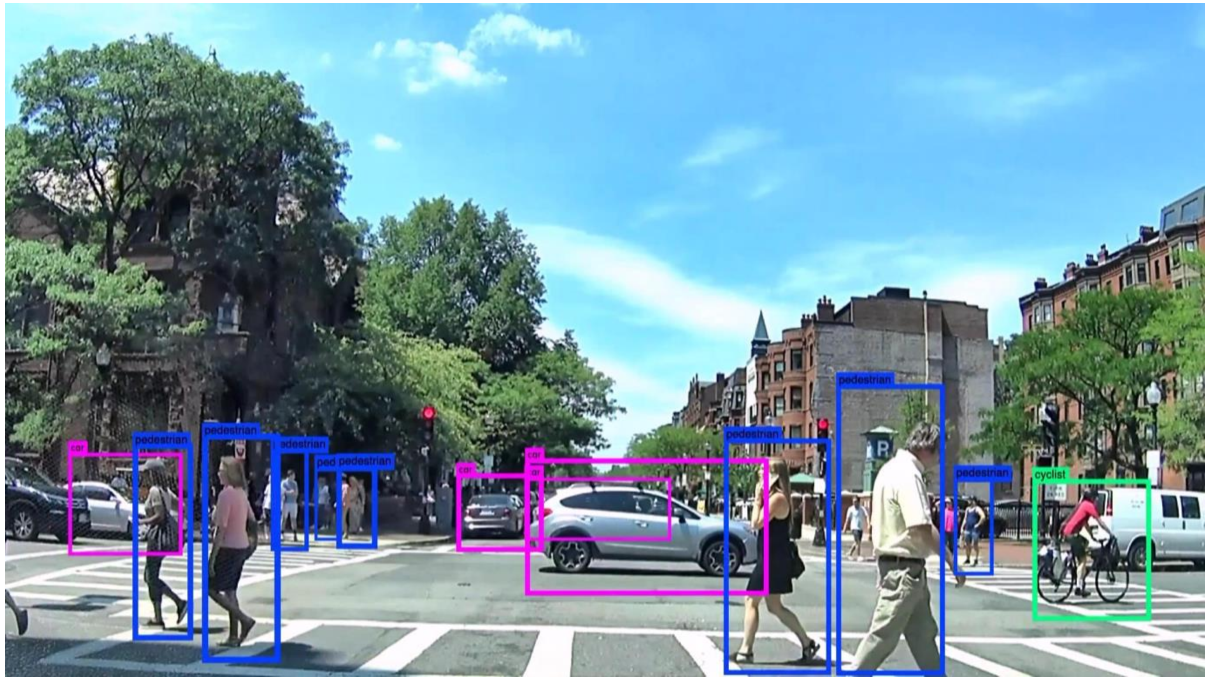
Models of representation learning dynamics

Andrew Saxe

Gatsby Unit & Sainsbury Wellcome Centre, UCL

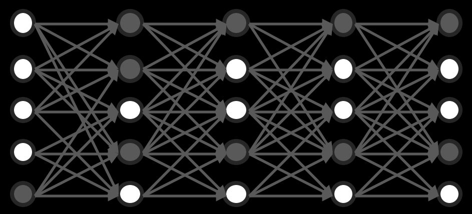

EDDIE
GILL


CURRY
30


car
pedestrian
pedestrian
pedestrian
pedestrian
car
pedestrian
pedestrian
pedestrian
cyclist


柯洁
00:4

Artificial Intelligence



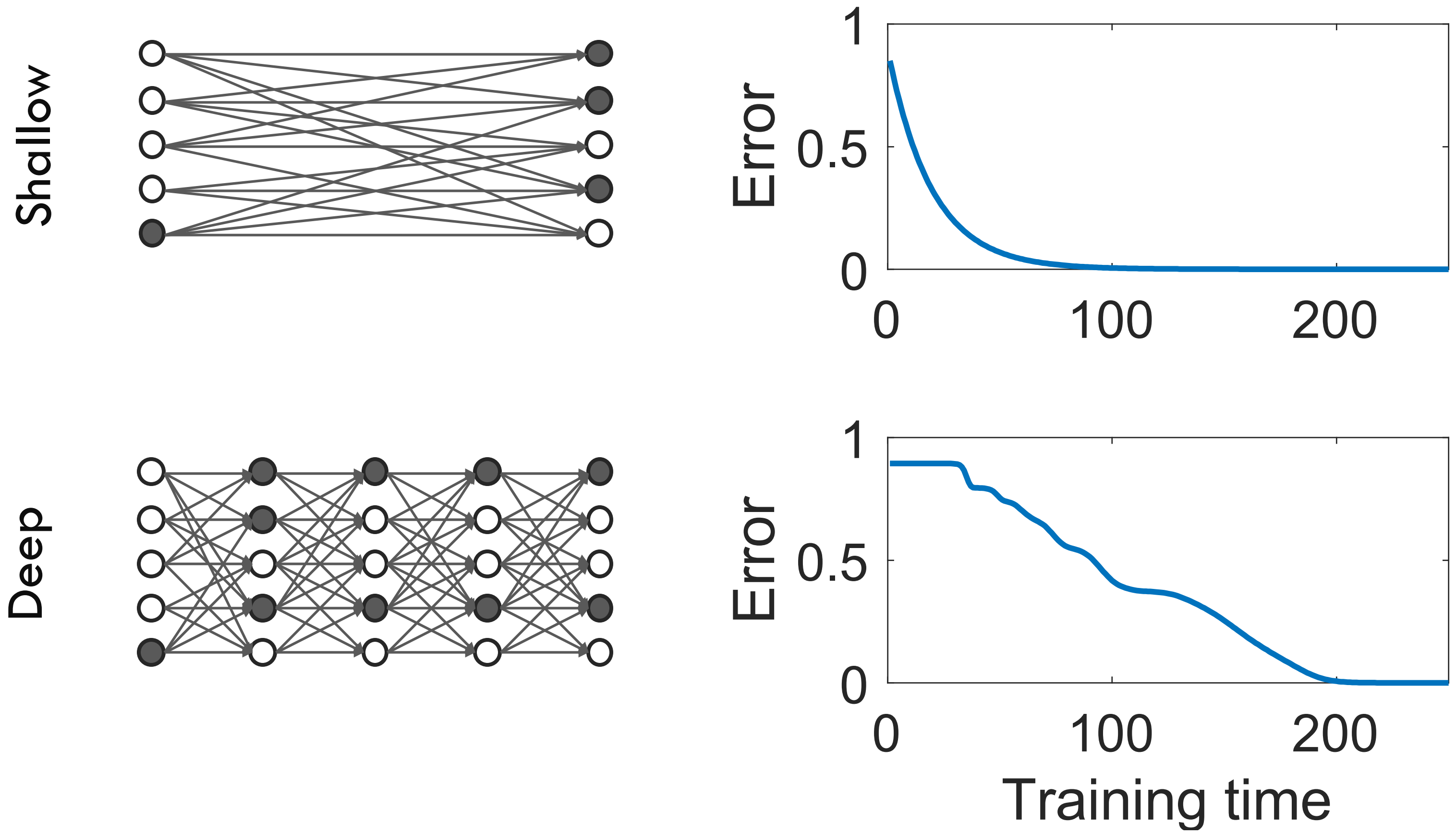
Neural Networks

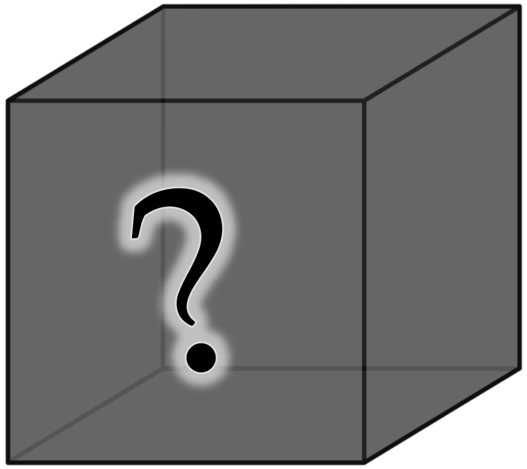
Brain & Mind

Today

1. Deep linear network dynamics
2. Nontrivial initializations: Lazy, rich, & beyond
3. Nonlinear networks & the neural race reduction

Depth complicates learning dynamics





L. Zdeborová. Understanding deep learning is also a job for physicists. *Nature Physics*, 2020

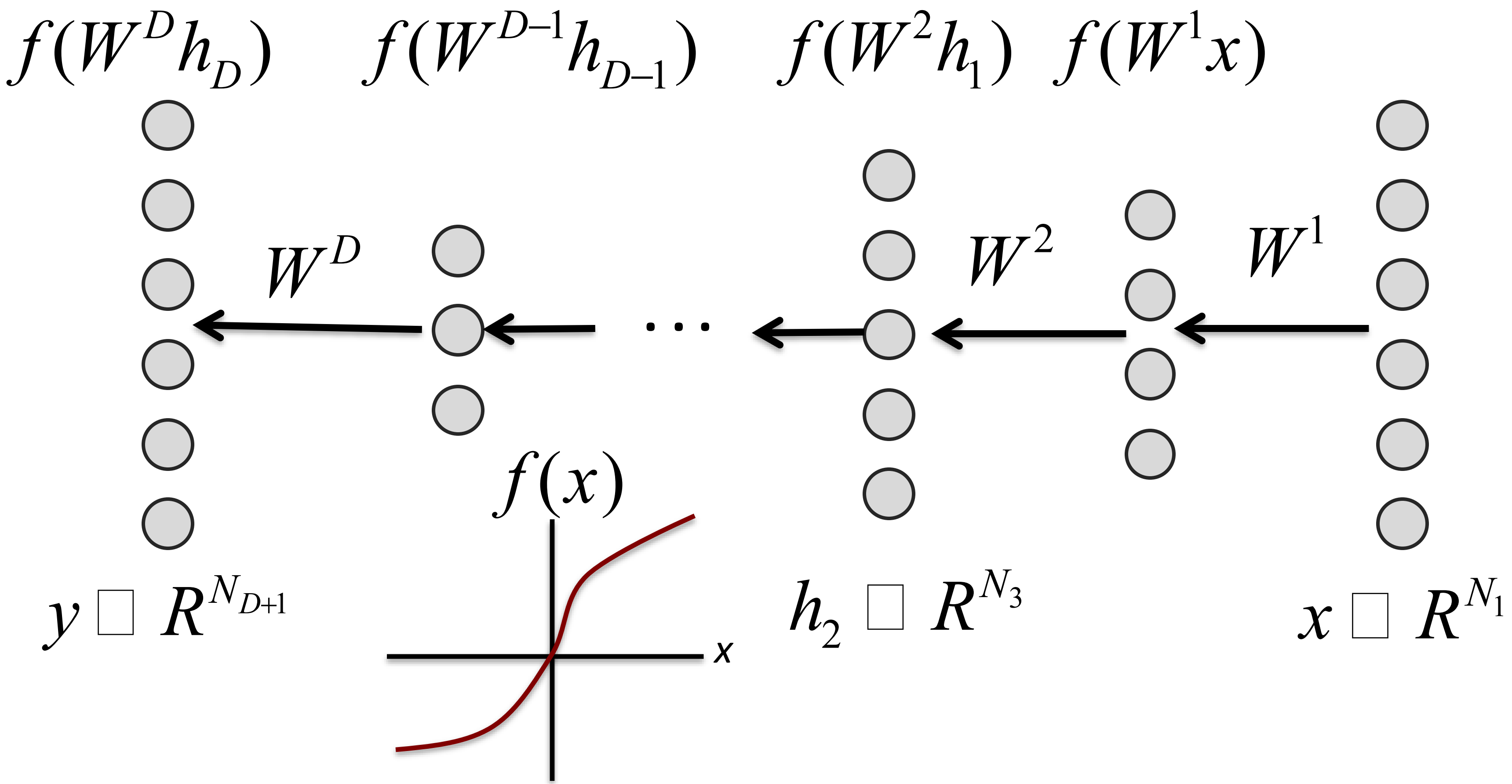
Credit: NYAS

Surrogate models

- Tackling these questions in full generality is challenging
- Instead, we can analyze a surrogate model that is simpler but retains key features of the full problem
- Particularly for brain sciences, crucial to have a minimal, tractable model
  - Conceptual clarity
  - Unambiguous predictions
  - Isolate contribution of depth, data statistics, nonlinearity

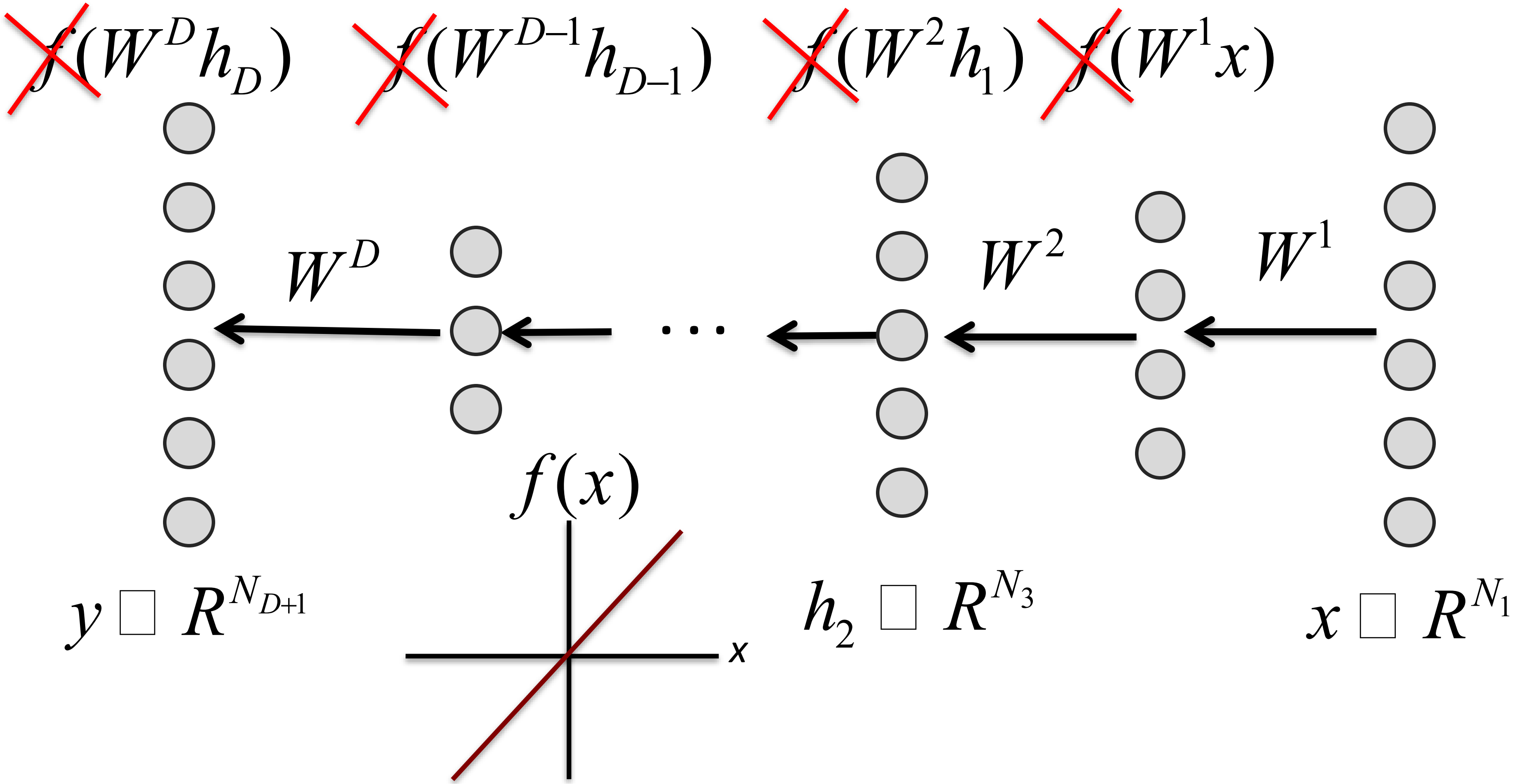
Deep network

- Little hope for a complete theory with arbitrary nonlinearities



Deep *linear* network

- Use a deep *linear* network as a starting point.



Gradient descent

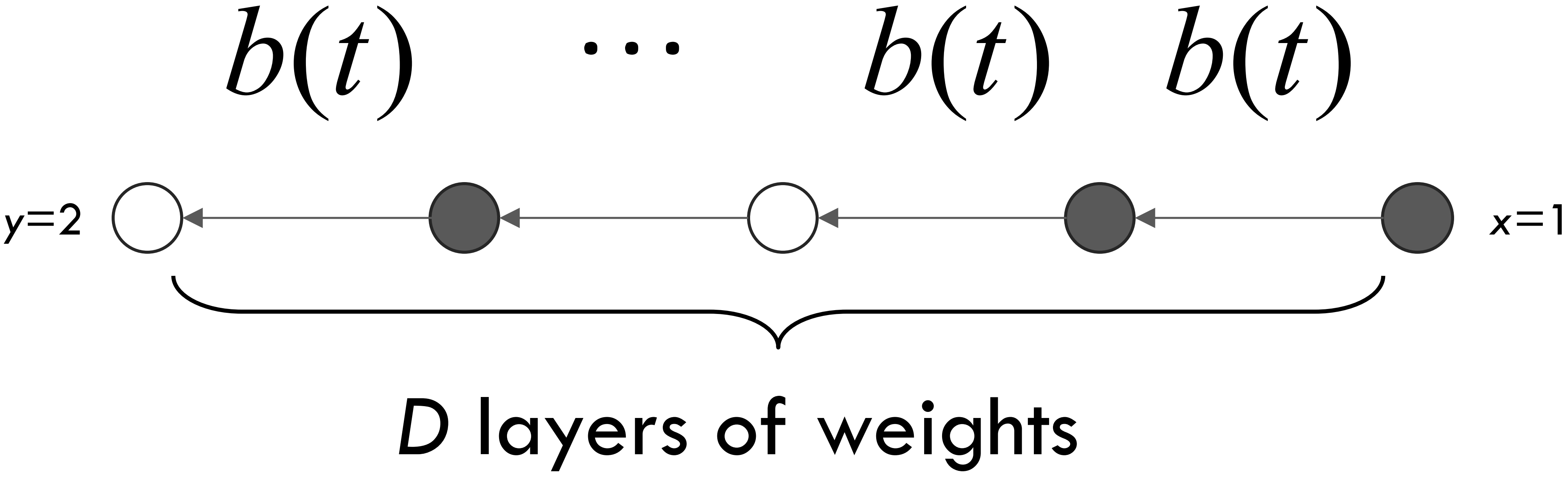
Mean squared error loss:

$$\min_{W_1,\cdots,W_D} \sum_{\mu} \left\| y^{\mu} - \left(\prod_{i=1}^{D} W^{i} \right) x^{\mu} \right\|^2$$

Gradient flow dynamics:

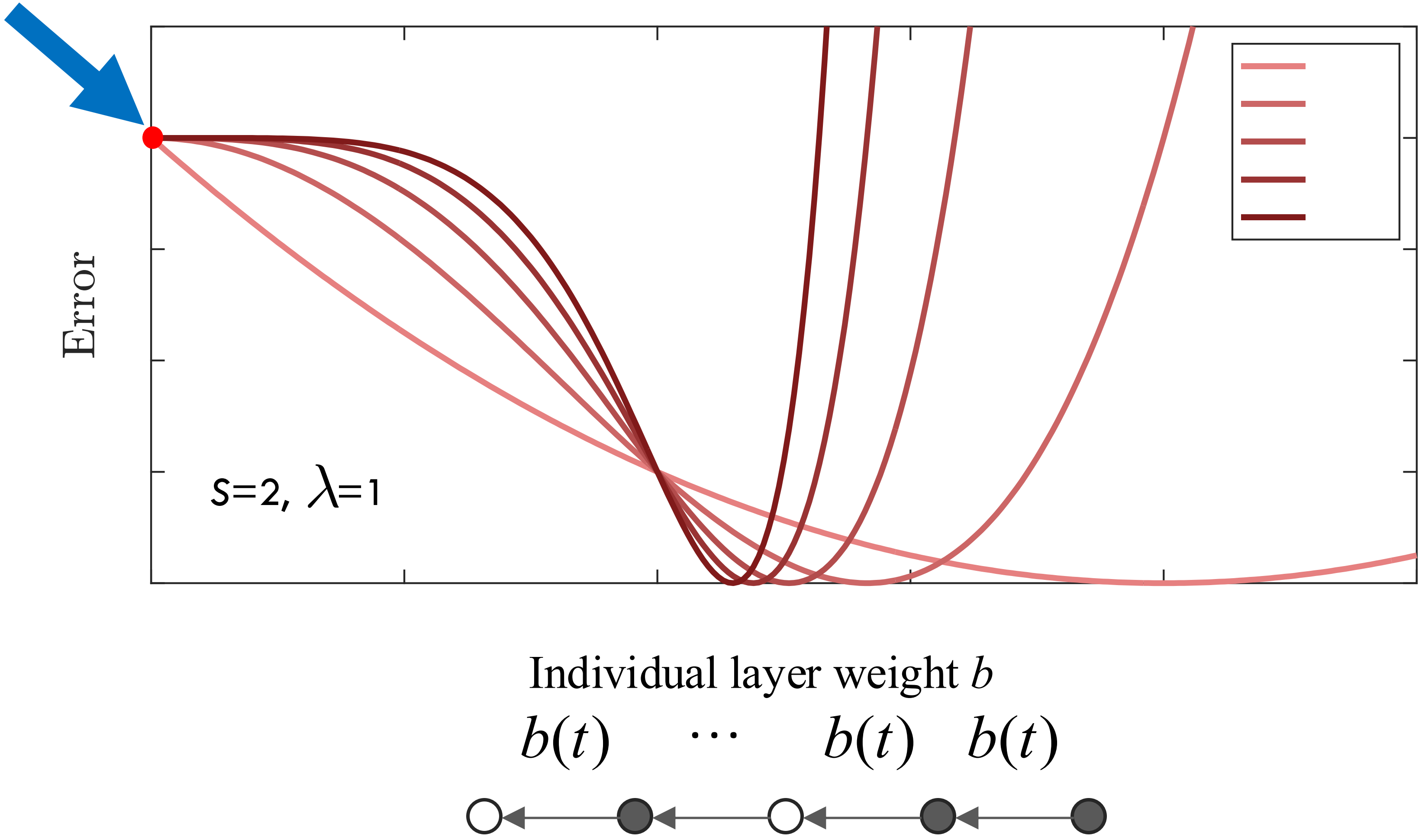
$$\tau \frac{d}{dt} W^{l} = \left(\prod_{i=l+1}^{D} W^{i} \right)^{T} \left[\Sigma^{yx} - \left(\prod_{i=1}^{D} W^{i} \right) \Sigma^{xx} \right] \left(\prod_{i=1}^{l-1} W^{i} \right)^{T} \qquad l = 1, \square, D$$

A linear chain

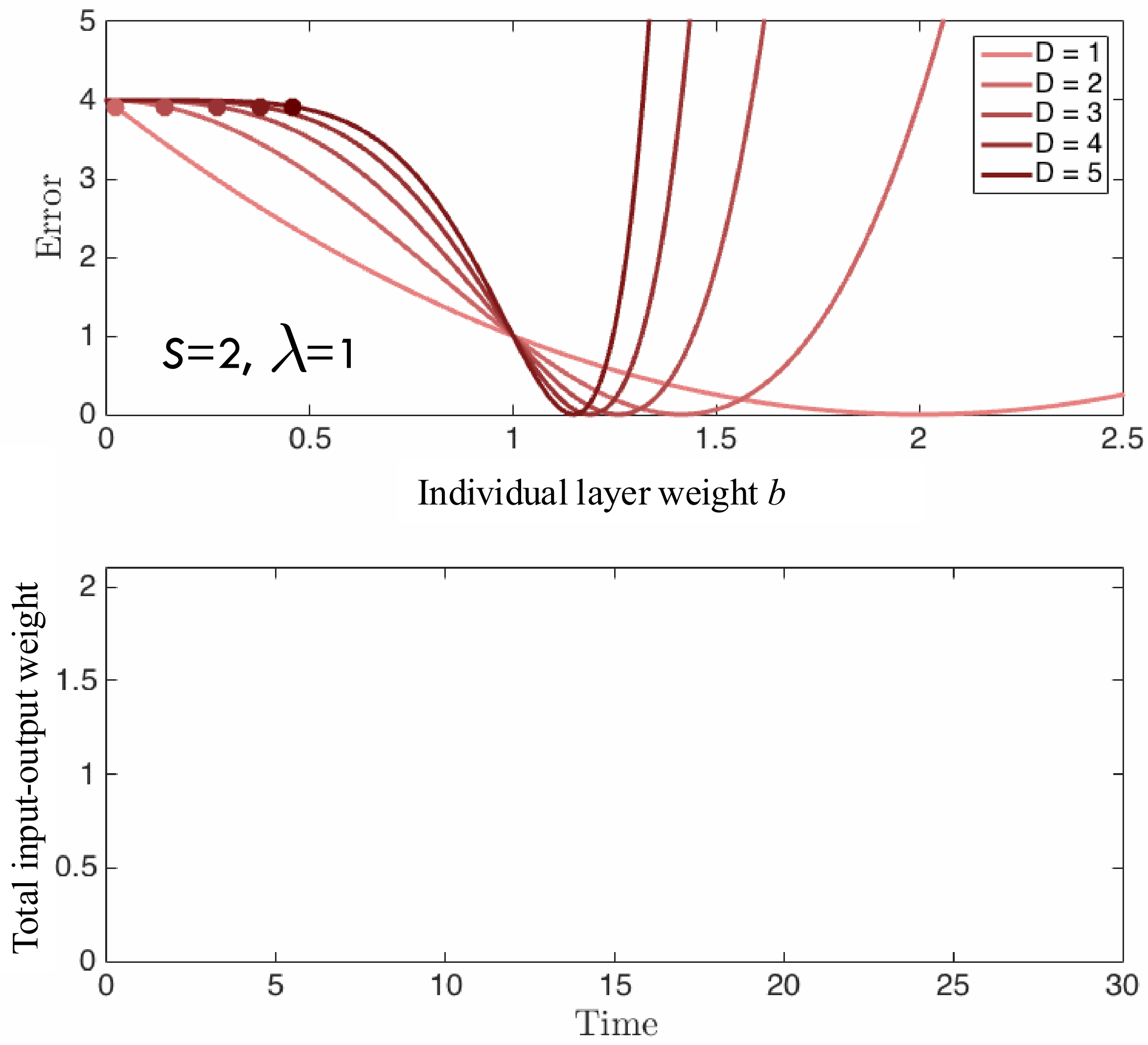


Error surface

Depth introduces a saddle point



Gradient descent dynamics



Analytic learning trajectory

Shallow (*D*=1): $$a(t) = \frac{s}{\lambda}\left(1 - e^{-t/\tau}\right) + a_0 e^{-t/\tau}$$

Deep (*D*=2): $$a(t) = \frac{s/\lambda}{1 - \left(1 - \frac{s}{\lambda a_0}\right) e^{-\frac{2st}{\tau}}}$$

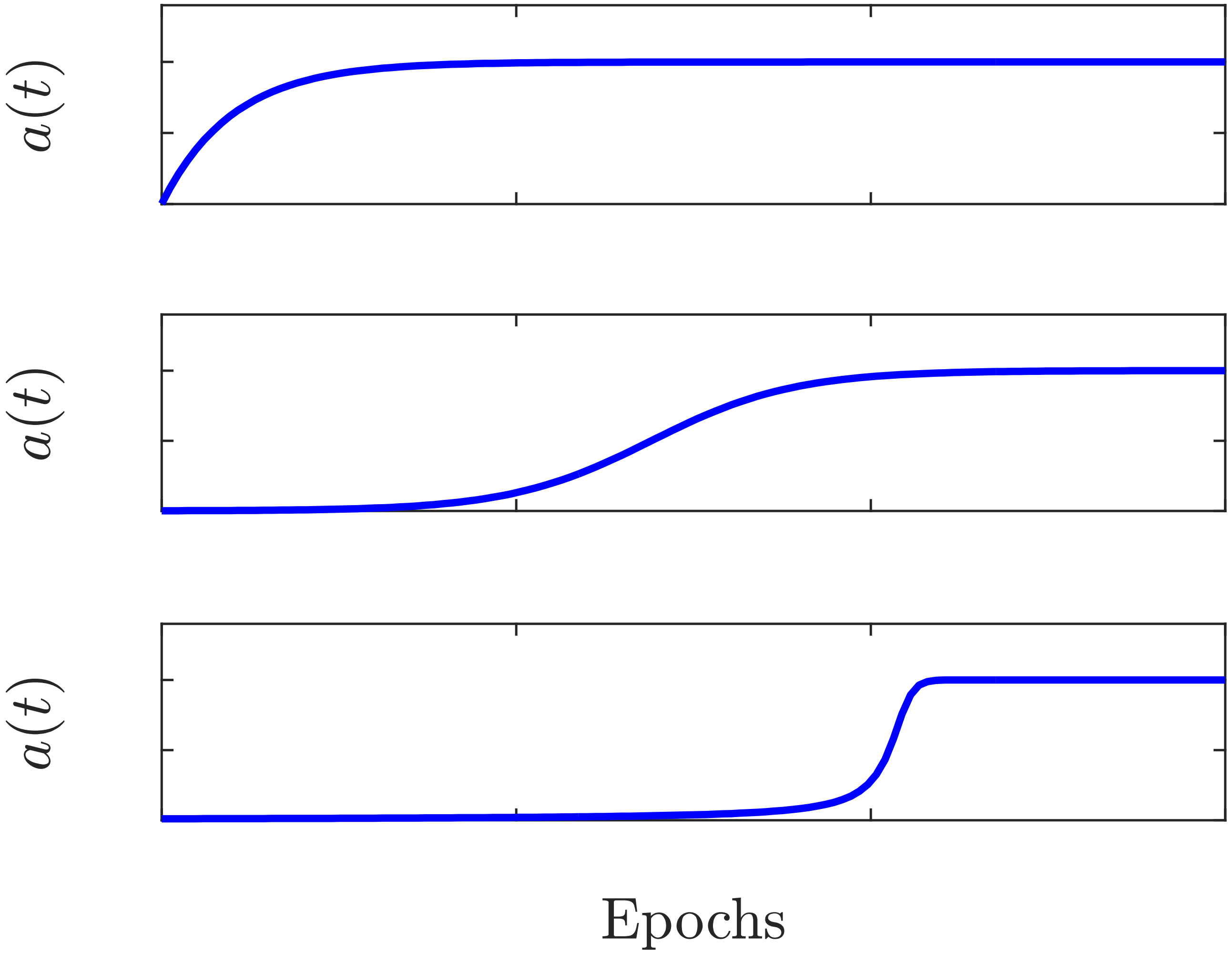
V. Deep (*D* □ □): $$a(t) = \frac{s/\lambda}{1 + W\left[\left(\frac{s}{\lambda a_0} - 1\right) e^{\frac{s}{\lambda a_0} - 1 - t/\tau}\right]}$$

