The Neural Race Reduction:
Dynamics of abstraction in gated networks

Andrew M. Saxe*[123], Shagun Sodhani*[2], Sam Lewallen[1]

[1]Gatsby Computational Neuroscience Unit & Sainsbury Wellcome Centre, UCL
[2]FAIR, Meta AI
[3]CIFAR Azrieli Global Scholar, CIFAR

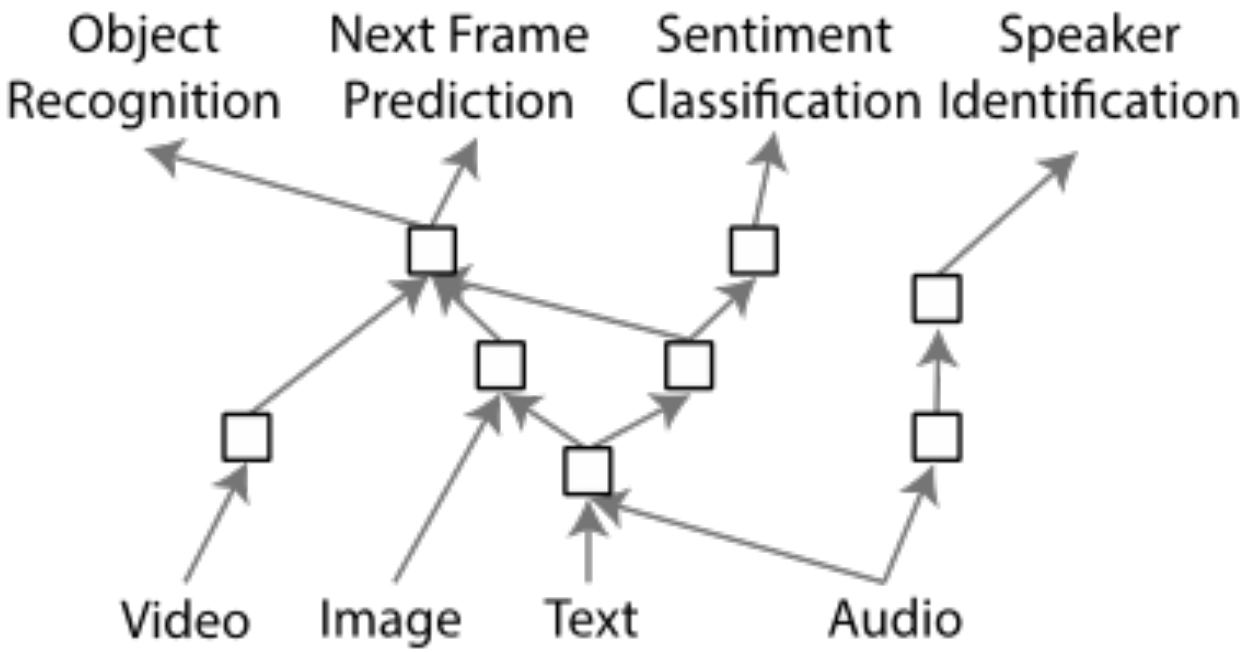
*Equal contributions

Shagun Sodhani*
(FAIR, Meta AI)

Sam Lewallen
(UCL)

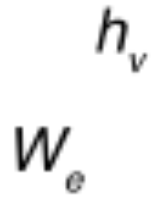
*Equal contribution

Mesoscale architecture

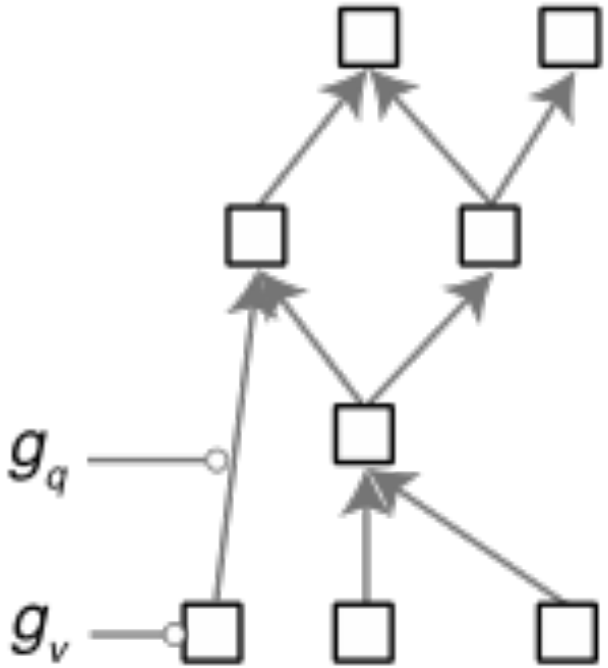


Gated Deep Linear Network

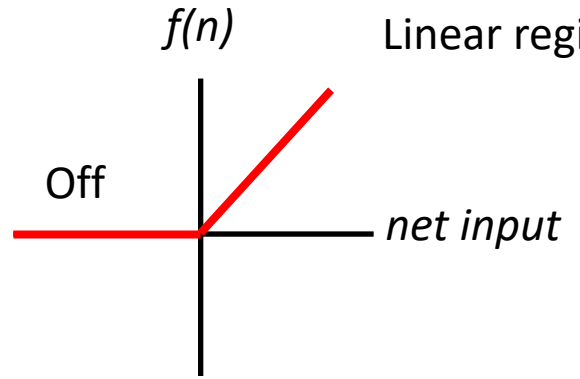
Arch graph Γ: nodes V, edges E



Gated Deep Linear Network



Gating interpretation: Relaxation of ReLU



$$h = \max(0, wx) = \underbrace{\text{step}(wx)}_{\text{Gate } g} wx$$

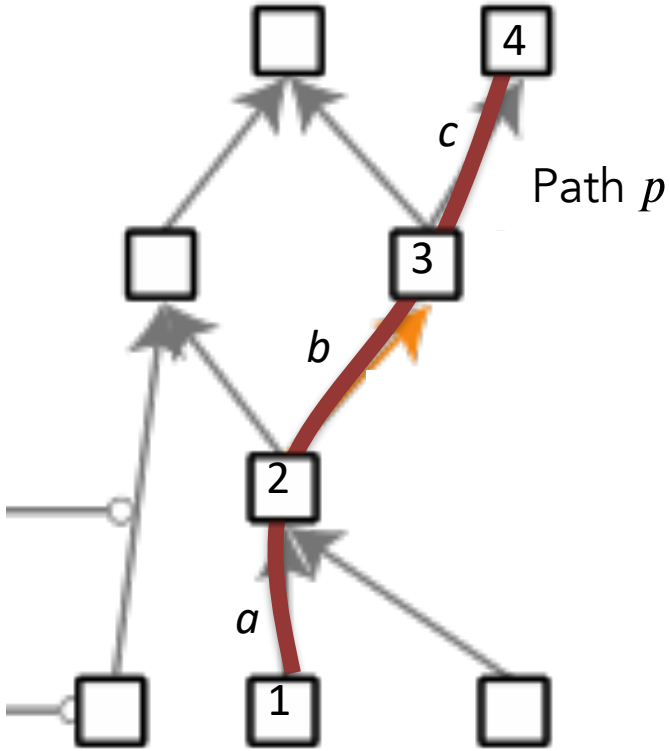
Gradient descent

Minimize L_2 loss $\mathscr{L}(\{W\}) = \left\langle \frac{1}{2} \sum_{v \in \mathrm{Out}(\Gamma)} \left\| y_v - h_v \right\|_2^2 \right\rangle_{x,y,g}$

