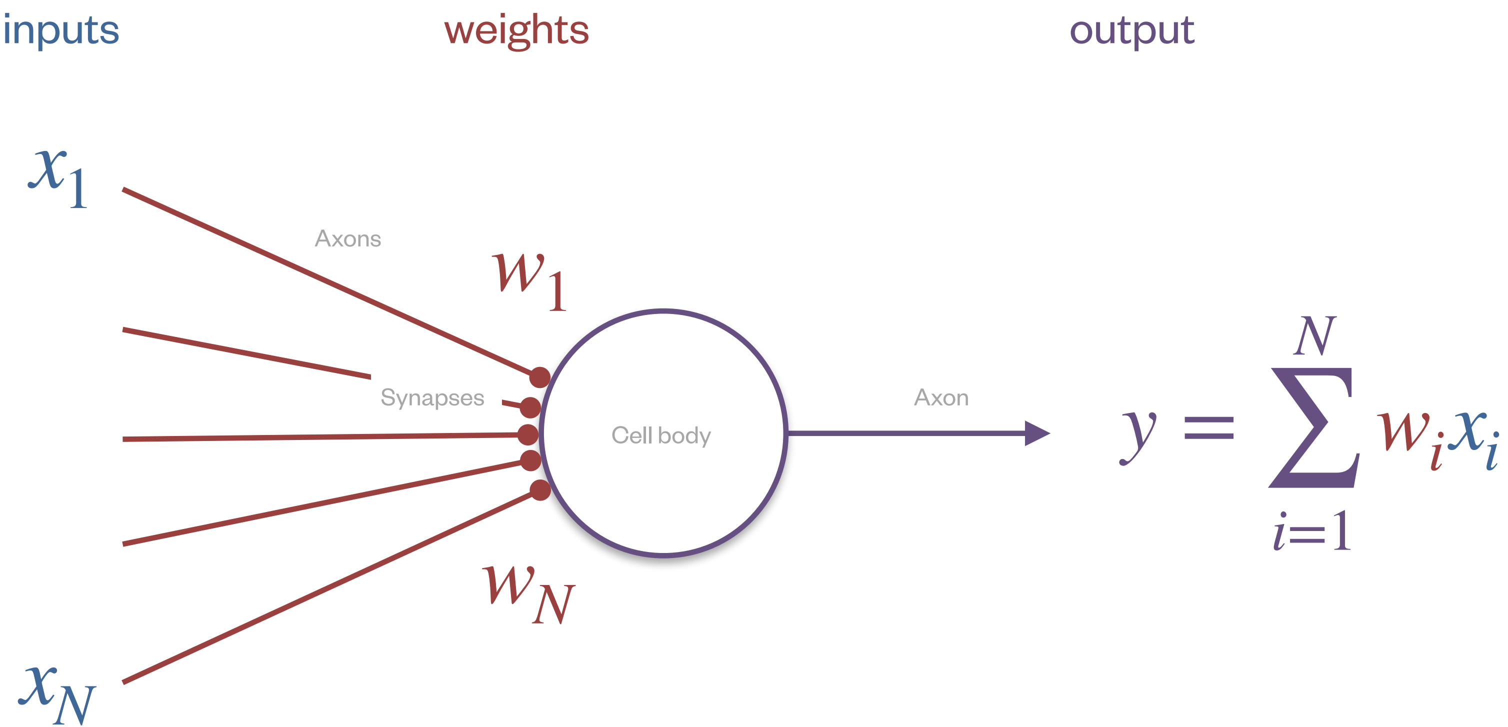
$$y = \sum_{i=1}^{N} w_i x_i$$



$$y = \sum_{i=1}^{N} w_i x_i$$

local, online synaptic plasticity rule:

$$\Delta w_i = f(w_i, x_i, y, \ldots)$$



$$y = \sum_{i=1}^{N} w_i x_i$$

local, online synaptic plasticity rule:

$$\Delta w_i = f(w_i, x_i, y, \ldots)$$

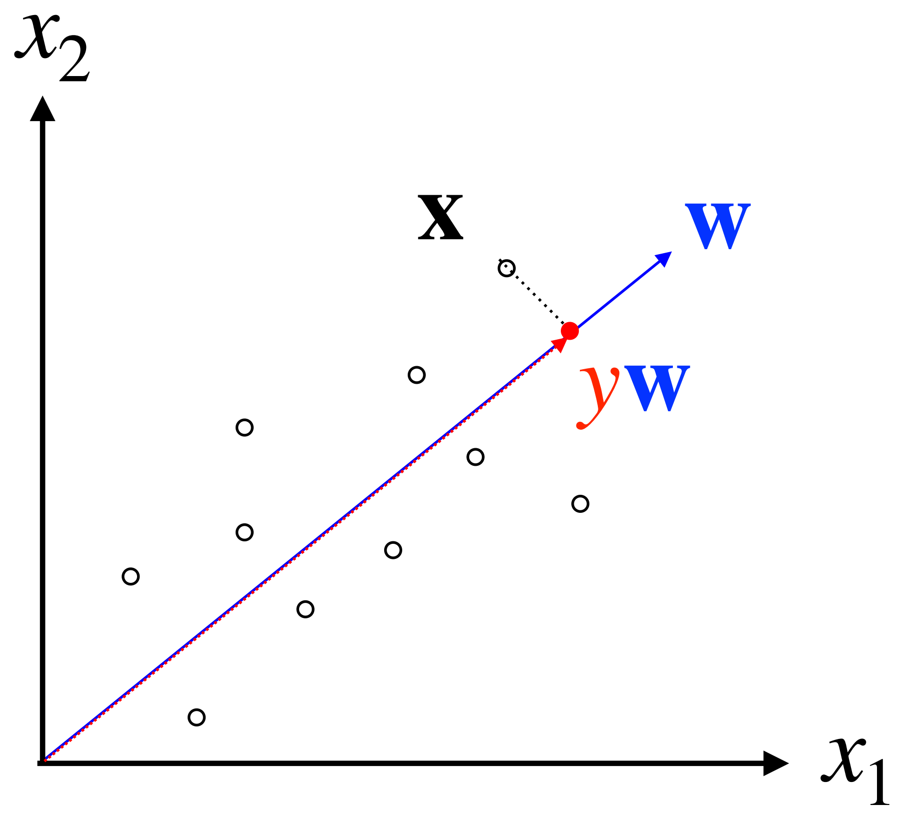
Goal: connect plasticity rules & learning objectives

Normative principle: maximize *information*

$$\min_{\mathbf{w}} \mathbb{E}\left[\|\mathbf{x} - \mathbf{w}\mathbf{w}^\top\mathbf{x}\|^2\right]$$

Oja's PCA model of a neuron

Oja 1982



Oja’s PCA model of a neuron

Normative principle: maximize *information*

$$\min_{\mathbf{w}} \mathbb{E}\left[\|\mathbf{x} - \mathbf{w}\mathbf{w}^{\top}\mathbf{x}\|^2\right]$$

$$y = \mathbf{w}^{\top}\mathbf{x}$$

$$\Delta w_i = \eta \left(\underbrace{y x_i}_{\text{Hebbian}} \underbrace{- y^2 w_i}_{\text{homeostatic}} \right)$$

Oja 1982

Oja’s PCA model of a neuron

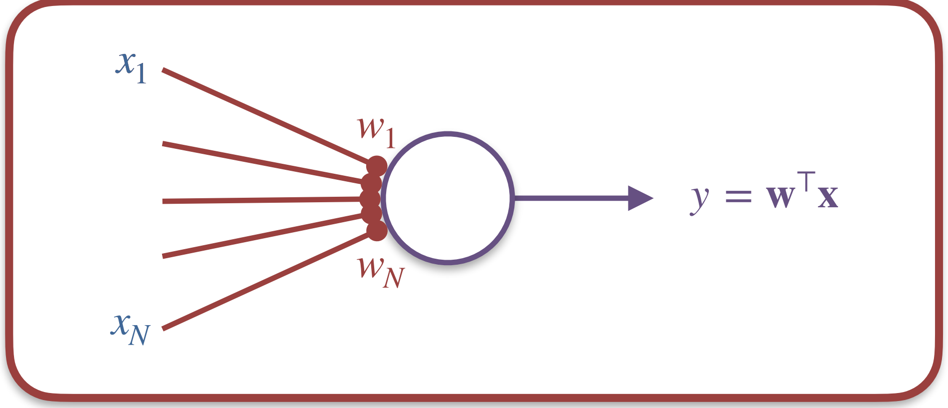
Oja 1982

Normative principle: maximize *information*

$$\min_{\mathbf{w}} \mathbb{E}\left[\|\mathbf{x} - \mathbf{w}\mathbf{w}^\top\mathbf{x}\|^2\right]$$

$$y = \mathbf{w}^\top\mathbf{x}$$

$$\Delta w_i = \eta \left(\underbrace{y x_i}_{\text{Hebbian}} \underbrace{- y^2 w_i}_{\text{homeostatic}}\right)$$



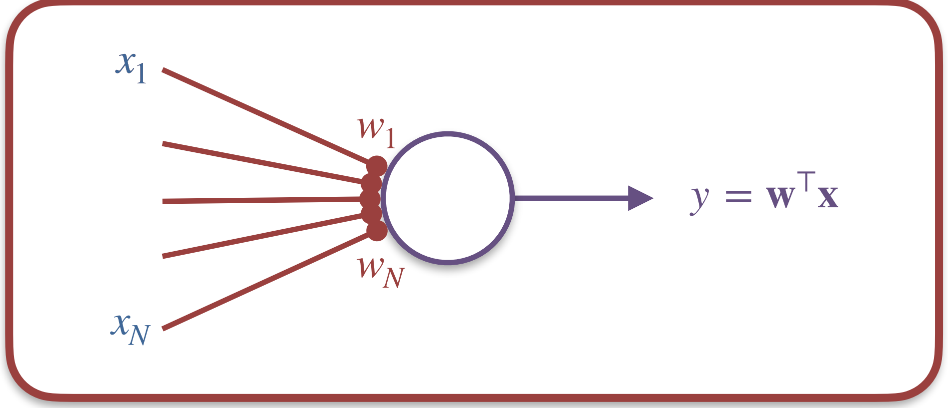
Oja's PCA model of a neuron

Normative principle: maximize *information*

$$\min_{\mathbf{w}} \mathbb{E}\left[\|\mathbf{x}-\mathbf{w}\mathbf{w}^\top\mathbf{x}\|^2\right]$$

$$y = \mathbf{w}^\top\mathbf{x}$$

$$\Delta w_i = \eta\left(\underbrace{y x_i}_{\text{Hebbian}} \underbrace{- y^2 w_i}_{\text{homeostatic}}\right)$$



Oja 1982

Neural algorithms wish list:

1. principled? maximizes info transmission (for Gaussian inputs)
2. sample efficient? matches info theoretic LB [Chou & Wang 2020]
3. resource efficient? local & online
4. free parameters? learning rate η
5. match data? predicts connectomic data [Chapochnikov et al. 2023]

Biological extensions of Oja's algorithm

Multichannel PCA (normative + local) [Pehlevan, Hu & Chklovskii 2015]

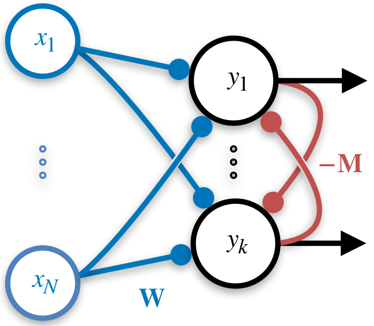
Manifold tiling [Sengupta et al. 2018]

Normative principle: maximize *information*

$$\min_{\{\mathbf{y}_t\}} \sum_{t=1}^{T} \sum_{t'=1}^{T} (\mathbf{y}_t^\top \mathbf{y}_{t'} - \mathbf{x}_t^\top \mathbf{x}_{t'})^2$$

Neural dynamics: $\mathbf{y} \leftarrow \mathbf{y} + \gamma(\mathbf{W}\mathbf{x} - \mathbf{M}\mathbf{y})$

$$\Delta\mathbf{W} \propto \underbrace{\mathbf{y}\mathbf{x}^\top}_{\text{Hebbian}} \underbrace{- \mathbf{W}}_{\text{homeostatic}} \qquad \Delta\mathbf{M} \propto \underbrace{\mathbf{y}\mathbf{y}^\top}_{\text{anti-Hebbian}} \underbrace{- \mathbf{M}}_{\text{homeostatic}}$$



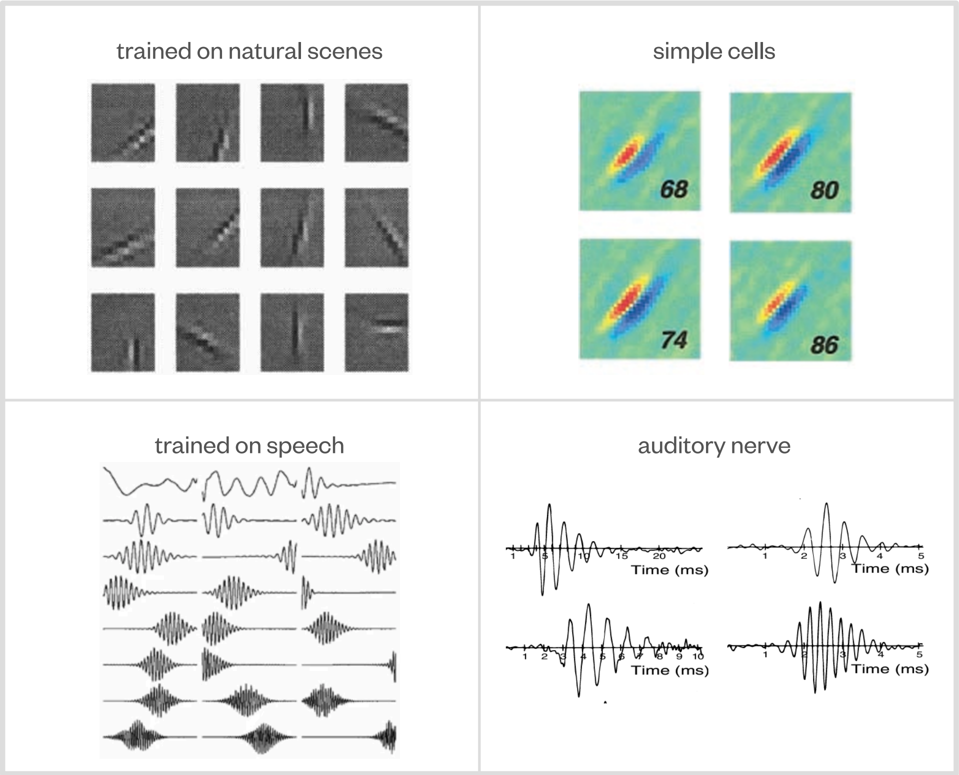
Limitation #1: Doesn’t account for other learning principles.