Analytic learning trajectory

Shallow (*D*=1):

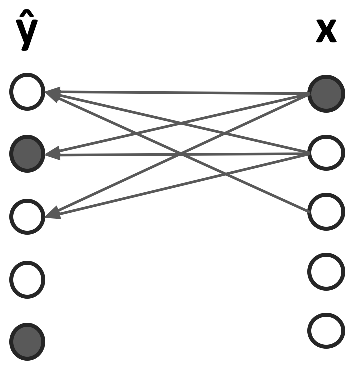
Deep (*D*=2):

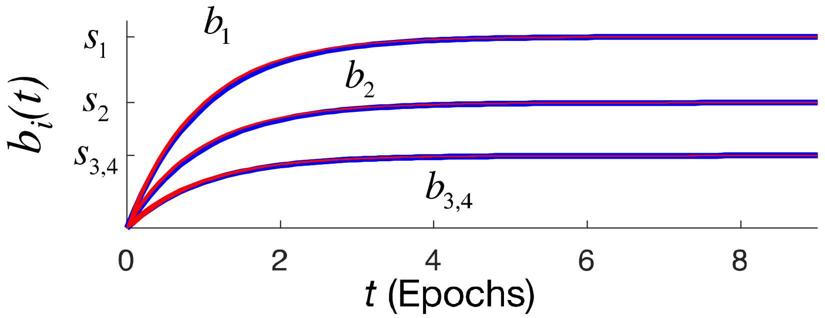
V. Deep (*D*):



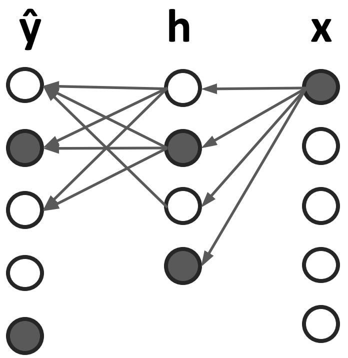
Full networks act like several 1D chains

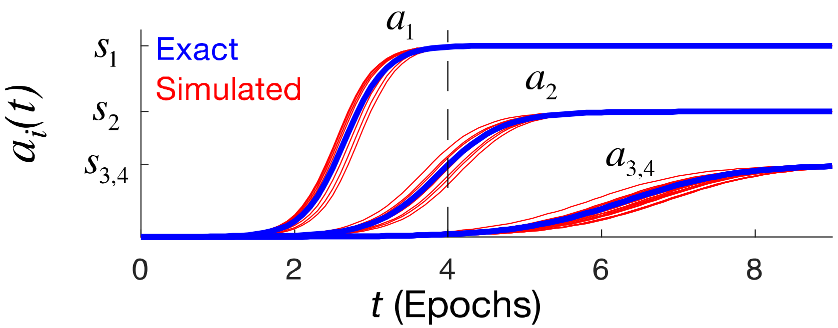
Shallow





Deep

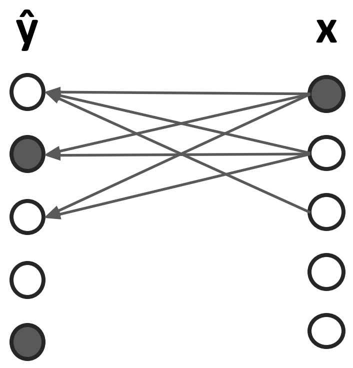


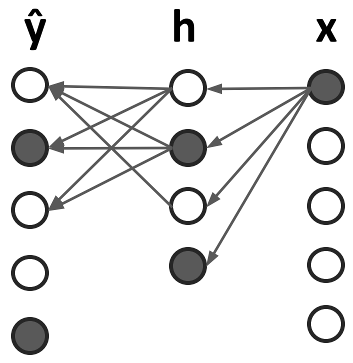


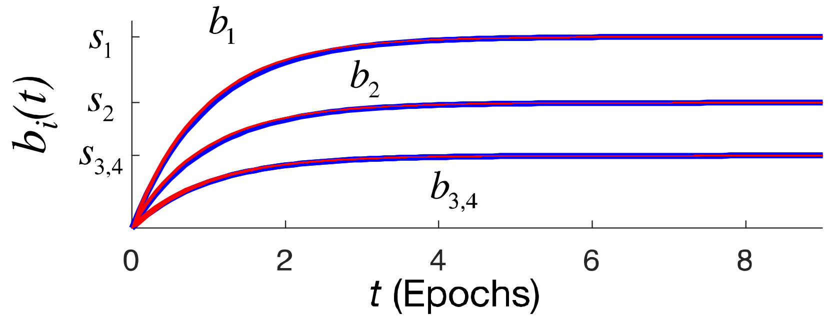
Depth introduces stage-like transitions

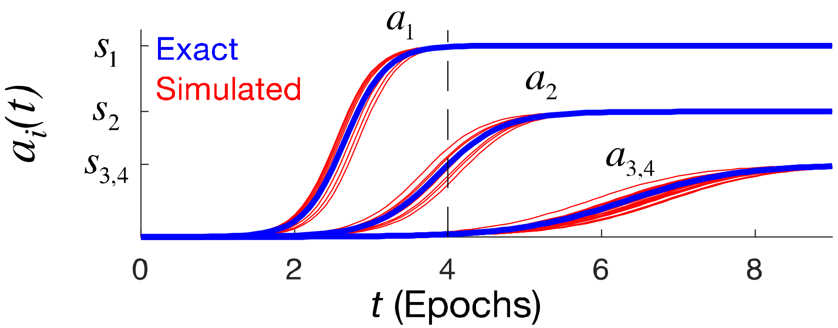
Shallow

Deep









Training speed

- How does training speed scale with depth?

- Time difference for deep net vs shallow net is

$$t_\infty - t_1 \approx O\left(\frac{1}{s b_0^D}\right)$$

t_D	epochs to train depth D network
b_0	Initial layer singular value
s	Minimum nonzero singular value
D	Depth

- Deep learning speed is highly sensitive to initial conditions

Effect of initialization

- **Small random weights** scale exponentially

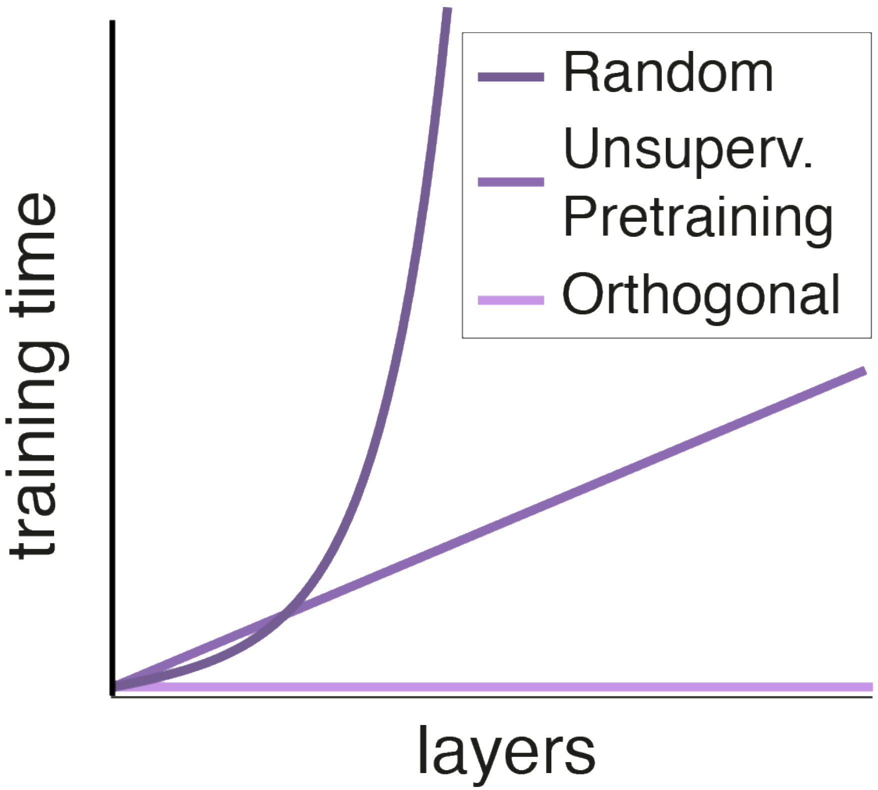
$$t_\infty - t_1 \approx O(1/b_0^D)$$

- **Pretraining + fine-tuning** scales linearly

$$t_\infty - t_1 \approx O(D/b_0^2)$$

- **Orthogonal initialization**: depth-independent

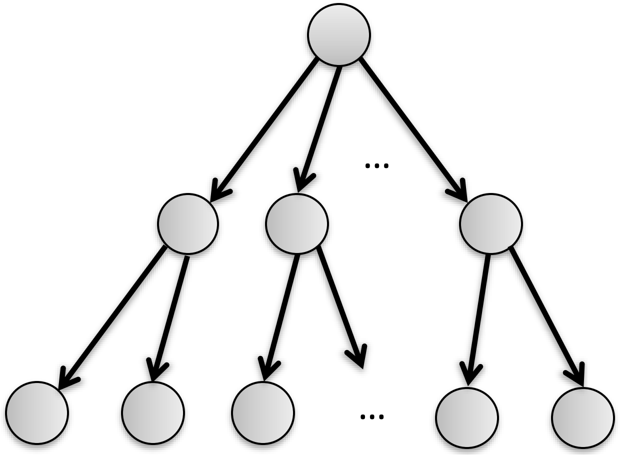
$$t_\infty - t_1 \approx O(1)$$



Connecting neural nets and graphical models

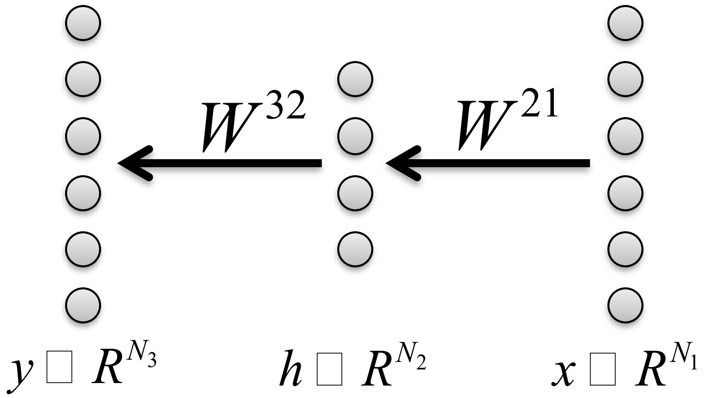
The "World":

Structured generative model



The "Learner":

Deep linear network



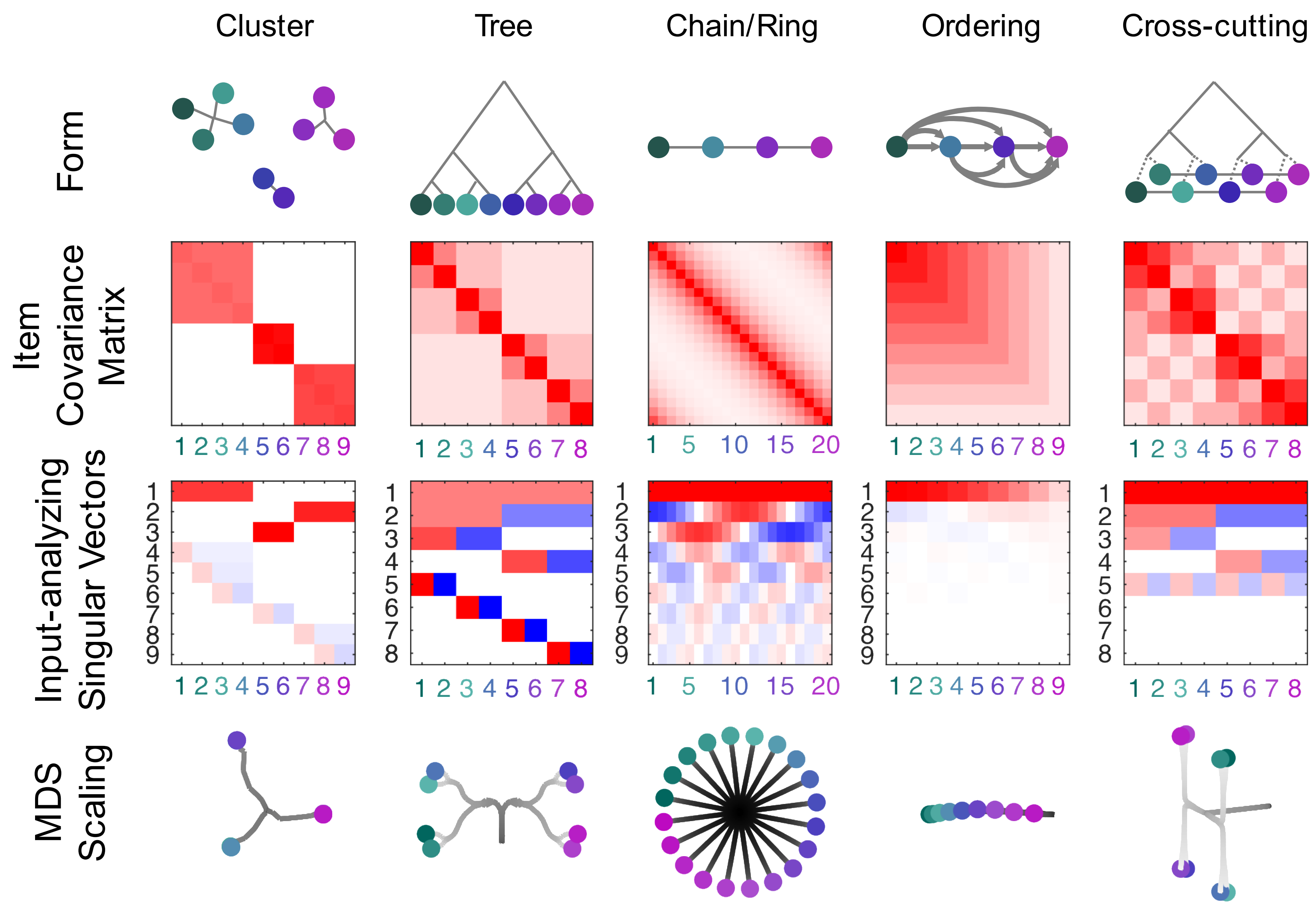
$$\{x^{\mu}, y^{\mu}\}, \mu = 1, \ldots, P.$$

Analytic link

- In the limit of many features, what matters to learning dynamics is SVD of correlation structure

- Can find this exactly for certain graphical models
  - Partitions
  - Trees
  - Grids/rings

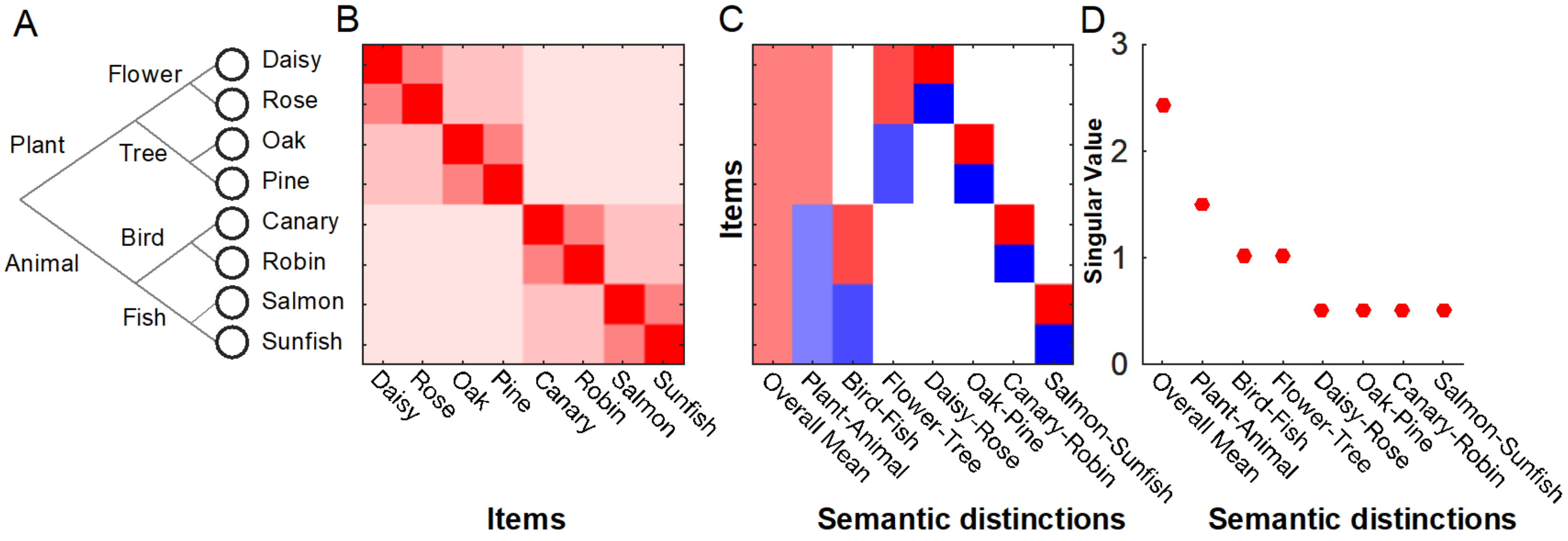
Learning diverse structures



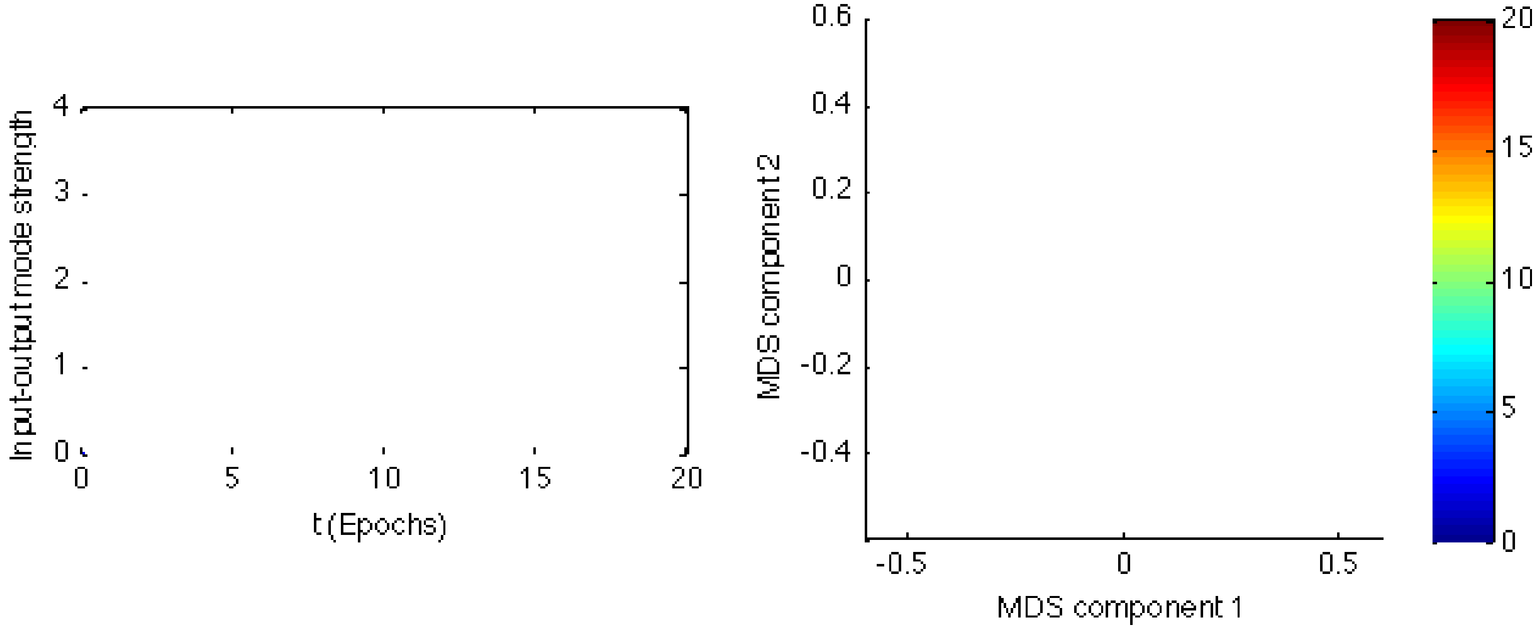
Progressive differentiation

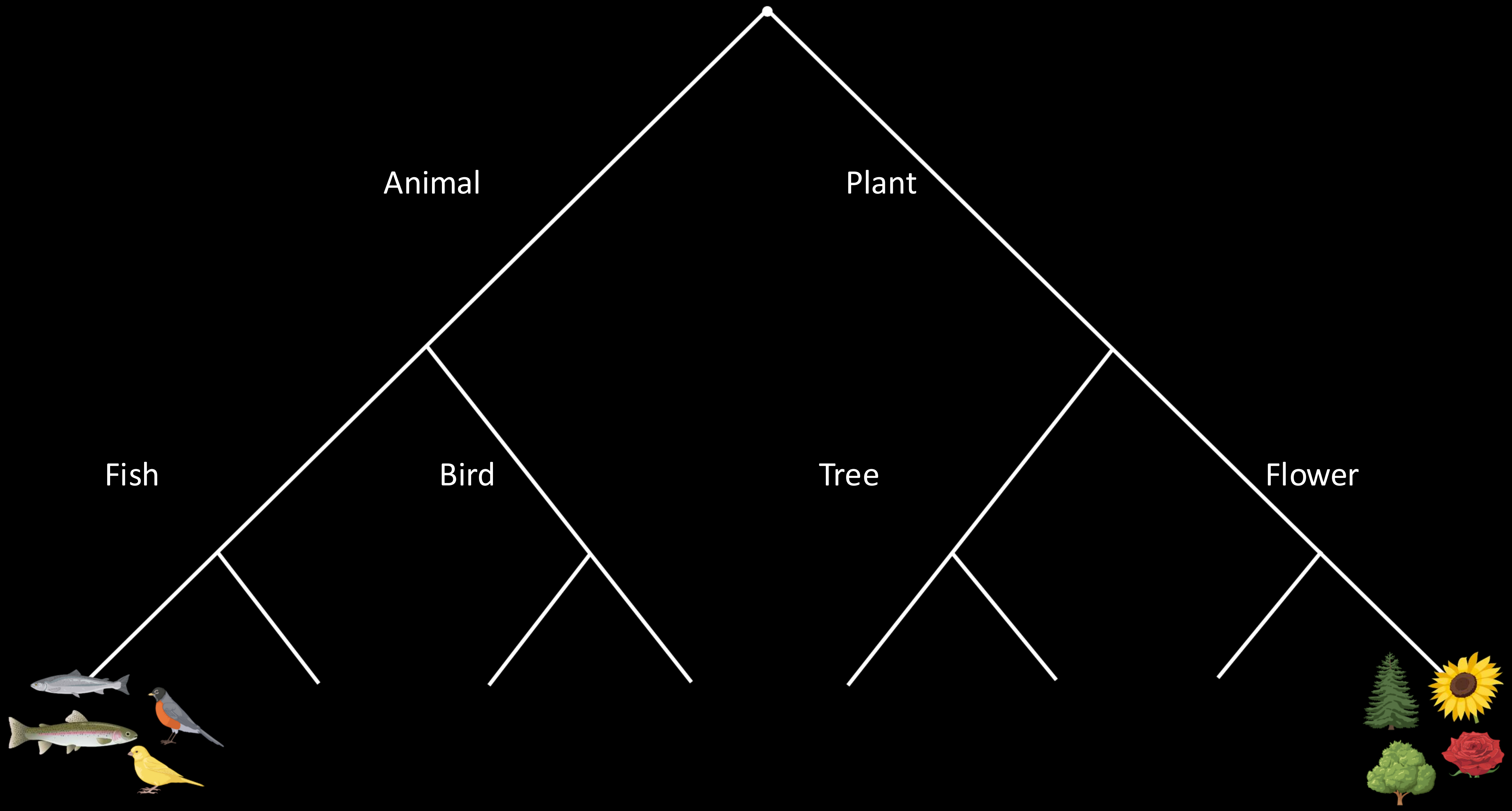
These networks **must** exhibit progressive differentiation:

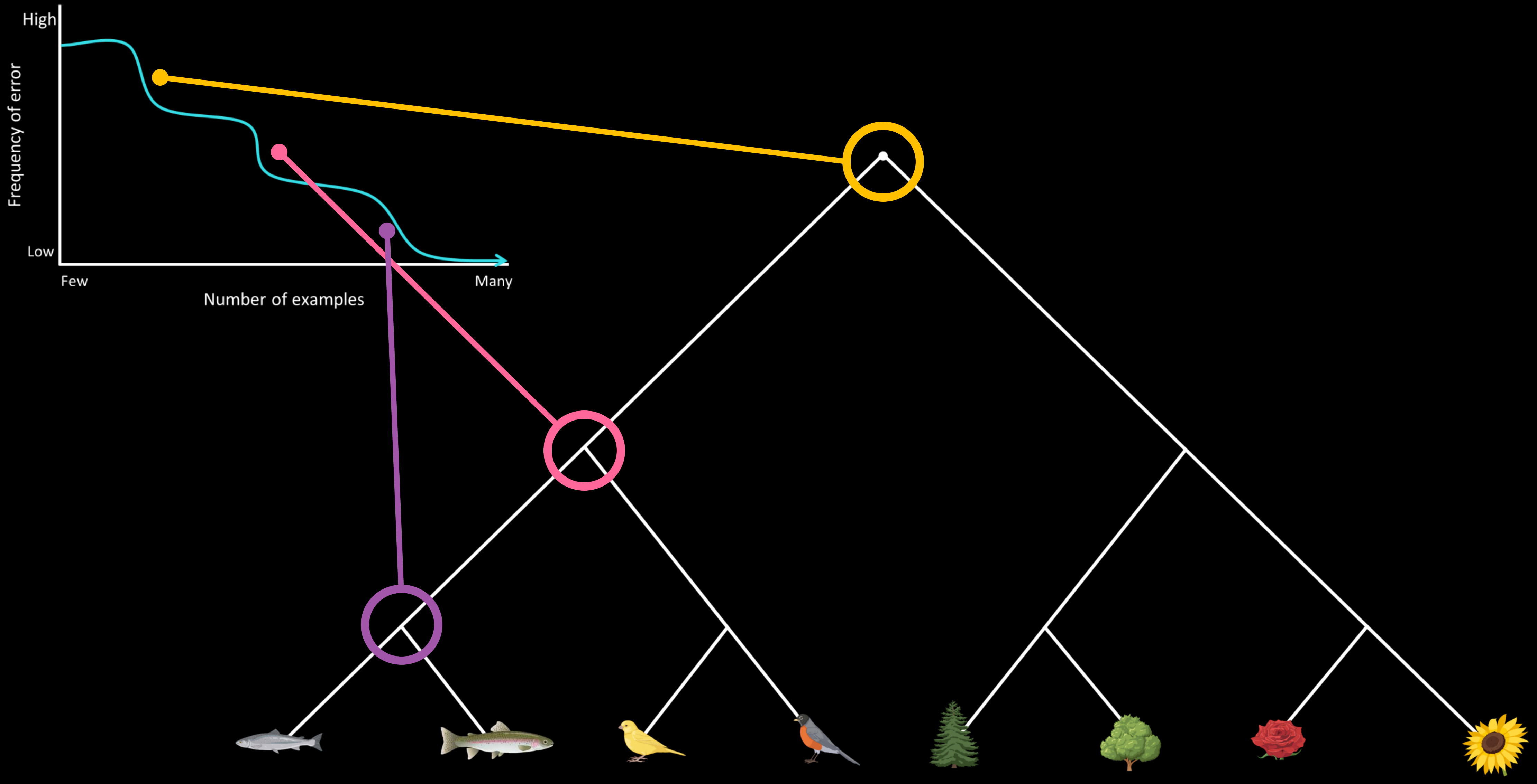
- Singular vectors mirror hierarchy
- Singular values decay with depth