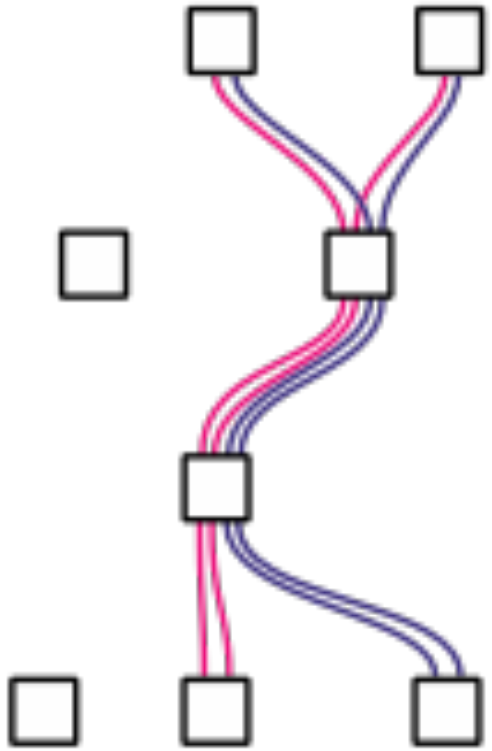
using gradient flow on the weights

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

Gradient descent



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

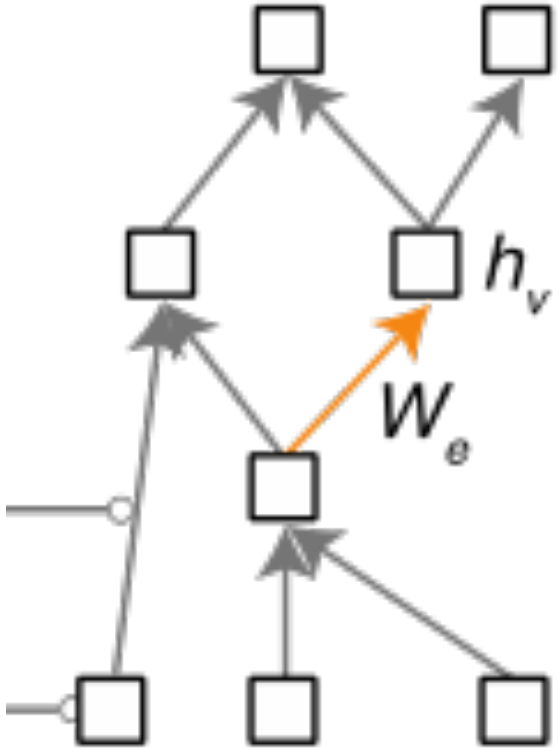
$$\mathcal{E}(p) = \Sigma^{yx}(p) - \sum_{j \in \mathcal{T}(p)} \mathrm{W_j} \Sigma^x(j,p)$$