Independent Component Analysis (ICA)

Normative principle: relevant features are *independent*

optimized filters

measured receptive fields



[Olshausen & Field 1997; Bell & Sejnowski 1995, 1997; van Hateren & van der Schaaf 1998; Hubel & Wiesel 1959; Ringach 2002; Lewicki 2002; de Boer & de Jongh 1978]

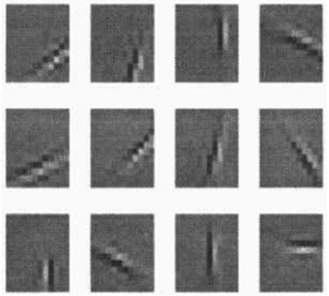
Independent Component Analysis (ICA)

Normative principle: relevant features are *independent*

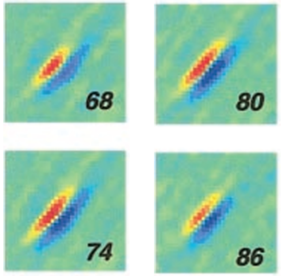
optimized filters

measured receptive fields

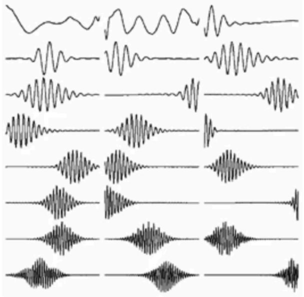
trained on natural scenes



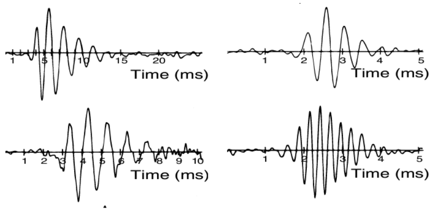
simple cells



trained on speech



auditory nerve



[Olshausen & Field 1997; Bell & Sejnowski 1995, 1997; van Hateren & van der Schaaf 1998; Hubel & Wiesel 1959; Ringach 2002; Lewicki 2002; de Boer & de Jongh 1978]

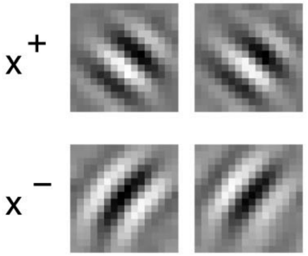
Slow Feature Analysis (SFA)

Normative principle: relevant features are *slow*

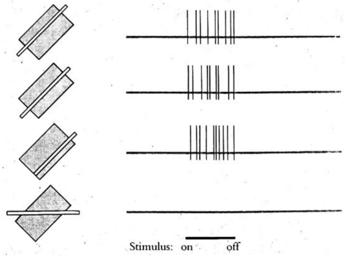
optimized filters

measured receptive fields

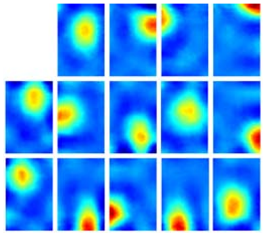
trained on natural scenes



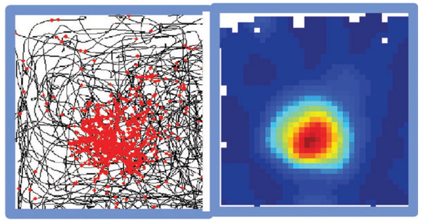
complex cells



trained on simulated visual inputs during virtual navigation



place cells



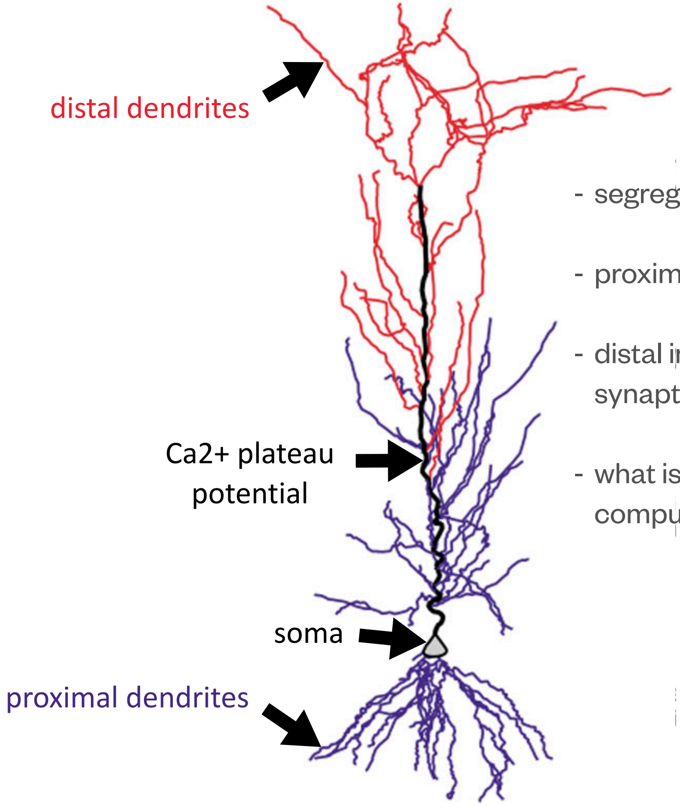
[Földiák 1991; Wiskott & Sejnowski 2002; Berkes et al. 2007; Franzius, Sprekeler & Wiskott 2007, Hubel & Wiesel 1968; O'Keefe & Dostrovsky 1971]

Limitation #1: Doesn’t account for other learning principles.

Limitation #2: Cannot explain multicompartmental neurons or non-Hebbian synaptic plasticity rules.

CA1 pyramidal neuron (axon omitted)

[Häusser & Mel 2003]

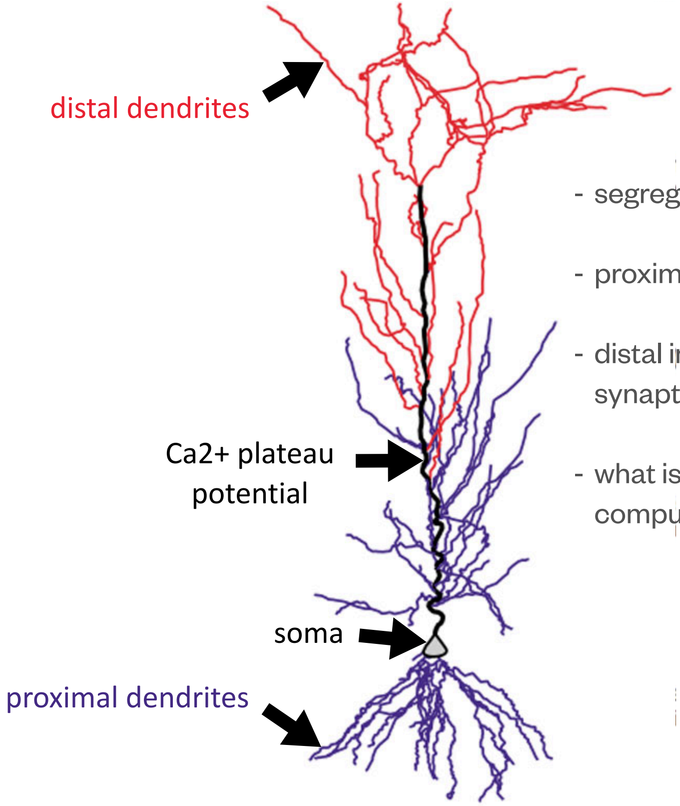


- segregated compartments
- proximal inputs: drive Na^+ spikes
- distal inputs: Ca^{2+} plateaus drive synaptic plasticity
- what is the pyramidal neuron computing?

Experimental: Schiller et al. 1997; Larkum et al. 1999; Takahashi & Magee 2009; Bittner et al. 2015; ... **Computational:** Körding & König 2001; Poirazi et al. 2003; Urbanczik & Senn 2014; Guergueiv et al. 2017; Whittington & Bogacz 2017; Sacramento et al. 2018; Milstein et al. 2021; ...

CA1 pyramidal neuron (axon omitted)

[Häusser & Mel 2003]

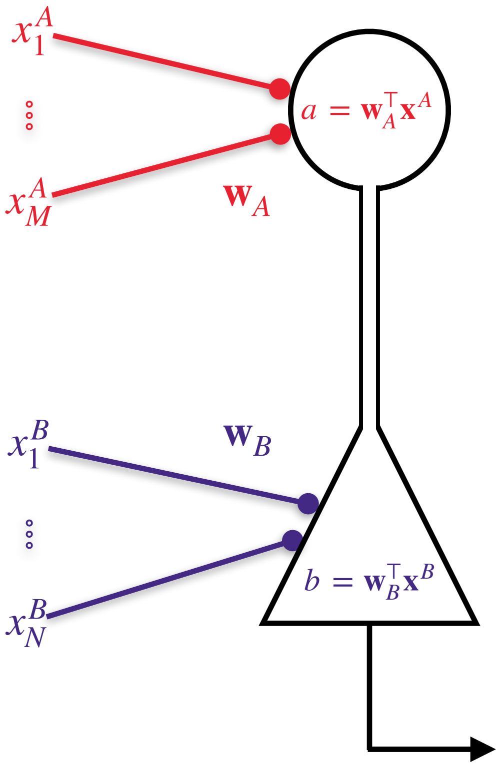


- segregated compartments
- proximal inputs: drive Na$^+$ spikes
- distal inputs: Ca^{2+} plateaus drive synaptic plasticity
- what is the pyramidal neuron computing?

Experimental: Schiller et al. 1997; Larkum et al. 1999; Takahashi & Magee 2009; Bittner et al. 2015; … **Computational:** Körding & König 2001; Poirazi et al. 2003; Urbanczik & Senn 2014; Guergueiv et al. 2017; Whittington & Bogacz 2017; Sacramento et al. 2018; Milstein et al. 2021; …

Canonical Correlation Analysis (CCA)

Normative principle: relevant features are *correlated*



$$y = a + b$$

Objective: maximize correlation between dendritic currents a and b

$$\max_{\mathbf{w}_A, \mathbf{w}_B} \mathbb{E}\left[(\mathbf{w}_A^\top \mathbf{x}^A)(\mathbf{w}_B^\top \mathbf{x}^B)\right]$$

subject to

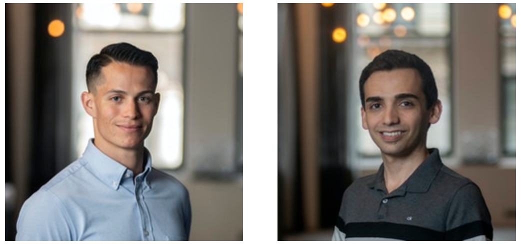
$$\mathbb{E}\left[(\mathbf{w}_A^\top \mathbf{x}^A)^2\right] = \mathbb{E}\left[(\mathbf{w}_B^\top \mathbf{x}^B)^2\right] = 1$$

Relation to:

- mutual information
- information bottleneck

[Hotelling 1936; Tishby, Pereira & Bialek 2000; Chechik et al. 2003; **Lipshutz** et al. 2021; Barreiro et al. 2024]

Goal: relate these learning principles (independence, slowness, correlation, etc.) to synaptic plasticity rules



A unified framework: symmetric generalized eigenvalue problems

$$\mathbf{A}\mathbf{w} = \lambda \mathbf{B}\mathbf{w}$$

$$\mathbf{A} = \mathbb{E}\left[\boldsymbol{\xi}_t \boldsymbol{\xi}_t^\top\right] \qquad \mathbf{B} = \mathbb{E}\left[\mathbf{B}_t\right]$$

Learning task	$\boldsymbol{\xi}_t$	$\mathbf{B}_t$
PCA	$\mathbf{x}_t$	$\mathbf{I}$
ICA	$\mathbf{x}_t$	$\lVert \mathbf{C}_X^{-1/2}\mathbf{x}_t \rVert^2 \mathbf{x}_t \mathbf{x}_t^\top$
SFA	$\mathbf{x}_t + \mathbf{x}_{t-1}$	$\mathbf{x}_t \mathbf{x}_t^\top$
CCA	$\begin{bmatrix} \mathbf{x}_t^A \\ \mathbf{x}_t^B \end{bmatrix}$	$\begin{bmatrix} \mathbf{x}_t^A \mathbf{x}_t^{A,\top} & \\ & \mathbf{x}_t^B \mathbf{x}_t^{B,\top} \end{bmatrix}$
contrastive PCA	$\delta_t \mathbf{x}_t$	$(1 - \delta_t)\mathbf{x}_t \mathbf{x}_t^\top$