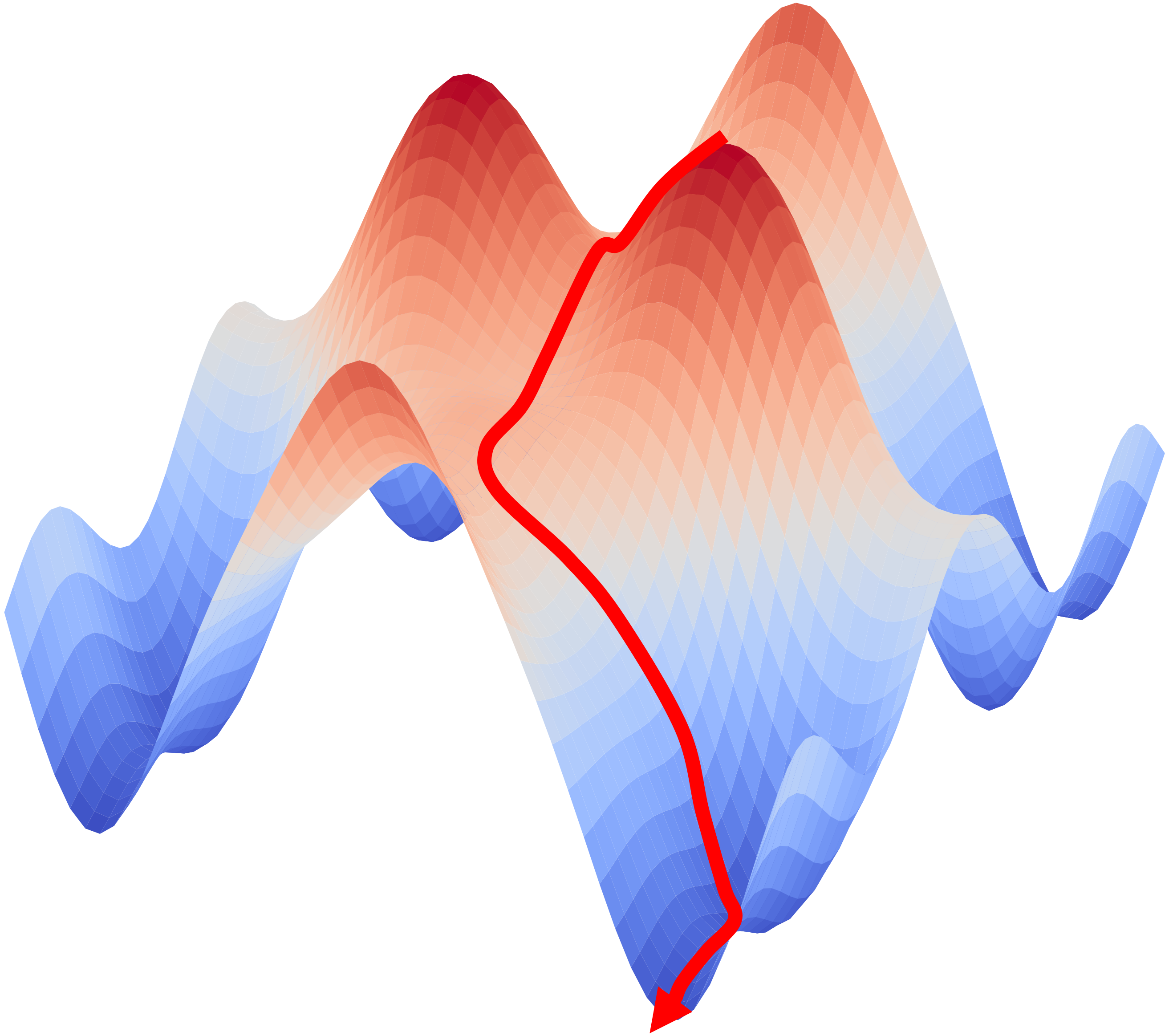
Progressive differentiation



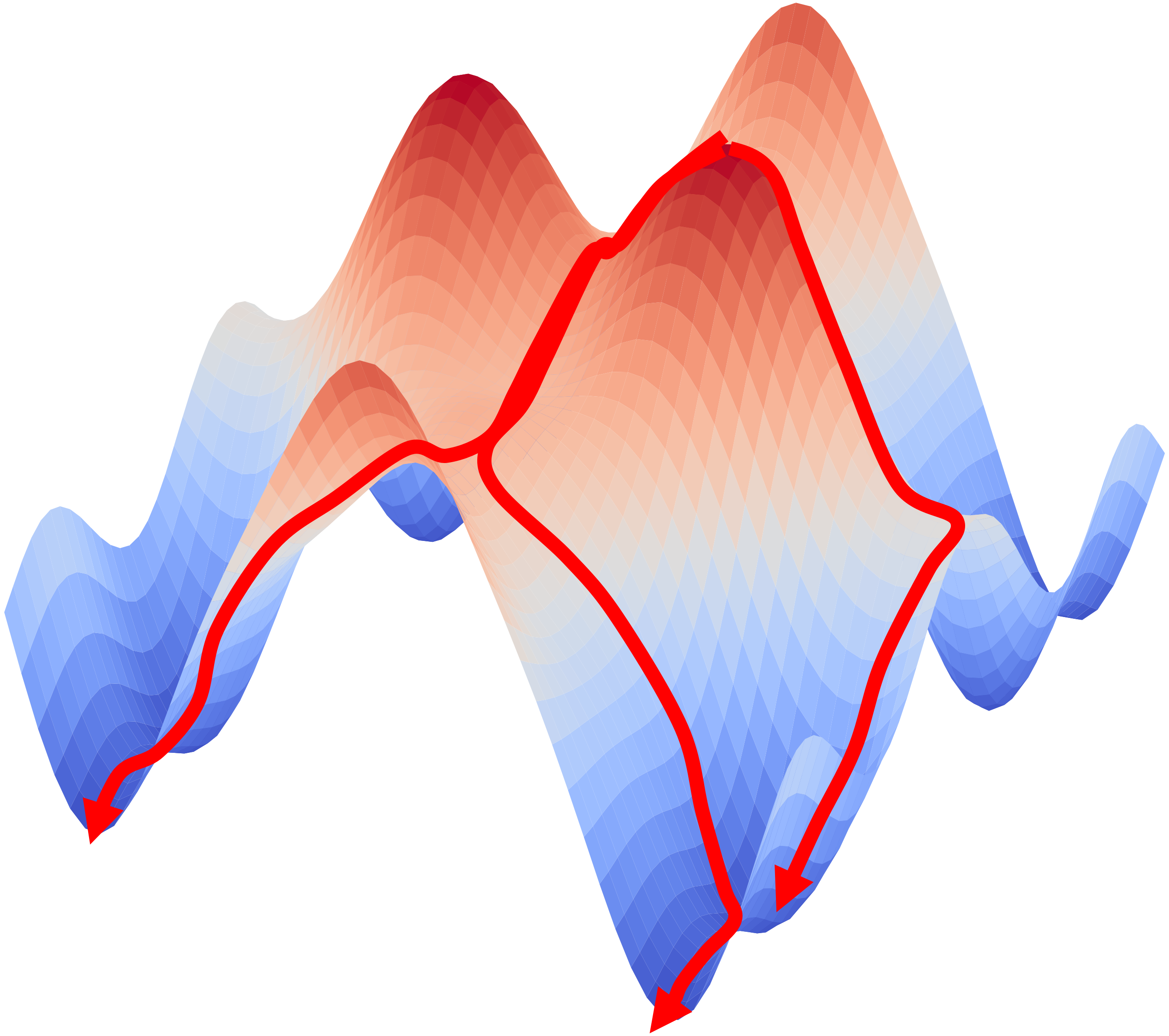

Animal
Plant
Fish
Bird
Tree
Flower


High
Frequency of error
Low
Few
Many
Number of examples

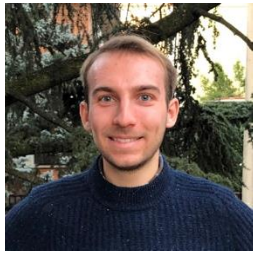
Depth introduces a hierarchy of saddle points



Individual variability amidst structure



Learning to make perceptual decisions from naïve to expert


Sam Liebana


Aeron Laffere


Chiara Toschi

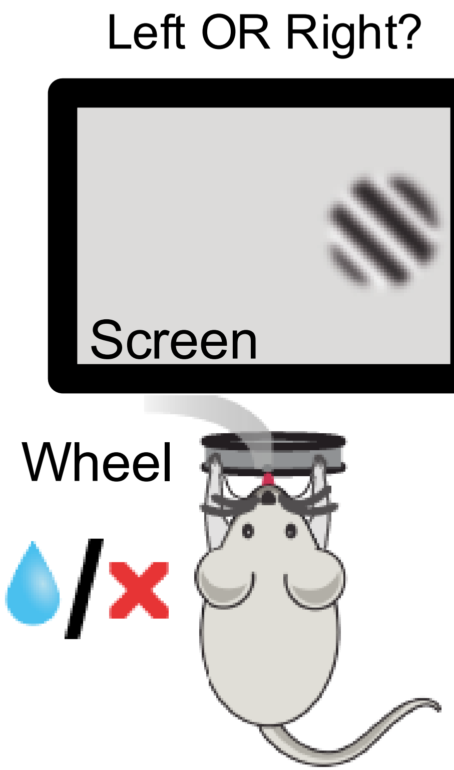

Peter
Zatka-Haas

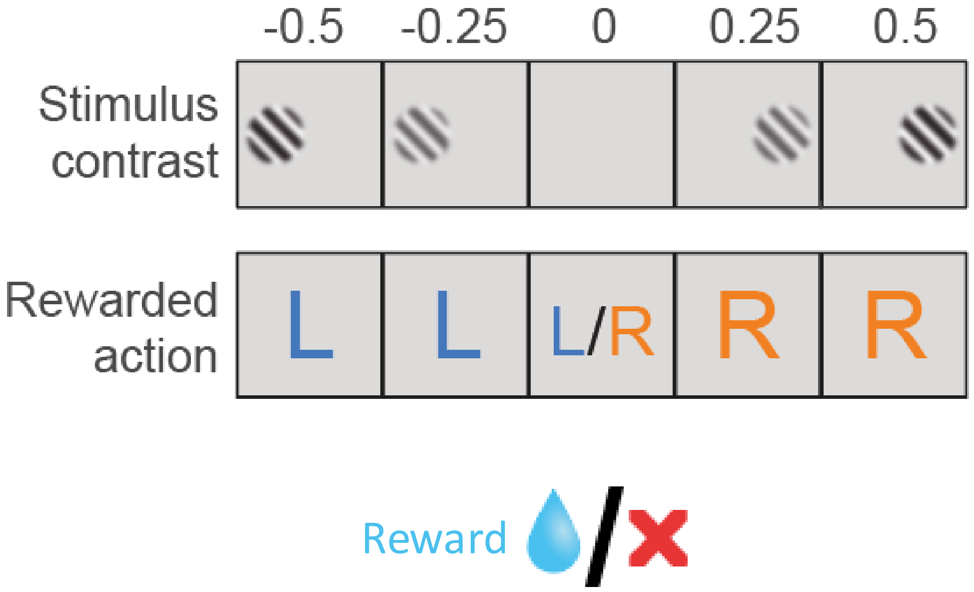

Louisa Schilling


Rafal Bogacz


Armin Lak

Learning to make perceptual decisions from naïve to expert

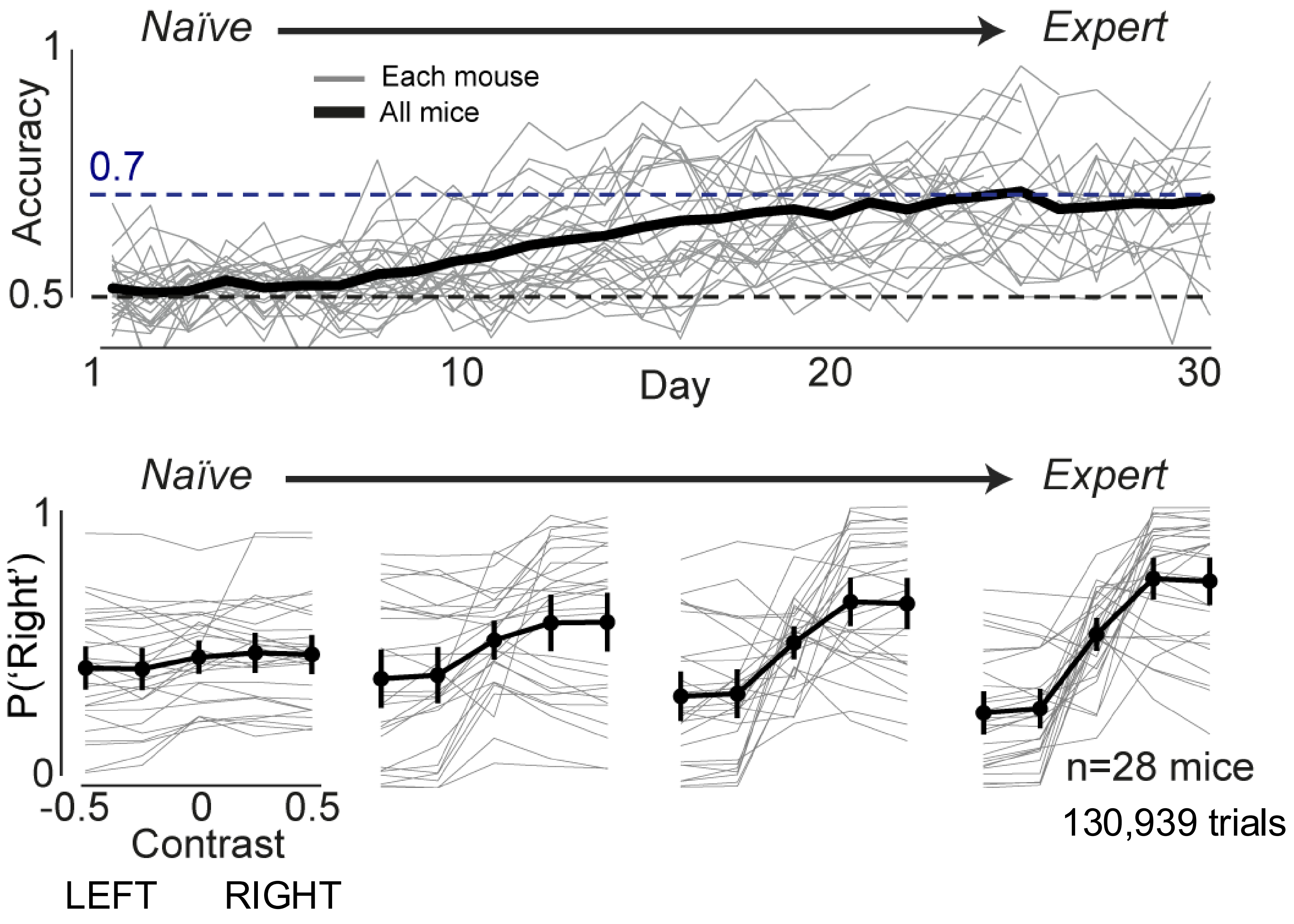




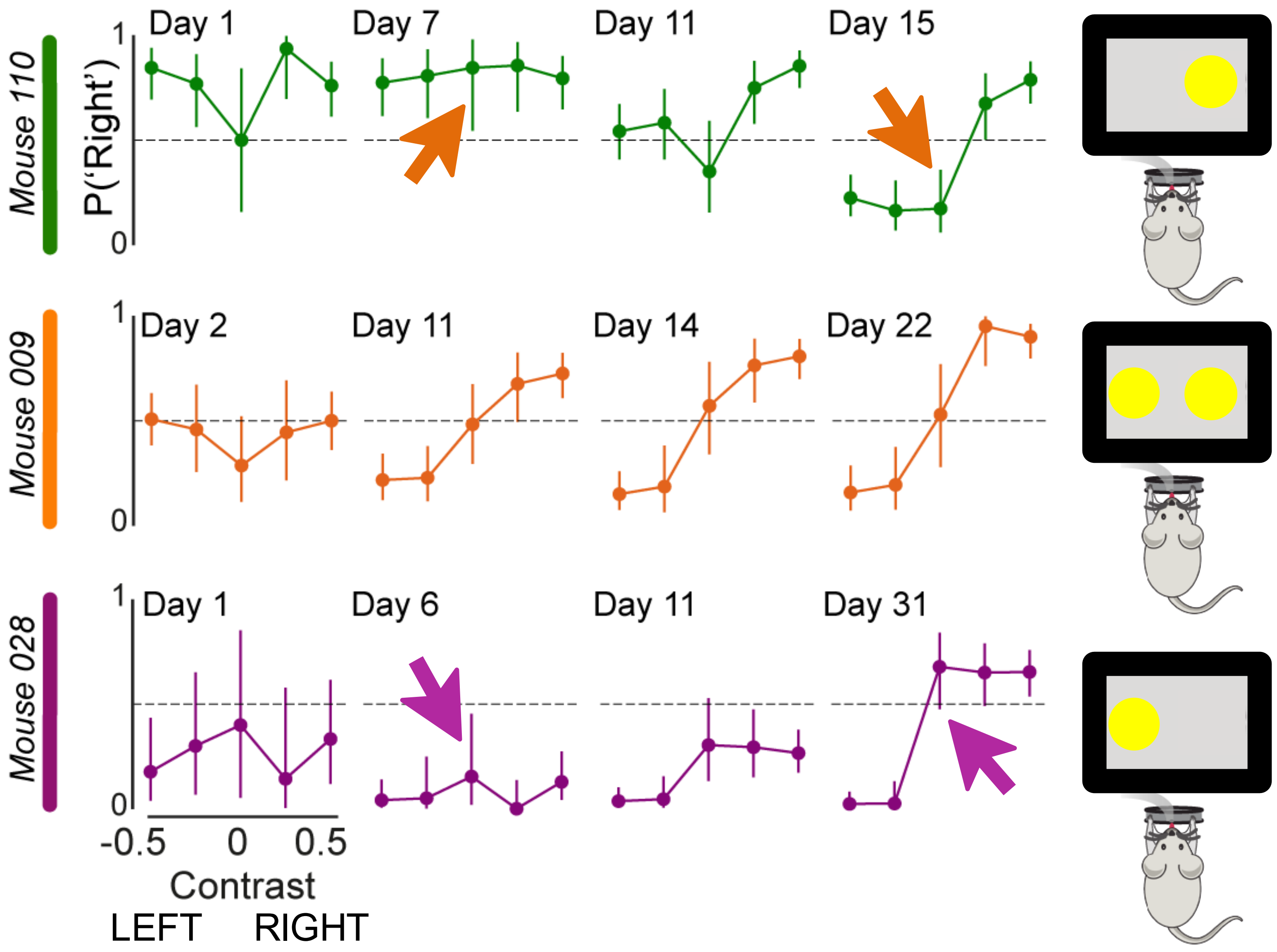
Full task from day 1 without any change over learning

Burgess, Lak*, Steinmetz*, Zatka-Haas* et al., Cell reports, 2017*

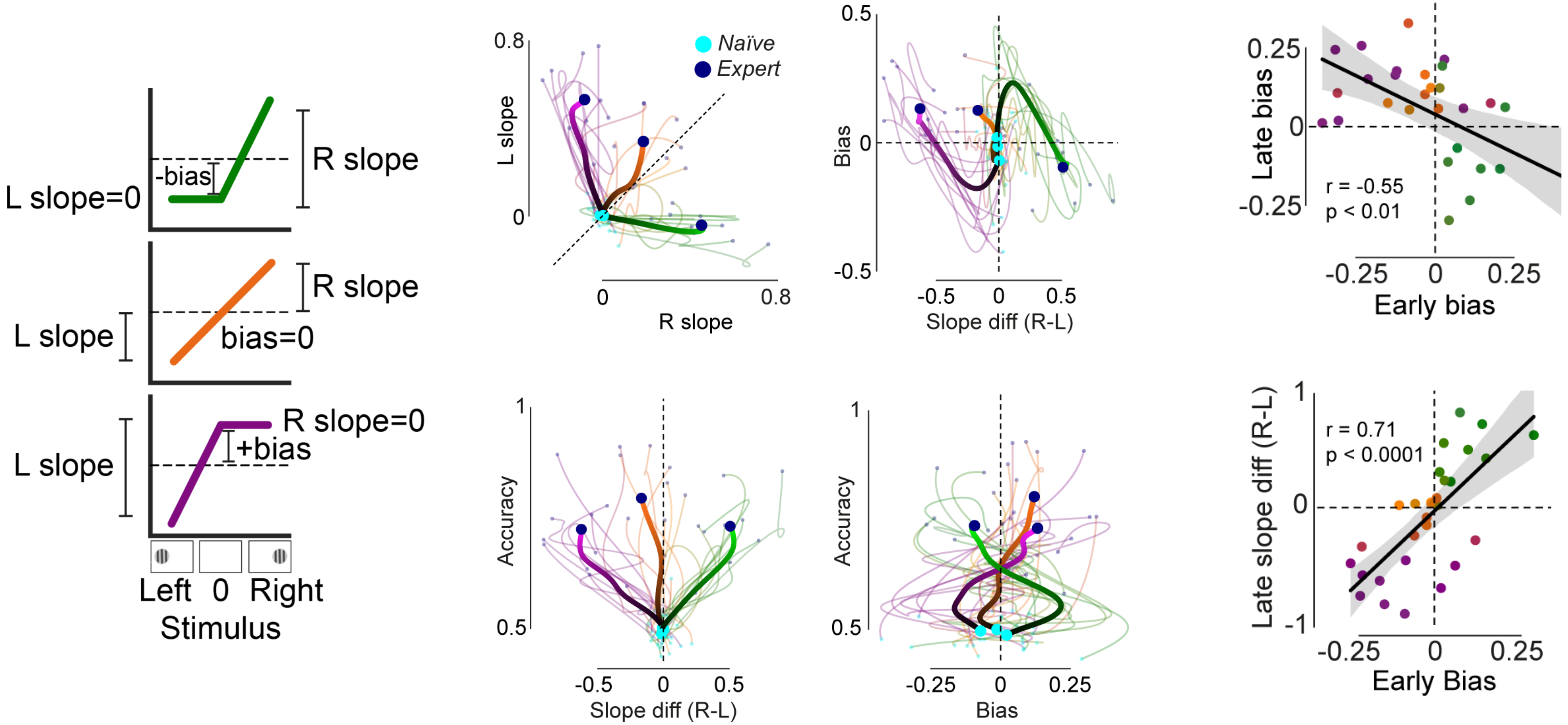
Learning to make perceptual decisions from naïve to expert



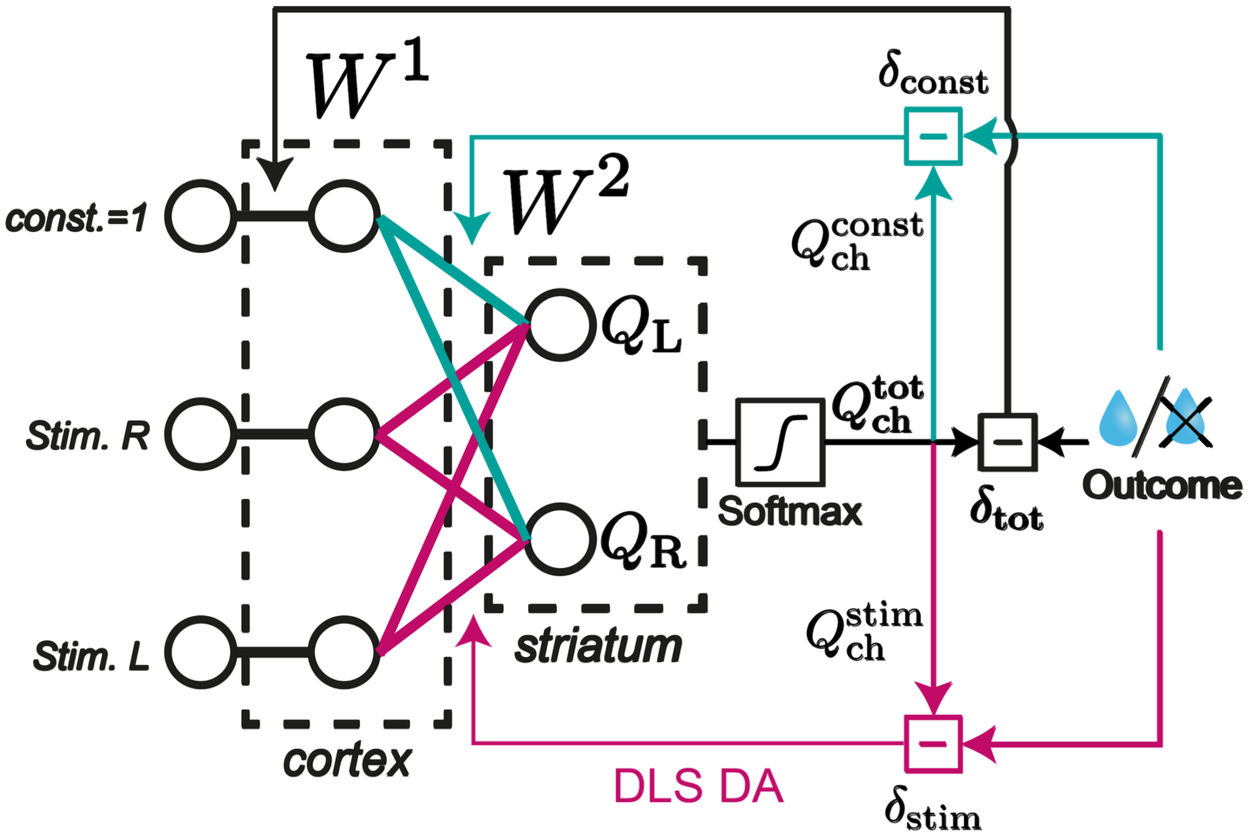
Mice exhibit diverse learning trajectories



Learning trajectories are individually diverse but systematic



A Deep RL Neural Network

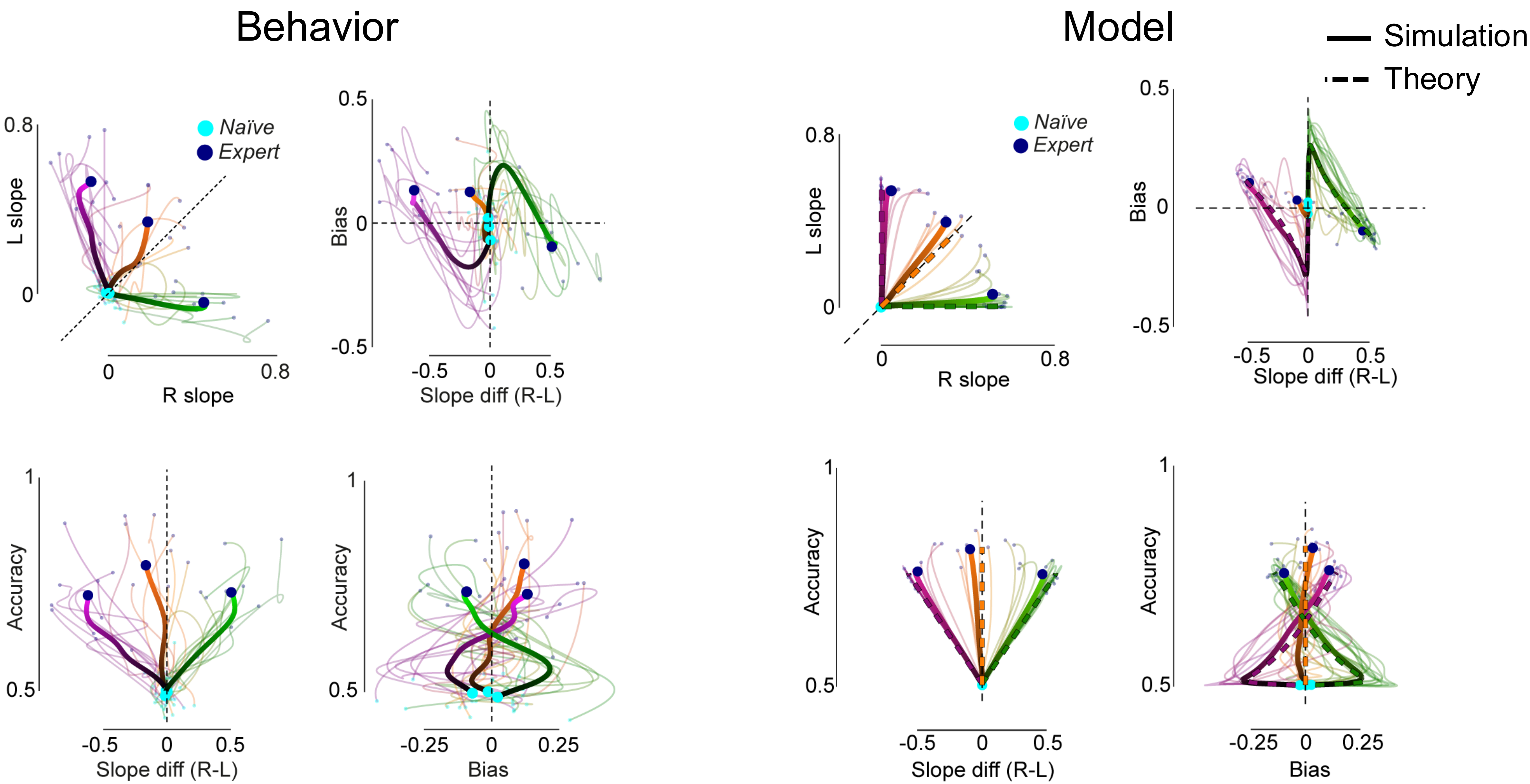


$$\mathcal{L}^{\mathrm{cortex}} = \frac{1}{2}\delta_{\mathrm{tot}}^2 = \frac{1}{2}(\mathrm{Rew} - Q_{\mathrm{ch}}^{\mathrm{tot}})^2$$