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

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

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

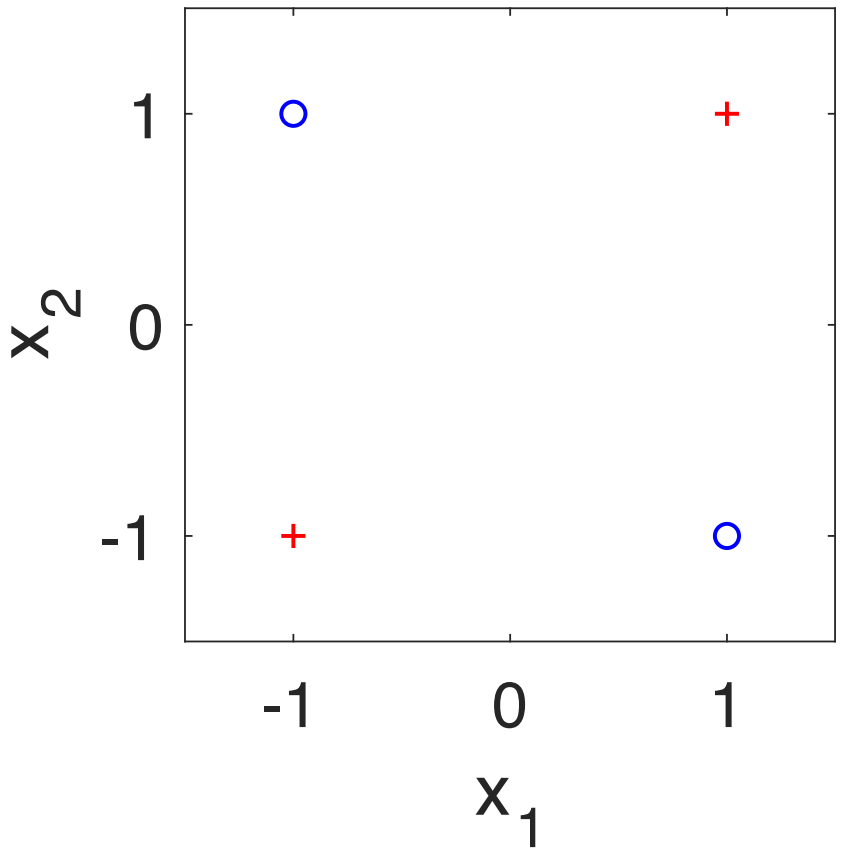
Intuition

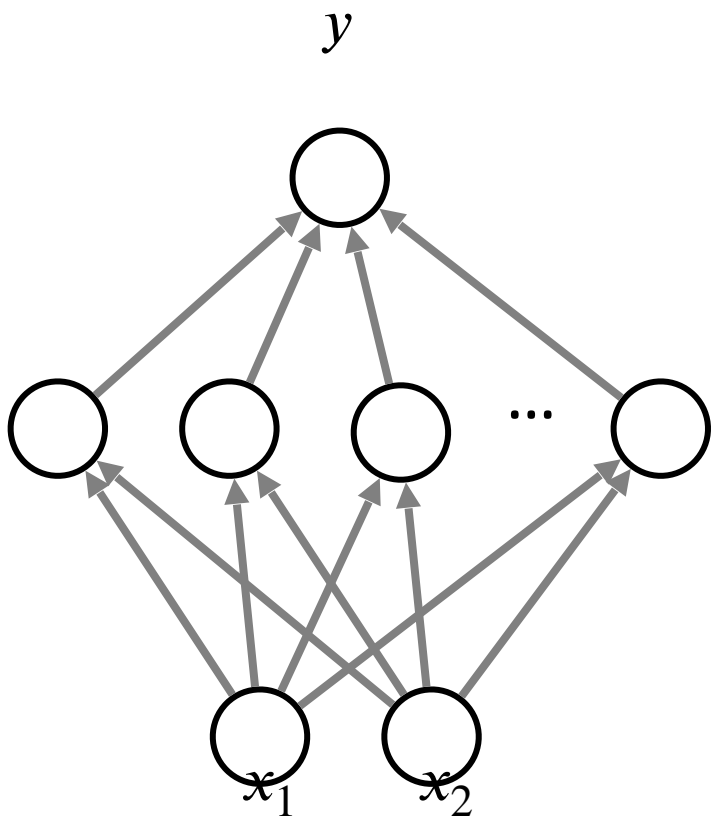


Each pathway behaves like a deep linear network

Gating controls the effective dataset for each pathway

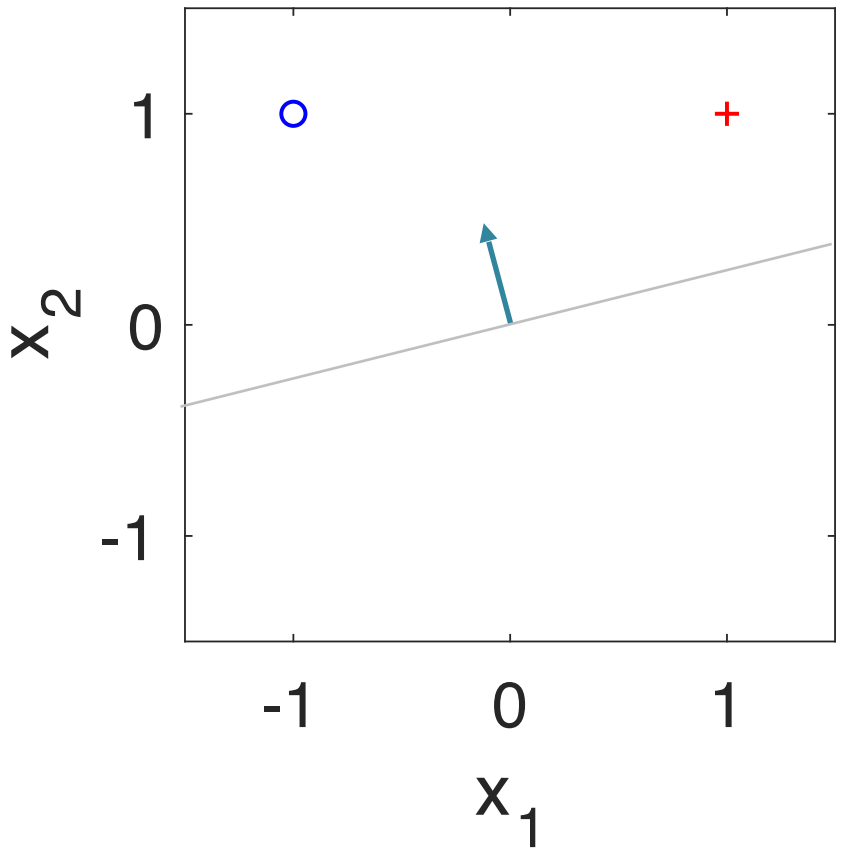
The XoR problem

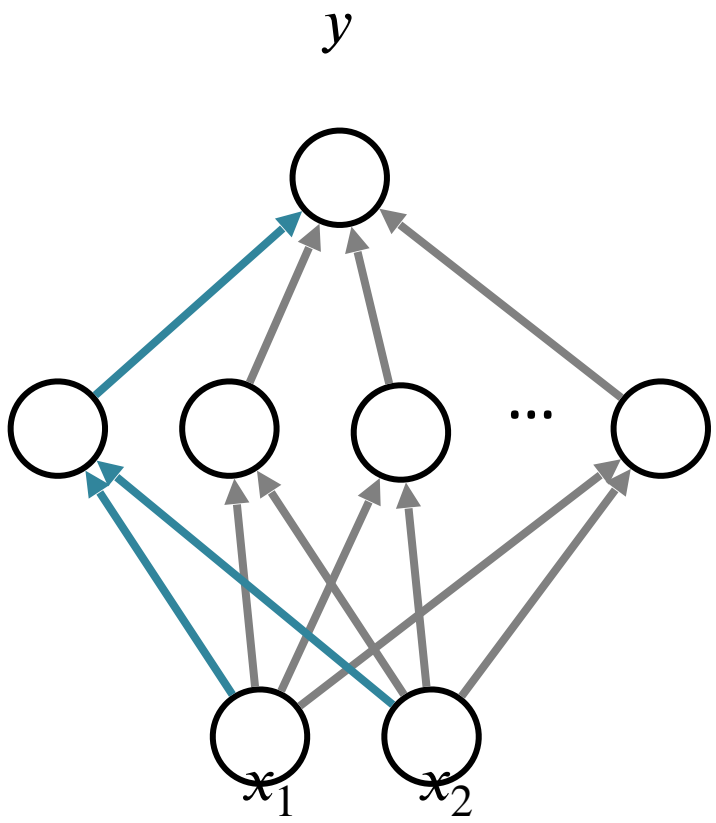




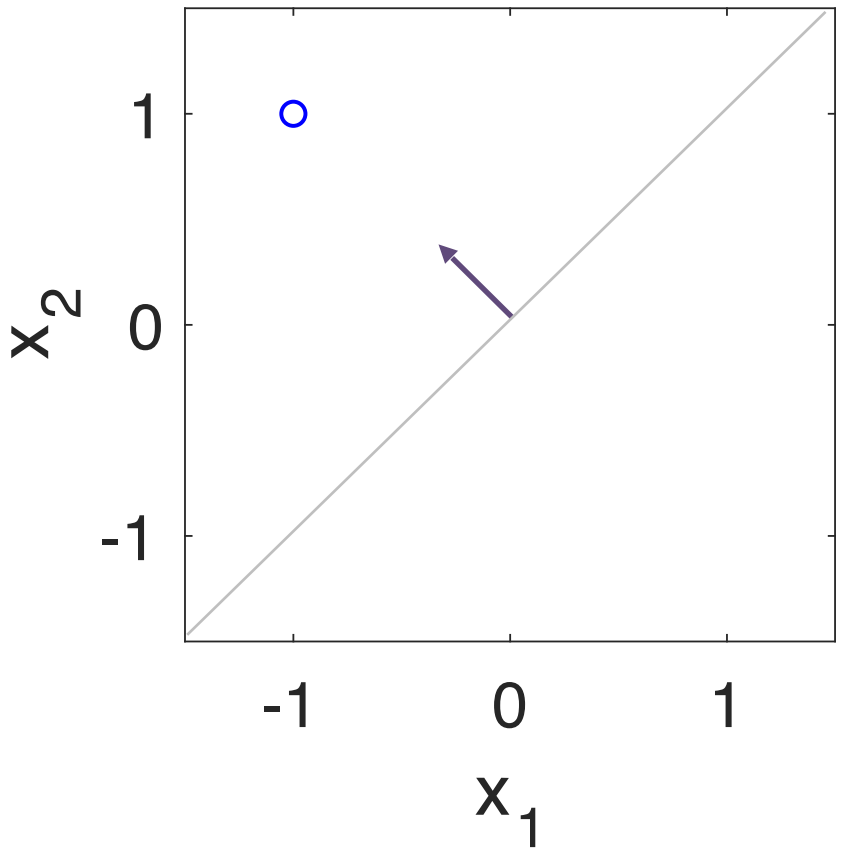
Minsky & Papert, 1969; Rumelhart, Hinton & Williams, 1986

