deep RL for flexible decision-making

fede carnevale
zador lab

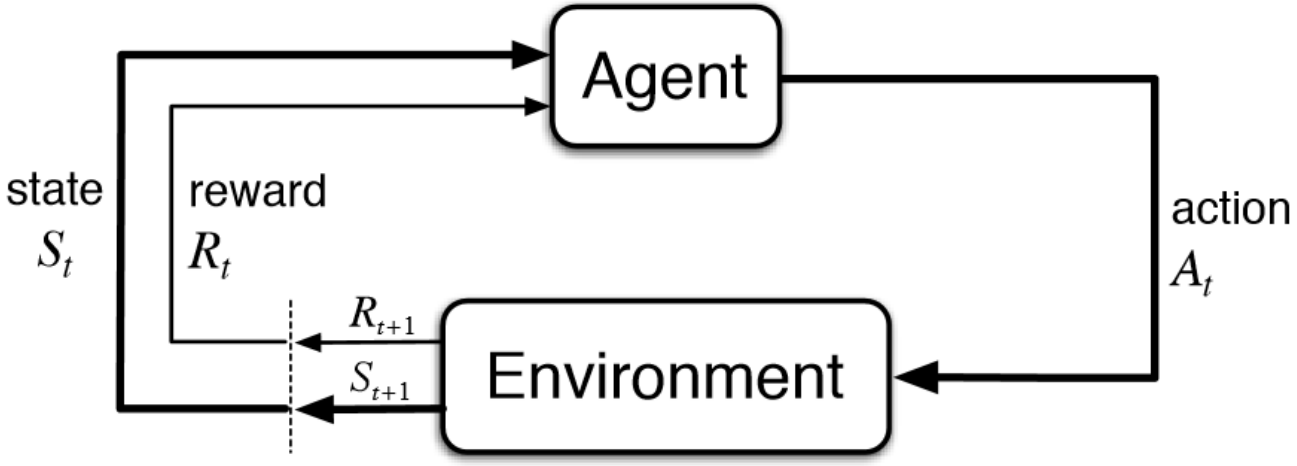
neuro in-house
September 29, 2017

Motivation

How do animals learn flexible behaviors that combine:

- external stimuli
- context, rules
- prior knowledge
- reward contingencies
- …

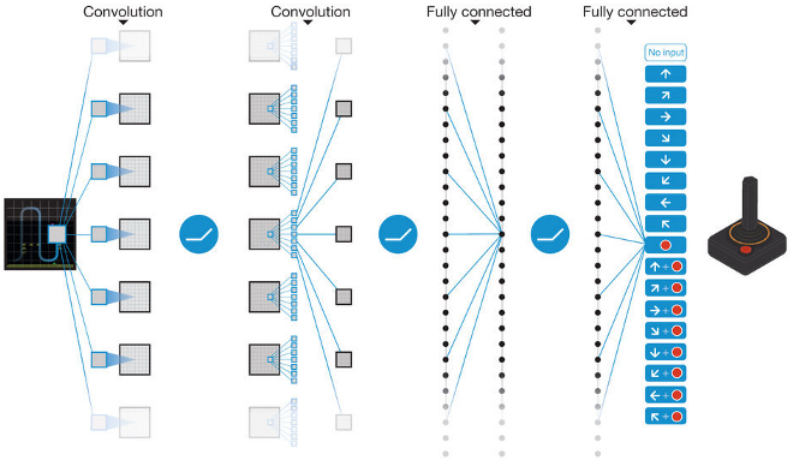
RL mathematical framework

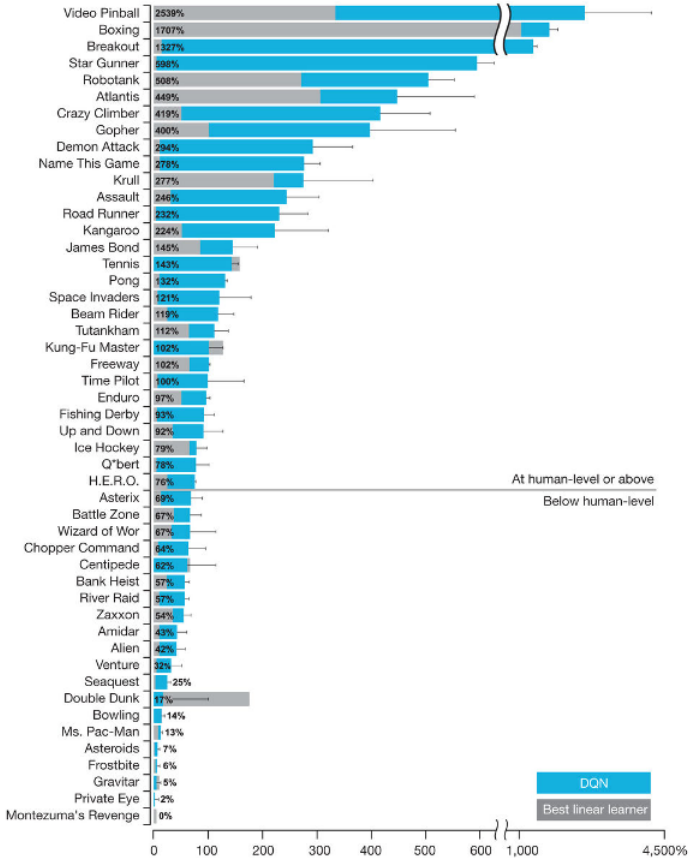


Policy: $\pi(a|s) = \mathbb{P}(A_t = a|S_t = s)$

Value: $V_\pi(s) = \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \ldots|S_t = s]$