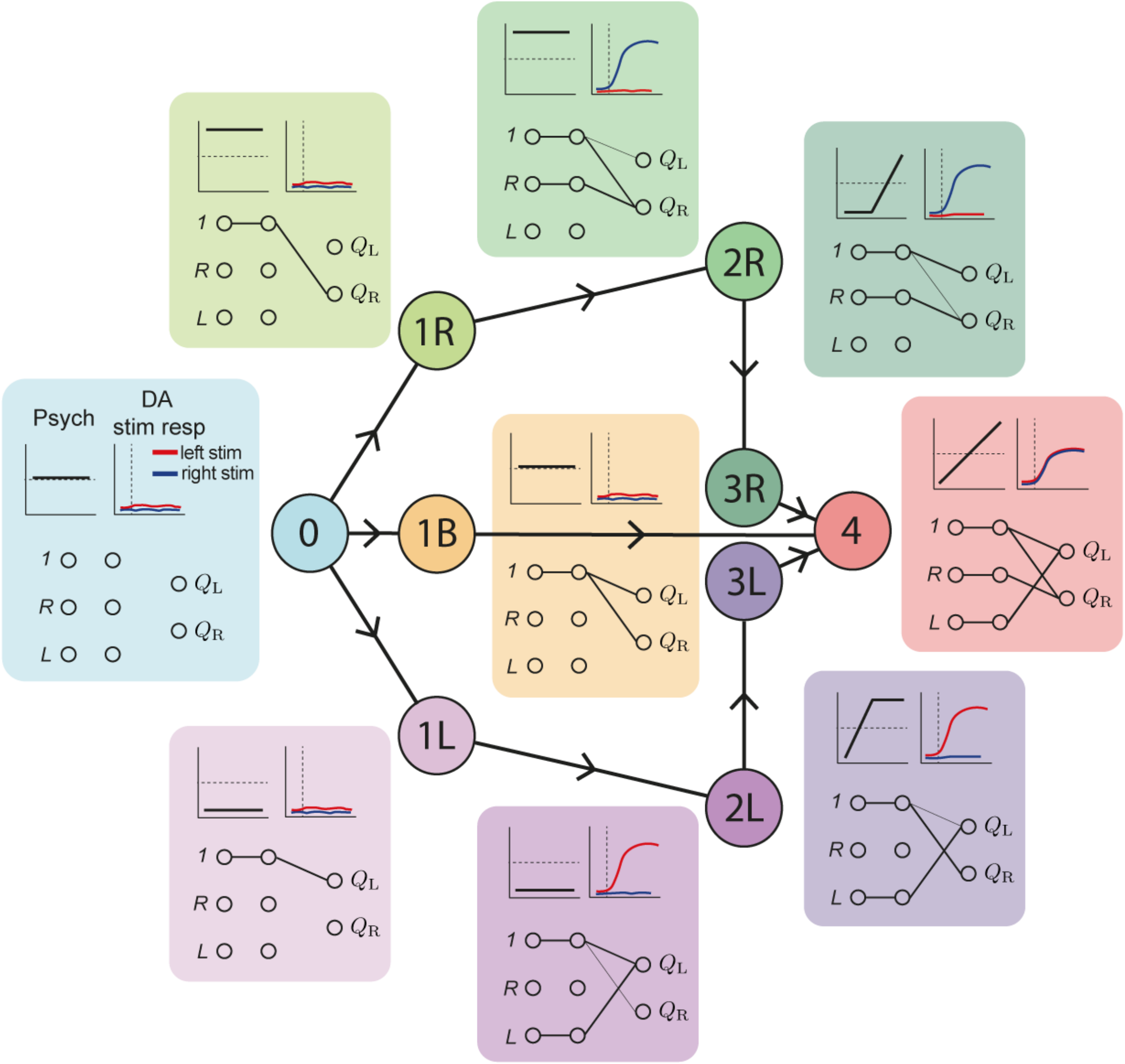
$$\mathcal{L}^{\mathrm{const}} = \frac{1}{2}\delta_{\mathrm{const}}^2 = \frac{1}{2}(\mathrm{Rew} - Q_{\mathrm{ch}}^{\mathrm{const}})^2$$

$$\mathcal{L}^{\mathrm{stim}} = \frac{1}{2}\delta_{\mathrm{stim}}^2 = \frac{1}{2}(\mathrm{Rew} - Q_{\mathrm{ch}}^{\mathrm{stim}})^2$$

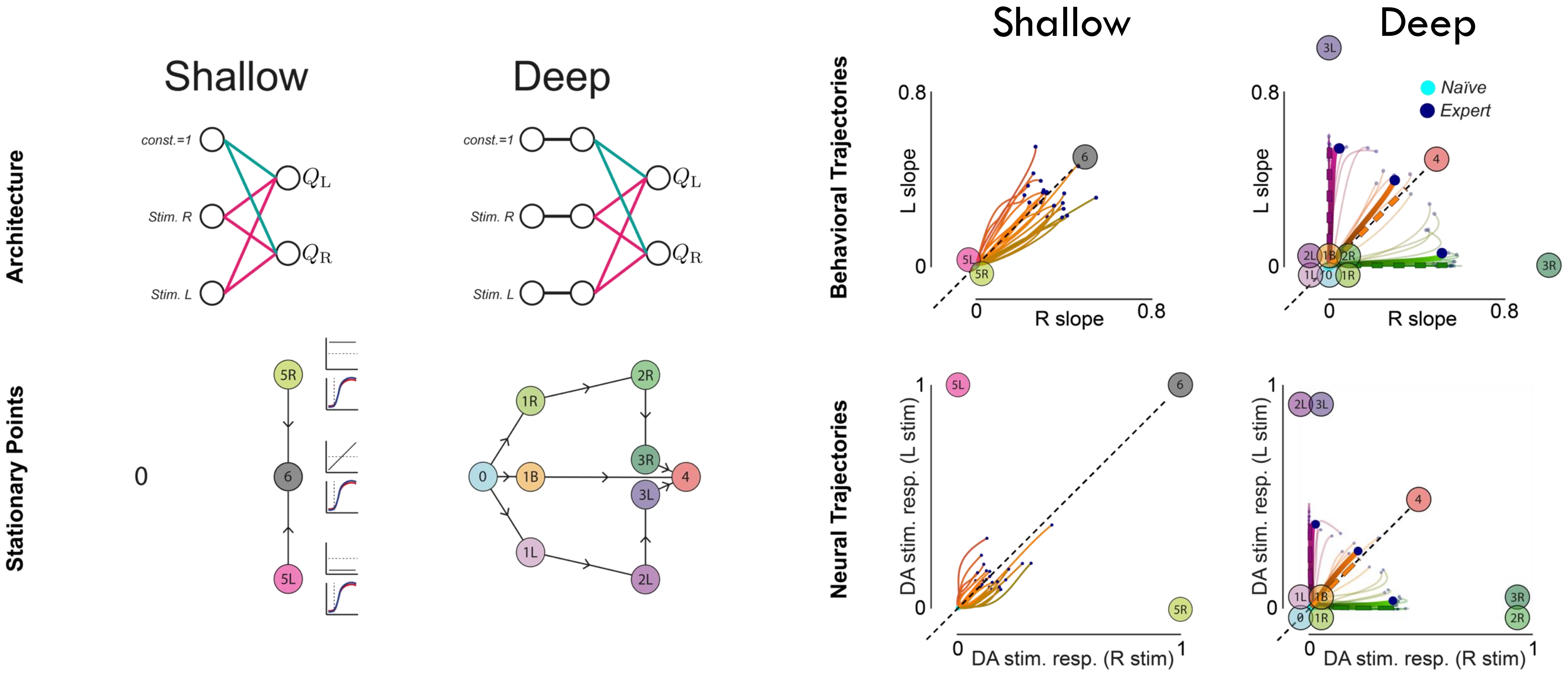
Model captures behavior



Dynamics pass near a hierarchy of saddle points



Saddle points arise through depth



Today

1. Deep linear network dynamics from *tabula rasa* initialization
2. Nontrivial initializations: Lazy, rich, & beyond
3. Nonlinear networks & the neural race reduction

Partitioned solution

$$\mathbf{QQ}^T(t) = \begin{pmatrix} \boldsymbol{Z_1}(t)\boldsymbol{A}^{-1}(t)\boldsymbol{Z_1}^T(t) & \boldsymbol{Z_1}(t)\boldsymbol{A}^{-1}(t)\boldsymbol{Z_2}^T(t) \\ \boldsymbol{Z_2}(t)\boldsymbol{A}^{-1}(t)\boldsymbol{Z_1}^T(t) & \boldsymbol{Z_2}(t)\boldsymbol{A}^{-1}(t)\boldsymbol{Z_2}^T(t) \end{pmatrix},$$

with the time-dependent variables $\boldsymbol{Z_1}(t) \in \mathbb{R}^{N_i \times N_h}$, $\boldsymbol{Z_2}(t) \in \mathbb{R}^{N_o \times N_h}$, *and* $\boldsymbol{A}(t) \in \mathbb{R}^{N_h \times N_h}$*:*

$$\boldsymbol{Z_1}(t) = \frac{1}{2}\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{G}} - \tilde{\boldsymbol{H}}\tilde{\boldsymbol{G}})e^{\tilde{\boldsymbol{S}}_\lambda \frac{t}{\tau}}\boldsymbol{B}^T - \frac{1}{2}\tilde{\boldsymbol{V}}(\tilde{\boldsymbol{G}} + \tilde{\boldsymbol{H}}\tilde{\boldsymbol{G}})e^{-\tilde{\boldsymbol{S}}_\lambda \frac{t}{\tau}}\boldsymbol{C}^T + \tilde{\mathbf{V}}_\perp e^{\boldsymbol{\lambda}_\perp \frac{t}{\tau}}\boldsymbol{D}^T, \quad (13)$$

$$\boldsymbol{Z_2}(t) = \frac{1}{2}\tilde{\boldsymbol{U}}(\tilde{\boldsymbol{G}} + \tilde{\boldsymbol{H}}\tilde{\boldsymbol{G}})e^{\tilde{\boldsymbol{S}}_\lambda \frac{t}{\tau}}\boldsymbol{B}^T + \frac{1}{2}\tilde{\boldsymbol{U}}(\tilde{\boldsymbol{G}} - \tilde{\boldsymbol{H}}\tilde{\boldsymbol{G}})e^{-\tilde{\boldsymbol{S}}_\lambda \frac{t}{\tau}}\boldsymbol{C}^T + \tilde{\mathbf{U}}_\perp e^{\boldsymbol{\lambda}_\perp \frac{t}{\tau}}\boldsymbol{D}^T, \quad (14)$$

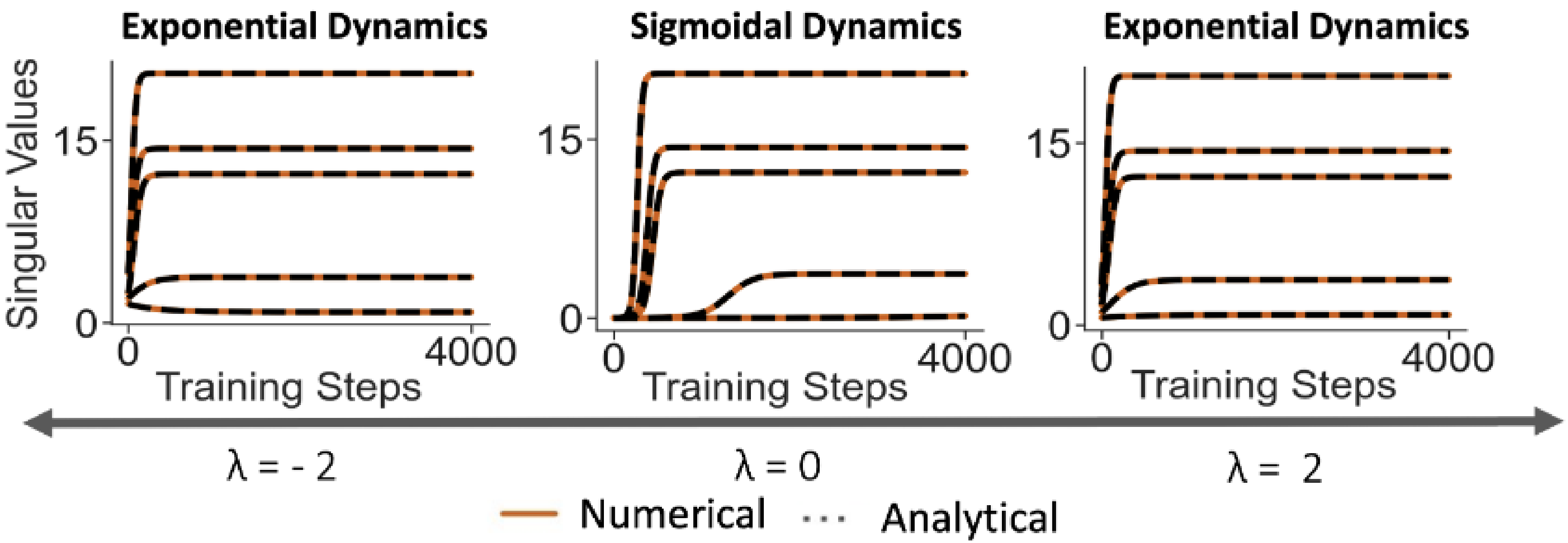
$$\boldsymbol{A}(t) = \mathbf{I} + \boldsymbol{B}\left(\frac{e^{2\tilde{\boldsymbol{S}}_\lambda \frac{t}{\tau}} - \mathbf{I}}{4\tilde{\boldsymbol{S}}_\lambda}\right)\boldsymbol{B}^T - \boldsymbol{C}\left(\frac{e^{-2\tilde{\boldsymbol{S}}_\lambda \frac{t}{\tau}} - \mathbf{I}}{4\tilde{\boldsymbol{S}}_\lambda}\right)\boldsymbol{C}^T + \boldsymbol{D}\left(\frac{e^{\boldsymbol{\lambda}_\perp \frac{t}{\tau}} - \mathbf{I}}{\boldsymbol{\lambda}_\perp}\right)\boldsymbol{D}^T. \quad (15)$$

and

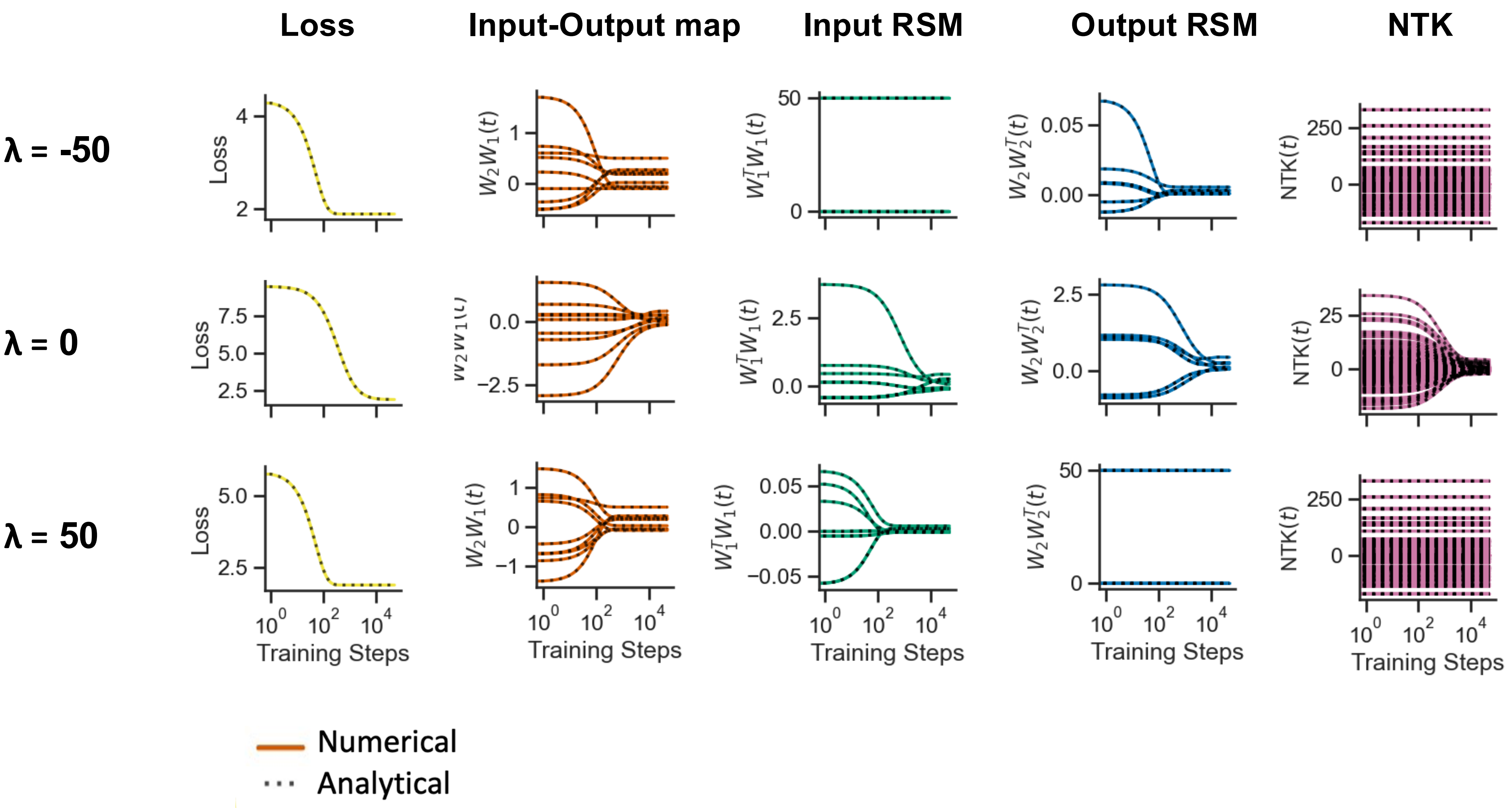
$$\tilde{\boldsymbol{S}}_\lambda = \sqrt{\tilde{\boldsymbol{S}}^2 + \frac{\lambda^2}{4}\mathbf{I}}, \quad \boldsymbol{\lambda}_\perp = \mathrm{sgn}(N_o - N_i)\frac{\lambda}{2}\mathbf{I}_{|N_o - N_i|}, \quad \tilde{\boldsymbol{H}} = \mathrm{sgn}(\lambda)\sqrt{\frac{\tilde{\boldsymbol{S}}_{\boldsymbol{\lambda}} - \tilde{\boldsymbol{S}}}{\tilde{\boldsymbol{S}}_{\boldsymbol{\lambda}} + \tilde{\boldsymbol{S}}}}, \quad \tilde{\boldsymbol{G}} = \frac{1}{\sqrt{\mathbf{I} + \tilde{\boldsymbol{H}}^2}}.$$

*where **B**, **C**, **D** are initialization-dependent matrices.*

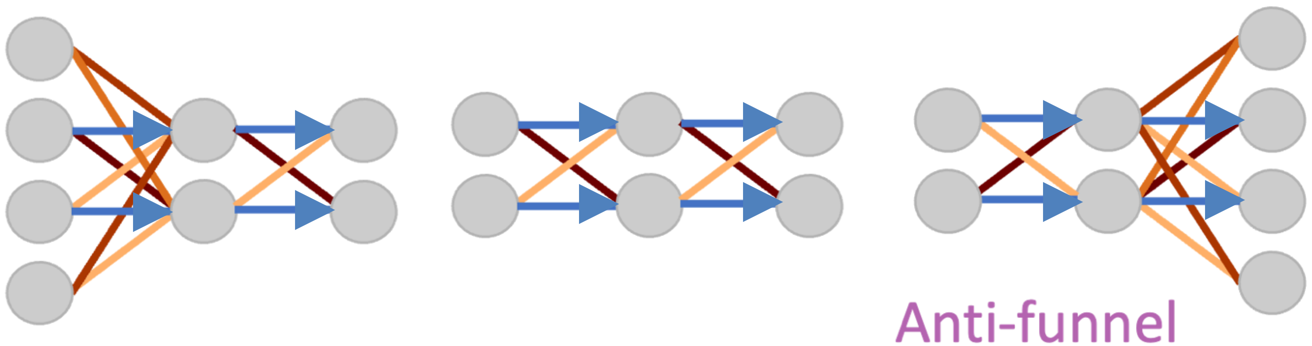
From exponential to sigmoidal dynamics



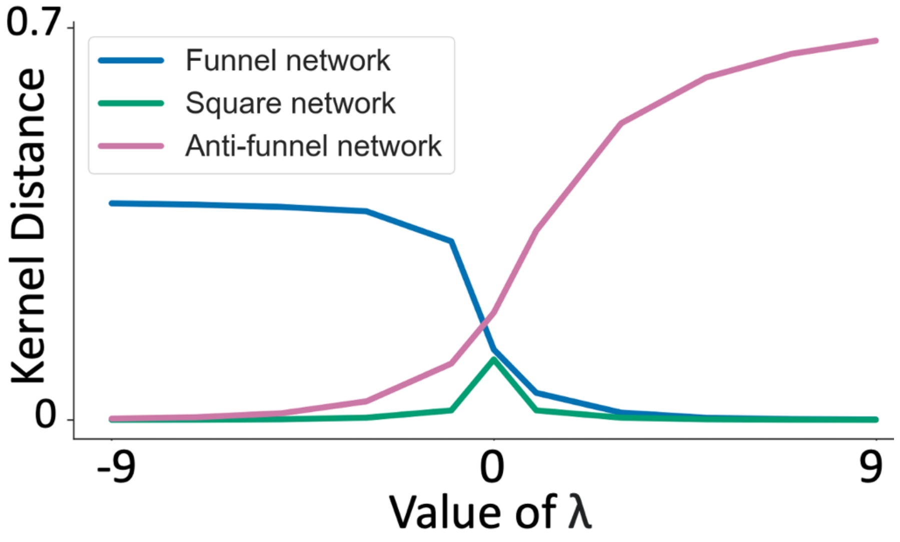
Rich and lazy learning



Architecture and learning regime

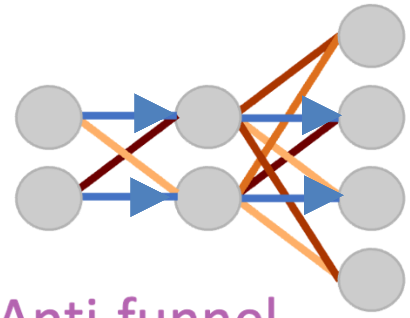


Funnel network Square network Anti-funnel network

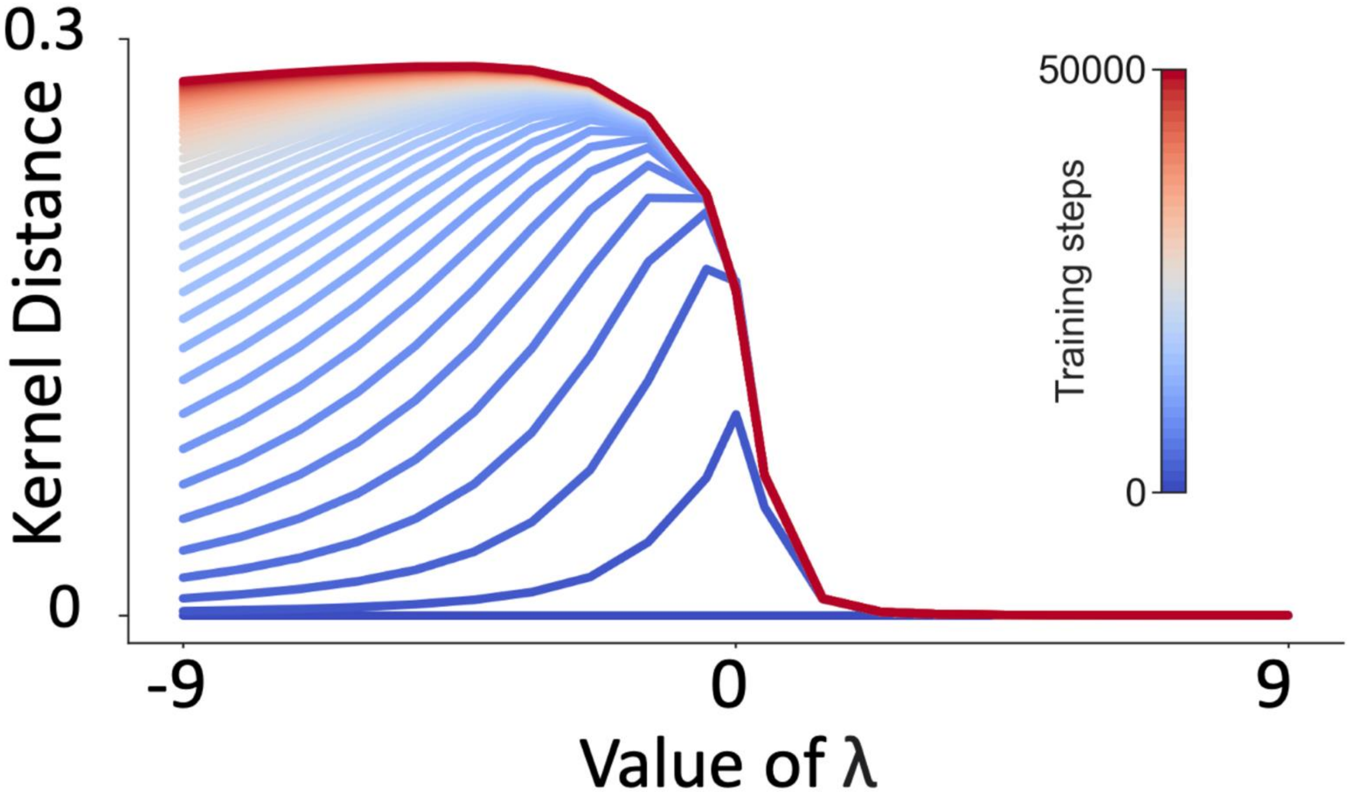


[Dominé and Anguita et al. 2024]

Delayed rich regime

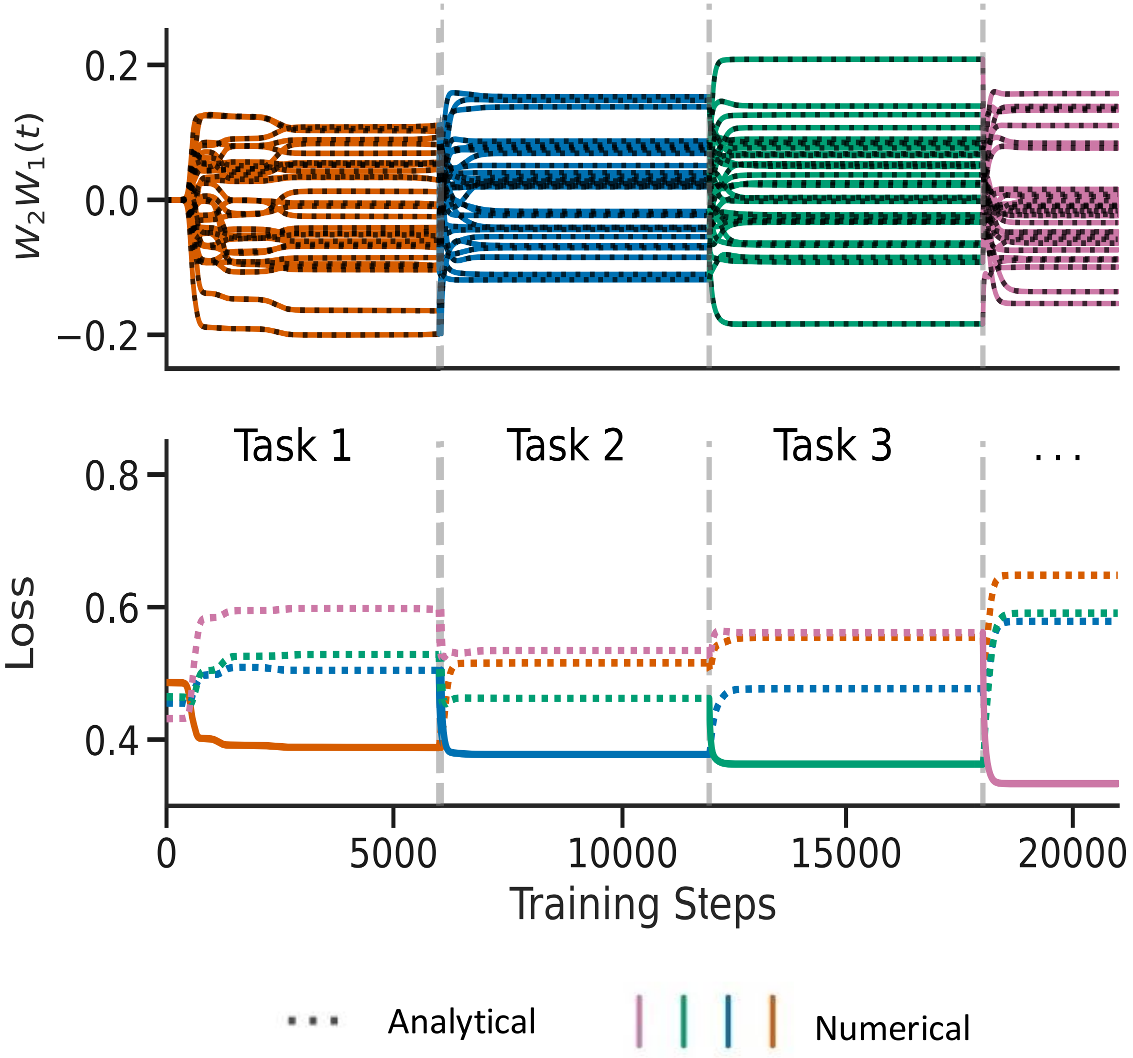


Anti-funnel network



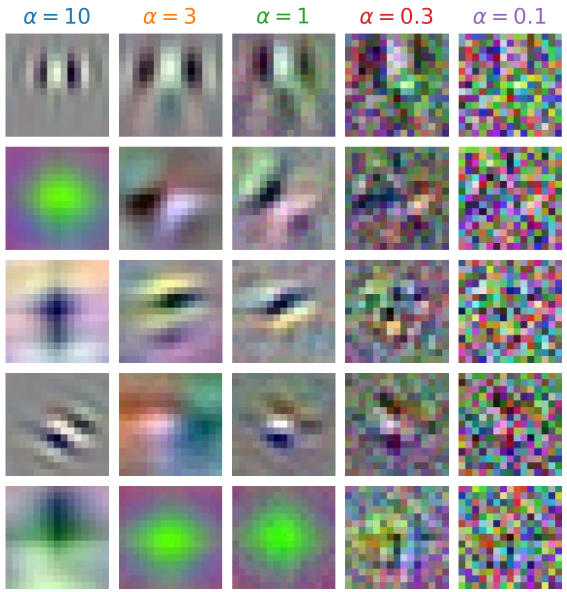
[Dominé and Anguita et al. 2024]