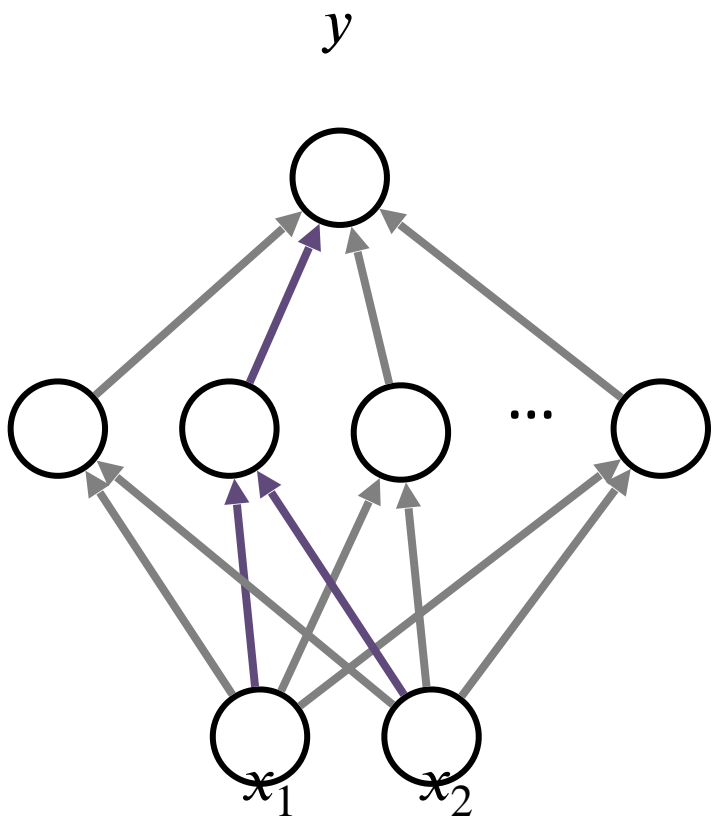
Gating dynamics



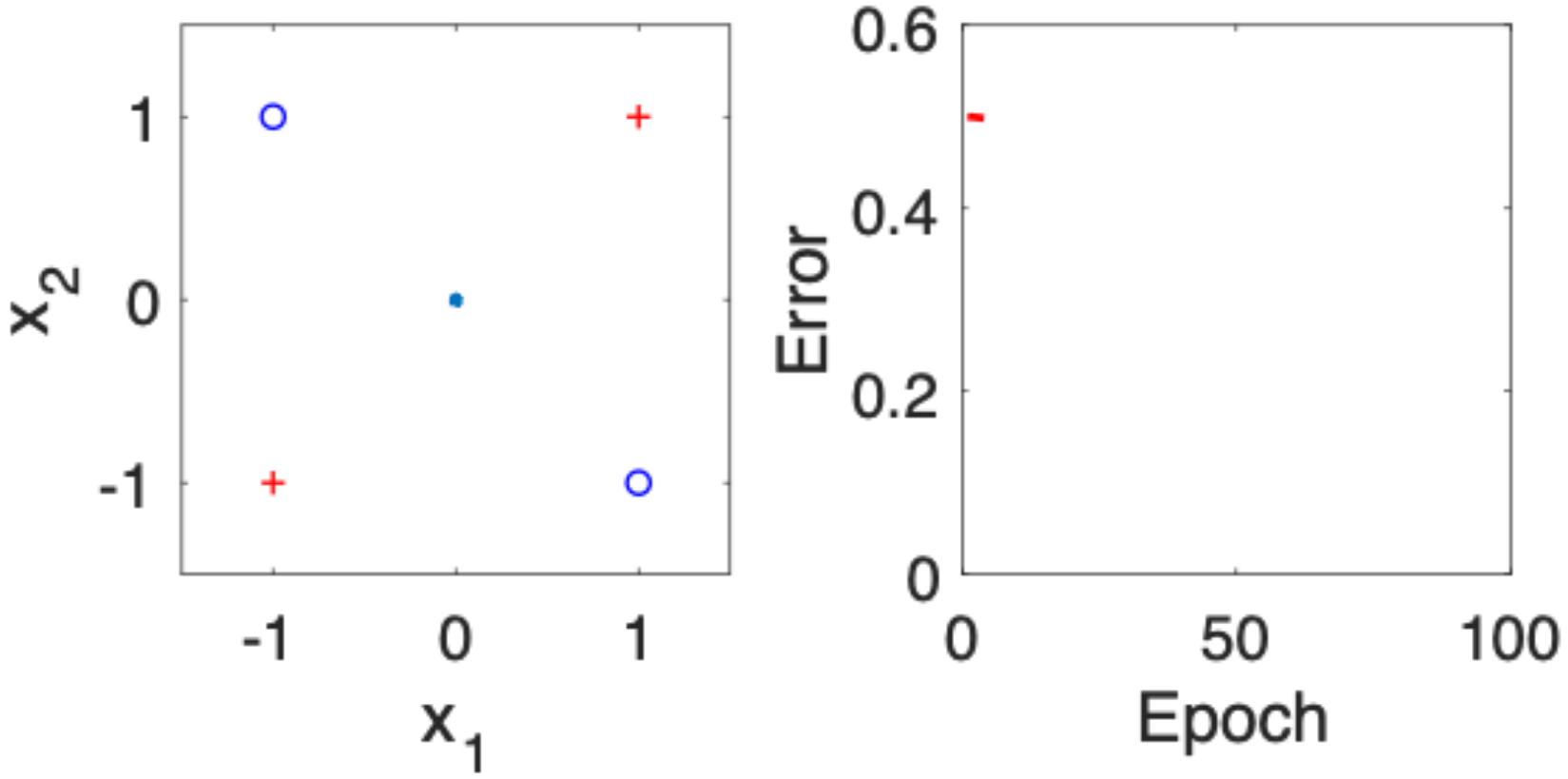


Gating dynamics