RL advances in the last years





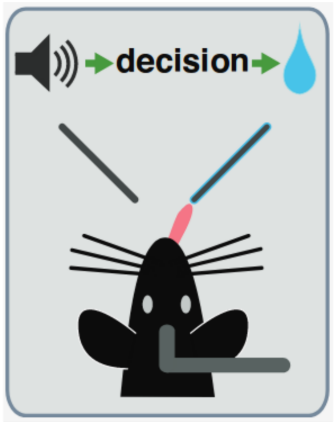
Mnih et al, 2014

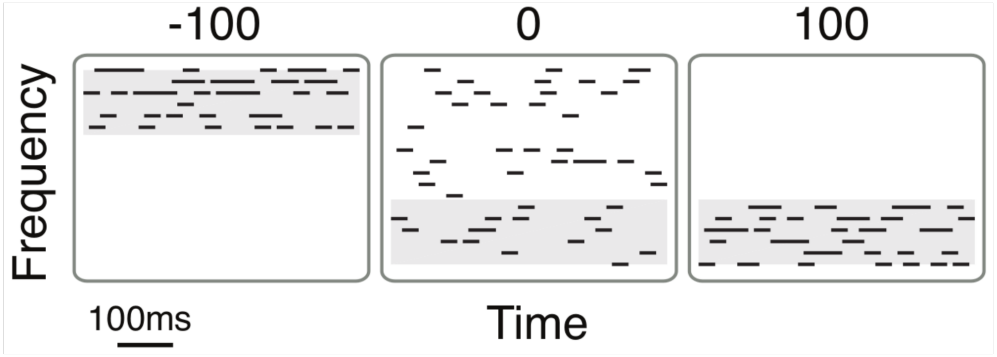
However,

- data inefficiency
- lack of flexibility

Animals, in contrast, are capable of learning flexible behaviors much more efficiently

Aki's task





Marbach and Zador, 2016

(S_t, A_t, R_t)

prior task:

Left:

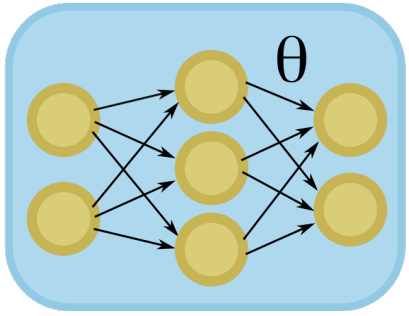
Right:

reward task:

Left:

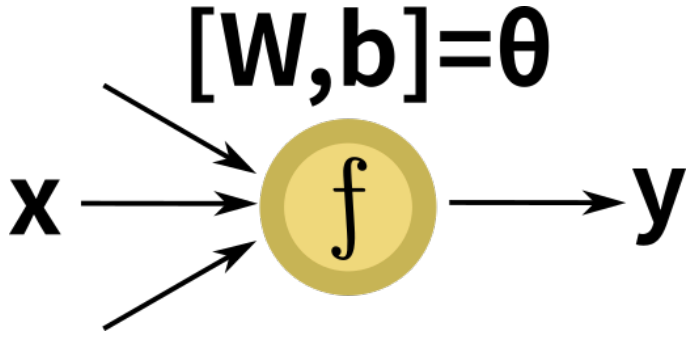
Right:

An RL agent for aki's task

$s^{(t)}$ → 

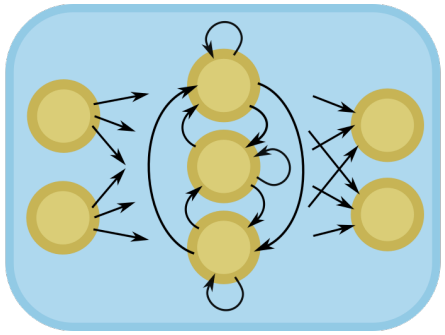
→ $\pi_\theta^{(t)}(s,a) = \mathbb{P}(a|s,\theta)$