Exact continual learning dynamics

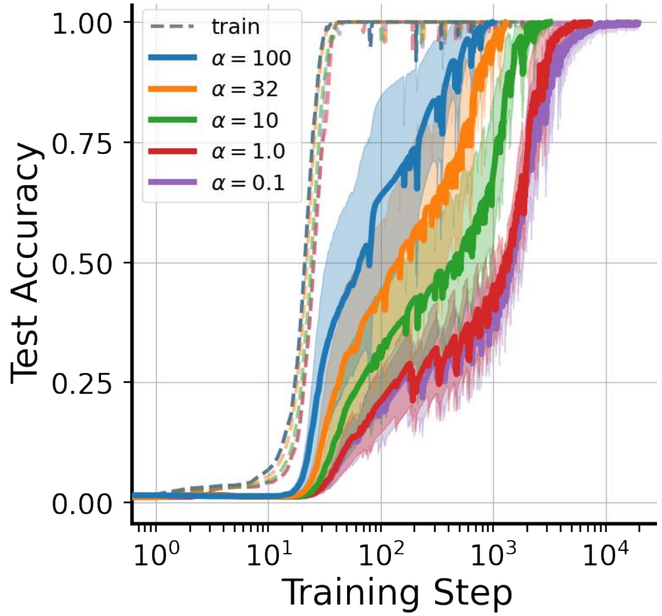


[Braun and Dominé et al. 2022, Dominé and Anguita et al. 2024]

Impact of relative scale initializations in practice



Promotes interpretability of early layers in CNNs



Decreases the time to grokking in modular arithmetic

Kunin et al. 2024

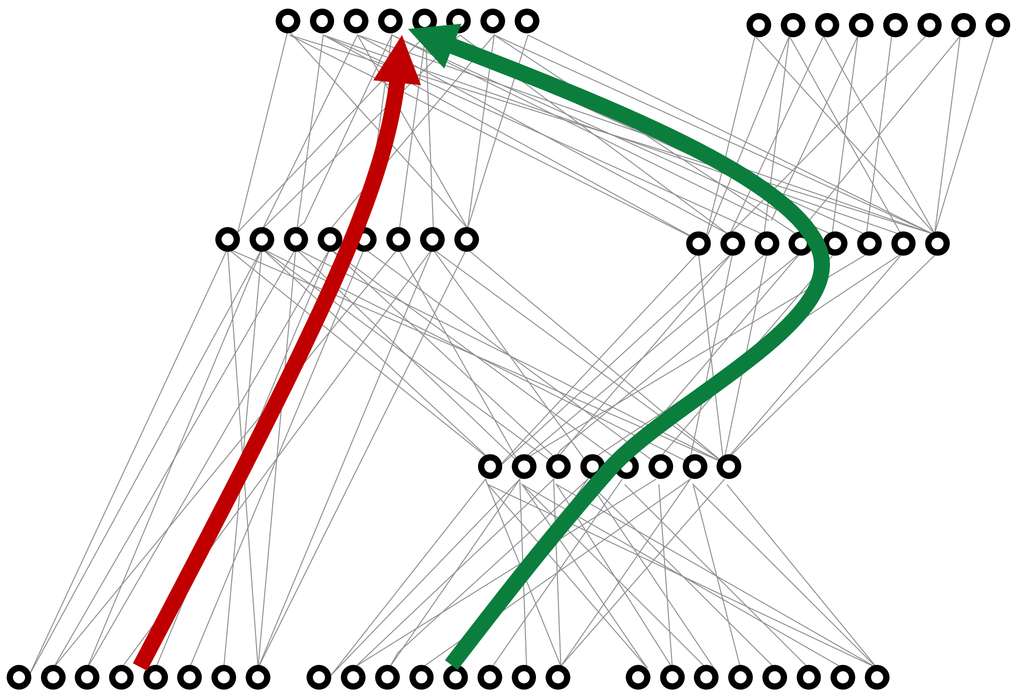
Today

1. Deep linear network dynamics from *tabula rasa* initialization

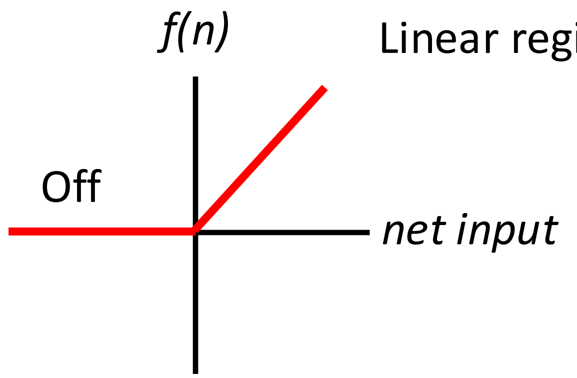
2. Nontrivial initializations: Lazy, rich, & beyond

3. Nonlinear networks & the neural race reduction

Gating: a simple view of nonlinearity



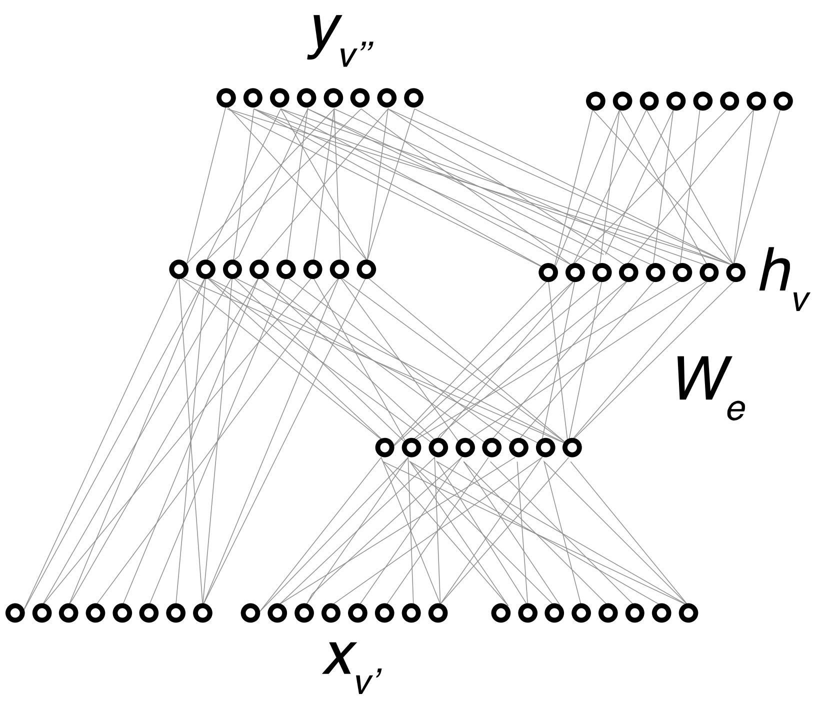
ReLU neural nonlinearity



When active, each pathway behaves like a deep linear network

Lippl et al., 2022; Saxe et al., 2022; Li & Sompolinsky, 2022

Gated Deep Linear Network



Arch graph Γ: nodes V, edges E

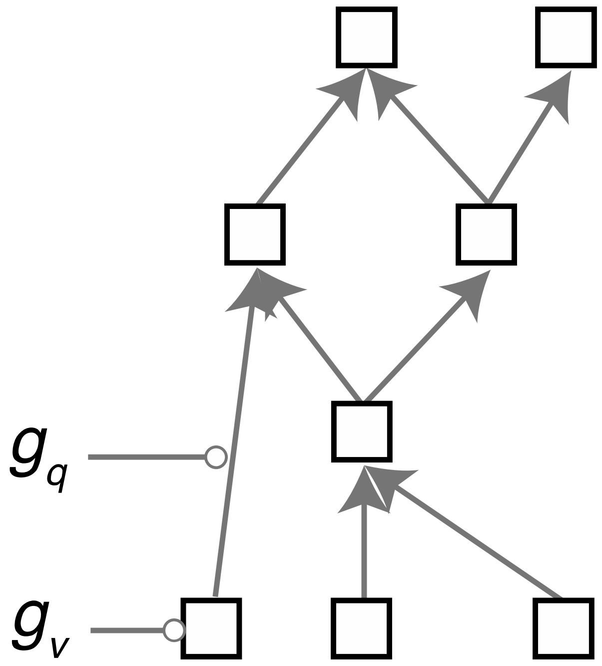
$y_{v''}$

h_v

W_e

$x_{v'}$

Gated Deep Linear Network



Forward propagation:

$$h_v = g_v \sum_{q \in E : t(q) = v} g_q W_q h_{s(q)}$$

$s(q)$: source node of edge q
$t(q)$: target node of edge q

Gradient descent

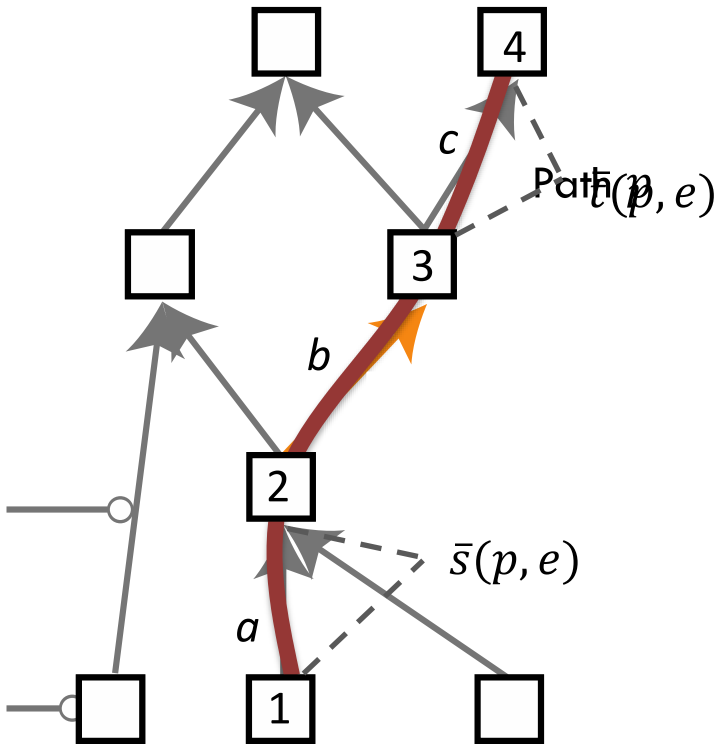
Minimize L_2 loss

$$\mathcal{L}(\{W\}) = \left\langle \frac{1}{2} \sum_{v \in \mathrm{Out}(\Gamma)} \|y_v - h_v\|_2^2 \right\rangle_{x,y,g}$$

using gradient flow on the weights

$$\tau \frac{d}{dt} W_e = -\frac{\partial \mathcal{L}(\{W\})}{\partial W_e} \quad \forall e \in E$$

Gradient descent



Path notation

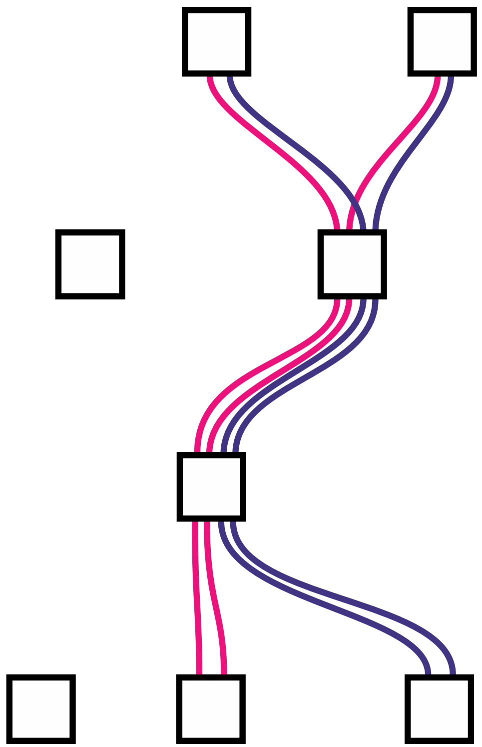
$$W_p = W_c W_b W_a$$

$$g_p = g_4 g_c g_3 g_b g_2 g_a g_1$$

$\bar{t}(p, e)$: target path of e

$\bar{s}(p, e)$: source path of e

Gradient descent



$$\tau \frac{d}{dt} W_e = \sum_{p \in \mathcal{P}(e)} W^T_{\bar{t}(p,e)} \mathcal{E}(p) W^T_{\bar{s}(p,e)}$$

$\mathcal{P}(e)$: All paths through e

$$\mathcal{E}(p) = \Sigma^{yx}(p) - \sum_{j \in \mathcal{T}(p)} \mathrm{W_j} \Sigma^{x}(j,p)$$

$\mathcal{T}(e)$: All paths terminating at same node as p

Correlation matrices

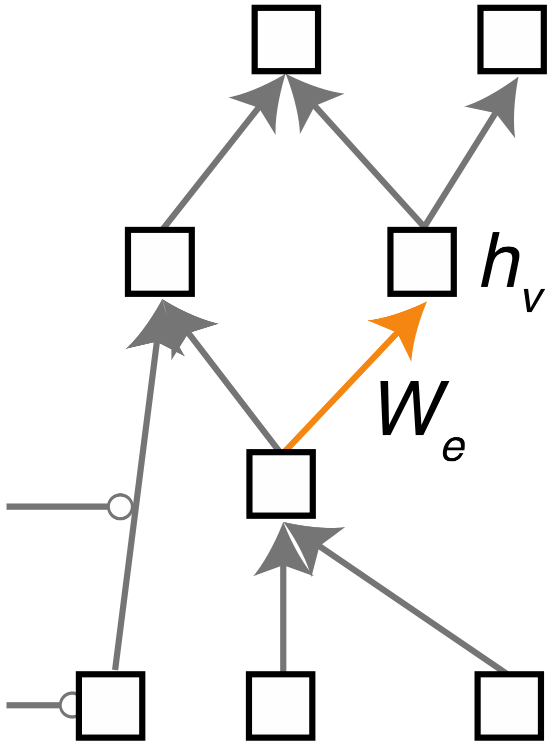
- Dynamics driven only by statistics:

$$\Sigma^{yx}(p) = \left\langle g_p y_{t(p)} x_{s(p)}^T \right\rangle_{y,x,g}$$

$$\Sigma^{x}(j,p) = \left\langle g_j x_{s(j)} x_{s(p)}^T g_p \right\rangle_{y,x,g}$$

- One correlation matrix per path

Intuition

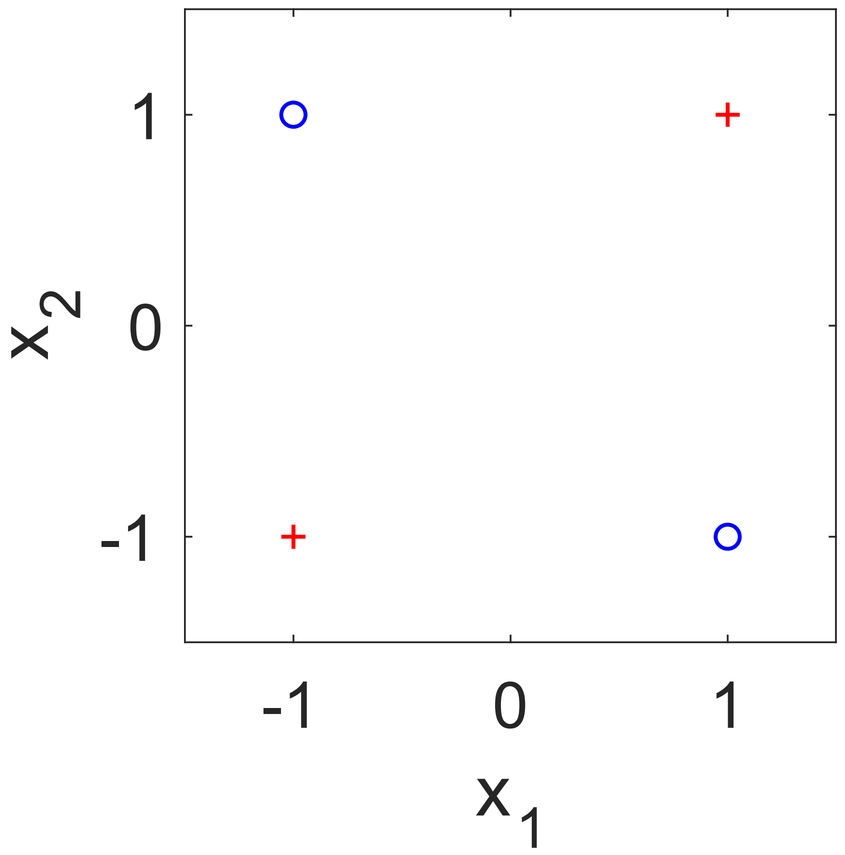


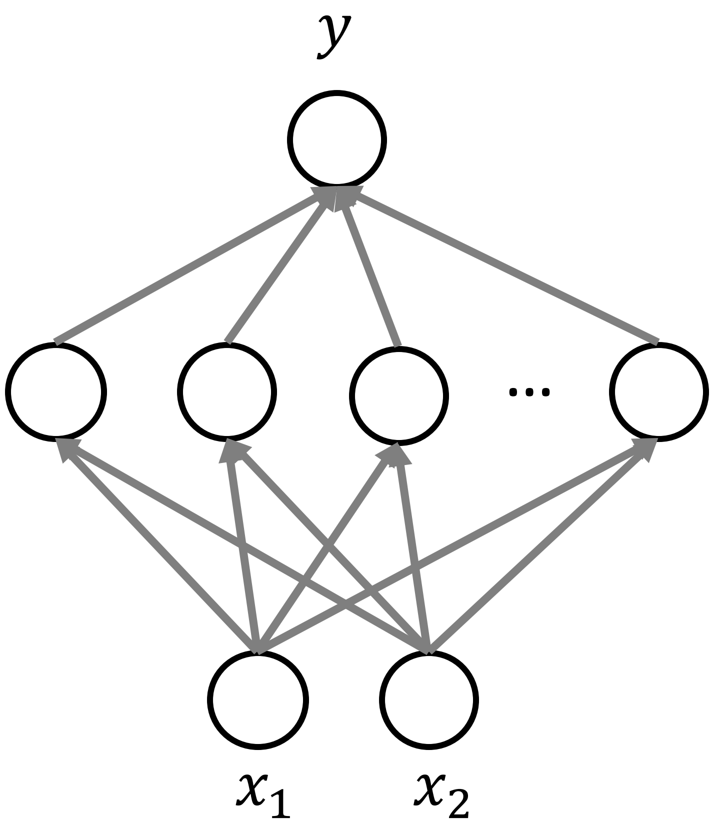
Each pathway behaves like a deep linear network

Gating controls the *effective dataset* for each pathway

All paths through an edge sum to determine dynamics

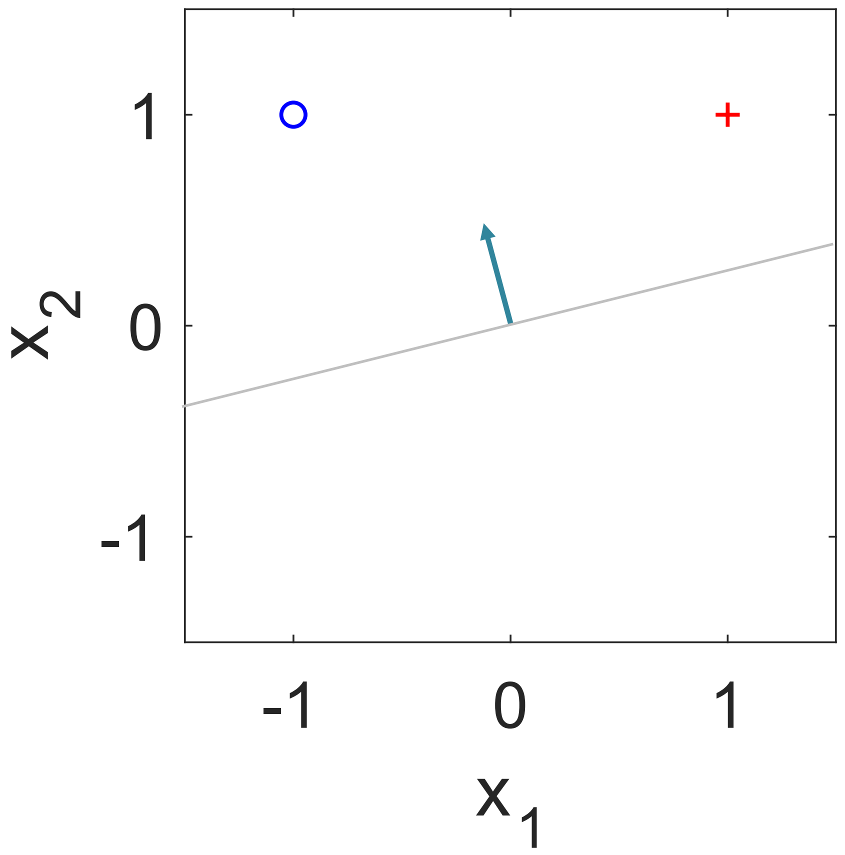
The XoR problem

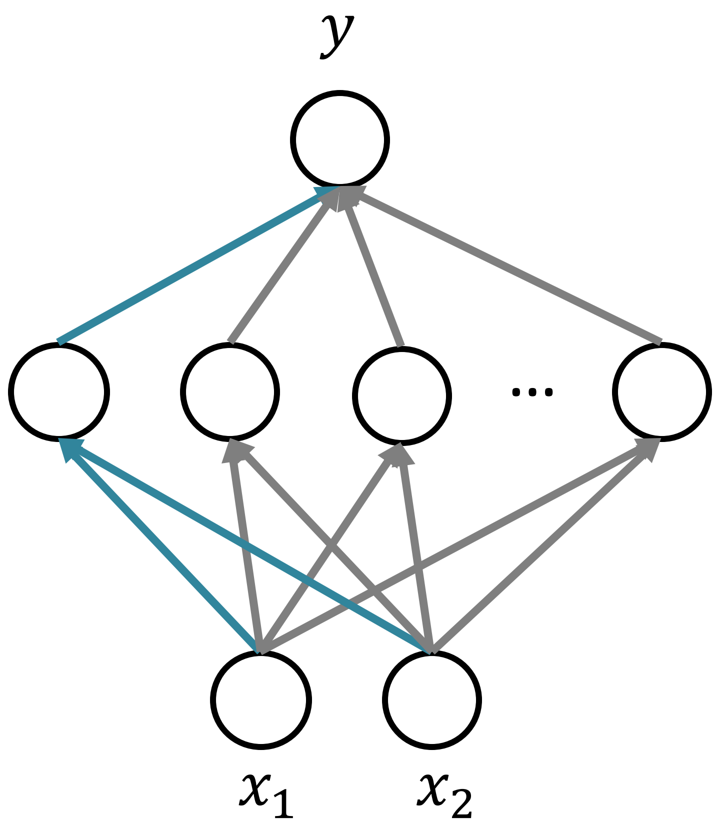




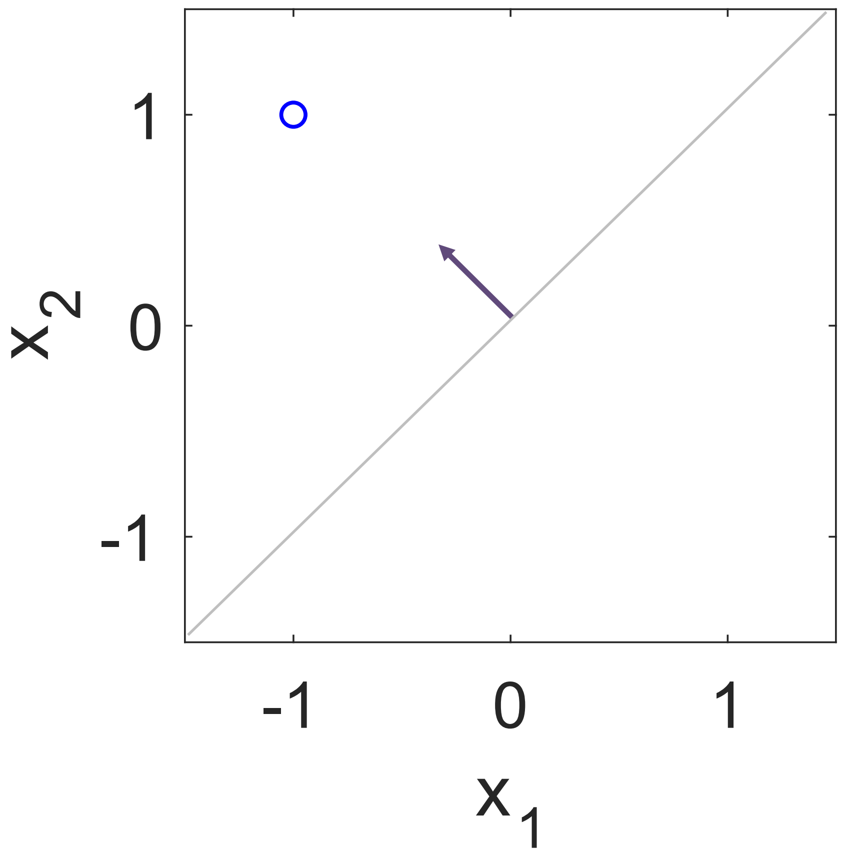
Minsky & Papert, 1969; Rumelhart, Hinton & Williams, 1986