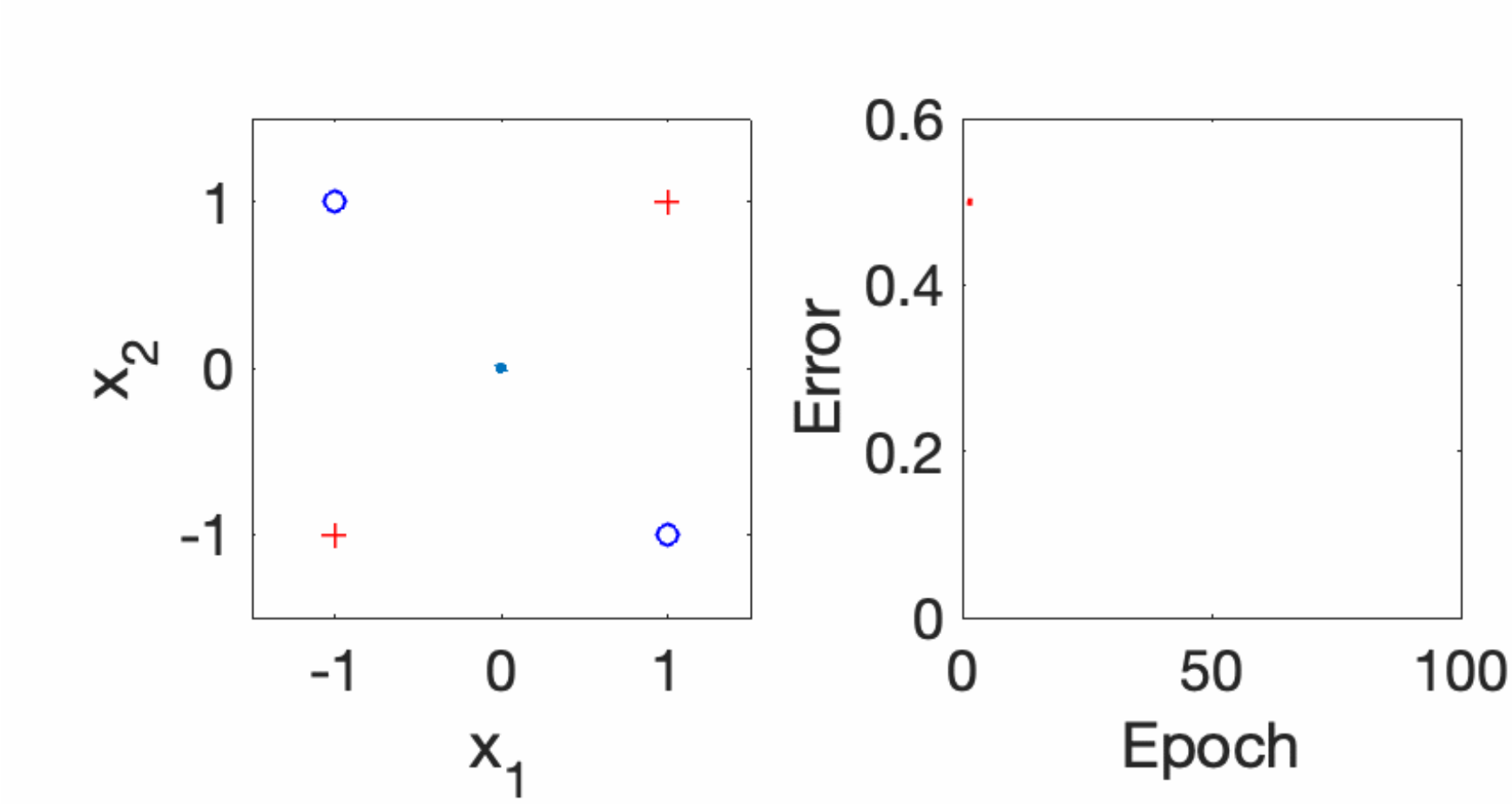




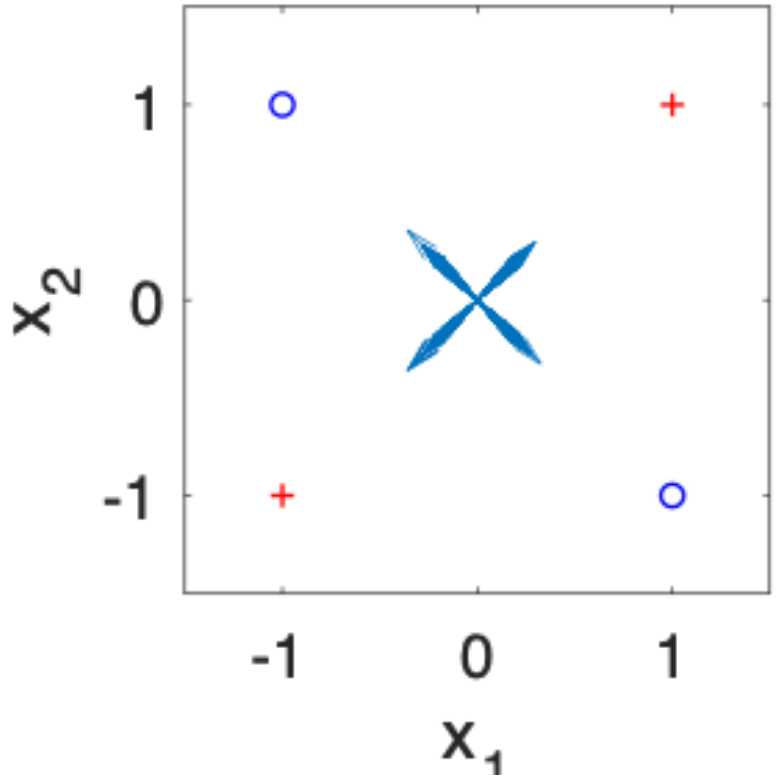
XoR Dynamics

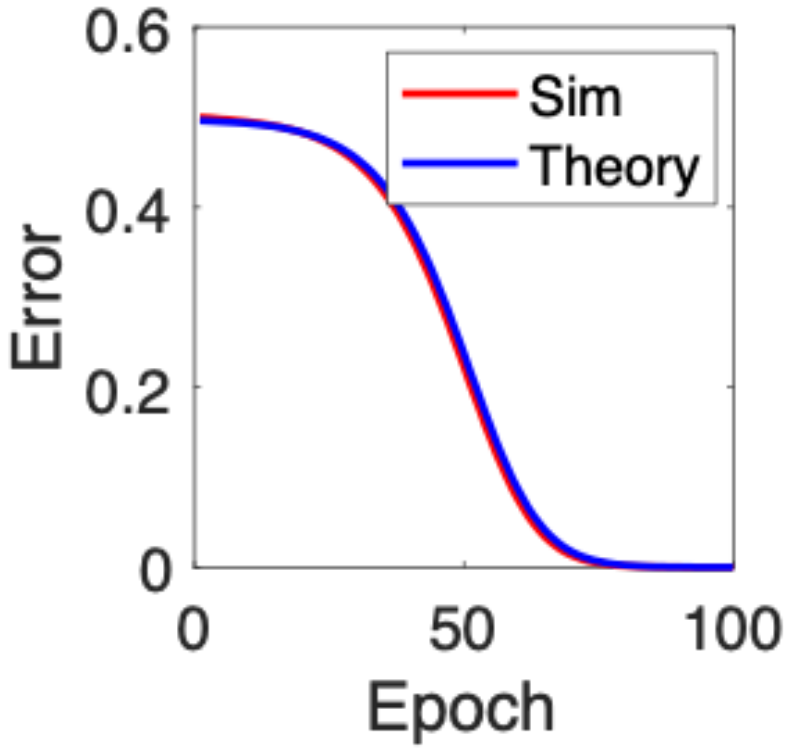


XoR Dynamics