Existing algorithms for solving generalized eigenvalue problems do not map onto biological NNs

[Arora et al. 2017, Bhatia et al. 2018]

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

$$\min_{\boldsymbol{\zeta}_t \in \mathbb{R}^k} \sum_{t=1}^{T} \sum_{t'=1}^{T} \left(\boldsymbol{\xi}_t^\top \mathbf{B}^{-1} \boldsymbol{\xi}_{t'} - \boldsymbol{\zeta}_t^\top \boldsymbol{\zeta}_{t'} \right)^2$$

opt. $\boldsymbol{\zeta}_t$ = proj. of $\boldsymbol{\xi}_t$ onto k-eigensubspace

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

$$\min_{\zeta_t \in \mathbb{R}^k} \sum_{t=1}^{T} \sum_{t'=1}^{T} \left(\underbrace{\boldsymbol{\xi}_t^\top \mathbf{B}^{-1} \boldsymbol{\xi}_{t'}}_{\text{input similarity}} - \zeta_t^\top \zeta_{t'} \right)^2$$

opt. ζ_t = proj. of $\boldsymbol{\xi}_t$ onto k-eigensubspace

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

$$\min_{\zeta_t \in \mathbb{R}^k} \sum_{t=1}^{T} \sum_{t'=1}^{T} \Big(\underset{\text{input similarity}}{\boldsymbol{\xi}_t^\top \mathbf{B}^{-1} \boldsymbol{\xi}_{t'}} - \underset{\text{output similarity}}{\boldsymbol{\zeta}_t^\top \boldsymbol{\zeta}_{t'}} \Big)^2$$

opt. $\boldsymbol{\zeta}_t$ = proj. of $\boldsymbol{\xi}_t$ onto k-eigensubspace

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

$$\min_{\boldsymbol{\zeta}_t\in\mathbb{R}^k}\sum_{t=1}^{T}\sum_{t'=1}^{T}\Big(\underbrace{\boldsymbol{\xi}_t^\top\mathbf{B}^{-1}\boldsymbol{\xi}_{t'}}_{\text{input similarity}}-\underbrace{\boldsymbol{\zeta}_t^\top\boldsymbol{\zeta}_{t'}}_{\text{output similarity}}\Big)^2$$

$$\min_{\boldsymbol{\zeta}_t}\sum_{t=1}^{T}\left[-2\boldsymbol{\zeta}_t^\top\left(\sum_{t'=1}^{T}\boldsymbol{\zeta}_{t'}\boldsymbol{\xi}_{t'}^\top\mathbf{B}^{-1}\right)\boldsymbol{\xi}_t+\boldsymbol{\zeta}_t^\top\left(\sum_{t'=1}^{T}\boldsymbol{\zeta}_{t'}\boldsymbol{\zeta}_{t'}^\top\right)\boldsymbol{\zeta}_t\right]$$

opt. $\boldsymbol{\zeta}_t$ = proj. of $\boldsymbol{\xi}_t$ onto k-eigensubspace

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

$$\min_{\zeta_t \in \mathbb{R}^k} \sum_{t=1}^{T} \sum_{t'=1}^{T} \Big(\underset{\text{input similarity}}{\boldsymbol{\xi}_t^\top \mathbf{B}^{-1} \boldsymbol{\xi}_{t'}} - \underset{\text{output similarity}}{\zeta_t^\top \zeta_{t'}} \Big)^2$$

opt. ζ_t = proj. of $\boldsymbol{\xi}_t$ onto k-eigensubspace

$$\min_{\zeta_t} \sum_{t=1}^{T} \left[-2\zeta_t^\top \underset{\text{input-output correlation}}{\left(\sum_{t'=1}^{T} \zeta_{t'} \boldsymbol{\xi}_{t'}^\top \mathbf{B}^{-1} \right)} \boldsymbol{\xi}_t + \zeta_t^\top \left(\sum_{t'=1}^{T} \zeta_{t'} \zeta_{t'}^\top \right) \zeta_t \right]$$

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

$$\min_{\zeta_t\in\mathbb{R}^k}\sum_{t=1}^{T}\sum_{t'=1}^{T}\left(\boldsymbol{\xi}_t^\top\mathbf{B}^{-1}\boldsymbol{\xi}_{t'}-\zeta_t^\top\zeta_{t'}\right)^2$$

input similarity

output similarity

opt. ζ_t = proj. of $\boldsymbol{\xi}_t$ onto k-eigensubspace

$$\min_{\zeta_t}\sum_{t=1}^{T}\left[-2\zeta_t^\top\left(\sum_{t'=1}^{T}\zeta_{t'}\boldsymbol{\xi}_{t'}^\top\mathbf{B}^{-1}\right)\boldsymbol{\xi}_t+\zeta_t^\top\left(\sum_{t'=1}^{T}\zeta_{t'}\zeta_{t'}^\top\right)\zeta_t\right]$$

input-output correlation

output-output correlation

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

$$\min_{\zeta_t \in \mathbb{R}^k} \sum_{t=1}^{T} \sum_{t'=1}^{T} \left(\boldsymbol{\xi}_t^\top \mathbf{B}^{-1} \boldsymbol{\xi}_{t'} - \zeta_t^\top \zeta_{t'} \right)^2$$

input similarity

output similarity

opt. ζ_t = proj. of $\boldsymbol{\xi}_t$ onto k-eigensubspace

$$\min_{\zeta_t} \sum_{t=1}^{T} \left[-2\zeta_t^\top \left(\sum_{t'=1}^{T} \zeta_{t'} \boldsymbol{\xi}_{t'}^\top \mathbf{B}^{-1} \right) \boldsymbol{\xi}_t + \zeta_t^\top \left(\sum_{t'=1}^{T} \zeta_{t'} \zeta_{t'}^\top \right) \zeta_t \right]$$

input-output correlation

output-output correlation

$$\min_{\mathbf{W}} \max_{\mathbf{M}} \frac{1}{T} \sum_{t=1}^{T} \min_{\zeta_t} \ell(\mathbf{W}, \mathbf{M}, \boldsymbol{\xi}_t, \mathbf{B}_t, \zeta_t)$$

$$\ell(\mathbf{W}, \mathbf{M}, \boldsymbol{\xi}_t, \mathbf{B}_t, \zeta_t) = -4\zeta_t^\top \mathbf{W} \boldsymbol{\xi}_t + 2\zeta_t^\top \mathbf{M} \zeta_t + 2\mathrm{Tr}(\mathbf{W}\mathbf{B}_t\mathbf{W}^\top) - \mathrm{Tr}(\mathbf{M}^2)$$

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

$$\min_{\zeta_t\in\mathbb{R}^k}\sum_{t=1}^{T}\sum_{t'=1}^{T}\left(\boldsymbol{\xi}_t^\top\mathbf{B}^{-1}\boldsymbol{\xi}_{t'}-\zeta_t^\top\zeta_{t'}\right)^2$$

input similarity

output similarity

opt. ζ_t = proj. of $\boldsymbol{\xi}_t$ onto k-eigensubspace

$$\min_{\zeta_t}\sum_{t=1}^{T}\left[-2\zeta_t^\top\left(\sum_{t'=1}^{T}\zeta_{t'}\boldsymbol{\xi}_{t'}^\top\mathbf{B}^{-1}\right)\boldsymbol{\xi}_t+\zeta_t^\top\left(\sum_{t'=1}^{T}\zeta_{t'}\zeta_{t'}^\top\right)\zeta_t\right]$$

input-output correlation

output-output correlation

$$\min_{\mathbf{W}}\max_{\mathbf{M}}\frac{1}{T}\sum_{t=1}^{T}\min_{\zeta_t}\ell(\mathbf{W},\mathbf{M},\boldsymbol{\xi}_t,\mathbf{B}_t,\zeta_t)$$

$$\ell(\mathbf{W},\mathbf{M},\boldsymbol{\xi}_t,\mathbf{B}_t,\zeta_t)=-4\zeta_t^\top\mathbf{W}\boldsymbol{\xi}_t+2\zeta_t^\top\mathbf{M}\zeta_t+2\mathrm{Tr}(\mathbf{W}\mathbf{B}_t\mathbf{W}^\top)-\mathrm{Tr}(\mathbf{M}^2)$$

feedforward weights

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

$$\min_{\zeta_t\in\mathbb{R}^k}\sum_{t=1}^{T}\sum_{t'=1}^{T}\Big(\underbrace{\boldsymbol{\xi}_t^\top\mathbf{B}^{-1}\boldsymbol{\xi}_{t'}}_{\text{input similarity}}-\underbrace{\zeta_t^\top\zeta_{t'}}_{\text{output similarity}}\Big)^2$$

opt. ζ_t = proj. of $\boldsymbol{\xi}_t$ onto k-eigensubspace

$$\min_{\zeta_t}\sum_{t=1}^{T}\left[-2\zeta_t^\top\underbrace{\left(\sum_{t'=1}^{T}\zeta_{t'}\boldsymbol{\xi}_{t'}^\top\mathbf{B}^{-1}\right)}_{\text{input-output correlation}}\boldsymbol{\xi}_t+\zeta_t^\top\underbrace{\left(\sum_{t'=1}^{T}\zeta_{t'}\zeta_{t'}^\top\right)}_{\text{output-output correlation}}\zeta_t\right]$$

$$\min_{\mathbf{W}}\max_{\mathbf{M}}\frac{1}{T}\sum_{t=1}^{T}\min_{\zeta_t}\ell(\mathbf{W},\mathbf{M},\boldsymbol{\xi}_t,\mathbf{B}_t,\zeta_t)$$

$$\ell(\mathbf{W},\mathbf{M},\boldsymbol{\xi}_t,\mathbf{B}_t,\zeta_t)=-4\zeta_t^\top\underbrace{\mathbf{W}}_{\text{feedforward weights}}\boldsymbol{\xi}_t+2\zeta_t^\top\underbrace{\mathbf{M}}_{\text{recurrent weights}}\zeta_t+2\mathrm{Tr}(\mathbf{W}\mathbf{B}_t\mathbf{W}^\top)-\mathrm{Tr}(\mathbf{M}^2)$$

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

$$\min_{\zeta_t\in\mathbb{R}^k}\sum_{t=1}^{T}\sum_{t'=1}^{T}\left(\boldsymbol{\xi}_t^\top\mathbf{B}^{-1}\boldsymbol{\xi}_{t'}-\zeta_t^\top\zeta_{t'}\right)^2$$

input

output

imilarity

ubspace

Optimizing ζ_t over $\mathbb{R}_+^k$ leads to Nonnegative Matrix Factorization

$$\min_{\zeta_t}\sum_{t=1}^{T}\left[-2\zeta_t^\top\left(\sum_{t'=1}^{T}\zeta_{t'}\boldsymbol{\xi}_{t'}^\top\mathbf{B}^{-1}\right)\boldsymbol{\xi}_t+\zeta_t^\top\left(\sum_{t'=1}^{T}\zeta_{t'}\zeta_{t'}^\top\right)\zeta_t\right]$$

input-output correlation

output-output correlation

$$\min_{\mathbf{W}}\max_{\mathbf{M}}\frac{1}{T}\sum_{t=1}^{T}\min_{\zeta_t}\ell(\mathbf{W},\mathbf{M},\boldsymbol{\xi}_t,\mathbf{B}_t,\zeta_t)$$

$$\ell(\mathbf{W},\mathbf{M},\boldsymbol{\xi}_t,\mathbf{B}_t,\zeta_t)=-4\zeta_t^\top\mathbf{W}\boldsymbol{\xi}_t+2\zeta_t^\top\mathbf{M}\zeta_t+2\mathrm{Tr}(\mathbf{W}\mathbf{B}_t\mathbf{W}^\top)-\mathrm{Tr}(\mathbf{M}^2)$$

feedforward weights

recurrent weights

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

$$\min_{\mathbf{W}} \max_{\mathbf{M}} \frac{1}{T} \sum_{t=1}^{T} \min_{\zeta_t} \ell(\mathbf{W}, \mathbf{M}, \boldsymbol{\xi}_t, \mathbf{B}_t, \zeta_t)$$

$$\ell(\mathbf{W}, \mathbf{M}, \boldsymbol{\xi}_t, \mathbf{B}_t, \zeta_t) = -4\zeta_t^\top \mathbf{W} \boldsymbol{\xi}_t + 2\zeta_t^\top \mathbf{M} \zeta_t + 2\mathrm{Tr}(\mathbf{W}\mathbf{B}_t\mathbf{W}^\top) - \mathrm{Tr}(\mathbf{M}^2)$$

feedforward weights

recurrent weights

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

$$\min_{\mathbf{W}} \max_{\mathbf{M}} \frac{1}{T} \sum_{t=1}^{T} \min_{\boldsymbol{\zeta}_t} \ell(\mathbf{W}, \mathbf{M}, \boldsymbol{\xi}_t, \mathbf{B}_t, \boldsymbol{\zeta}_t)$$

$$\ell(\mathbf{W}, \mathbf{M}, \boldsymbol{\xi}_t, \mathbf{B}_t, \boldsymbol{\zeta}_t) = -4\boldsymbol{\zeta}_t^\top \underset{\text{feedforward weights}}{\mathbf{W}} \boldsymbol{\xi}_t + 2\boldsymbol{\zeta}_t^\top \underset{\text{recurrent weights}}{\mathbf{M}} \boldsymbol{\zeta}_t + 2\mathrm{Tr}(\mathbf{W}\mathbf{B}_t\mathbf{W}^\top) - \mathrm{Tr}(\mathbf{M}^2)$$

```
input $\{(\boldsymbol{\xi}_t, \mathbf{B}_t)\}$; parameters $\gamma > 0$ and $0 < \eta < \tau$
initialize $\mathbf{W} \in \mathbb{R}^{k \times n}$ and $\mathbf{M} \in \mathbb{S}^k_{++}$
for $t = 1, 2, \ldots$ do
    repeat
        $\boldsymbol{\zeta}_t \leftarrow \boldsymbol{\zeta}_t + \gamma(\mathbf{W}\boldsymbol{\xi}_t - \mathbf{M}\boldsymbol{\zeta}_t)$    // recurrent neural dynamics
    until convergence
    $\mathbf{W} \leftarrow \mathbf{W} + 2\eta(\boldsymbol{\zeta}_t\boldsymbol{\xi}_t^\top - \mathbf{W}\mathbf{B}_t)$    // feedforward synaptic updates
    $\mathbf{M} \leftarrow \mathbf{M} + \frac{\eta}{\tau}(\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t^\top - \mathbf{M})$    // recurrent synaptic updates
end for
```