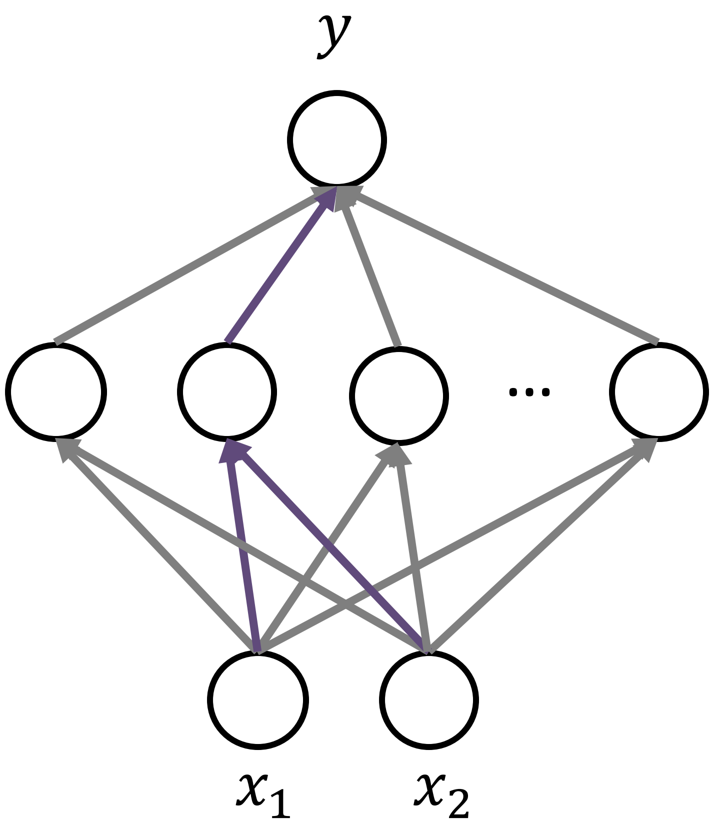
Gated dynamics



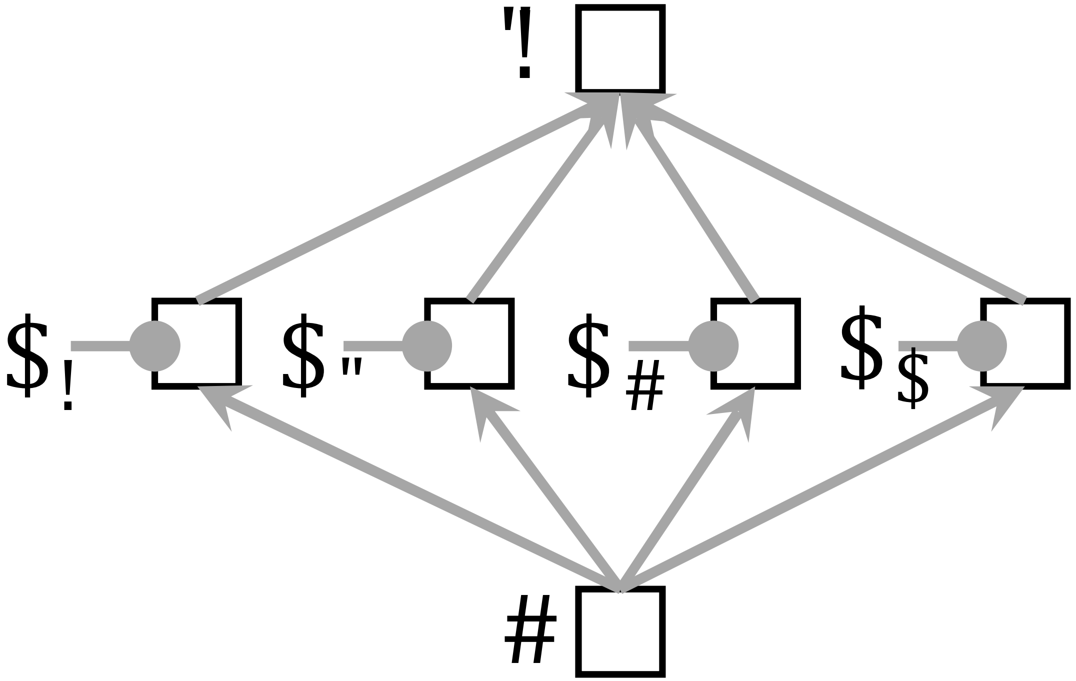


Gated dynamics



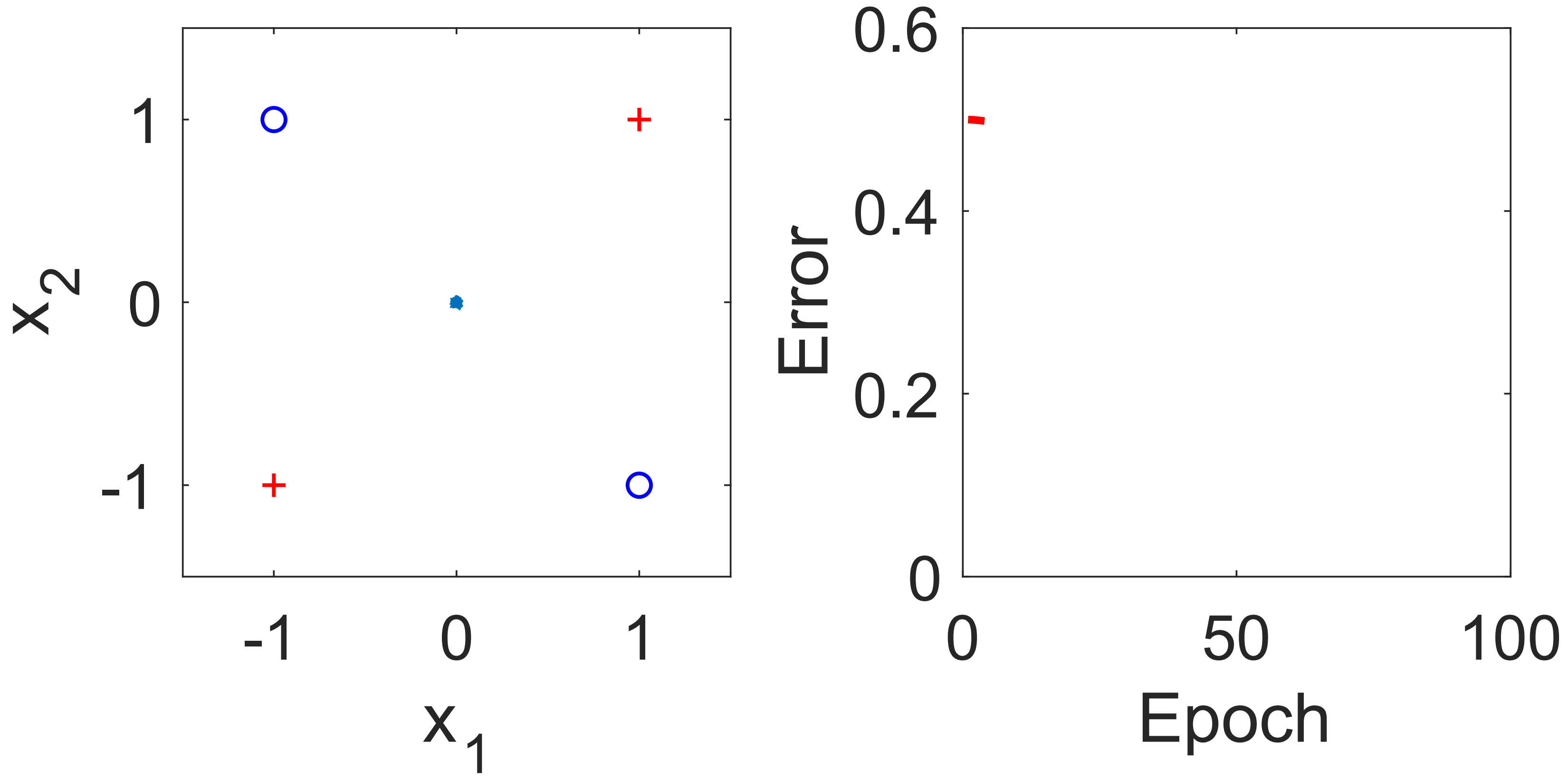


Gated DLN on XoR

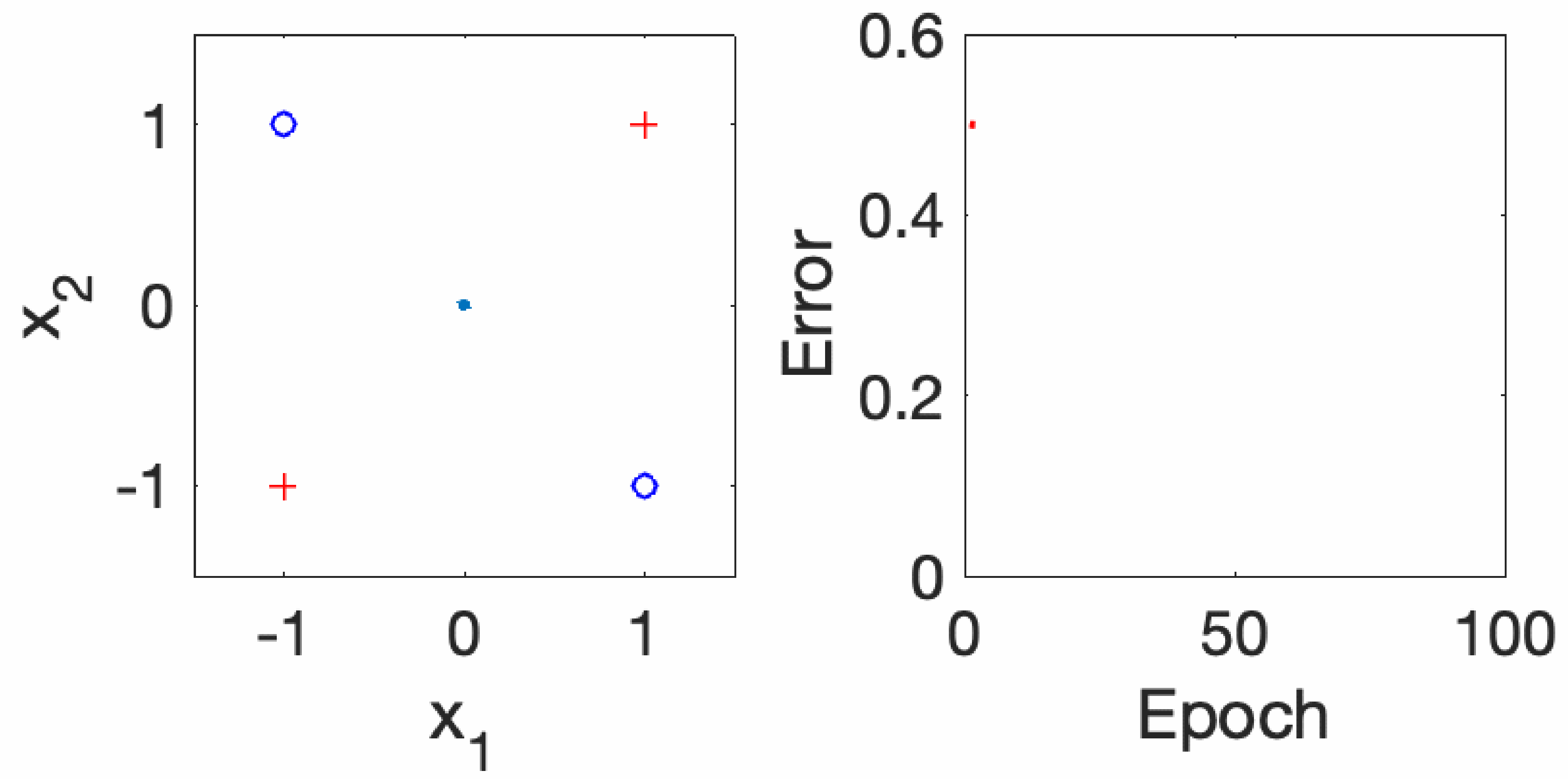


$$g_i = \begin{cases} 1 \text{ on example } i \\ 0 \text{ otherwise} \end{cases}$$

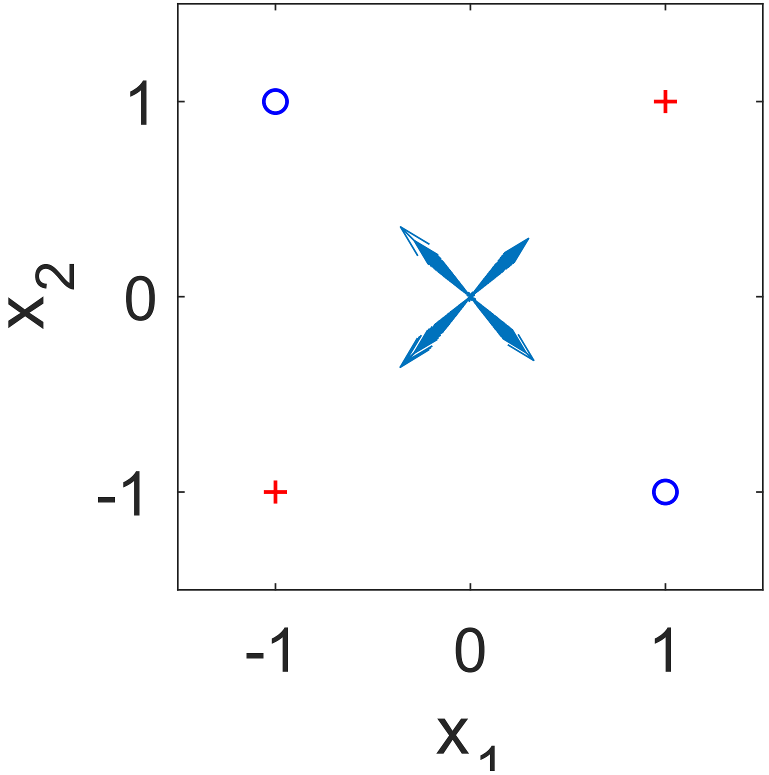
XoR Dynamics

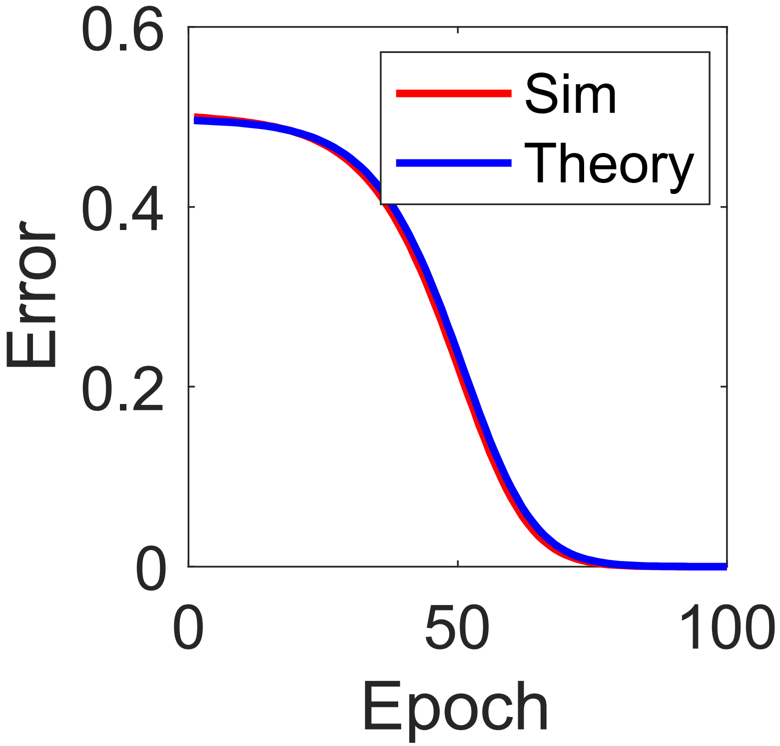


XoR Dynamics



XoR Dynamics





Reduction and (occasionally) exact solutions

"Decoupled" initialization:

$$W_e(t) = R_{t(e)} B_e(t) R_{s(e)}^T \quad \forall e$$

Mutually diagonalizable correlations:

$$\begin{aligned} \Sigma^{yx}(p) &= U_{t(p)} S(p) V_{s(p)}^T \\ \Sigma^{x}(j,p) &= V_{s(j)} D(j,p) V_{s(p)}^T \end{aligned}$$

Reduction:

$$\tau \frac{d}{dt} B_e = \sum_{p \in P(e)} B_{p \setminus e} \left[S(p) - \sum_{j \in T(t(p))} B_j D(j,p) \right]$$

Assumptions & caveats

- Reduction exact for GDLNs

- Also exact for ReLU networks under the assumptions:
  - Gates on each example match the activity set
  - No neurons switch their activity set
  - Initial weights are decoupled

- Can approximate ReLU networks with small random weights, but not always

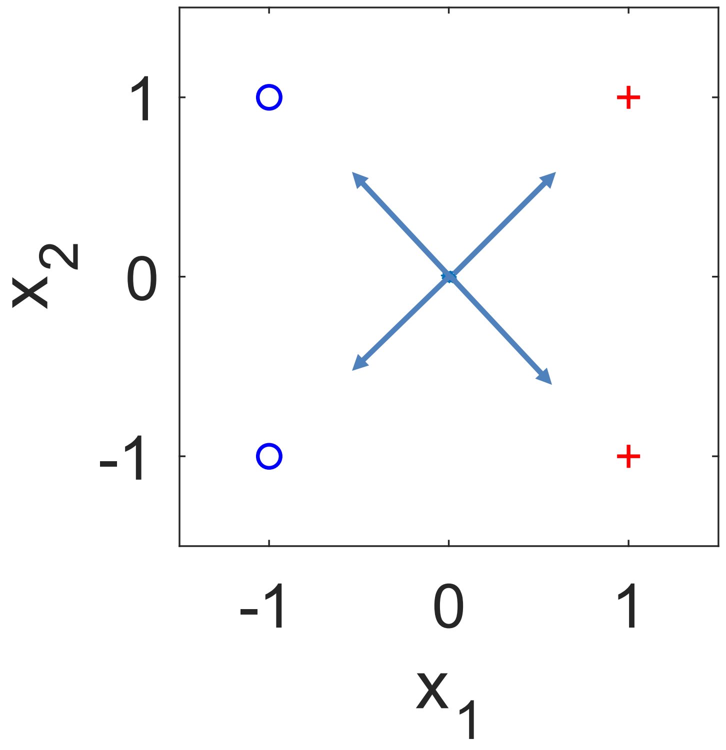
The neural race reduction

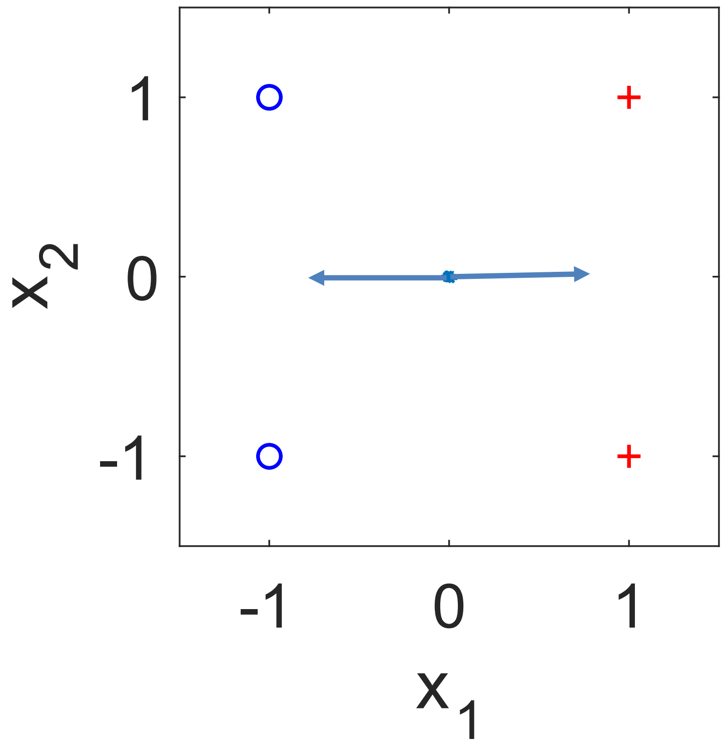
- In a large network with many pathways, these compete to reduce the global error

- A pathway's learning speed depends on:
  - Effective dataset (larger input-output correlation faster)
  - Pathway depth (deeper generally slower)
  - Initialization (larger/imbalanced generally faster)
  - Edge sharing (more pathways through edge generally faster)

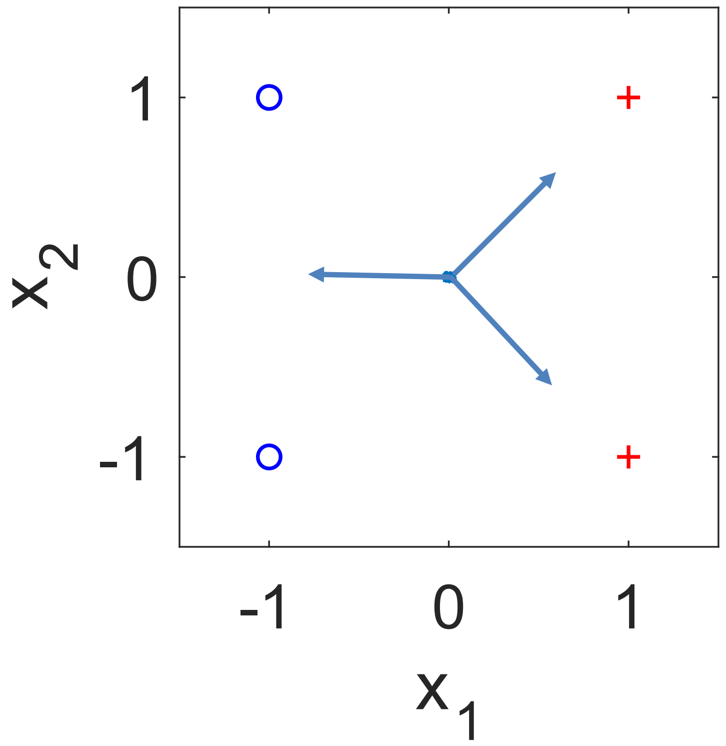
- The fastest pathways can dominate the solution

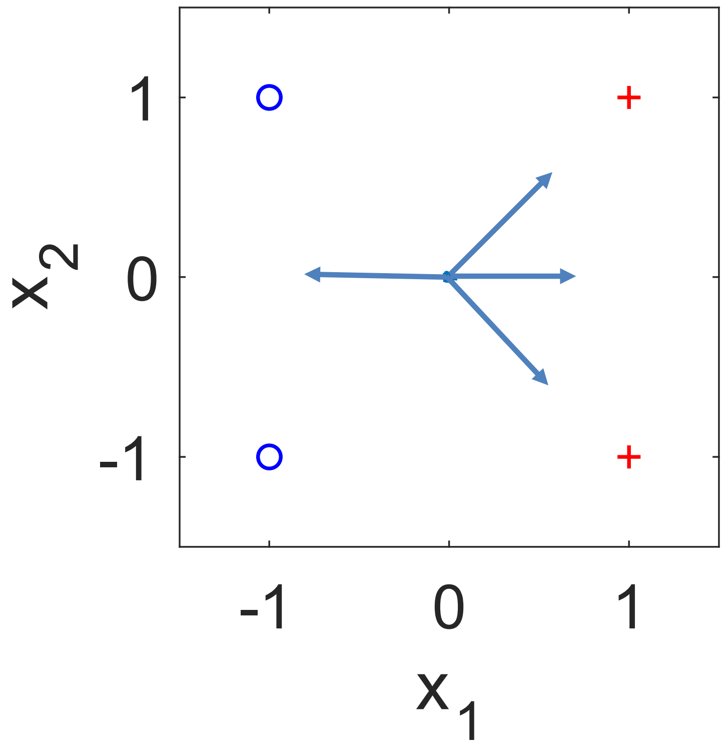
Which gating structures?





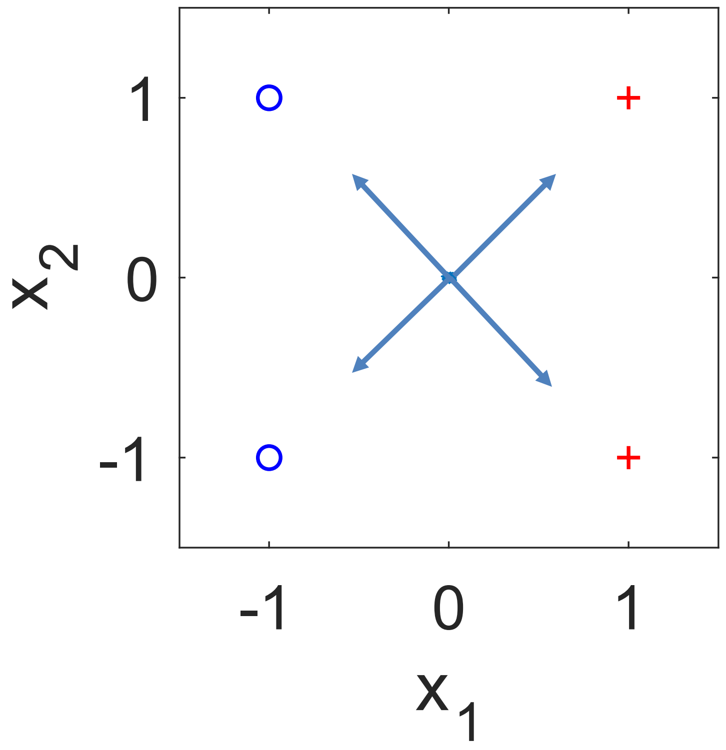
Which gating structures?

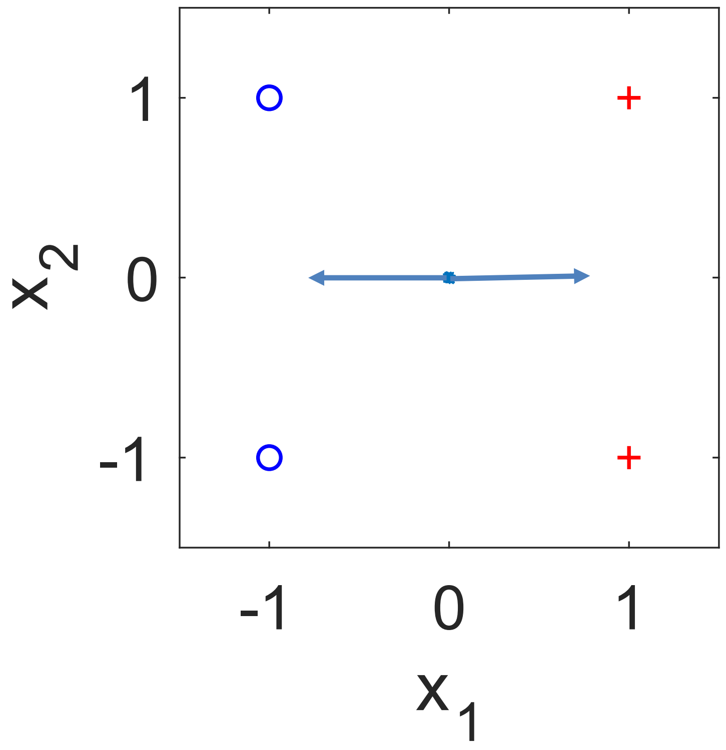




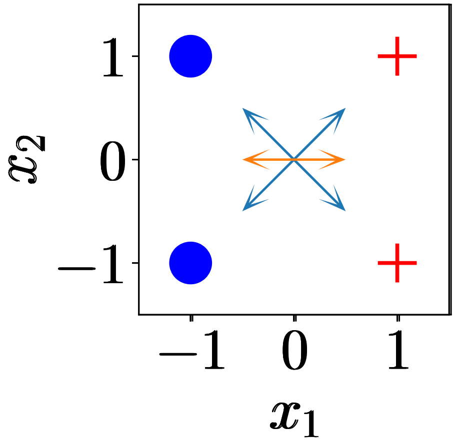
Neural Race Reduction

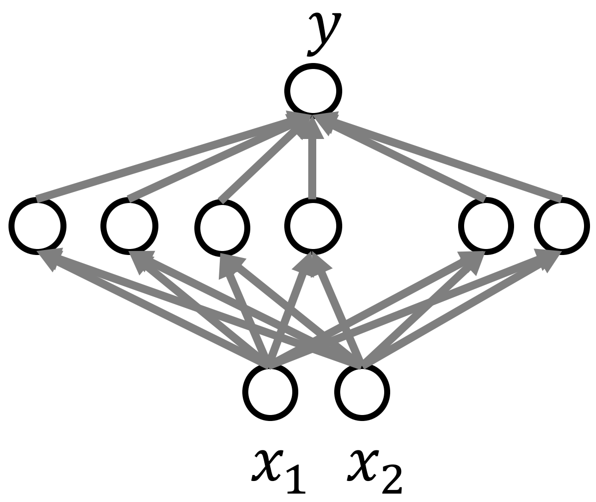
- Each gating scheme yields a distinct effective dataset and deep linear network trajectory
- The one which learns fastest dominates the solution

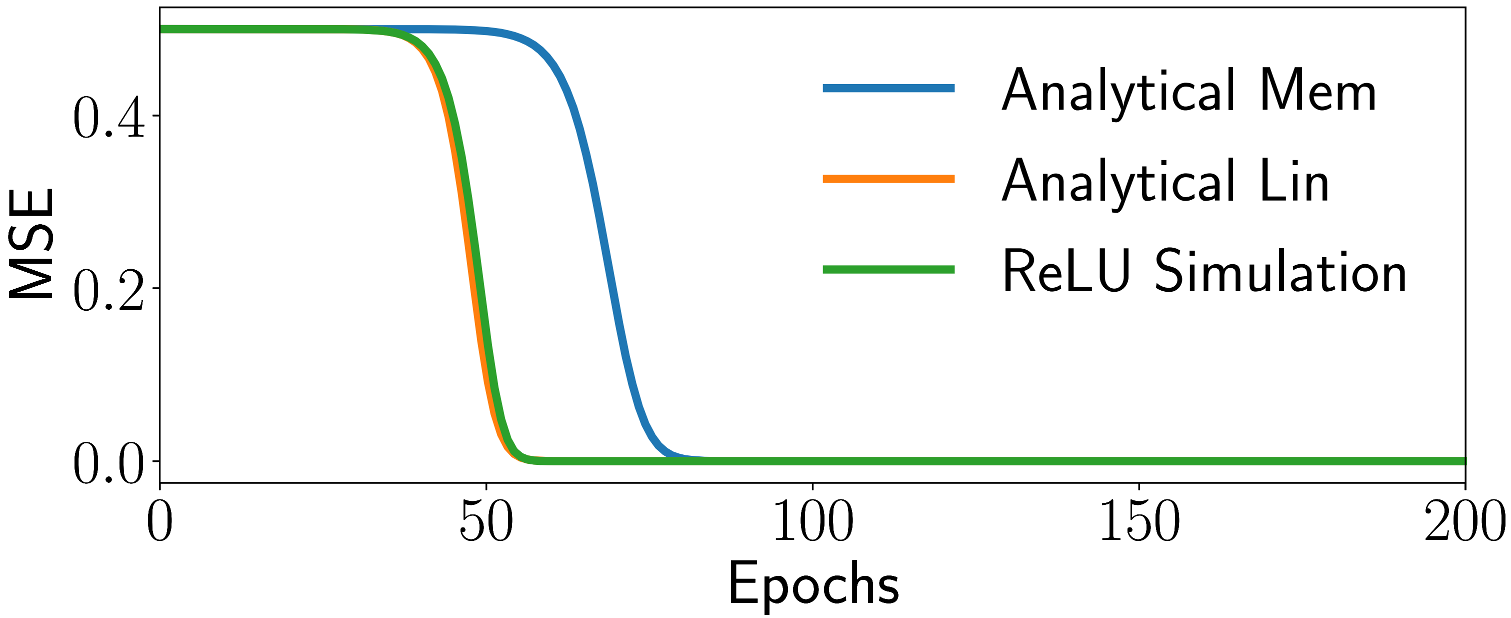




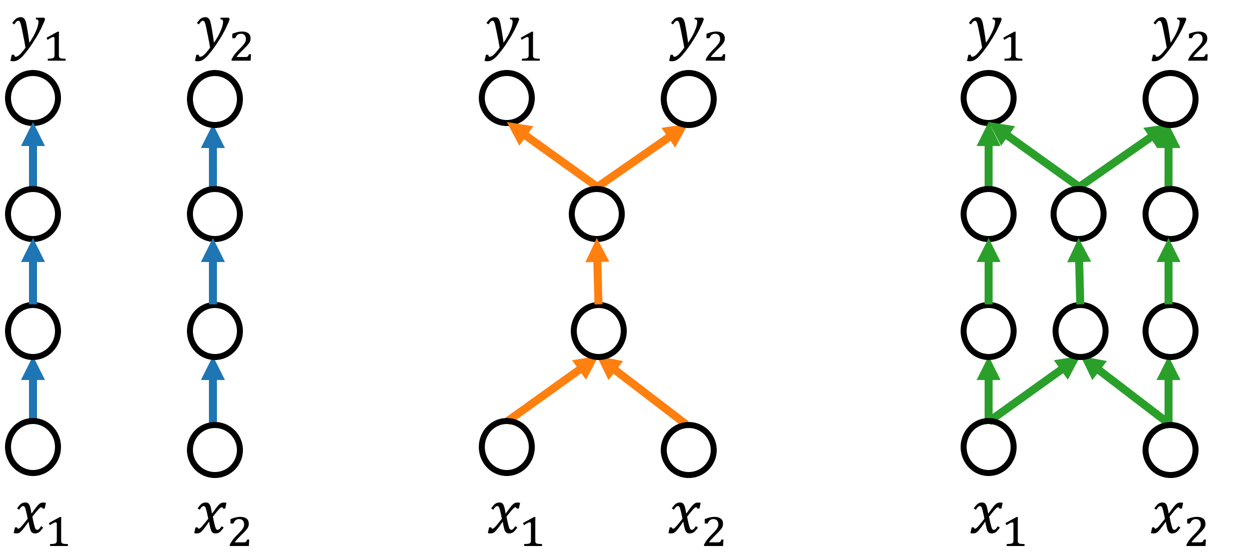
The neural race: stronger input-output correlations