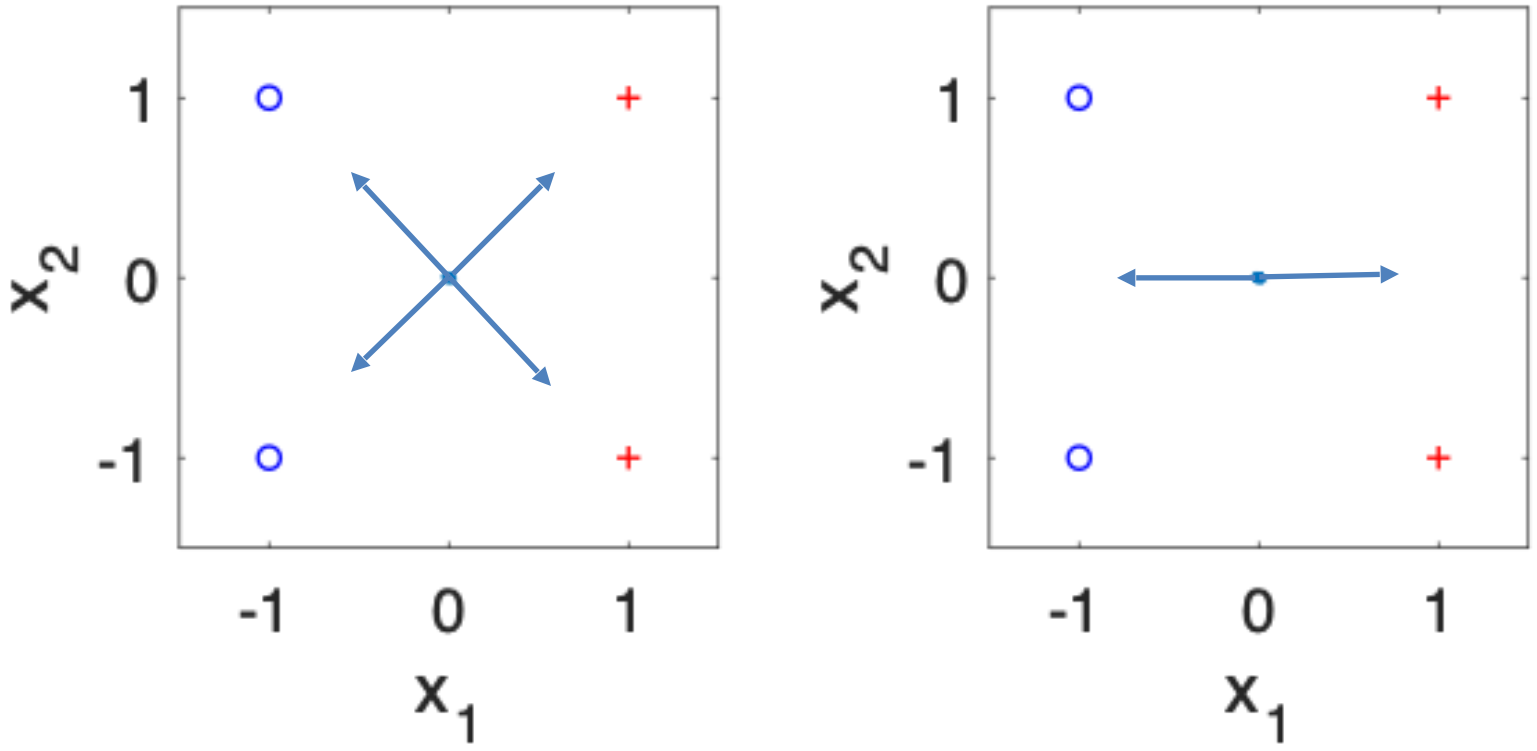


XoR Dynamics



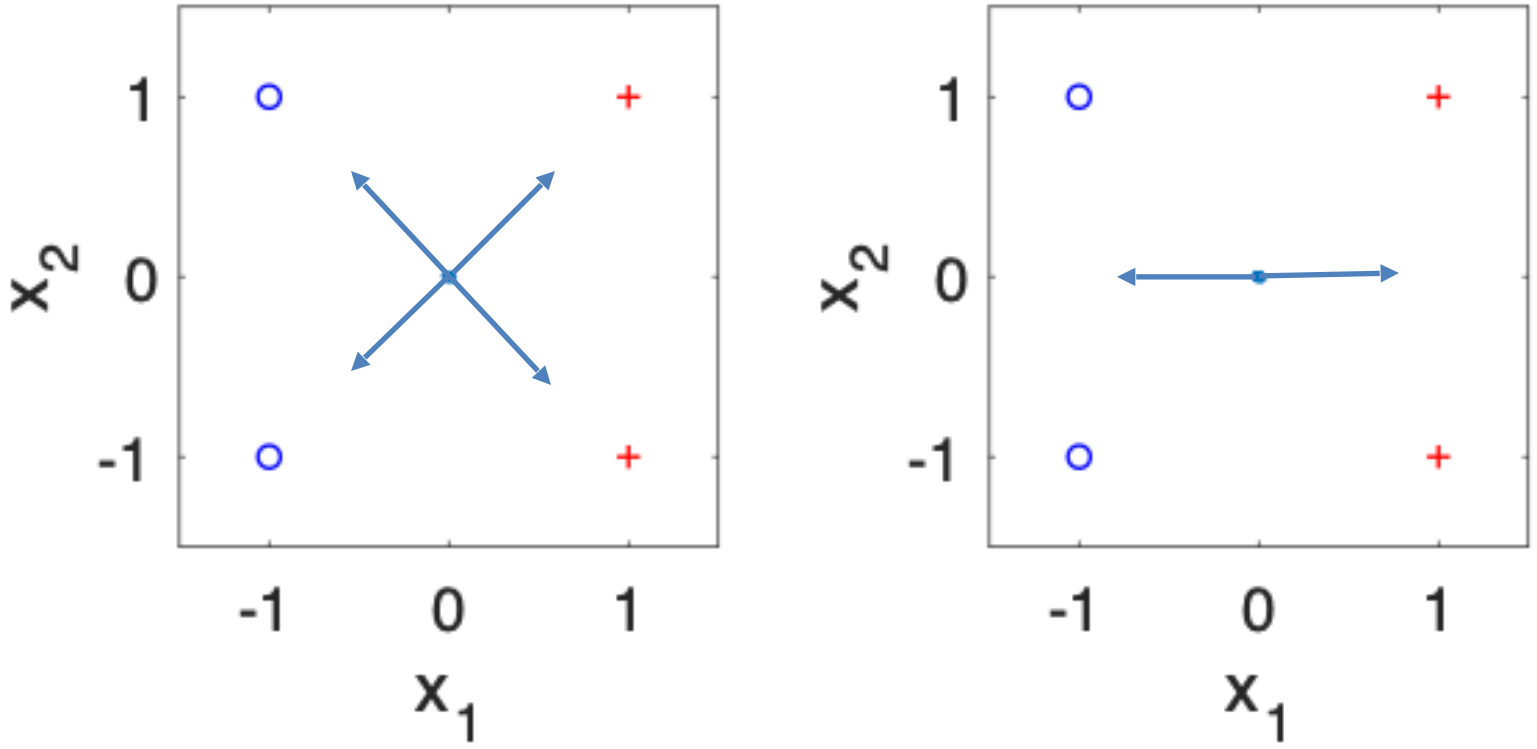


Which gating structures?