Find θ that maximizes $J(\theta) = \mathbb{E}_{\pi_\theta}[R]$



$\mathbf{y} = f(\mathbf{W}\mathbf{x} + \mathbf{b})$

Recurrent Neural Networks

$S^{(t)}$ →  → $\pi_\theta^{(t)}(s,a) = \mathbb{P}(a|s,\theta)$

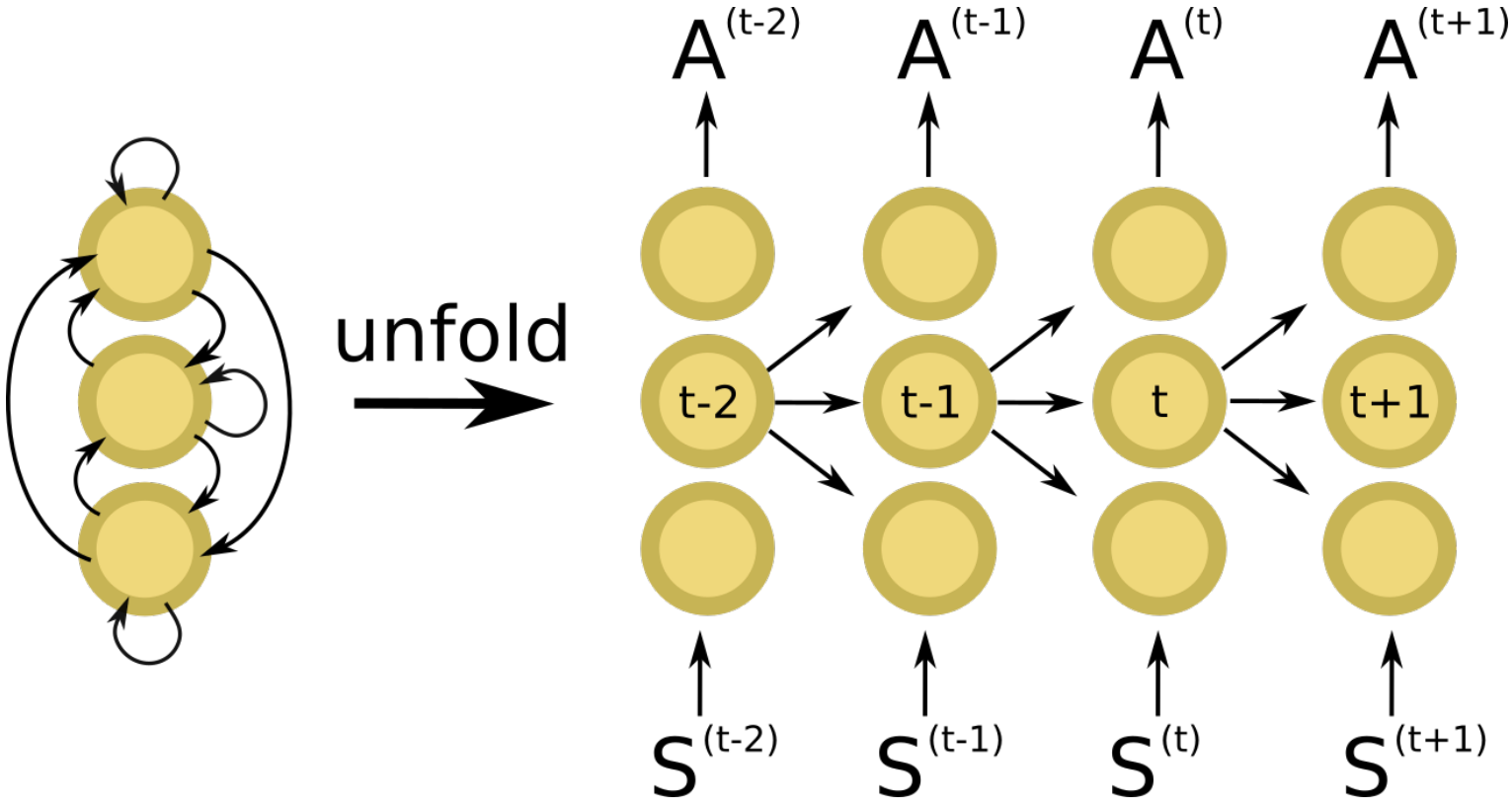
$$\mathbf{x}^{(t)} = \mathbf{W}\mathbf{h}^{(t-1)} + \mathbf{U}\mathbf{s}^{(t)} + \mathbf{b}$$

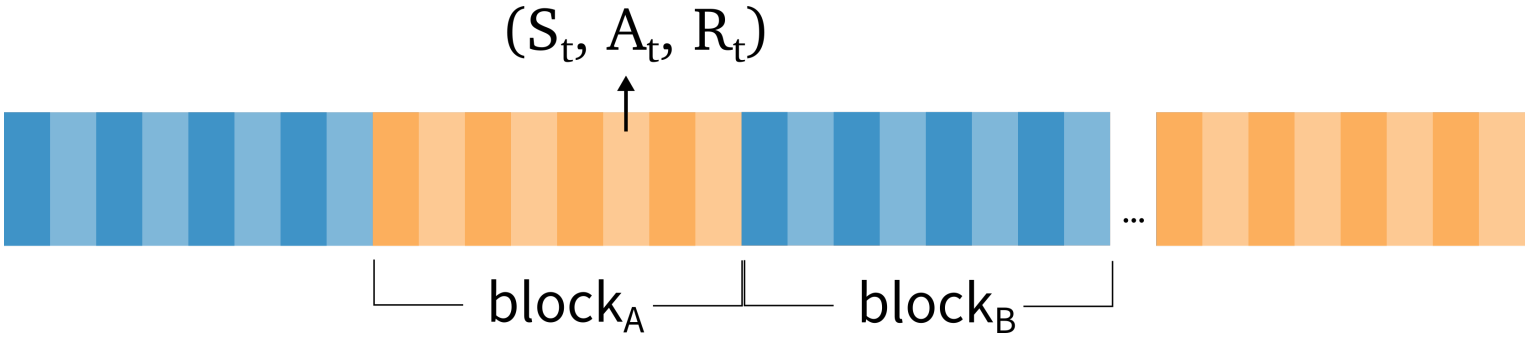
$$\mathbf{h}^{(t)} = \tanh\left(\mathbf{x}^{(t)}\right)$$