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

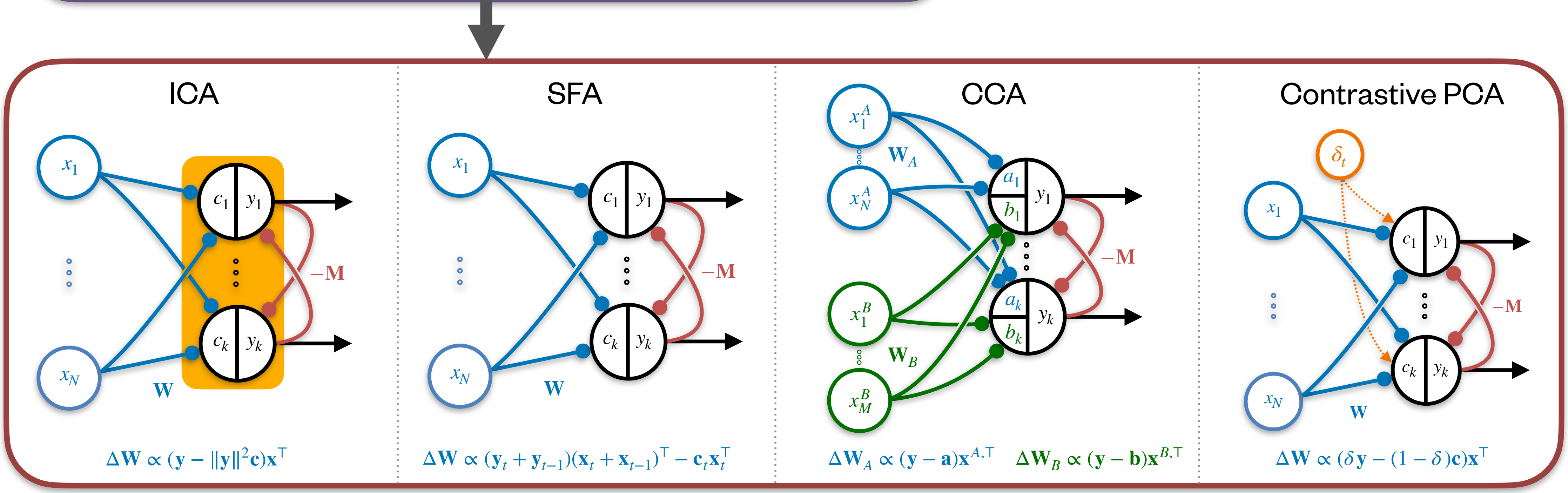
```
input $\{(\boldsymbol{\xi}_t, \mathbf{B}_t)\}$; parameters $\gamma > 0$ and $0 < \eta < \tau$
initialize $\mathbf{W} \in \mathbb{R}^{k \times n}$ and $\mathbf{M} \in \mathbb{S}^k_{++}$
for $t = 1, 2, \ldots$ do
    repeat
        $\boldsymbol{\zeta}_t \leftarrow \boldsymbol{\zeta}_t + \gamma(\mathbf{W}\boldsymbol{\xi}_t - \mathbf{M}\boldsymbol{\zeta}_t)$    // recurrent neural dynamics
    until convergence
    $\mathbf{W} \leftarrow \mathbf{W} + 2\eta(\boldsymbol{\zeta}_t \boldsymbol{\xi}_t^\top - \mathbf{W}\mathbf{B}_t)$    // feedforward synaptic updates
    $\mathbf{M} \leftarrow \mathbf{M} + \frac{\eta}{\tau}(\boldsymbol{\zeta}_t \boldsymbol{\zeta}_t^\top - \mathbf{M})$    // recurrent synaptic updates
end for
```

Learning task	$\boldsymbol{\xi}_t$	$\mathbf{B}_t$
PCA	$\mathbf{x}_t$	$\mathbf{I}$
ICA	$\mathbf{x}_t$	$\Vert\mathbf{C}_X^{-1/2}\mathbf{x}_t\Vert^2\mathbf{x}_t\mathbf{x}_t^\top$
SFA	$\mathbf{x}_t + \mathbf{x}_{t-1}$	$\mathbf{x}_t\mathbf{x}_t^\top$
CCA	$\begin{bmatrix}\mathbf{x}_t^A \\ \mathbf{x}_t^B\end{bmatrix}$	$\begin{bmatrix}\mathbf{x}_t^A\mathbf{x}_t^{A,\top} & \\ & \mathbf{x}_t^B\mathbf{x}_t^{B,\top}\end{bmatrix}$
contrastive PCA	$\delta_t\mathbf{x}_t$	$(1 - \delta_t)\mathbf{x}_t\mathbf{x}_t^\top$

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

```
input $\{(\boldsymbol{\xi}_t, \mathbf{B}_t)\}$; parameters $\gamma > 0$ and $0 < \eta < \tau$
initialize $\mathbf{W} \in \mathbb{R}^{k \times n}$ and $\mathbf{M} \in \mathbb{S}^k_{++}$
for $t = 1, 2, \ldots$ do
    repeat
        $\boldsymbol{\zeta}_t \leftarrow \boldsymbol{\zeta}_t + \gamma(\mathbf{W}\boldsymbol{\xi}_t - \mathbf{M}\boldsymbol{\zeta}_t)$    // recurrent neural dynamics
    until convergence
    $\mathbf{W} \leftarrow \mathbf{W} + 2\eta(\boldsymbol{\zeta}_t \boldsymbol{\xi}_t^\top - \mathbf{W}\mathbf{B}_t)$    // feedforward synaptic updates
    $\mathbf{M} \leftarrow \mathbf{M} + \frac{\eta}{\tau}(\boldsymbol{\zeta}_t \boldsymbol{\zeta}_t^\top - \mathbf{M})$    // recurrent synaptic updates
end for
```

Learning task	$\boldsymbol{\xi}_t$	$\mathbf{B}_t$
PCA	$\mathbf{x}_t$	$\mathbf{I}$
ICA	$\mathbf{x}_t$	$\lVert \mathbf{C}_X^{-1/2}\mathbf{x}_t \rVert^2 \mathbf{x}_t \mathbf{x}_t^\top$
SFA	$\mathbf{x}_t + \mathbf{x}_{t-1}$	$\mathbf{x}_t \mathbf{x}_t^\top$
CCA	$\begin{bmatrix} \mathbf{x}_t^A \\ \mathbf{x}_t^B \end{bmatrix}$	$\begin{bmatrix} \mathbf{x}_t^A \mathbf{x}_t^{A,\top} & \\ & \mathbf{x}_t^B \mathbf{x}_t^{B,\top} \end{bmatrix}$
contrastive PCA	$\delta_t \mathbf{x}_t$	$(1 - \delta_t)\mathbf{x}_t \mathbf{x}_t^\top$



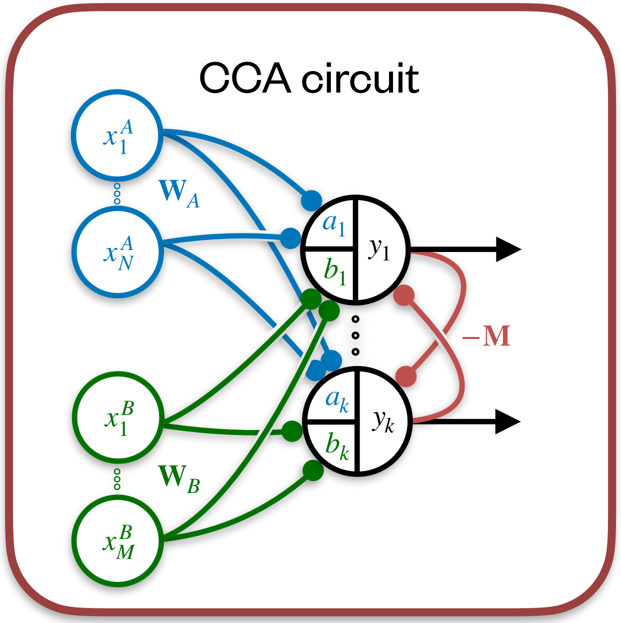
$\Delta \mathbf{W} \propto (\mathbf{y} - \lVert \mathbf{y} \rVert^2 \mathbf{c})\mathbf{x}^\top$

$\Delta \mathbf{W} \propto (\mathbf{y}_t + \mathbf{y}_{t-1})(\mathbf{x}_t + \mathbf{x}_{t-1})^\top - \mathbf{c}_t \mathbf{x}_t^\top$

$\Delta \mathbf{W}_A \propto (\mathbf{y} - \mathbf{a})\mathbf{x}^{A,\top}$ $\Delta \mathbf{W}_B \propto (\mathbf{y} - \mathbf{b})\mathbf{x}^{B,\top}$

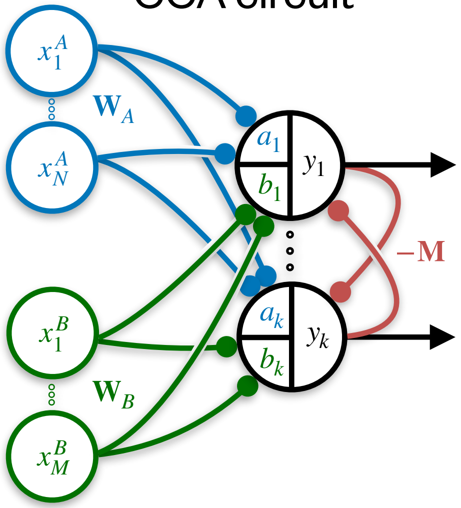
$\Delta \mathbf{W} \propto (\delta \mathbf{y} - (1 - \delta)\mathbf{c})\mathbf{x}^\top$

Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

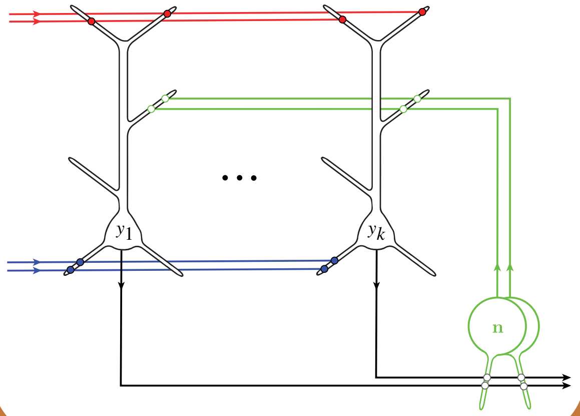


Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

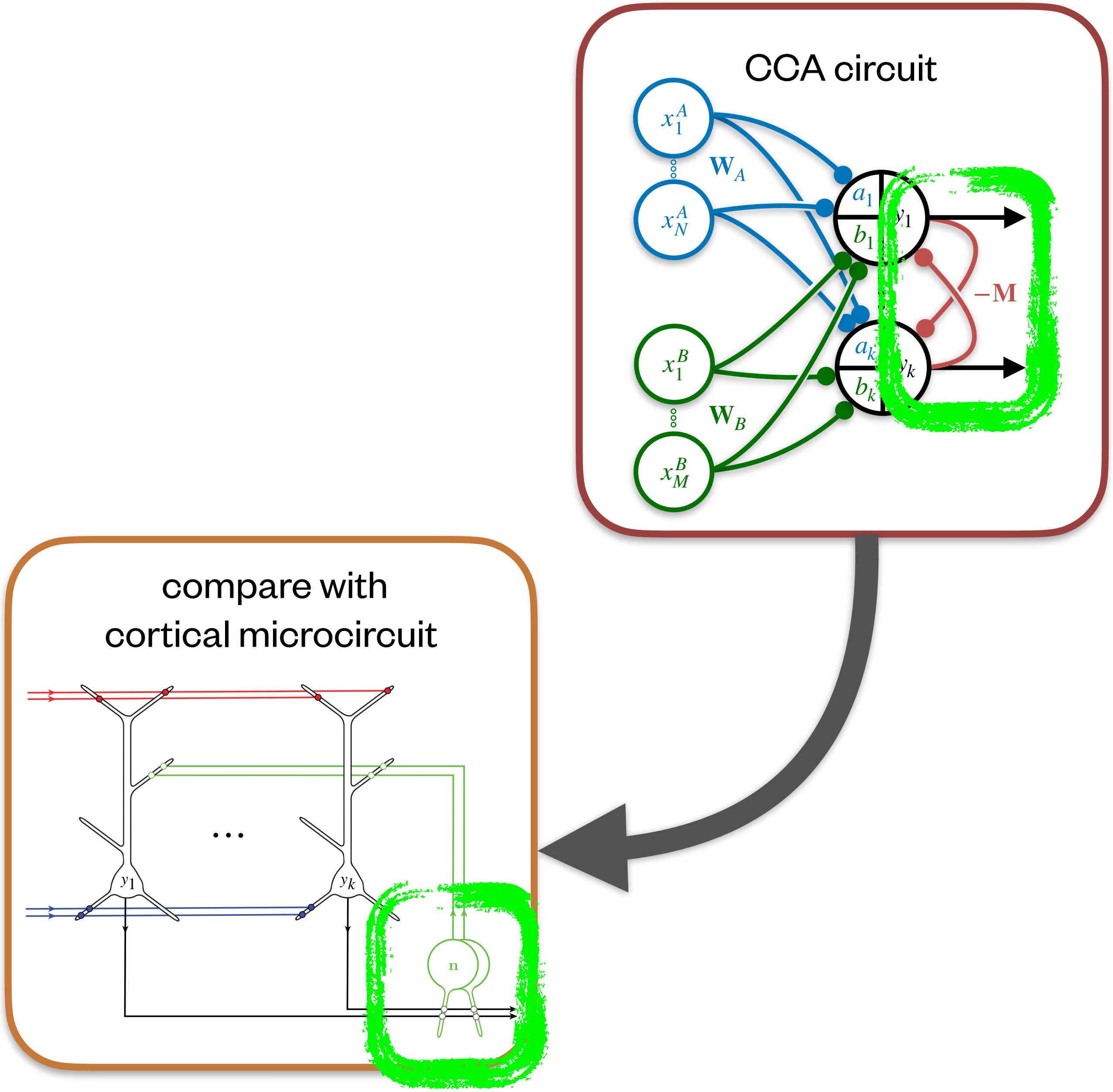
CCA circuit



compare with
cortical microcircuit



Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023


CCA circuit

compare with
cortical microcircuit

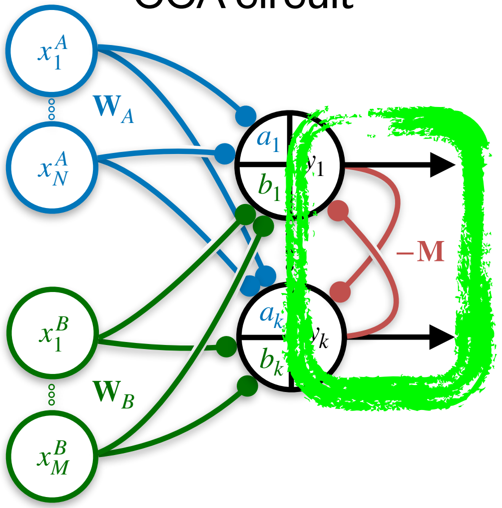
Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

CCA + output decorrelation

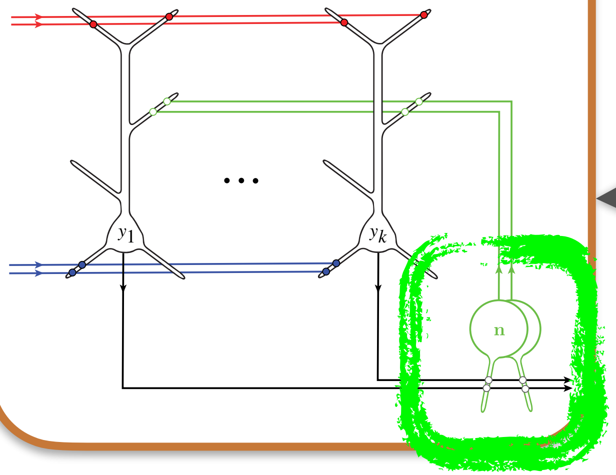
$$\min_{\mathbf{W}_a,\mathbf{W}_b}\max_{\mathbf{W}_n}\sum_{t=1}^{T}\min_{\mathbf{y}_t}\max_{\mathbf{n}_t}\ell(\mathbf{W}_a,\mathbf{W}_b,\mathbf{Q},\mathbf{x}^A,\mathbf{x}^B,\mathbf{y},\mathbf{n})$$

$$\begin{aligned}\ell(\mathbf{W}_a,\mathbf{W}_b,\mathbf{Q},\mathbf{x}^A,\mathbf{x}^B,\mathbf{y},\mathbf{n}) = \mathbf{y}^\top\mathbf{y}-\mathbf{n}^\top\mathbf{n}&\\ -2\mathbf{y}^\top\mathbf{W}_a\mathbf{x}^A+\mathrm{Tr}\left(\mathbf{W}_a\mathbf{W}_a^\top\right)&\\ -2\mathbf{y}^\top\mathbf{W}_b\mathbf{x}^B+\mathrm{Tr}\left(\mathbf{W}_b\mathbf{W}_b^\top\right)&\\ +2\mathbf{n}^\top\mathbf{Q}\mathbf{y}-\mathrm{Tr}\left(\mathbf{Q}\mathbf{Q}^\top\right)&\end{aligned}$$