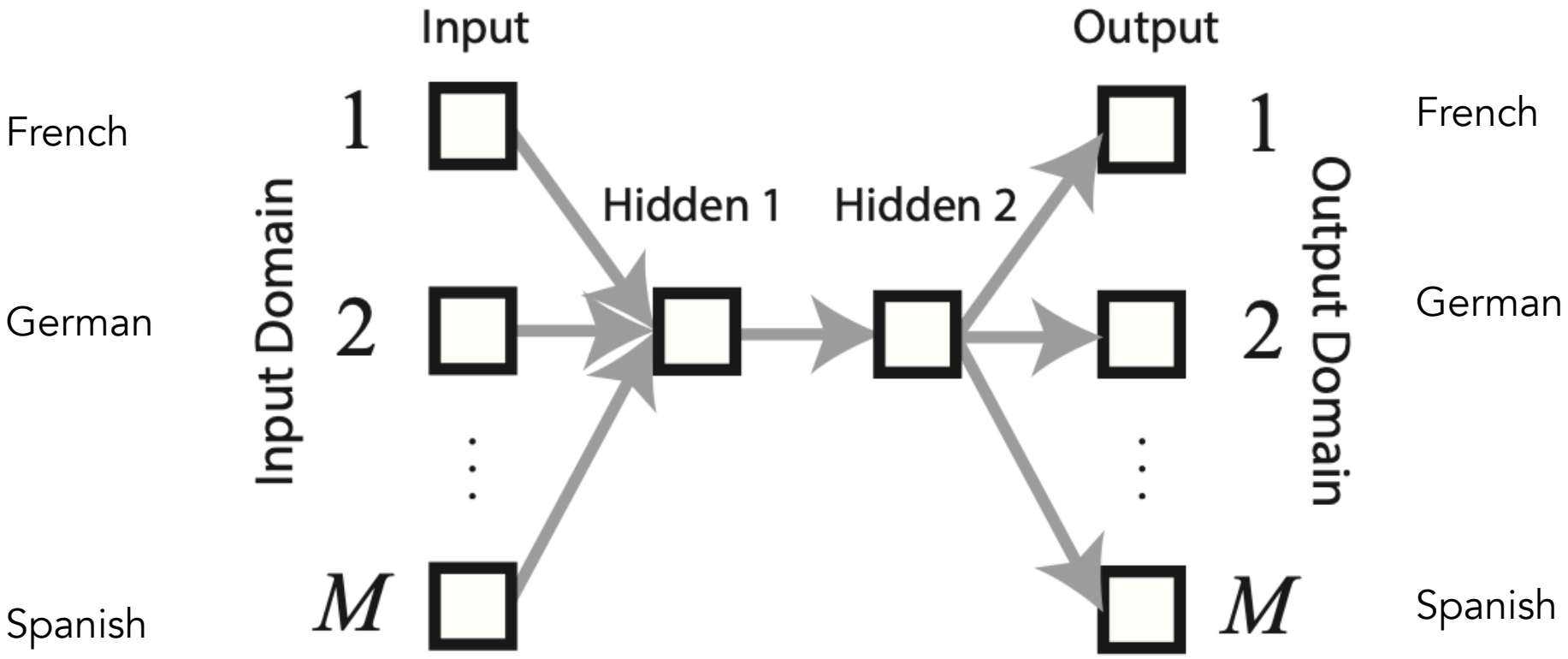
Neural Race Reduction

- Different gating schemes yield different effective datasets and deep linear network trajectories
- The ones which learn fastest dominate the solution



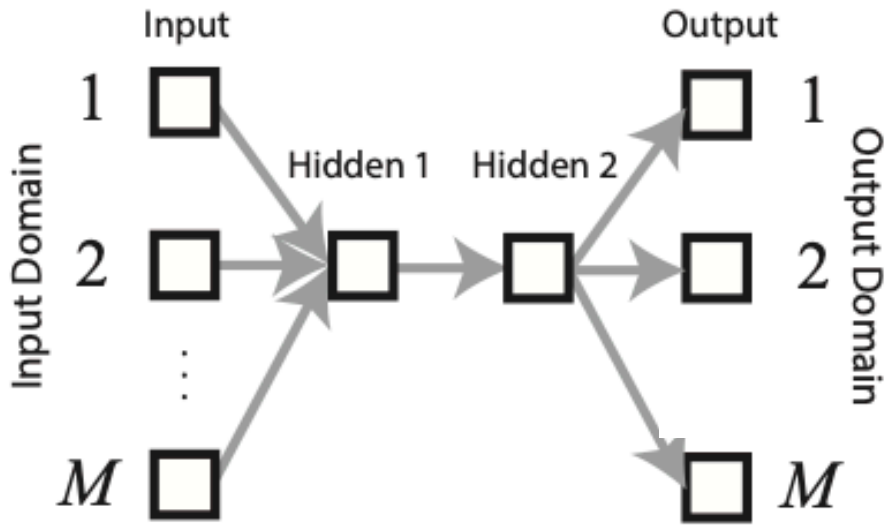
C.f. Lottery Ticket Hypothesis, Frankle & Carbin, 2019