$$\mathbf{o}^{(t)} = \mathbf{V}\mathbf{h}^{(t)} + \mathbf{c}$$

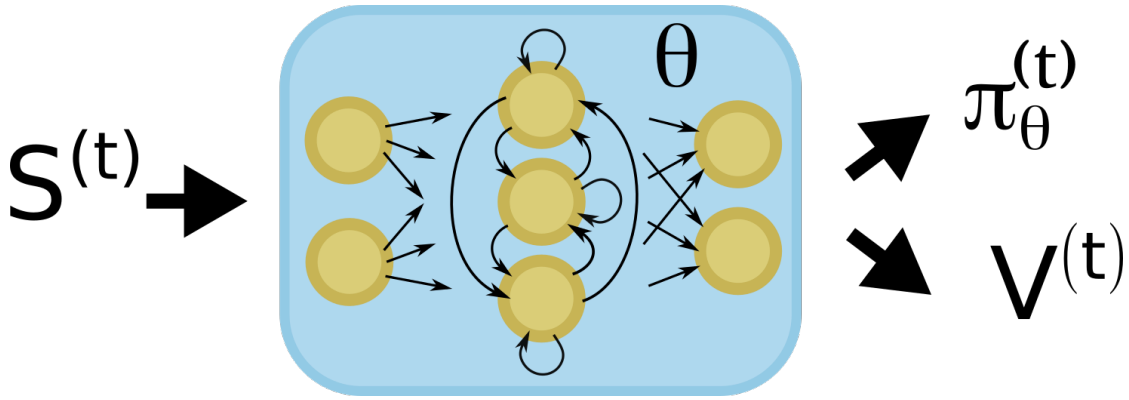
Recurrent connections allow previous inputs to persist (memory) and affect the output.

Recurrent Neural Networks





Training procedure

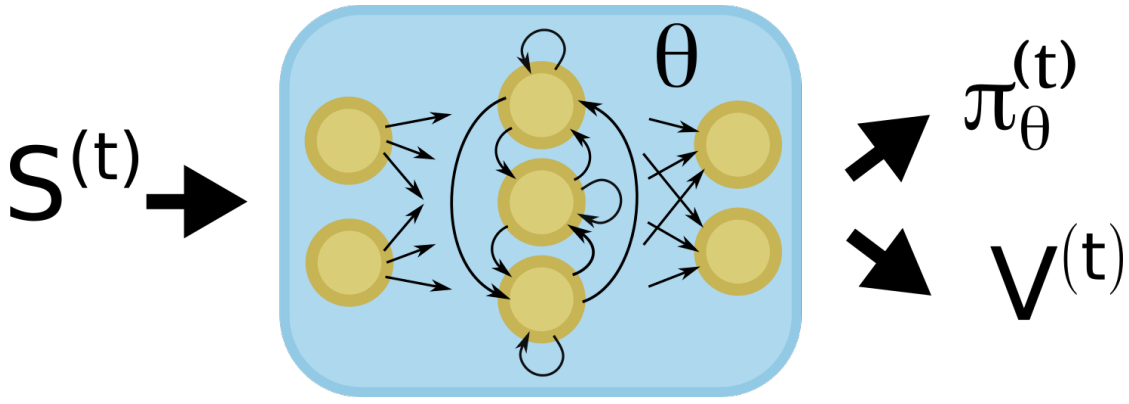


Find θ that maximizes $J(\theta) = \mathbb{E}_{\pi_\theta}[R]$

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}[\nabla_\theta log\, \pi_\theta(s,a)\, V^{\pi_\theta}(s)]$$

$$\theta \leftarrow \theta + \alpha \nabla_\theta log\, \pi_\theta(s,a)\, V^{\pi_\theta}(s)$$