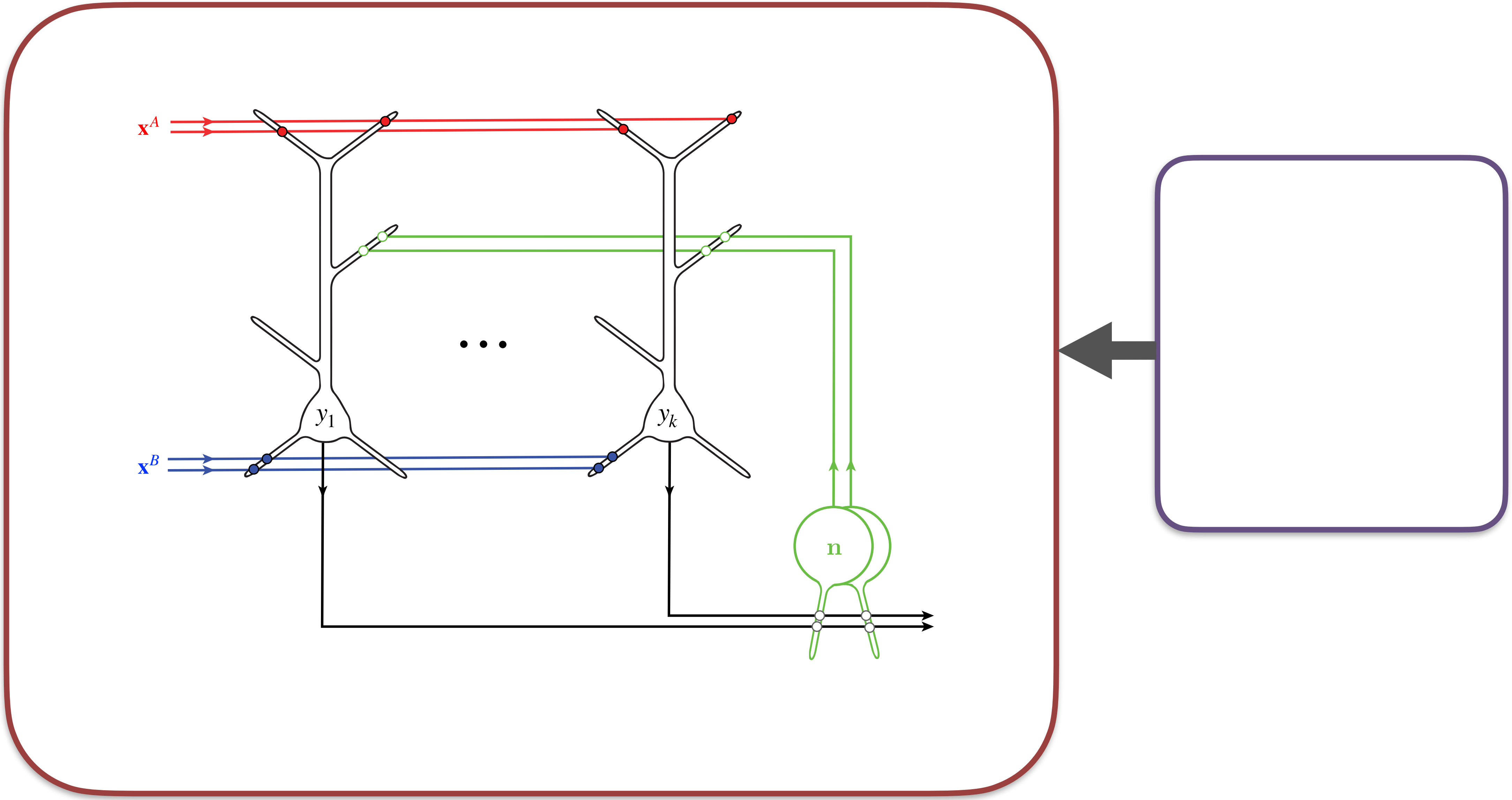
CCA circuit



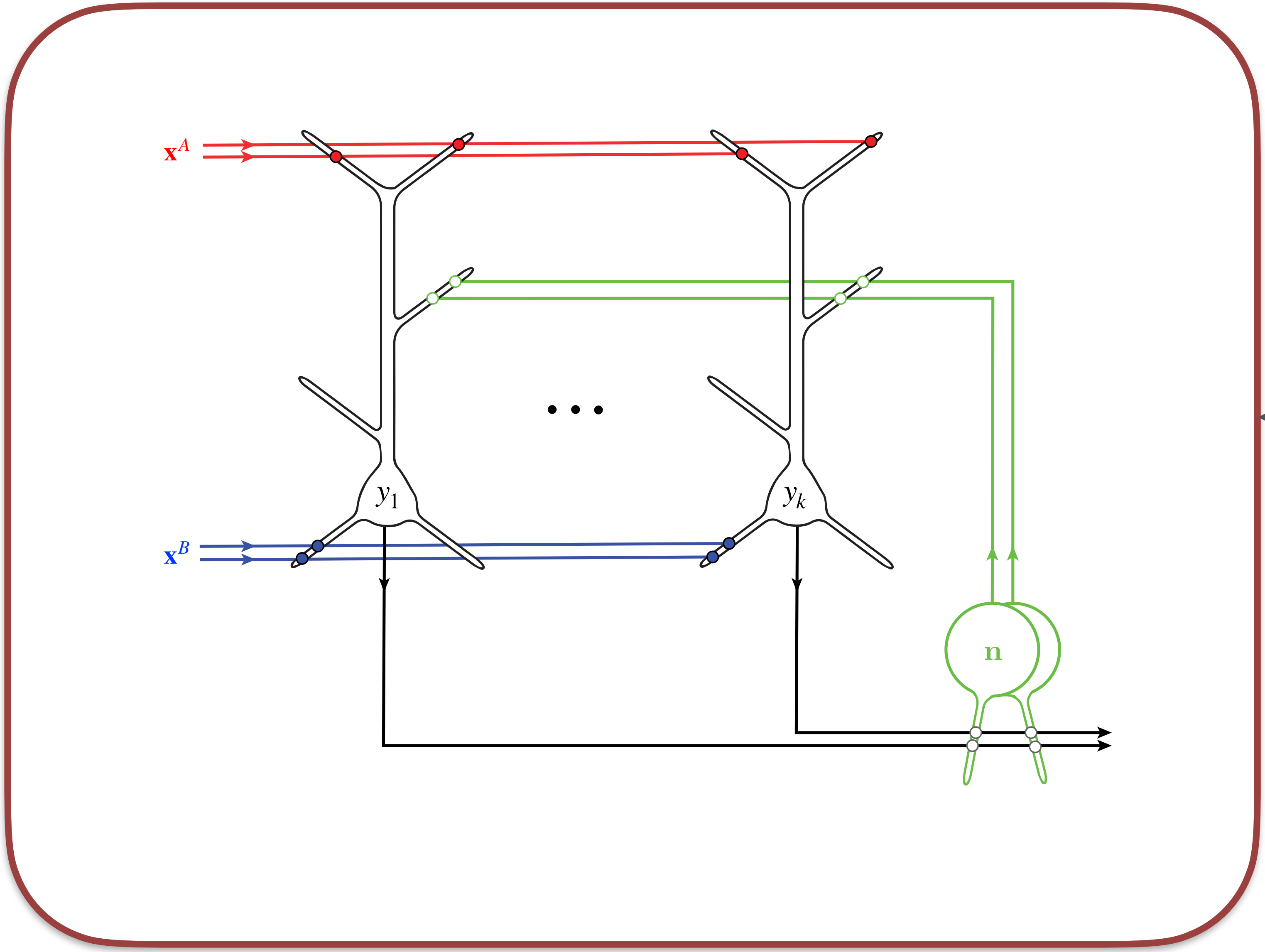
compare with cortical microcircuit



Lipshutz*, Bahroun*, Golkar* et al. *PRX Life* 2023

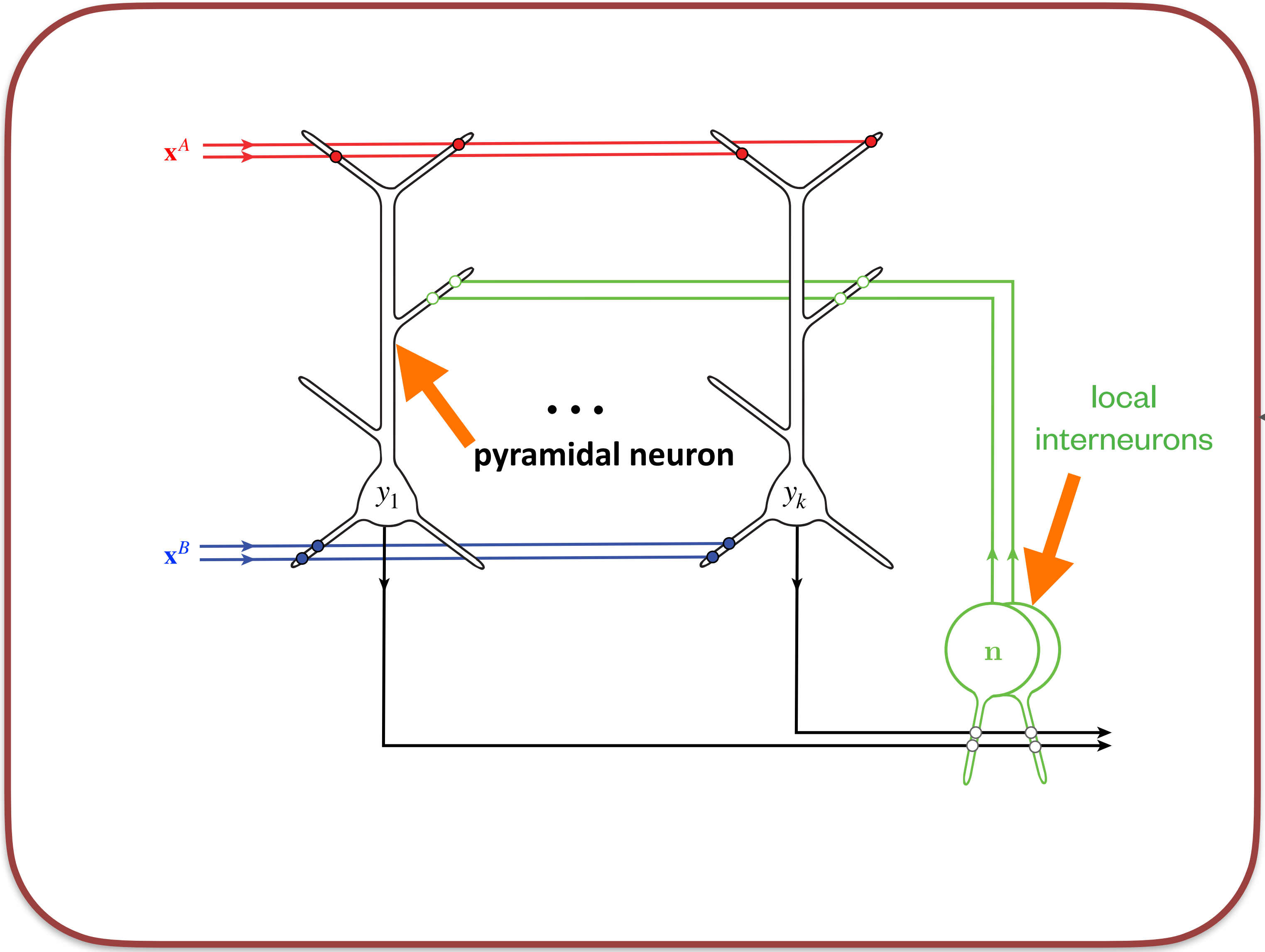


Golkar, Lipshutz et al. *NeurIPS* 2020; Lipshutz et al. *Neural Computation* 2021



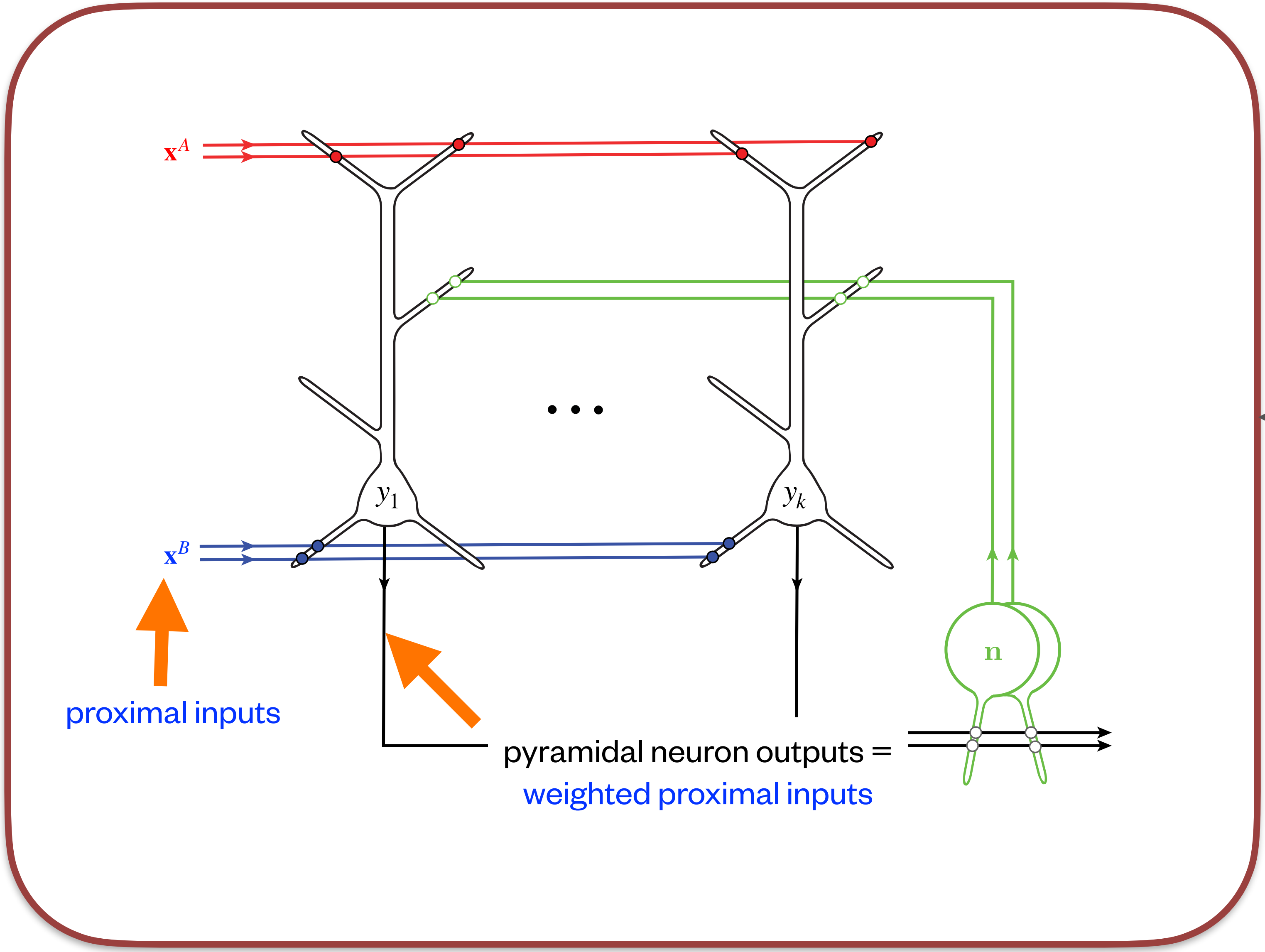
CCA algorithm

Golkar, Lipshutz et al. *NeurIPS* 2020; Lipshutz et al. *Neural Computation* 2021



CCA algorithm

Golkar, Lipshutz et al. *NeurIPS* 2020; Lipshutz et al. *Neural Computation* 2021



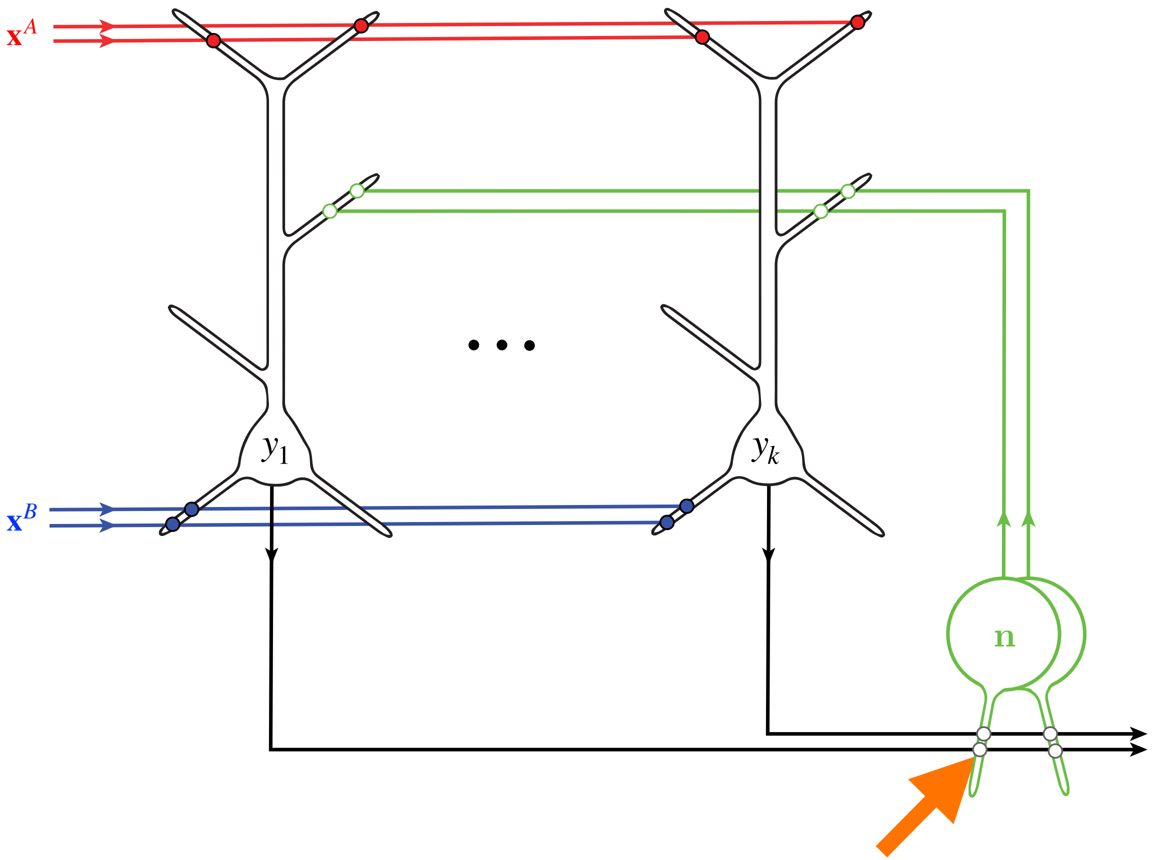
proximal inputs

pyramidal neuron outputs =
weighted proximal inputs

CCA algorithm

1. $\mathbf{y} = \mathbf{W}_B \mathbf{x}^B$

Golkar, Lipshutz et al. *NeurIPS* 2020; Lipshutz et al. *Neural Computation* 2021

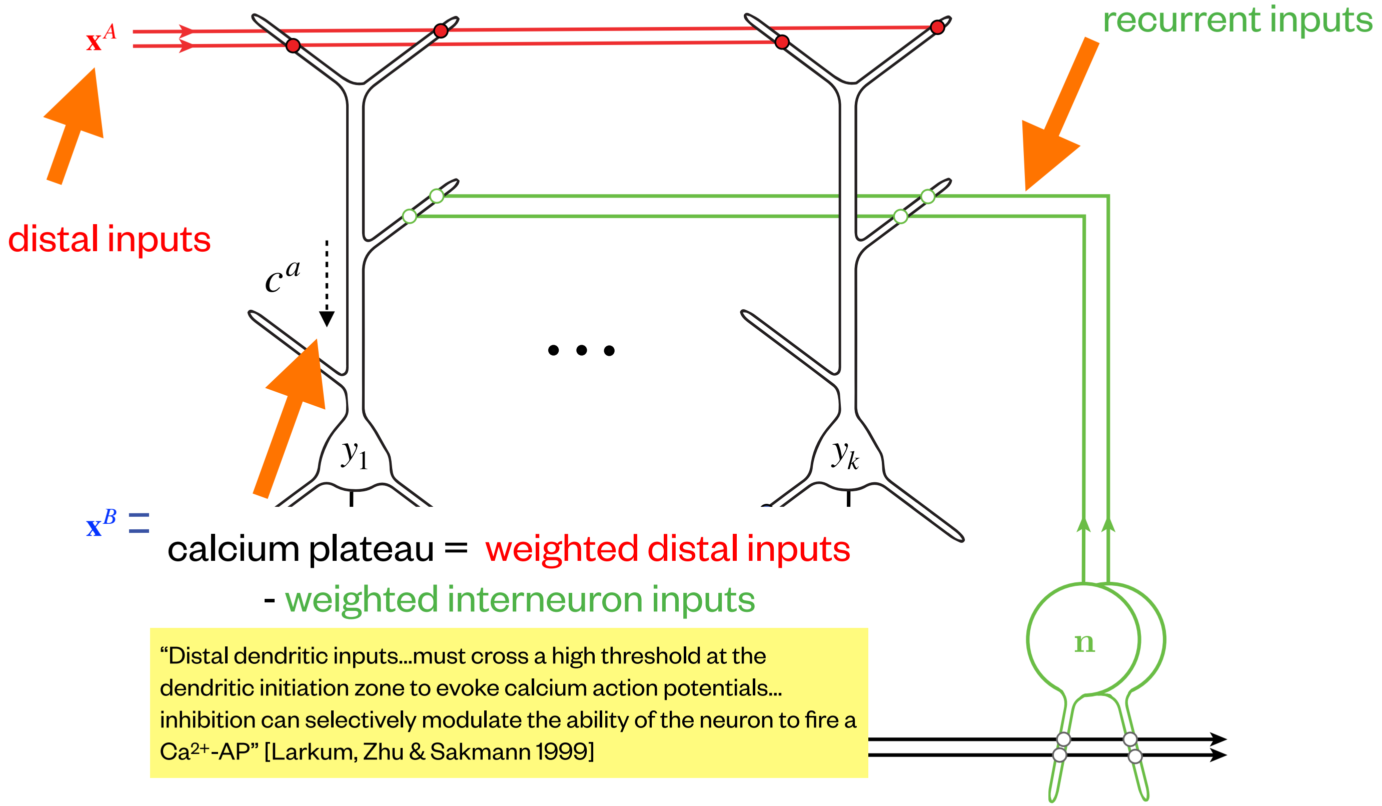


input = weighted pyramidal neuron activities

CCA algorithm

1. $\mathbf{y} = \mathbf{W}_B\mathbf{x}^B$
2. $\mathbf{n} = \mathbf{Q}^\top\mathbf{y}$