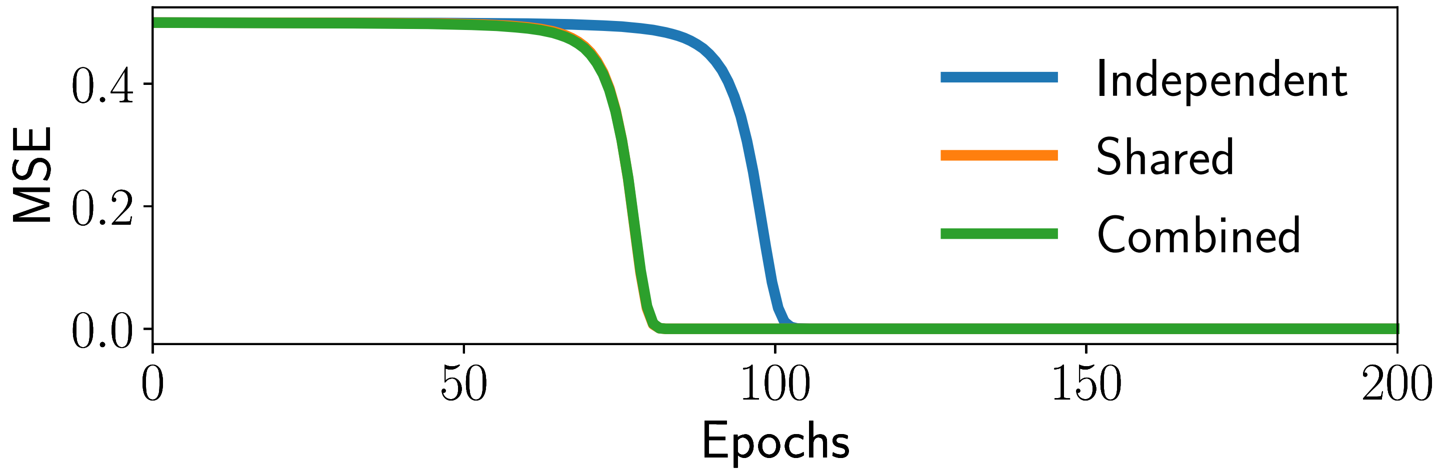




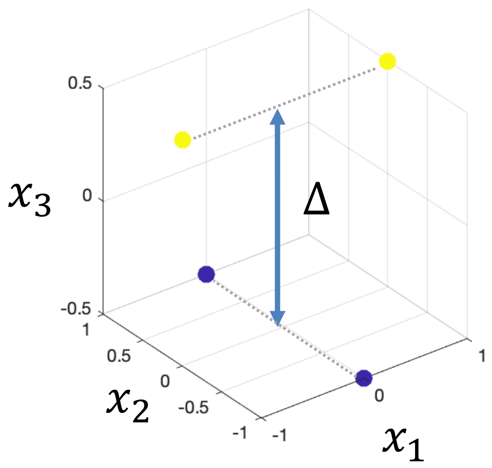


The neural race: edge sharing

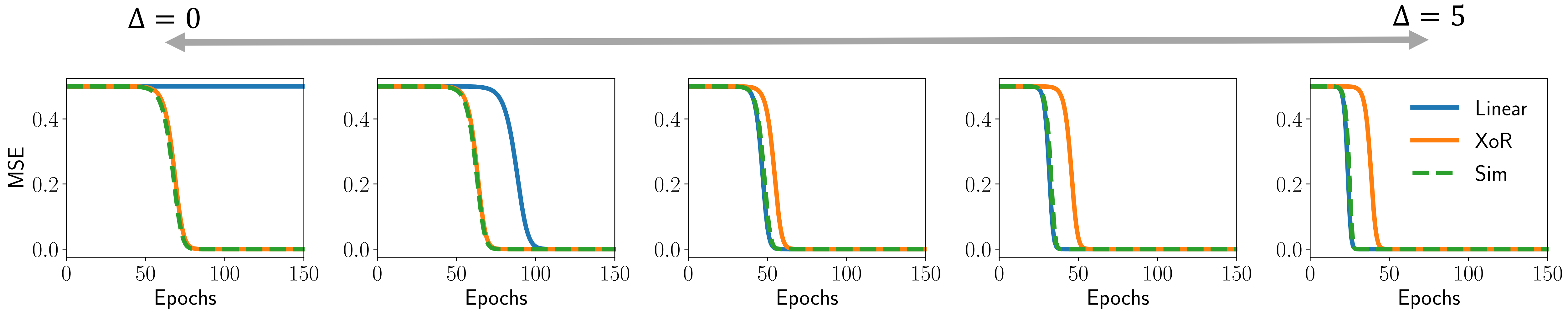




Example: transition to nonlinearity

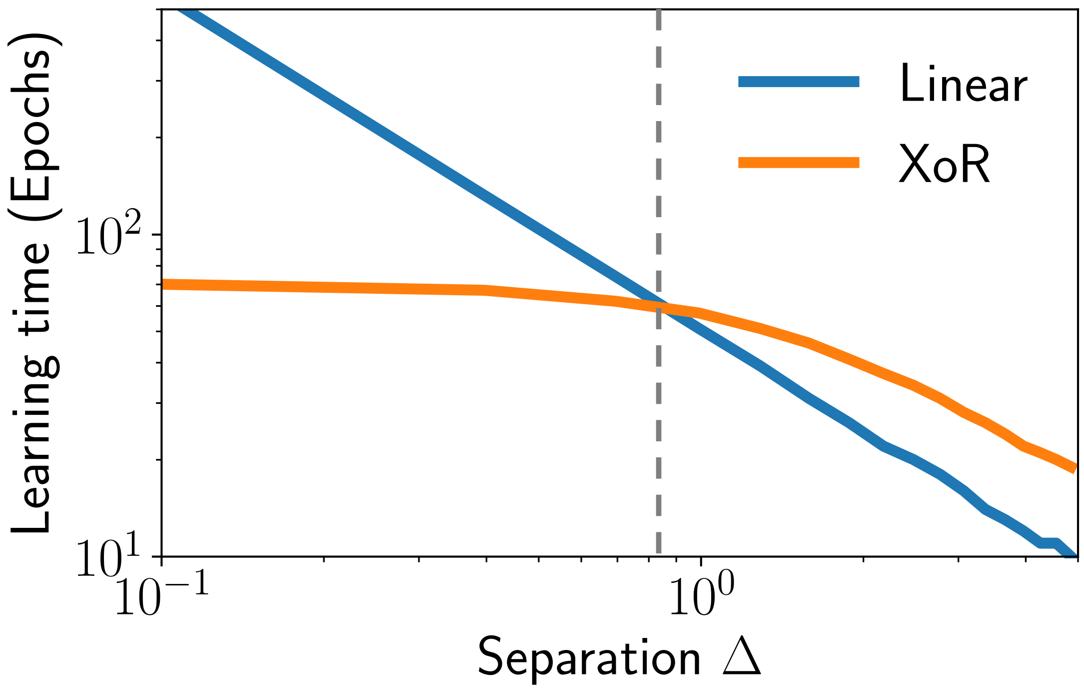


linearly separable with margin Δ for any $\Delta > 0$,
collapses to XoR at $\Delta = 0$



Example: transition to nonlinearity

$\Delta_{\text{switch}} = \sqrt{2/3}$

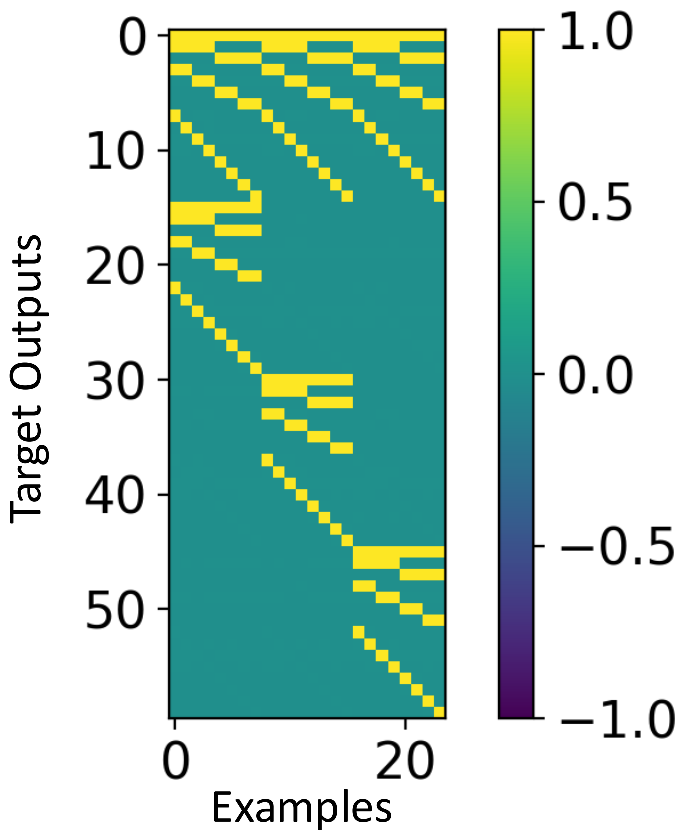


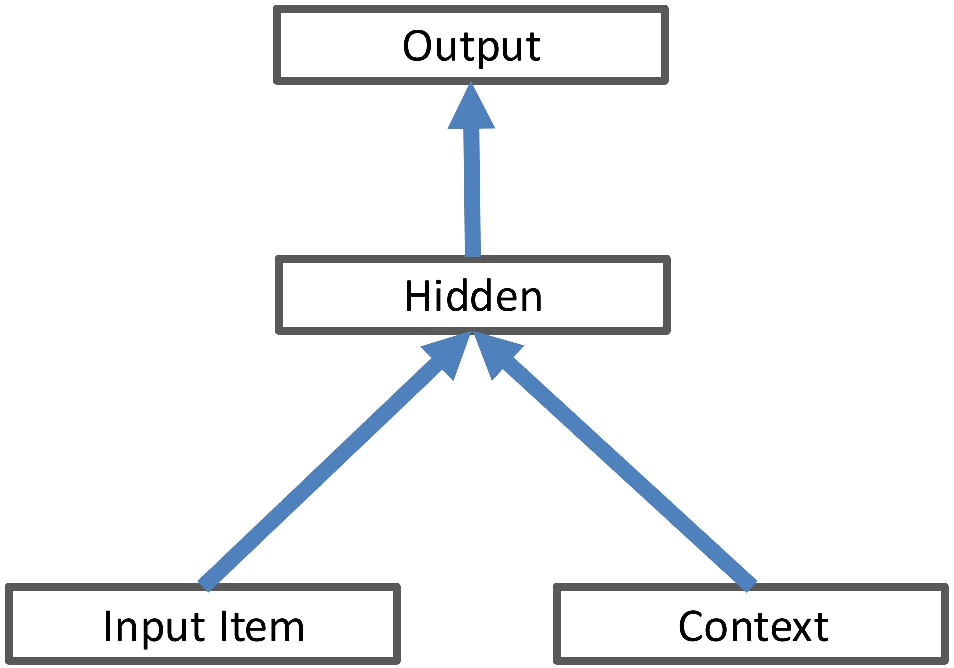
$s_{lin} = \Delta/2$

$$s_{XoR} = \frac{\sqrt{2 + \Delta^2}}{P}$$

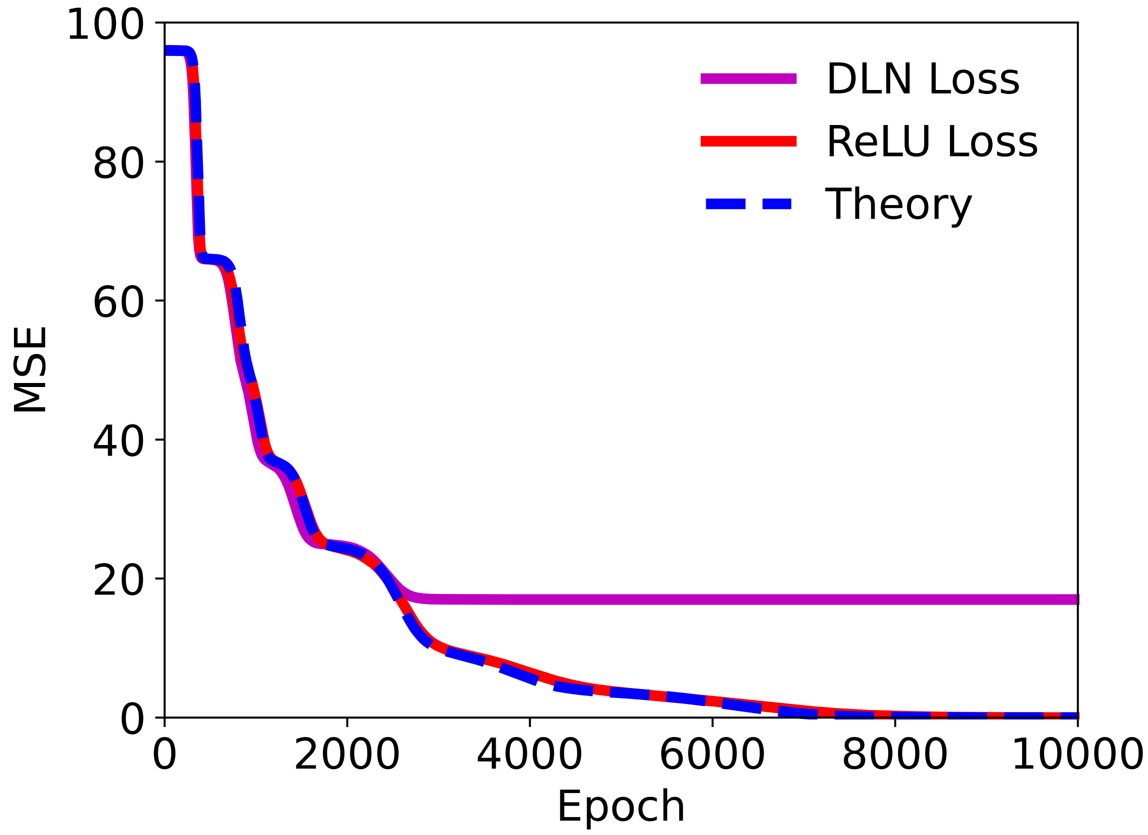
Nonlinear representations emerge before they are strictly necessary