Training procedure



Algorithm Reinforce

Initialize θ arbitrarily

for each block $([s_1, a_1, r_1], ..., [s_T, a_T, r_T])$ **do**

for $t = 1$ to T **do**

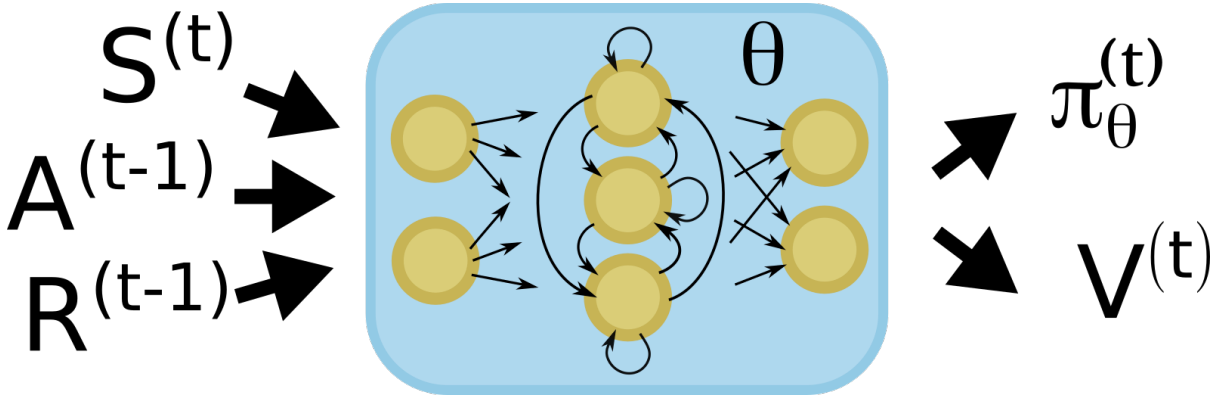
$\delta_v(s_t) = r_t + \gamma V(s_t) - V(s_{t-1})$

$\theta \leftarrow \theta + \alpha \nabla_\theta log\, \pi_\theta(s_t, a_t) \delta_v(s_t)$