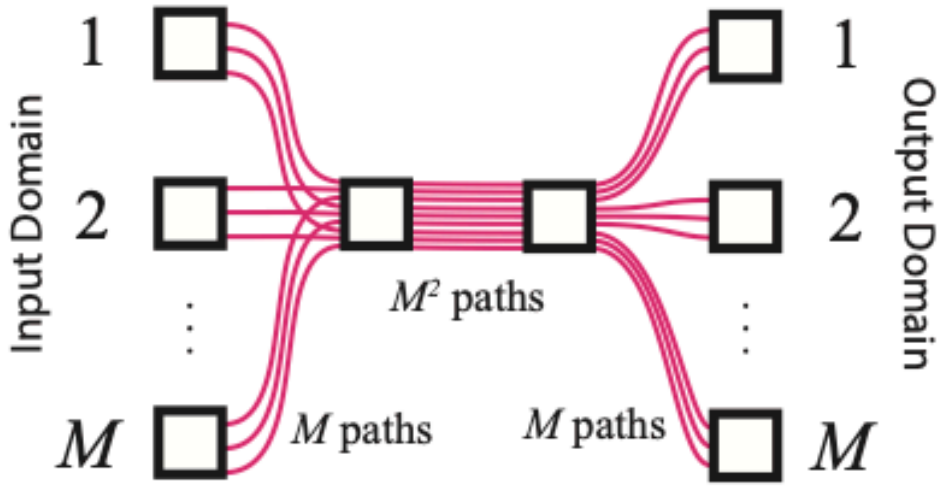
Routing Network



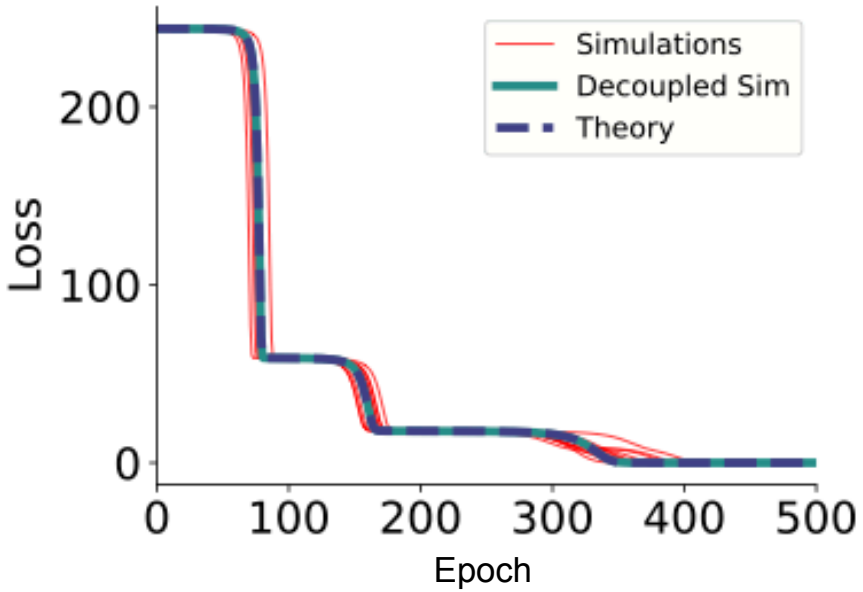
Each domain has distinctive input/outputs but similar underlying structural form

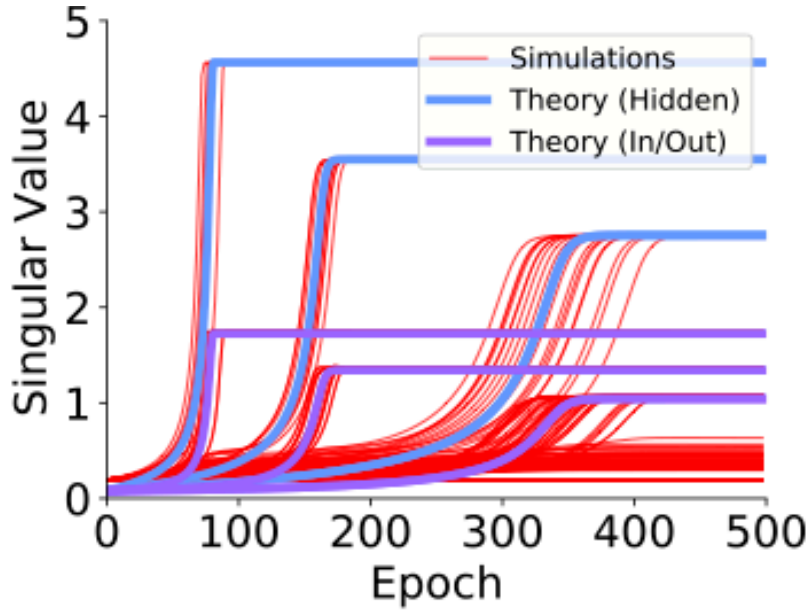
Pathway structure



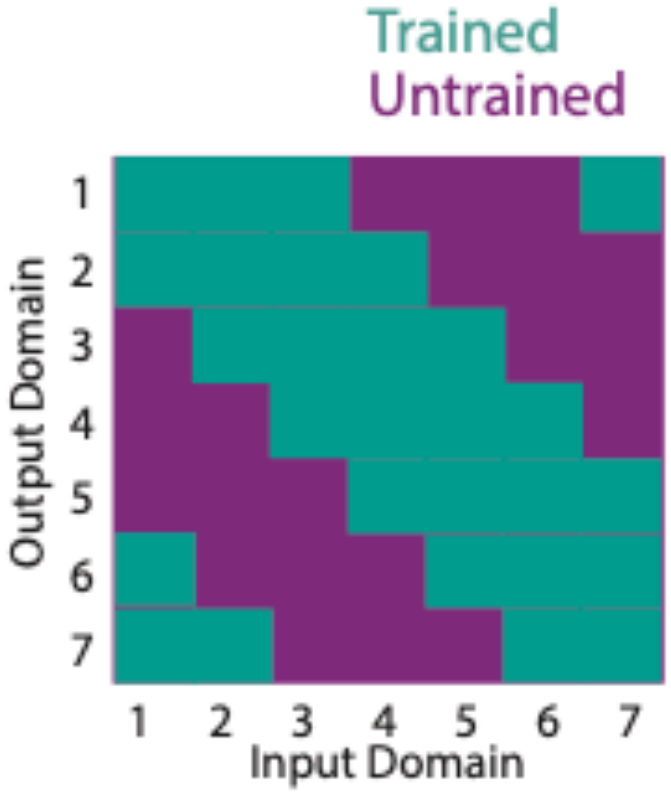


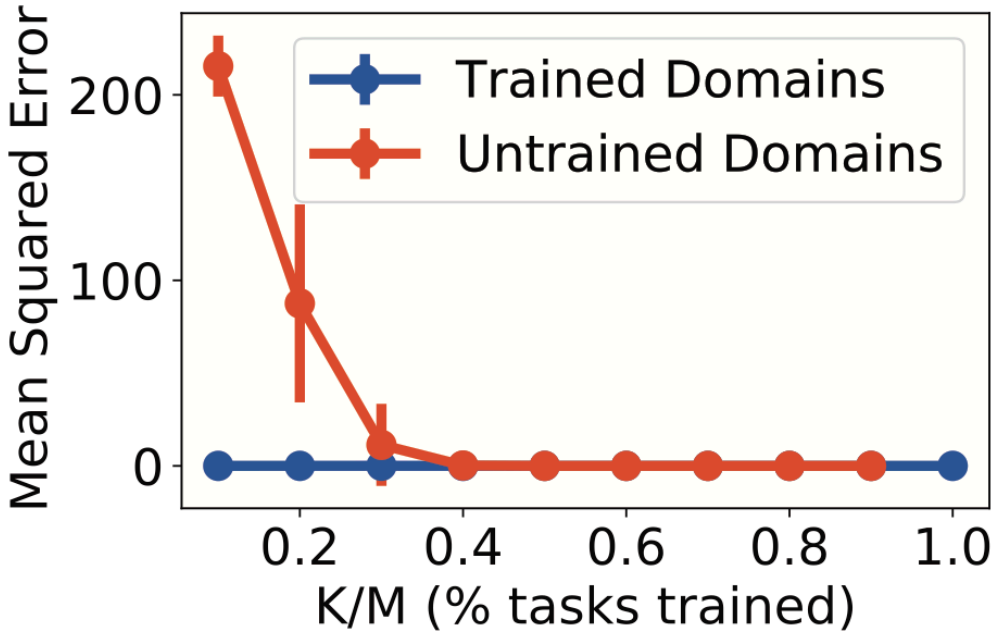
Dynamics of abstraction





Systematic generalization





Can translate between unseen domain pairs by switching gating

Summary

- Gated deep linear networks provide a surrogate model for studying nonlinear representation learning, the effect of architecture, and generalization

- Dynamics take the form of a **neural race**, with different pathways viewing different effective datasets

- Winning pathways dominate the solution