Context-dependent Processing

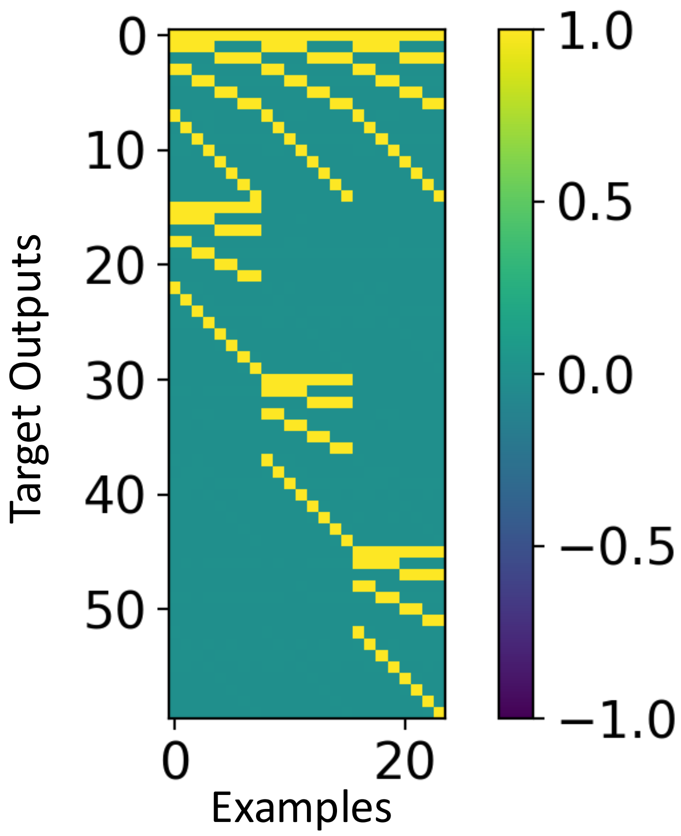


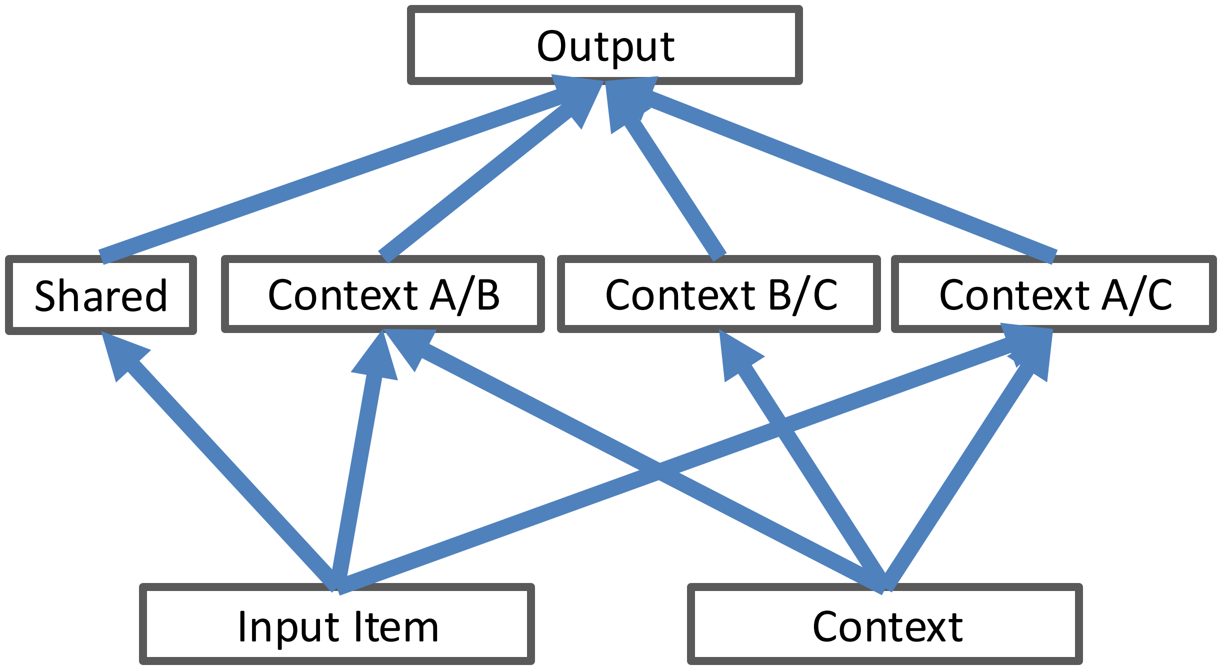


Context-dependent Processing

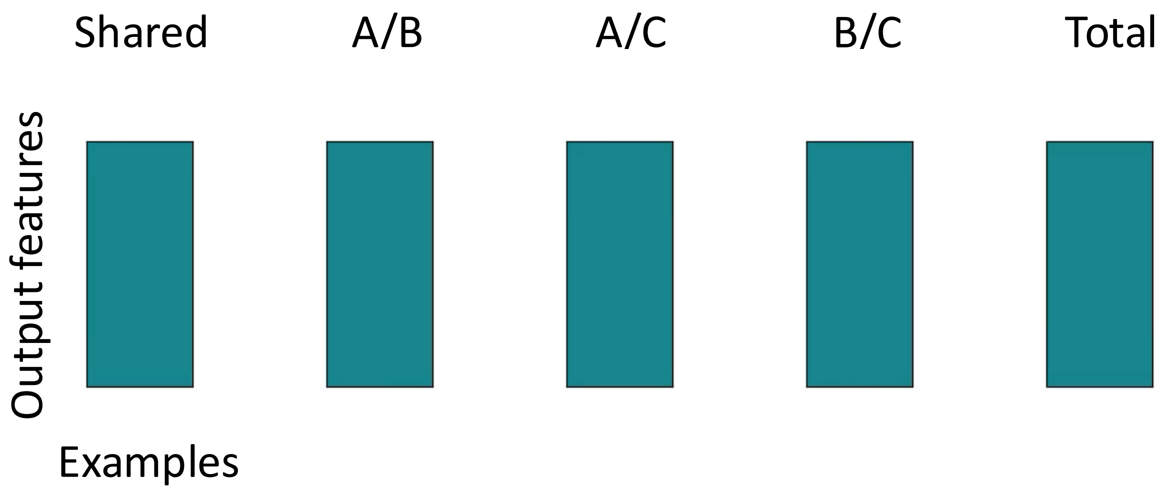


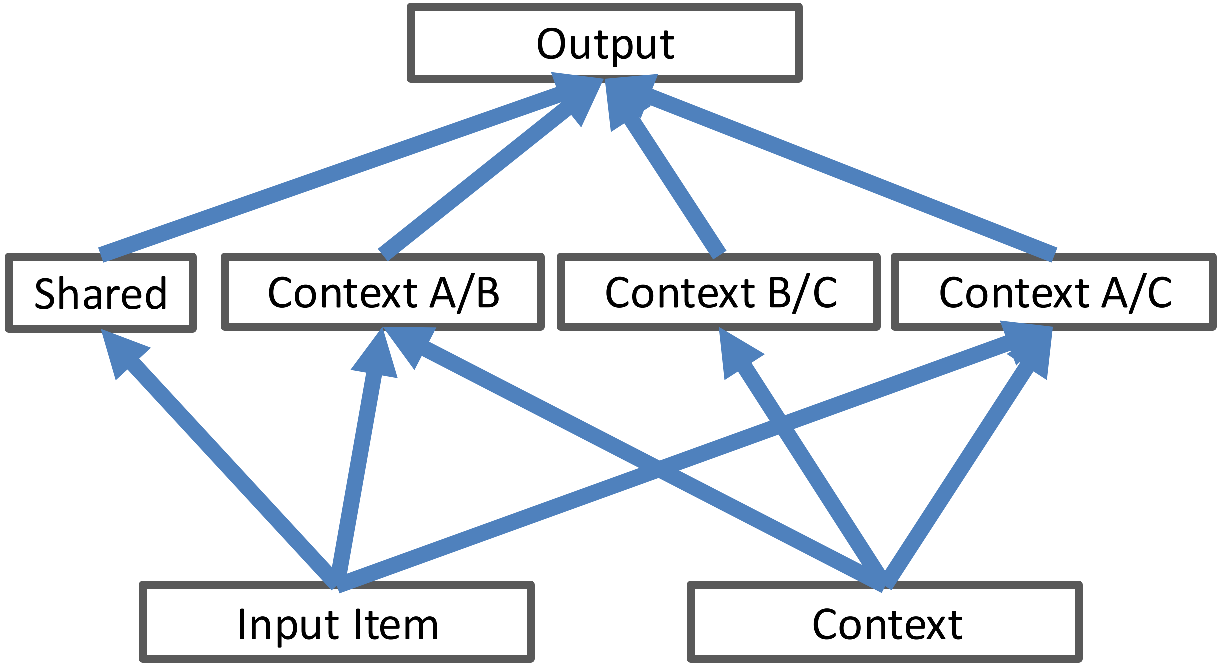
Context-dependent Processing



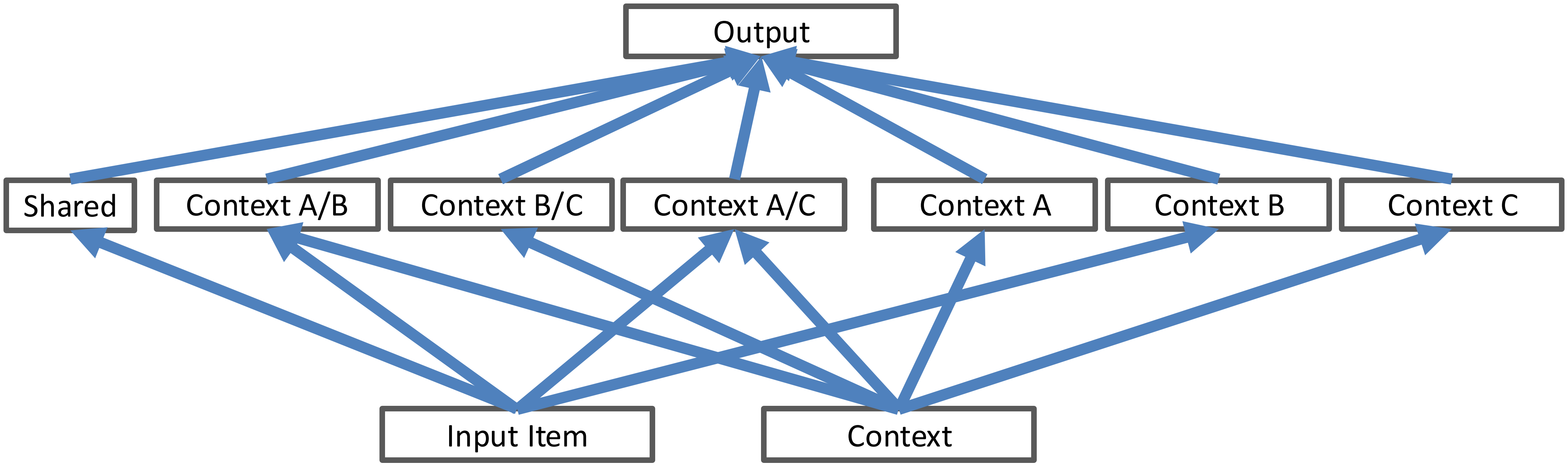


Context-dependent Processing

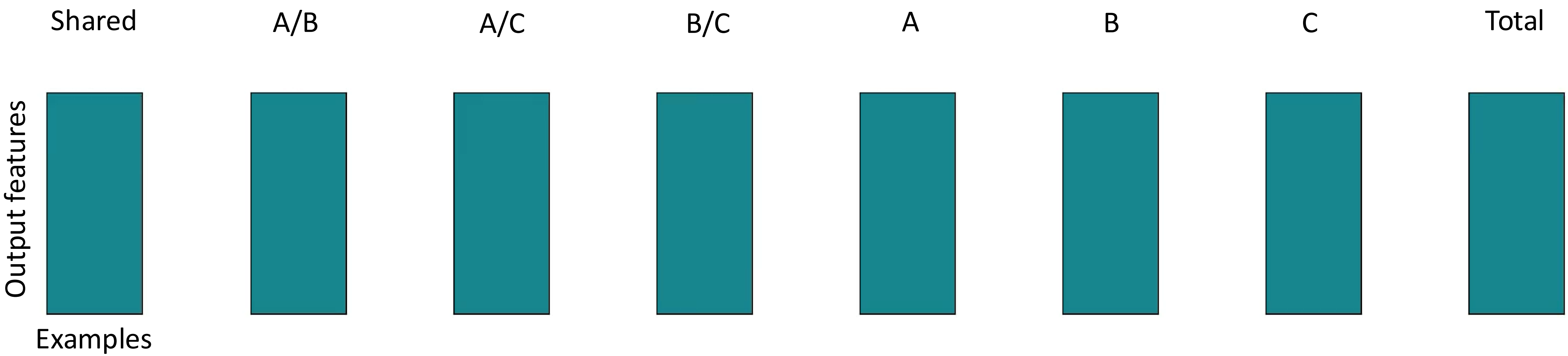




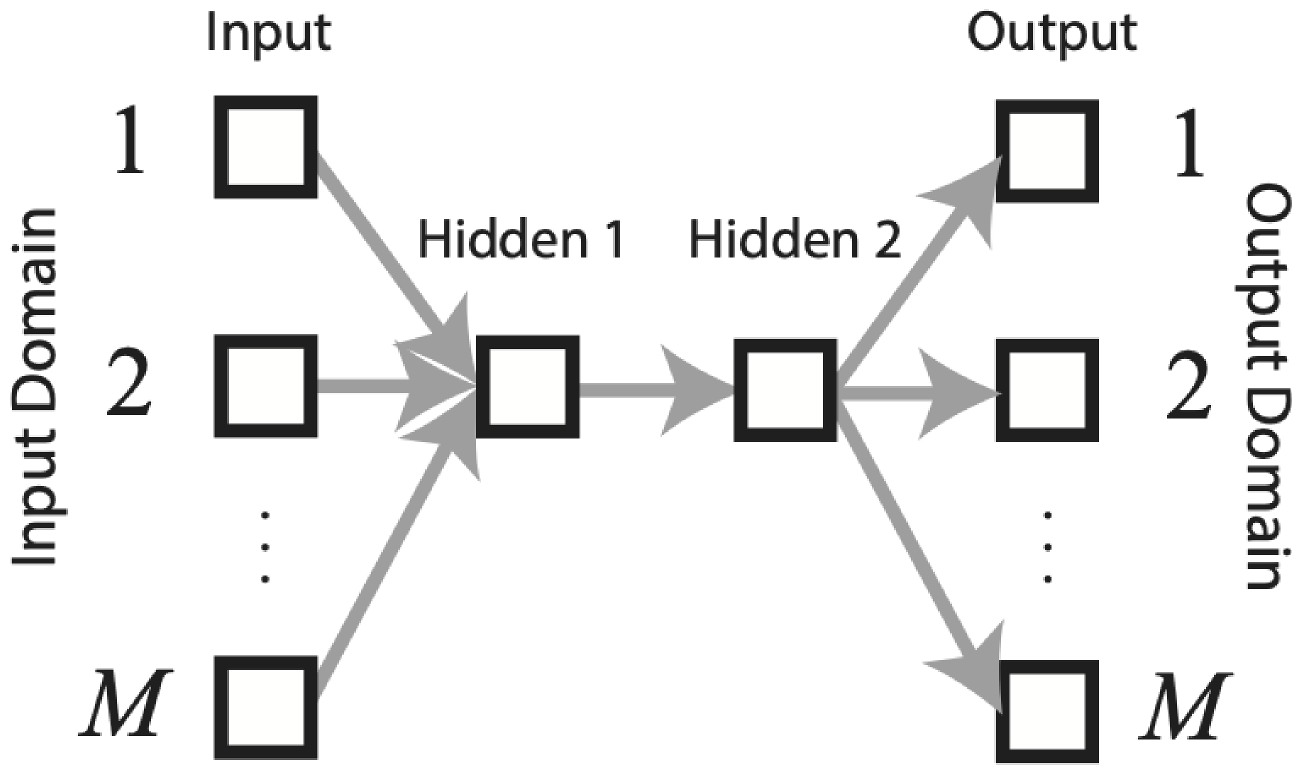
Context-dependent processing



Context-dependent processing



Example: Routing network

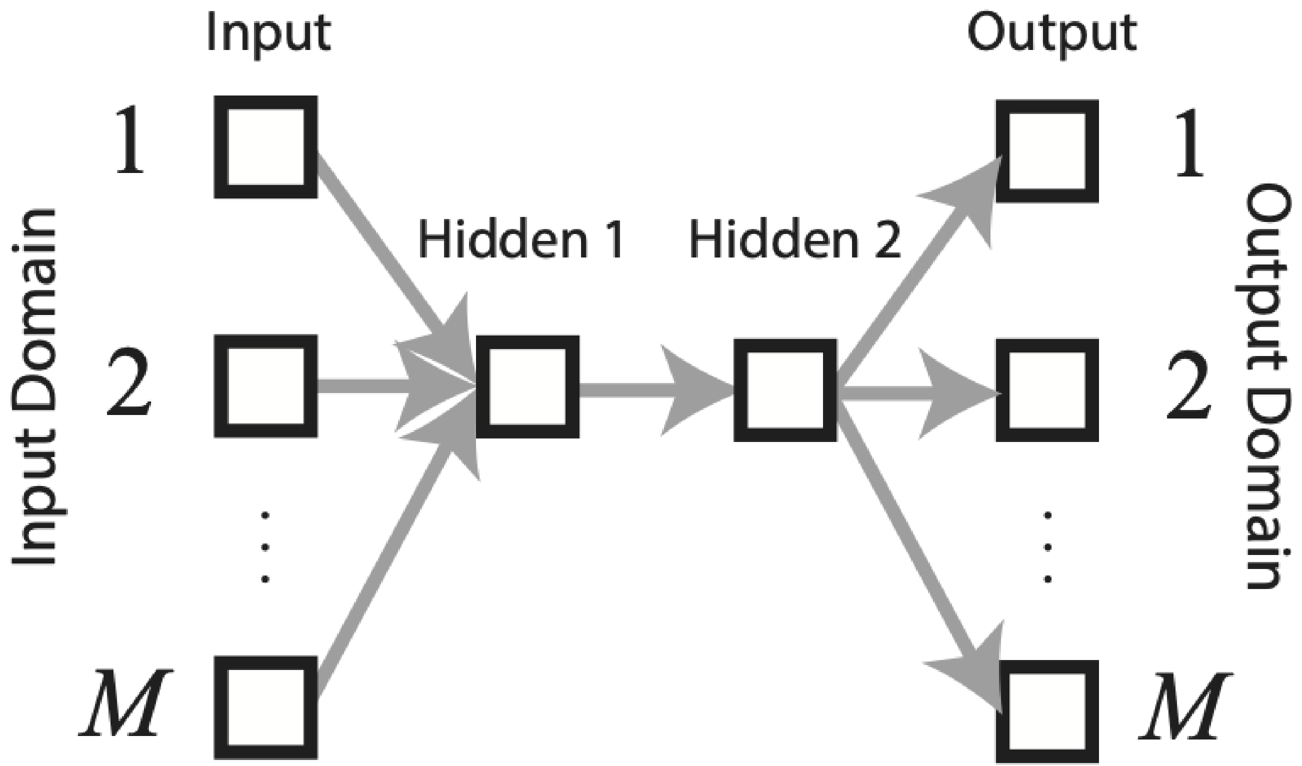


Ex: multilingual translation

French

German

Spanish



French

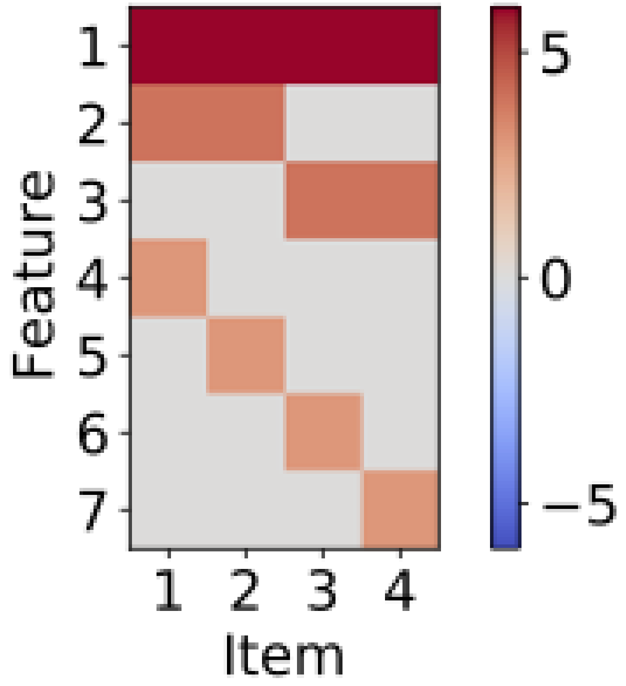
German

Spanish

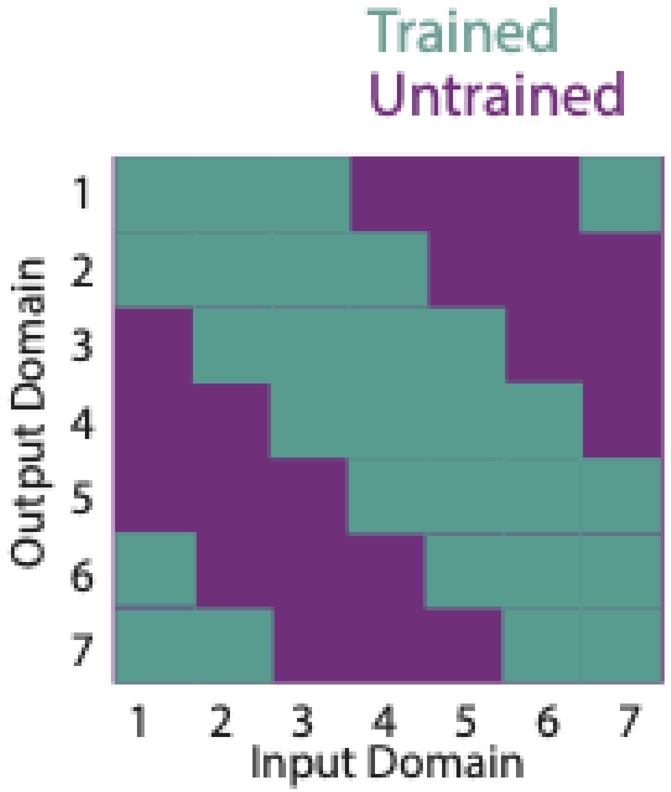
Each domain has distinctive inputs/outputs but similar underlying structural form

Dataset

Simple hierarchical dataset for each domain



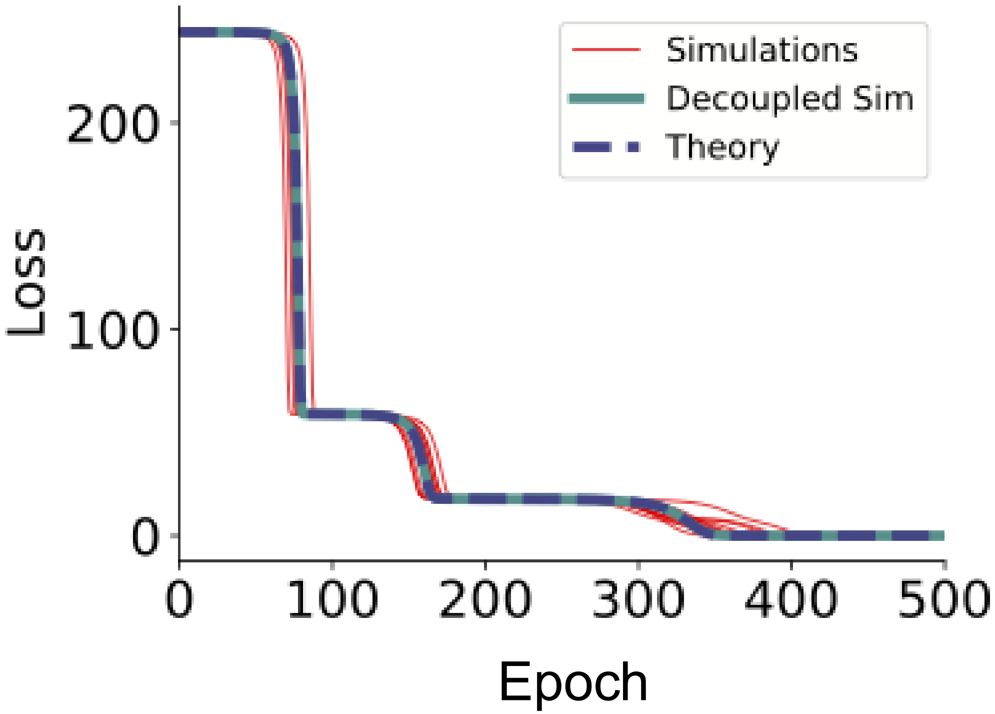
Subset of trained domain pairs

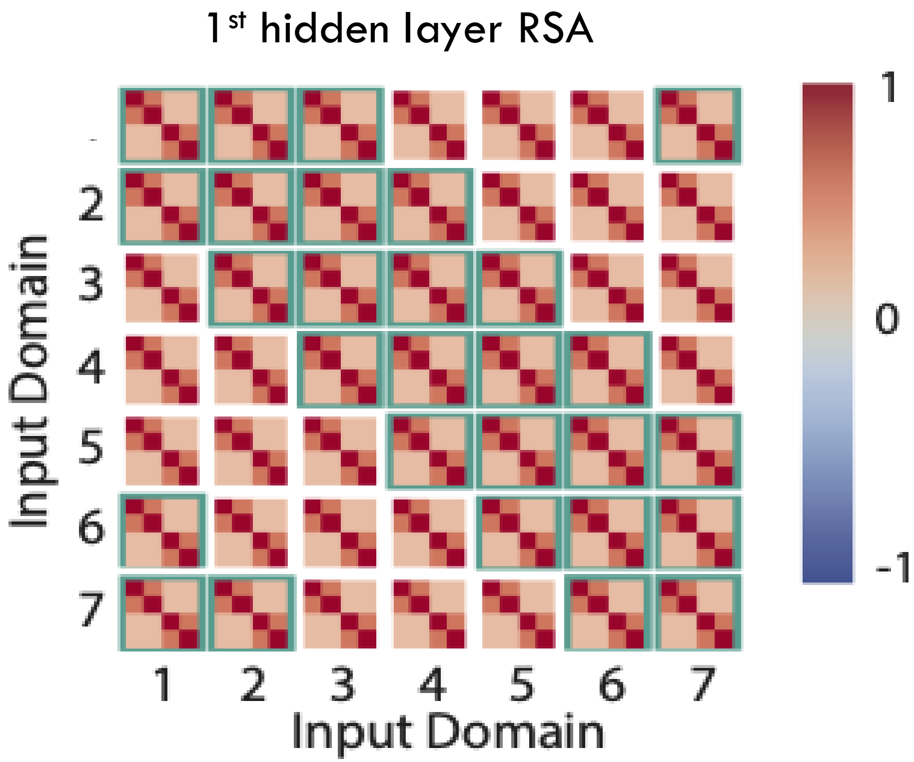


M: # domains

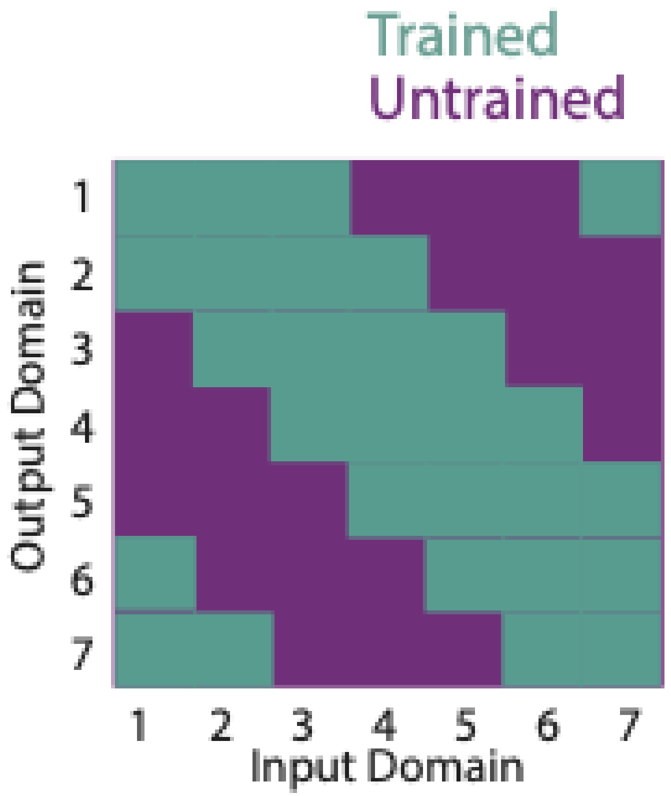
K: # trained output domains per input domain

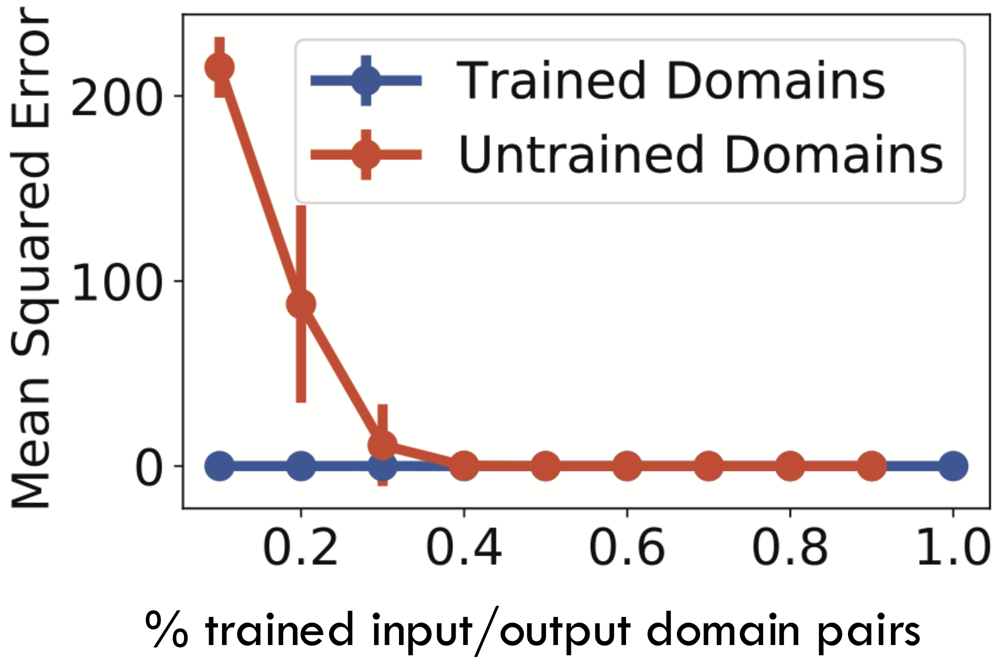
Dynamics of abstraction





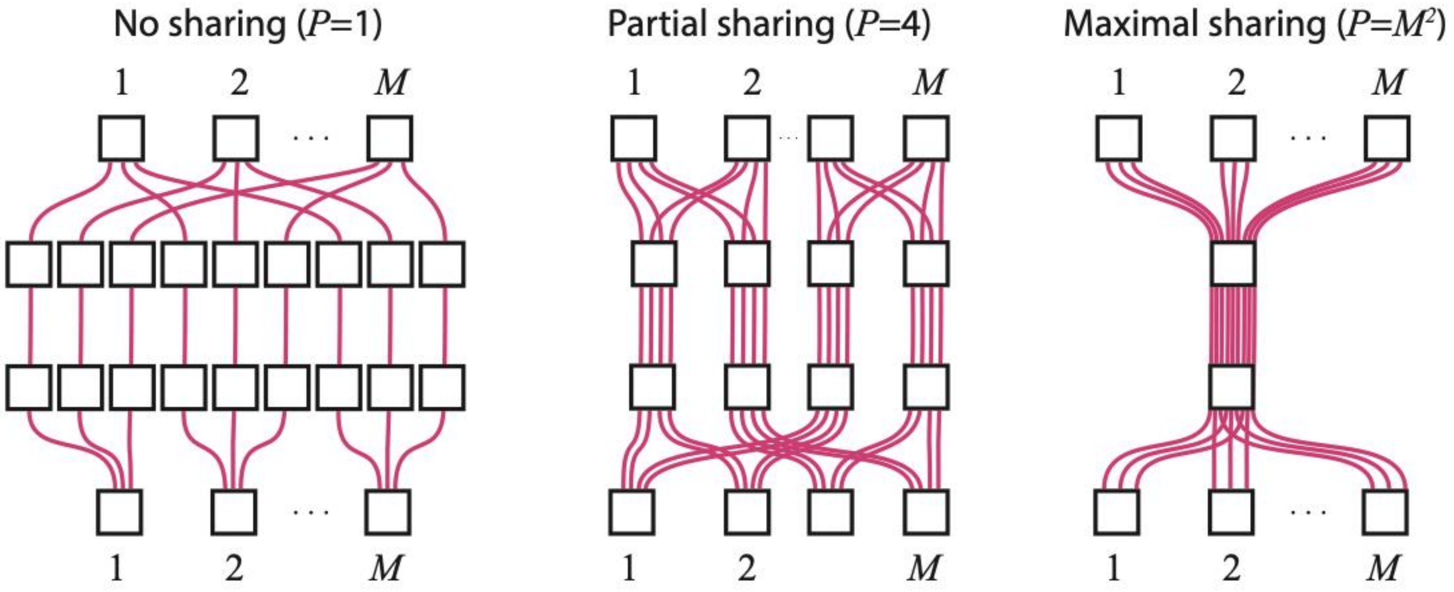
Systematic generalization

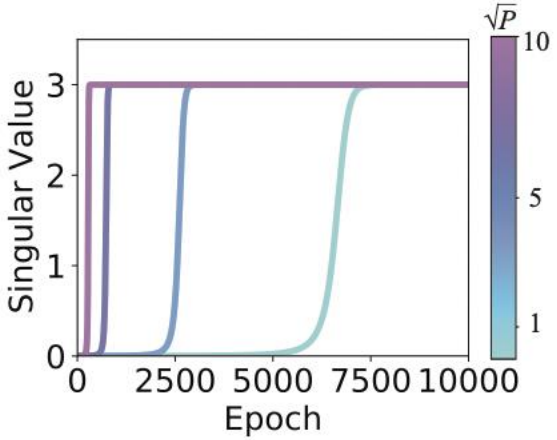


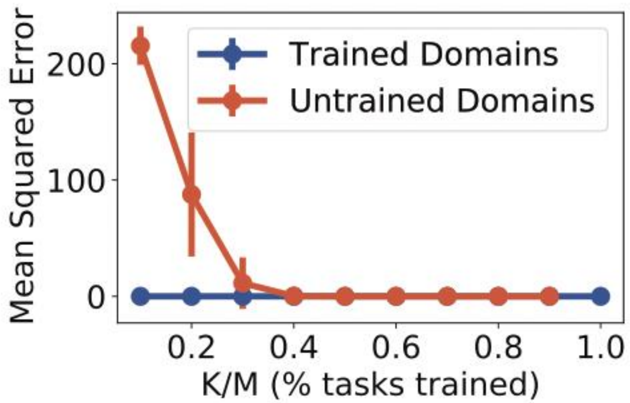


Saxe*, Sodhani*, & Lewallen, ICML 2022

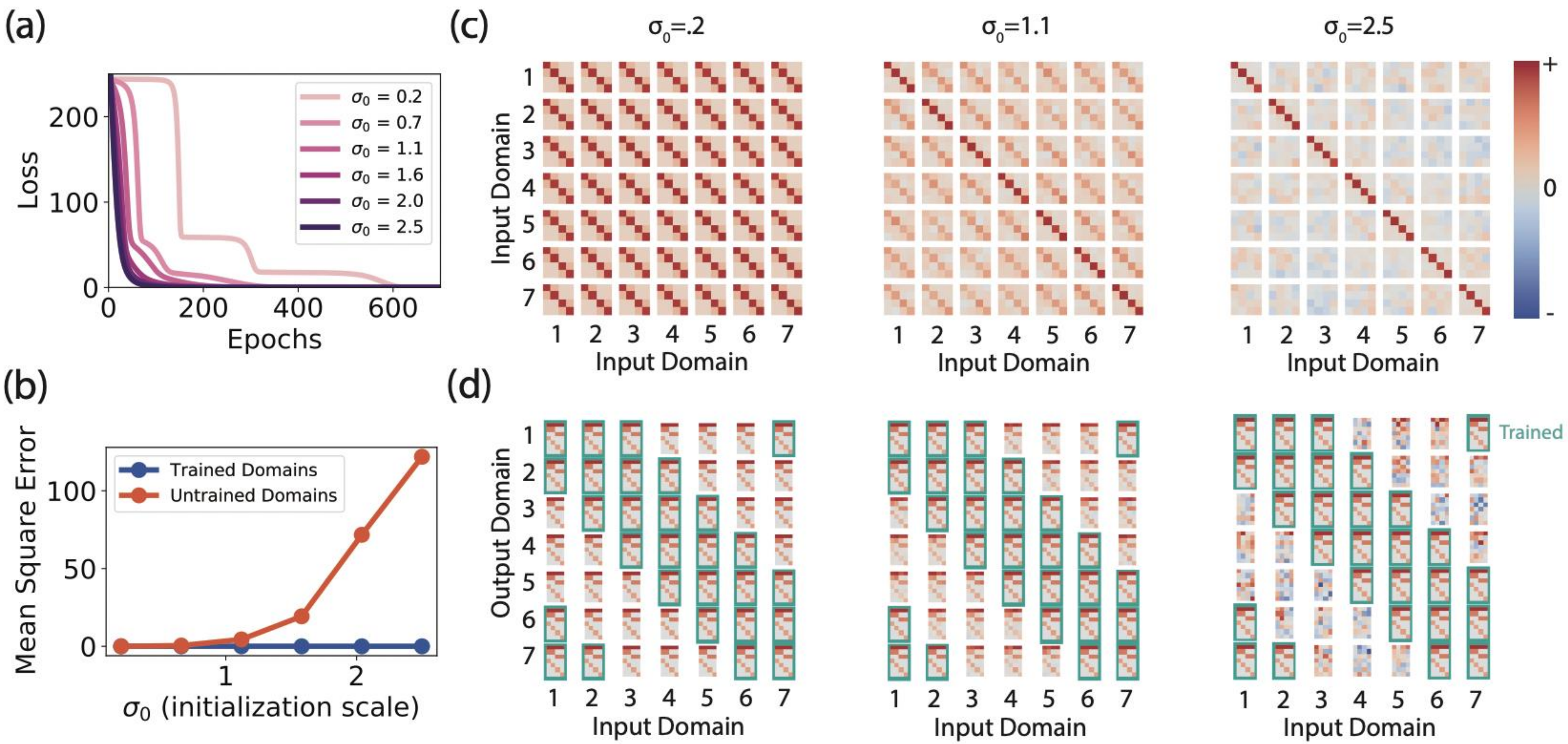
Race dynamics favor shared structure



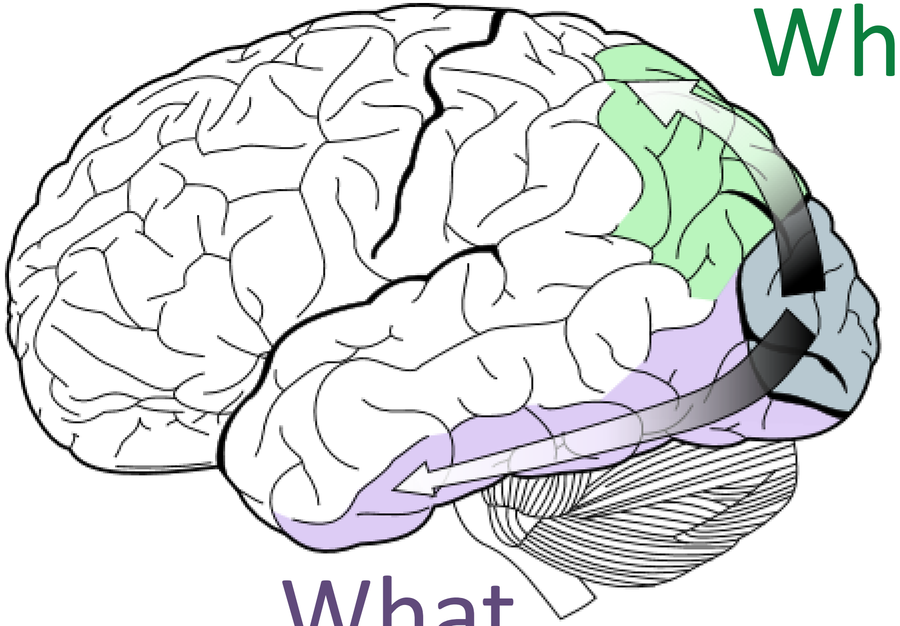


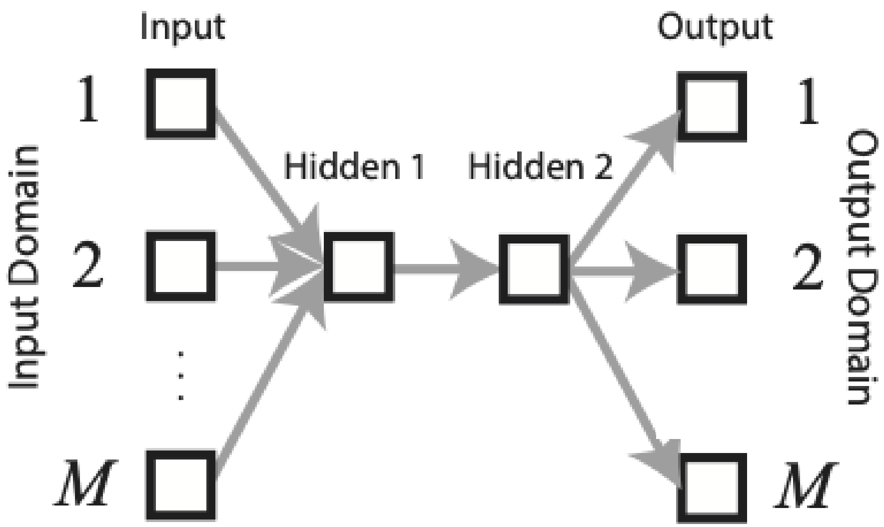


Initialization dependence: rich vs lazy learning



Factorization & principle of convergence





Selket, https://commons.wikimedia.org/w/index.php?curid=1679336

Rogers & McClelland, 2003

Multipotential representation learning

- Animals can recombine their existing knowledge to exploit new opportunities
- In machine learning systems, this ability can emerge at scale (e.g., in context learning)
- What are the factors that give rise to *multipotential* representations?

Conclusion & outlook

- Depth introduces a hierarchy of saddle points into the loss landscape, yielding a quasi-systematic progression through stages
- Initialization determines whether these saddle points influence dynamics, yielding several learning regimes
- In nonlinear networks, pathways race to explain the dataset

References

- Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, *116*(23), 11537–11546.
- Liebana Garcia et al. (2025). Dopamine encodes deep network teaching signals for individual learning trajectories. *Cell.*
- Braun, L., Dominé, C.C.J., Fitzgerald, J., Saxe, A.M. (2022) Exact learning dynamics of deep linear networks with prior knowledge. In *NeurIPS.*
- Dominé, C.C.J., Anguita, N., Proca, A. M., Braun, L., Kunin, D., Mediano, P. A. M., & Saxe, A. M. (2025). From lazy to rich: Exact learning dynamics in deep linear networks. In *ICLR.*
- Saxe*, A. M., Sodhani*, S., & Lewallen, S. (2022). The Neural Race Reduction: Dynamics of Abstraction in Gated Networks. In *ICML*. *Equal contribution.
- Jarvis, D., Klein, R., Rosman, B., Saxe, A.M. (2025). Make haste slowly: A theory of emergent structured mixed selectivity in feature learning ReLU networks. In *ICLR.*

Aaditya Singh
Anika Lowe
Basile Confavreux
Clementine Domine
Cris Holobetz
Devon Jarvis
Erin Grant
Jin Lee
Jirko Rubruck
Lukas Braun
Nishil Patel
Rodrigo Carrasco Davis
Rachel Swanson
Sam Lewallen
Sam Liebana
Sarah Armstrong
Sebastian Lee
Stefano Sarao Mannelli
Tyler Boyd-Meredith
Verena Klar
Victor Pedrosa