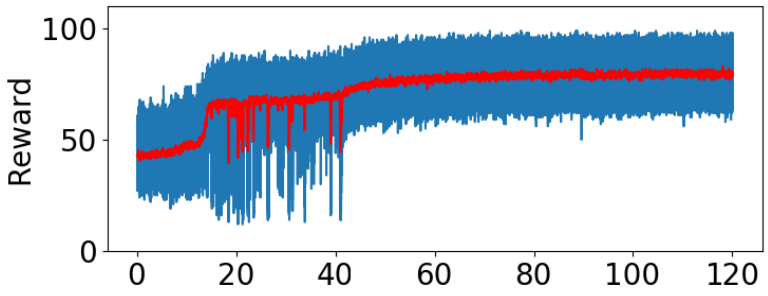
end for

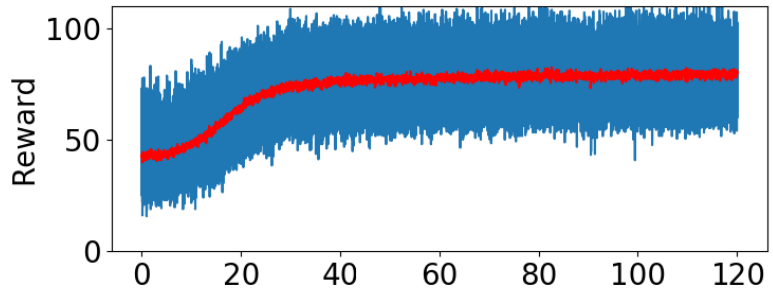
end for

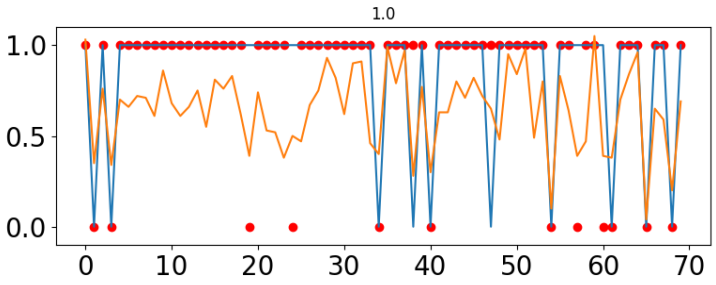
Training procedure

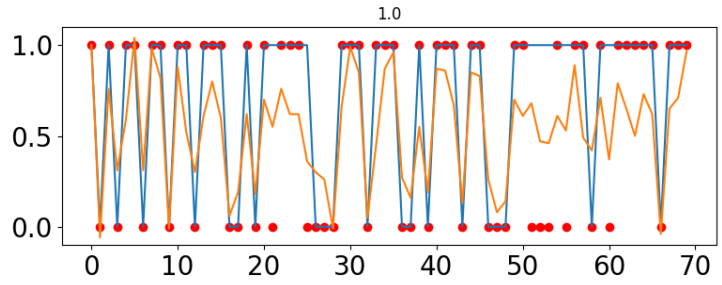


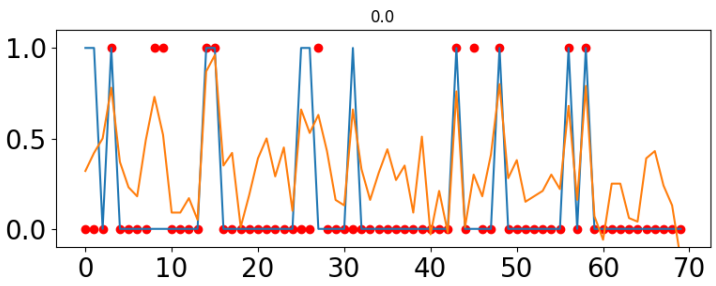
The agent solves both tasks

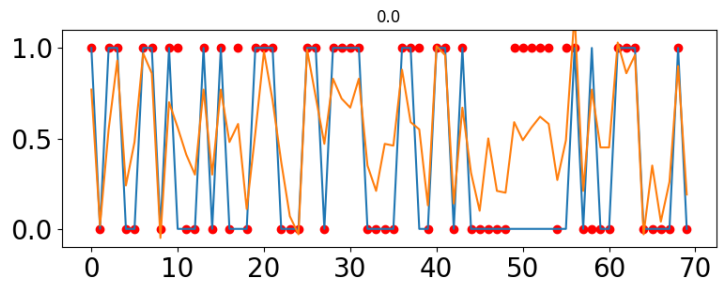