Golkar, Lipshutz et al. *NeurIPS* 2020; Lipshutz et al. *Neural Computation* 2021



$\mathbf{x}^B$ = calcium plateau = weighted distal inputs

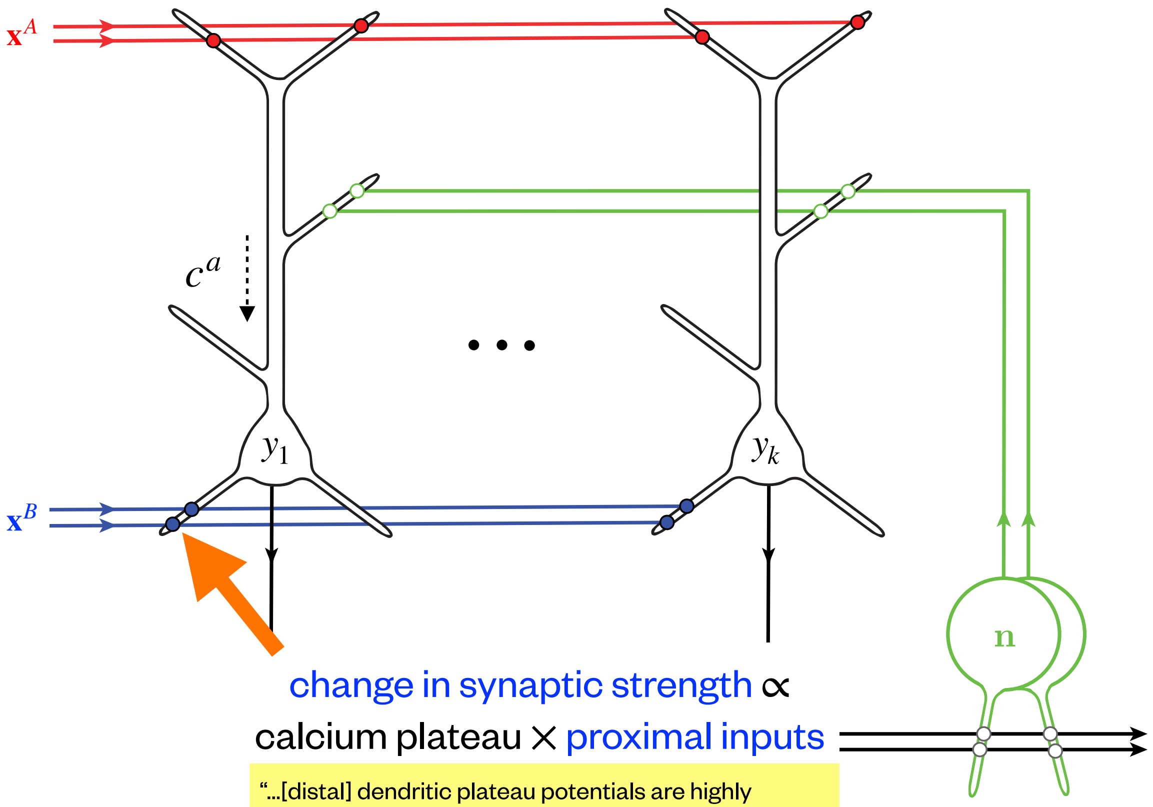
- weighted interneuron inputs

> “Distal dendritic inputs…must cross a high threshold at the dendritic initiation zone to evoke calcium action potentials… inhibition can selectively modulate the ability of the neuron to fire a Ca^{2+}-AP” [Larkum, Zhu & Sakmann 1999]

CCA algorithm

1. $\mathbf{y} = \mathbf{W}_B \mathbf{x}^B$
2. $\mathbf{n} = \mathbf{Q}^\top \mathbf{y}$
3. $\mathbf{c}^a = \mathbf{W}_A \mathbf{x}^A - \mathbf{Q}\mathbf{n}$

Golkar, Lipshutz et al. *NeurIPS* 2020; Lipshutz et al. *Neural Computation* 2021



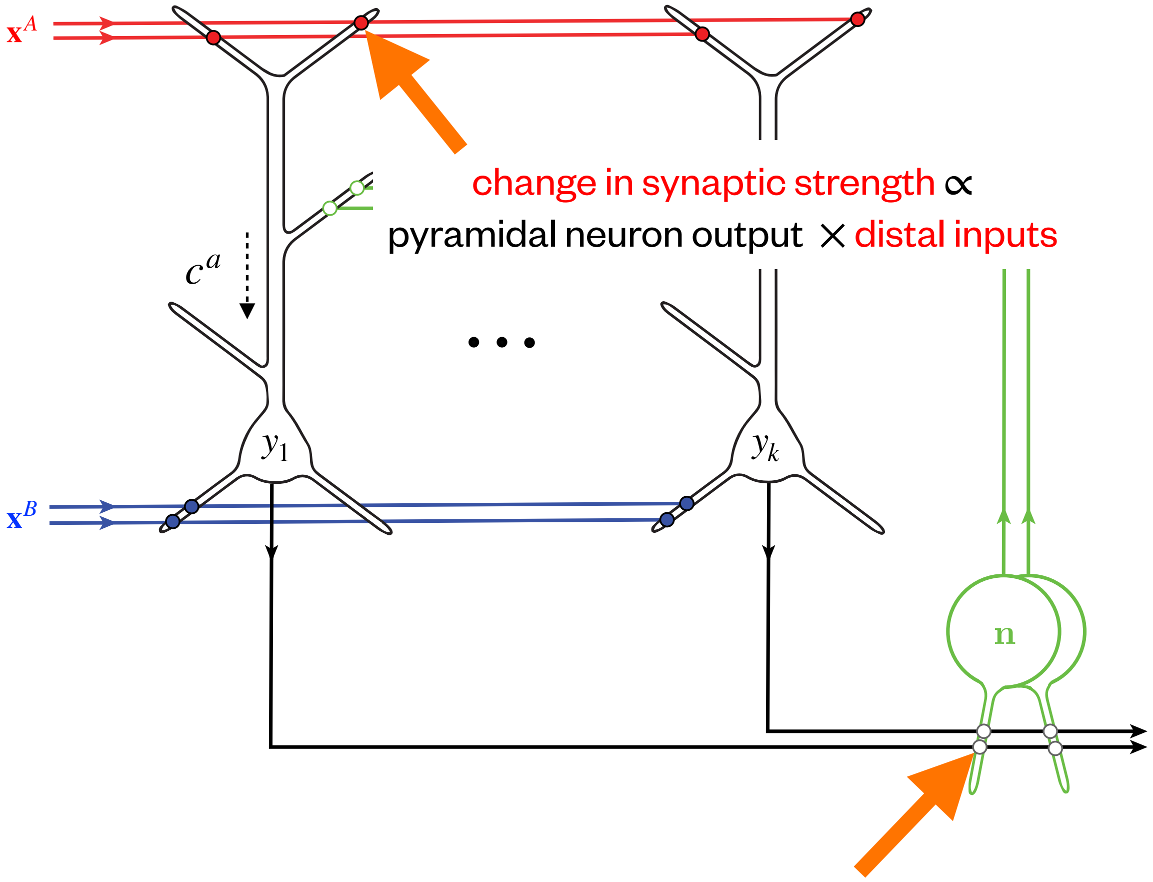
change in synaptic strength ∝
calcium plateau × proximal inputs

"…[distal] dendritic plateau potentials are highly effective induction mechanisms for long-term plasticity at the [proximal] synapse…" [Takahashi & Magee 2009]

CCA algorithm

1. $\mathbf{y} = \mathbf{W}_B \mathbf{x}^B$
2. $\mathbf{n} = \mathbf{Q}^\top \mathbf{y}$
3. $\mathbf{c}^a = \mathbf{W}_A \mathbf{x}^A - \mathbf{Q}\mathbf{n}$
4. $\mathbf{W}_B \leftarrow \mathbf{W}_B + \eta \mathbf{c}^a \mathbf{x}^{B,\top}$

Golkar, Lipshutz et al. *NeurIPS* 2020; Lipshutz et al. *Neural Computation* 2021



change in synaptic strength ∝ pyramidal neuron output × distal inputs

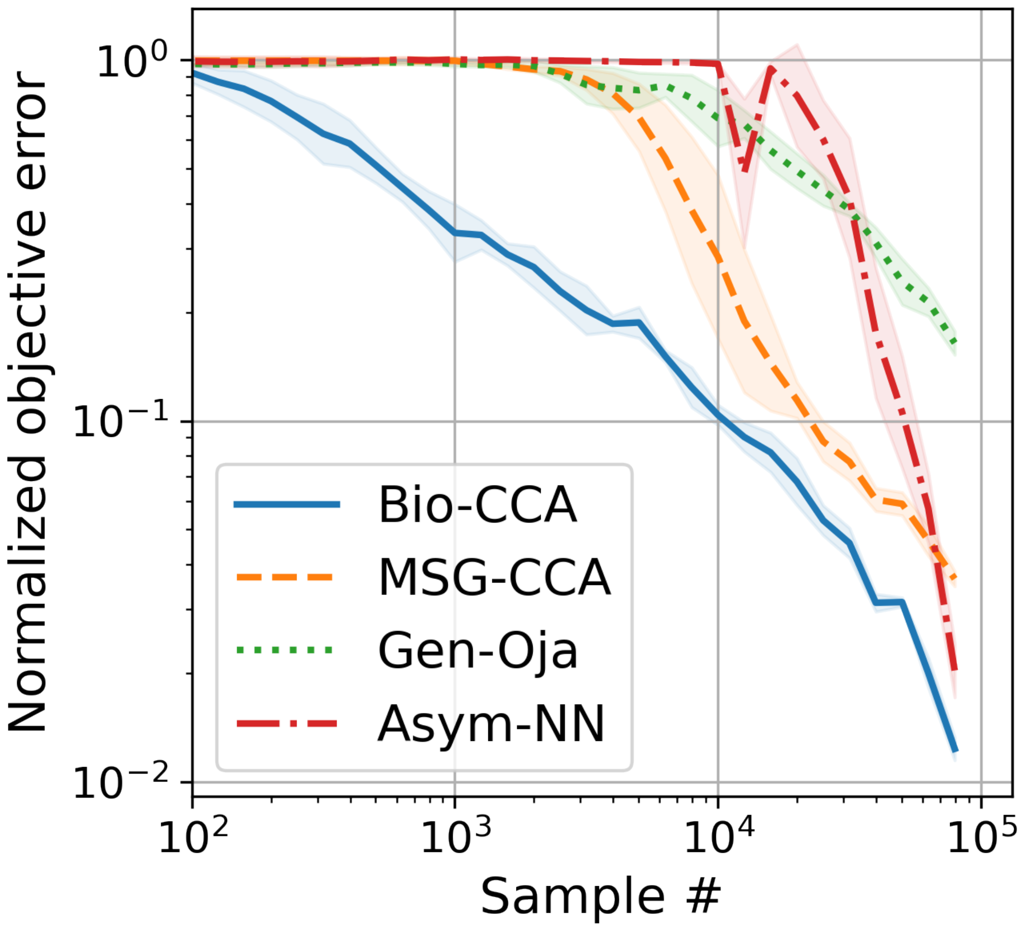
change in synaptic strength ∝ inputs × outputs

CCA algorithm

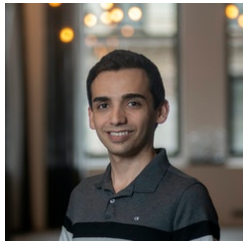
1. $\mathbf{y} = \mathbf{W}_B\mathbf{x}^B$
2. $\mathbf{n} = \mathbf{Q}^\top\mathbf{y}$
3. $\mathbf{c}^a = \mathbf{W}_A\mathbf{x}^A - \mathbf{Q}\mathbf{n}$
4. $\mathbf{W}_B \leftarrow \mathbf{W}_B + \eta\mathbf{c}^a\mathbf{x}^{B,\top}$
5. $\mathbf{W}_A \leftarrow \mathbf{W}_A + \eta(\mathbf{y}\mathbf{x}^{A,\top} - \mathbf{W}_A)$
6. $\mathbf{Q} \leftarrow \mathbf{Q} + \frac{\eta}{\tau}(\mathbf{y}\mathbf{n}^\top - \mathbf{Q})$

Golkar, Lipshutz et al. *NeurIPS* 2020; Lipshutz et al. *Neural Computation* 2021