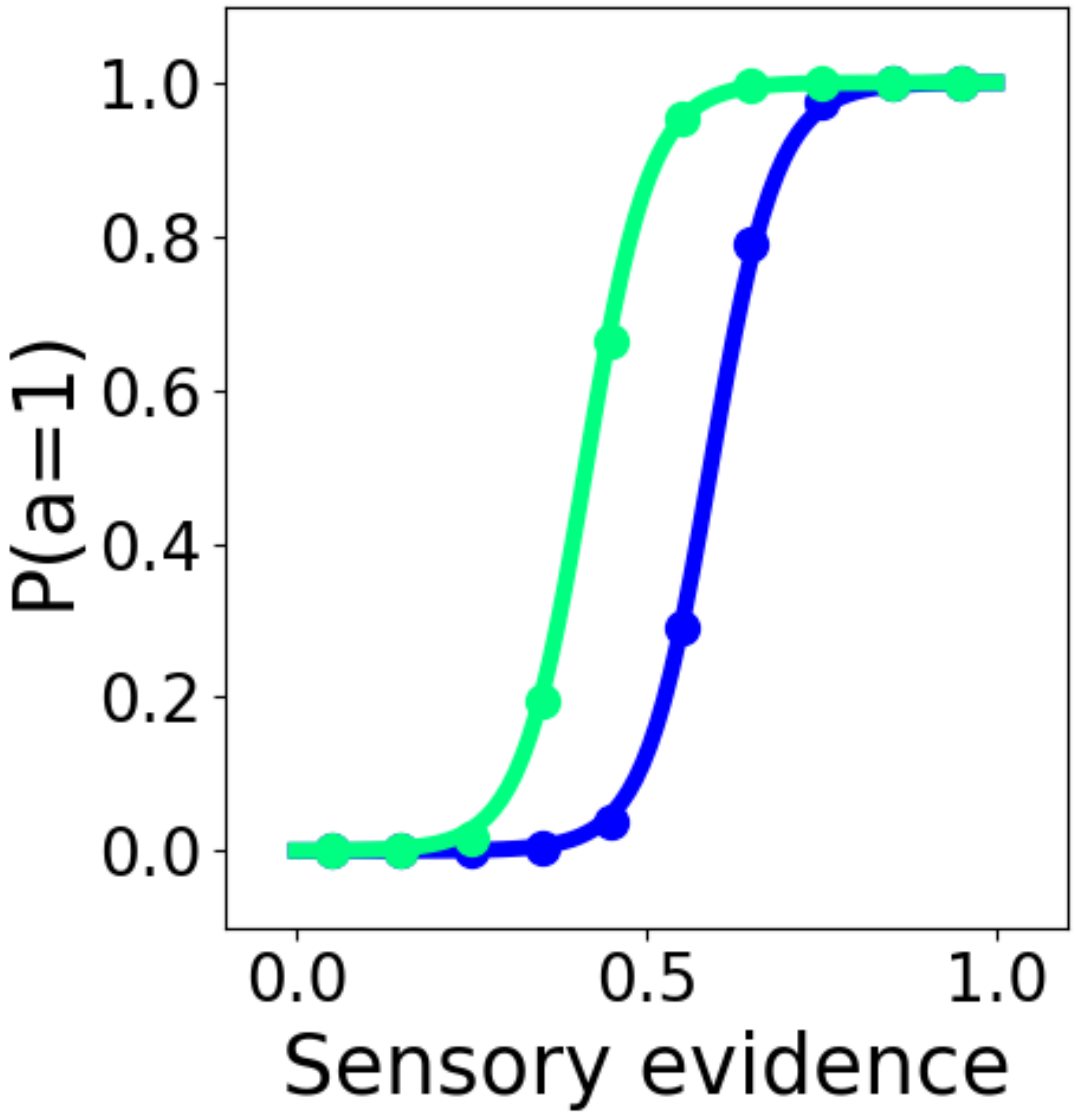




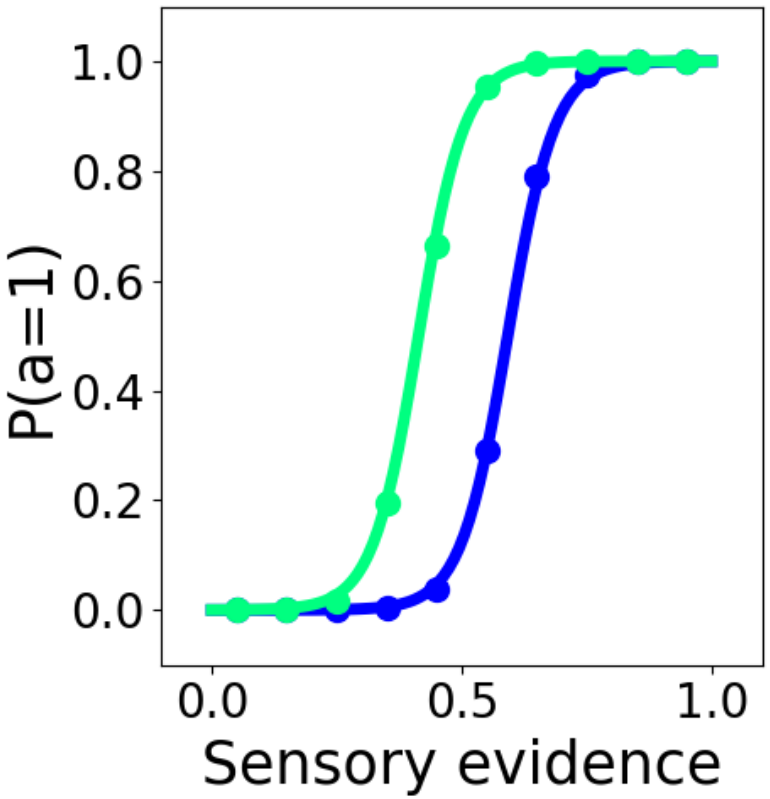


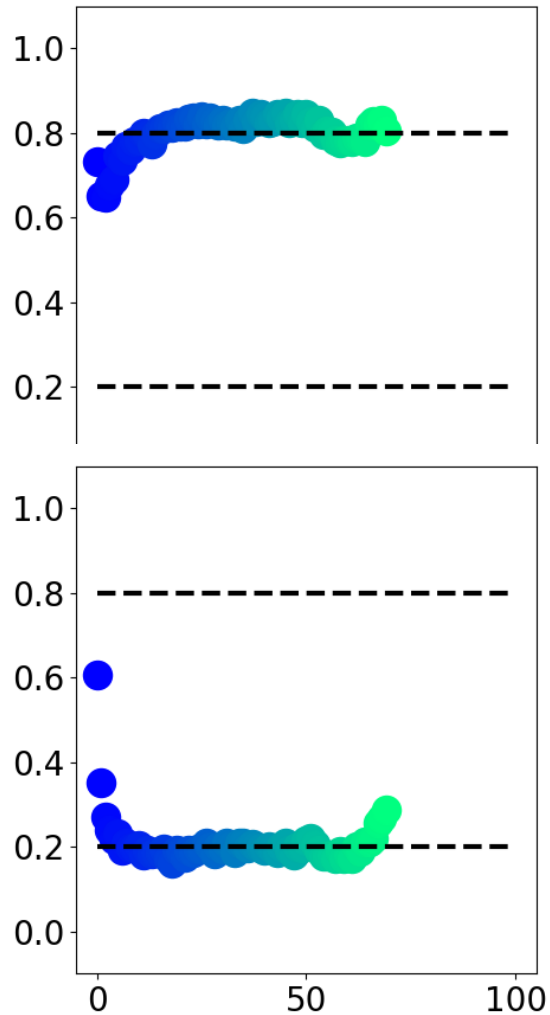


Asym prior/reward biases choice

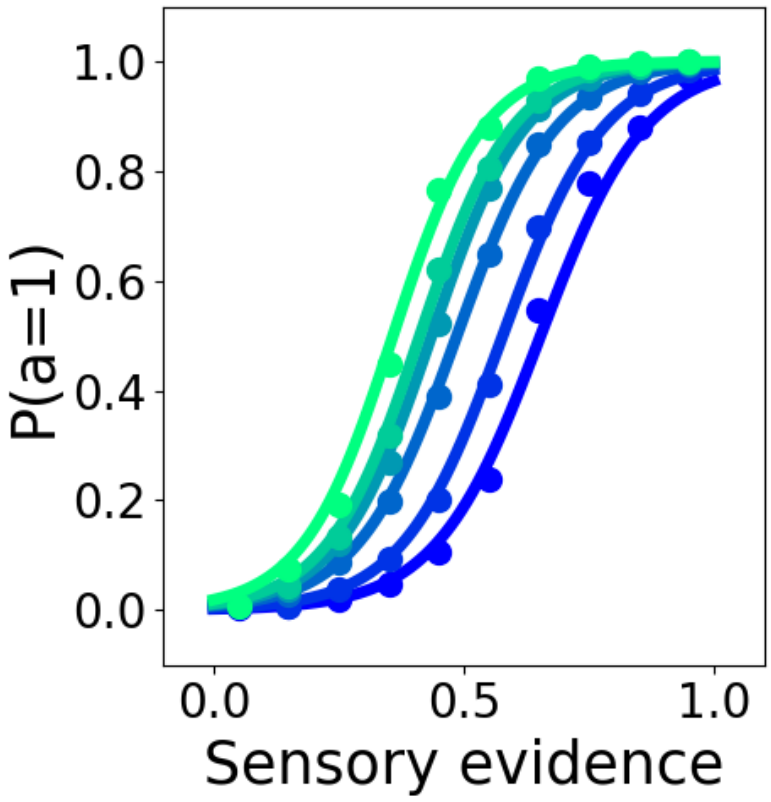


Bias develops within block

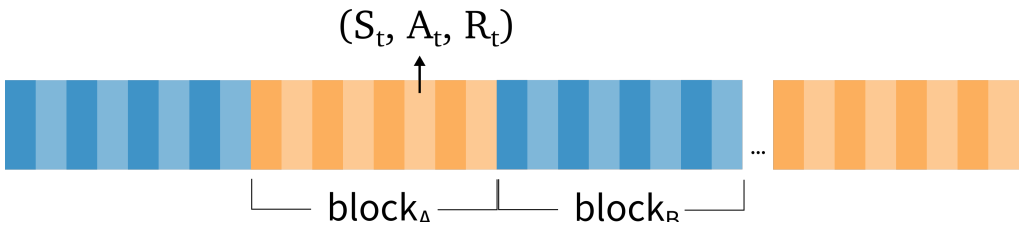