Empirical evidence that the algorithm is sample efficient



Lipshutz et al. *Neural Computation* 2021

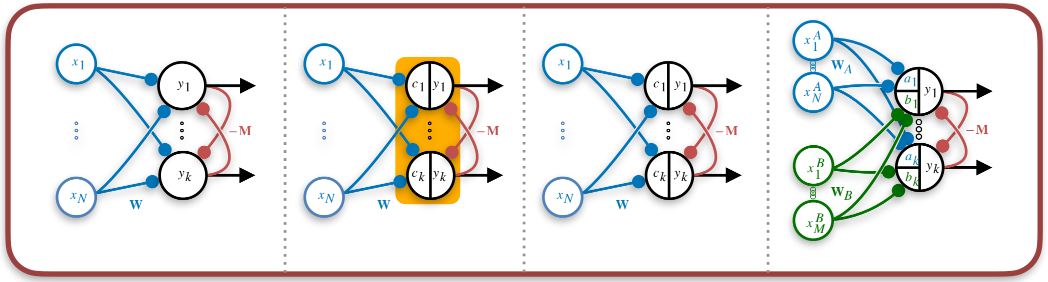


Interim summary

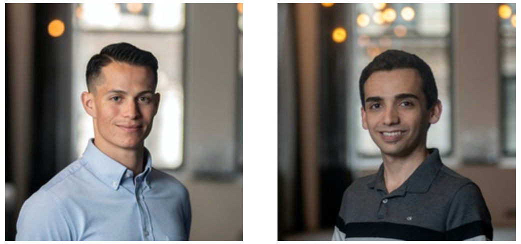
Proposed a general framework for relating learning principles to synaptic plasticity rules

$$\mathbf{A}\mathbf{w} = \lambda\mathbf{B}\mathbf{w}$$

```
input $\{(\boldsymbol{\xi}_t, \mathbf{B}_t)\}$; parameters $0 < \eta < \tau$
initialize $\mathbf{W} \in \mathbb{R}^{k\times n}$ and $\mathbf{M} \in \mathbb{S}^k_{++}$
for $t = 1, 2, \ldots$ do
    repeat
        $\boldsymbol{\zeta}_t \leftarrow \boldsymbol{\zeta}_t + \gamma(\mathbf{W}\boldsymbol{\xi}_t - \mathbf{M}\boldsymbol{\zeta}_t)$
    until convergence
    $\mathbf{W} \leftarrow \mathbf{W} + 2\eta(\boldsymbol{\zeta}_t\boldsymbol{\xi}_t^\top - \mathbf{W}\mathbf{B}_t)$
    $\mathbf{M} \leftarrow \mathbf{M} + \frac{\eta}{\tau}(\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t^\top - \mathbf{M})$
end for
```



Lipshutz et al. *NeurIPS* 2020; Golkar, Lipshutz et al. *NeurIPS* 2020; Lipshutz et al. *Neur. Comp.* 2021; Bahroun et al. 2021; Lipshutz et al. *Biol. Cyber.* 2022; Golkar, Lipshutz et al. *ICASSP* 2023; Lipshutz et al. *PRX Life* 2023



Interim summary

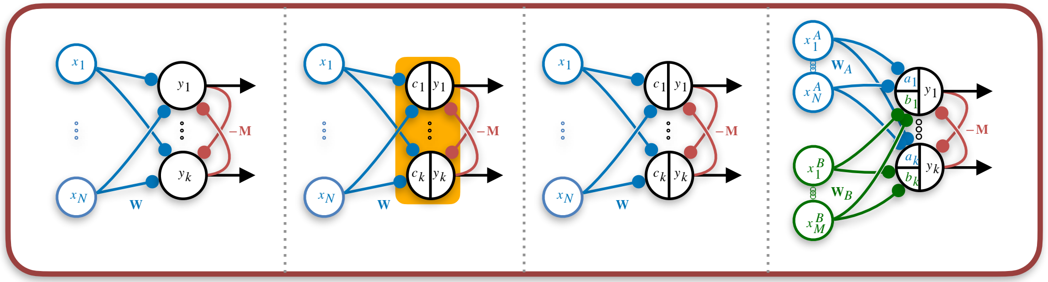
Proposed a general framework for relating learning principles to synaptic plasticity rules

Neural algorithms wish list:

- principled? yes
- free parameters? 2
- sample efficient? yes (empirical)
- resource efficient? local & online
- match data? consistent with observations in the cortical microcircuit

$$\mathbf{Aw} = \lambda\mathbf{Bw}$$

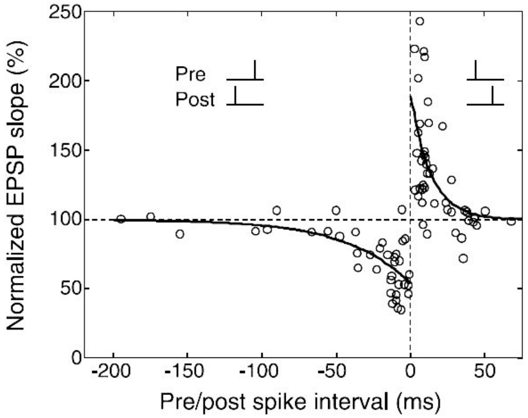
```
input $\{(\boldsymbol{\xi}_t, \mathbf{B}_t)\}$; parameters $0 < \eta < \tau$
initialize $\mathbf{W} \in \mathbb{R}^{k\times n}$ and $\mathbf{M} \in \mathbb{S}^k_{++}$
for $t = 1, 2, \ldots$ do
    repeat
        $\boldsymbol{\zeta}_t \leftarrow \boldsymbol{\zeta}_t + \gamma(\mathbf{W}\boldsymbol{\xi}_t - \mathbf{M}\boldsymbol{\zeta}_t)$
    until convergence
    $\mathbf{W} \leftarrow \mathbf{W} + 2\eta(\boldsymbol{\zeta}_t\boldsymbol{\xi}_t^\top - \mathbf{W}\mathbf{B}_t)$
    $\mathbf{M} \leftarrow \mathbf{M} + \frac{\eta}{\tau}(\boldsymbol{\zeta}_t\boldsymbol{\zeta}_t^\top - \mathbf{M})$
end for
```



Lipshutz et al. *NeurIPS* 2020; Golkar, Lipshutz et al. *NeurIPS* 2020; Lipshutz et al. *Neur. Comp.* 2021; Bahroun et al. 2021; Lipshutz et al. *Biol. Cyber.* 2022; Golkar, Lipshutz et al. *ICASSP* 2023; Lipshutz et al. *PRX Life* 2023

What about spike-timing-dependent plasticity (STDP)?

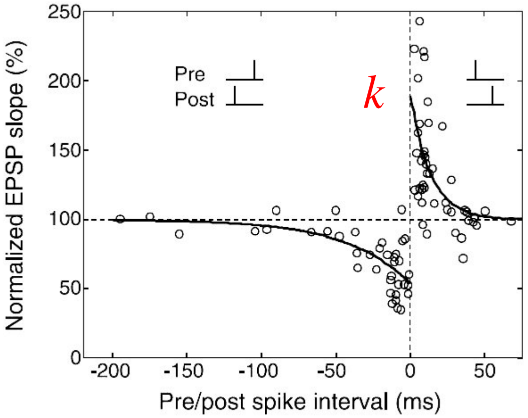
STDP in rat hippocampus



[**Bi & Poo, 1998**; Feldman 2012]

What about spike-timing-dependent plasticity (STDP)?

STDP in rat hippocampus



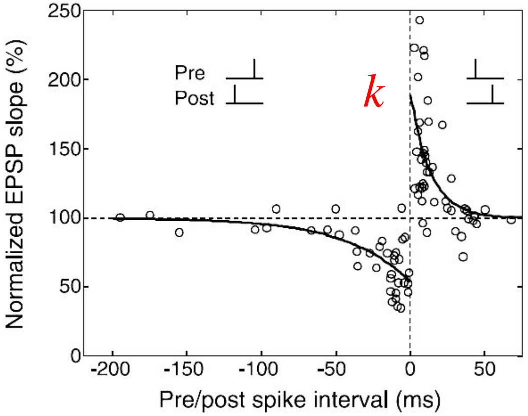
[**Bi & Poo, 1998**; Feldman 2012]

$$\mathbf{A}\mathbf{w} = \lambda \mathbf{B}\mathbf{w}$$

$$\mathbf{A} = \sum_{t,\tau} k(t-\tau)\mathbf{x}_t\mathbf{x}_\tau^\top, \quad \mathbf{B} = \sum_t \mathbf{x}_t\mathbf{x}_t^\top$$

What about spike-timing-dependent plasticity (STDP)?

STDP in rat hippocampus



[**Bi & Poo, 1998**; Feldman 2012]

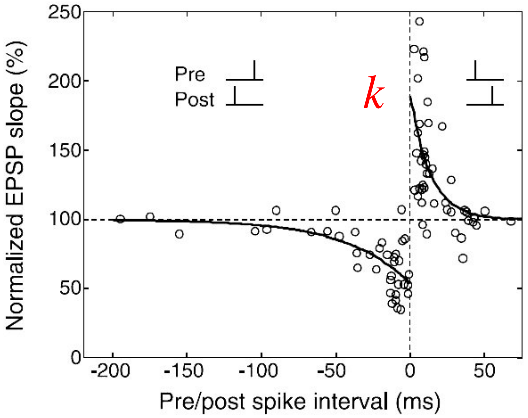
$$\mathbf{A}\mathbf{w} = \lambda \mathbf{B}\mathbf{w}$$

$$\mathbf{A} = \sum_{t,\tau} k(t-\tau)\mathbf{x}_t \mathbf{x}_\tau^\top, \quad \mathbf{B} = \sum_t \mathbf{x}_t \mathbf{x}_t^\top$$

- $\mathbf{A}$ is not **symmetric**, function of kernel / time reversibility
- eigenvectors are complex-valued
- dynamic receptive fields with speed $\propto \mathrm{Im}(\lambda_1)$

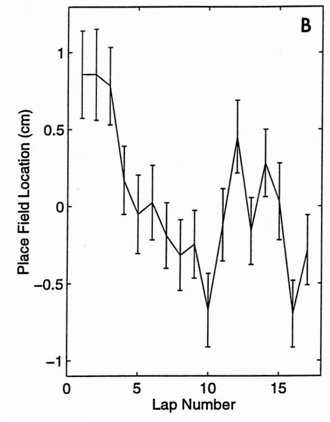
What about spike-timing-dependent plasticity (STDP)?

STDP in rat hippocampus



[**Bi & Poo, 1998**; Feldman 2012]

dynamic RFs



[**Mehta et al. 1997**; Dong et al. 2021]

$$\mathbf{A}\mathbf{w} = \lambda \mathbf{B}\mathbf{w}$$

$$\mathbf{A} = \sum_{t,\tau} k(t-\tau)\mathbf{x}_t\mathbf{x}_\tau^\top, \quad \mathbf{B} = \sum_t \mathbf{x}_t\mathbf{x}_t^\top$$

- $\mathbf{A}$ is not **symmetric**, function of kernel / time reversibility
- eigenvectors are complex-valued
- dynamic receptive fields with speed $\propto \mathrm{Im}(\lambda_1)$

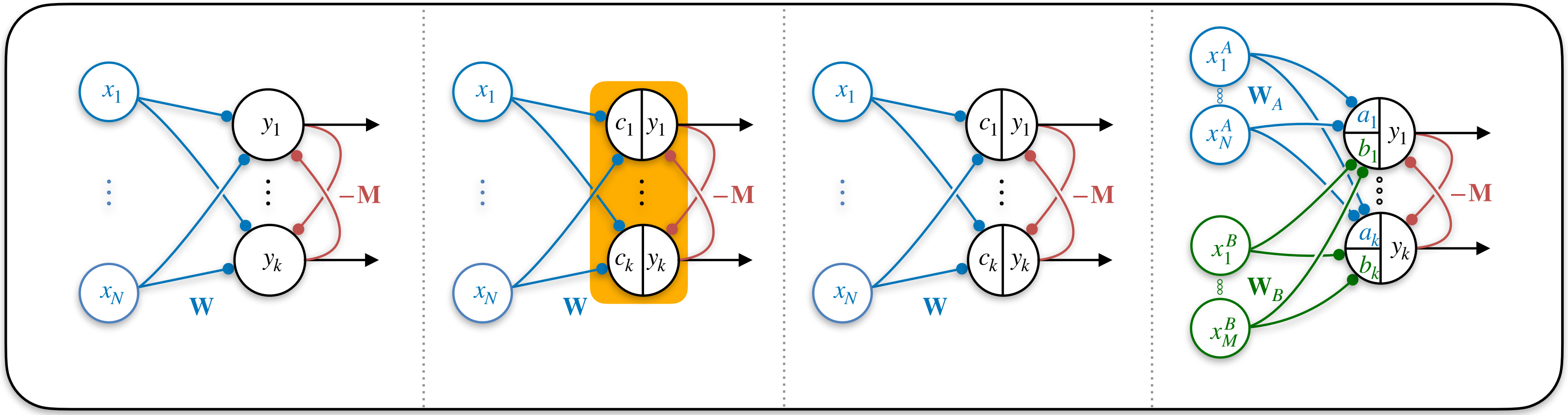
Goal: concise mathematical descriptions of the statistical learning algorithms that support sensory processing.

Where do we go from here?

Wish list for biological statistical learning algorithms:

1. sample efficient? optimize for sensory statistics
2. resource efficient? spiking?
3. no free parameters? match learning rates to environment
4. matches neural data? hierarchical, nonlinear processing, feedback, recurrence

Thank you.



Flatiron Institute
Yanis Bahroun
Siavash Golkar
Charles Windolf
Tiberiu Tesileanu
Anirvan Sengupta
Dmitri Chklovskii