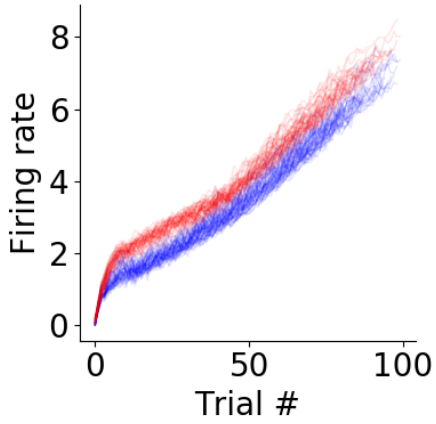


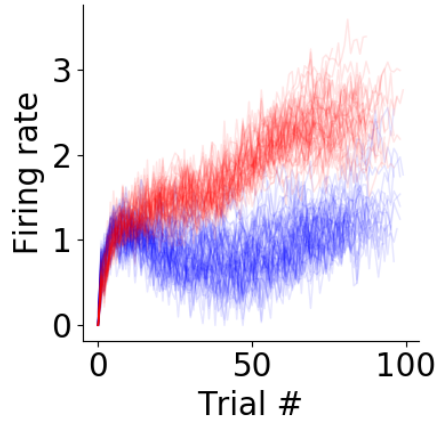
The agent can generalize to new priors/rewards

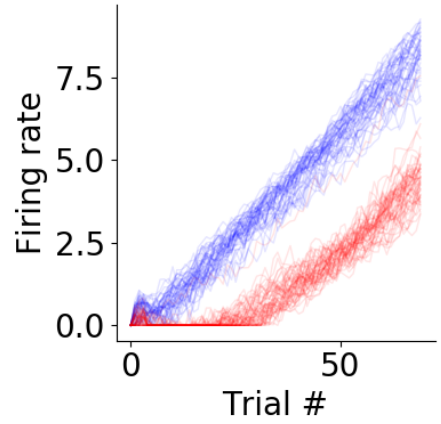


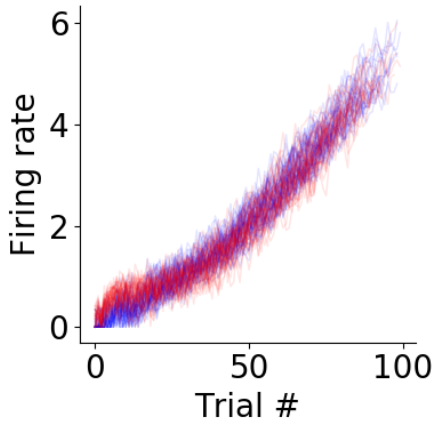
Firing rates in time

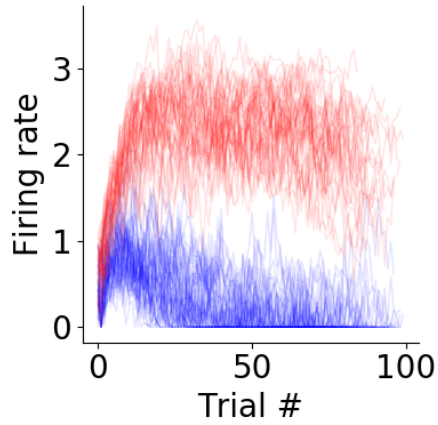


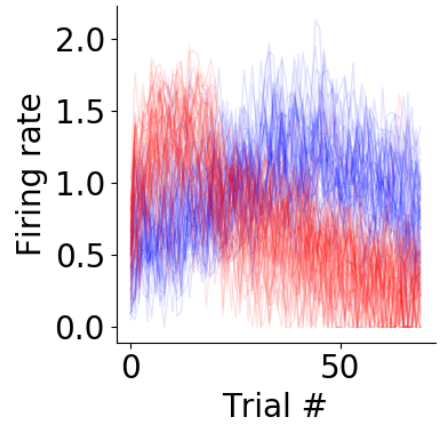




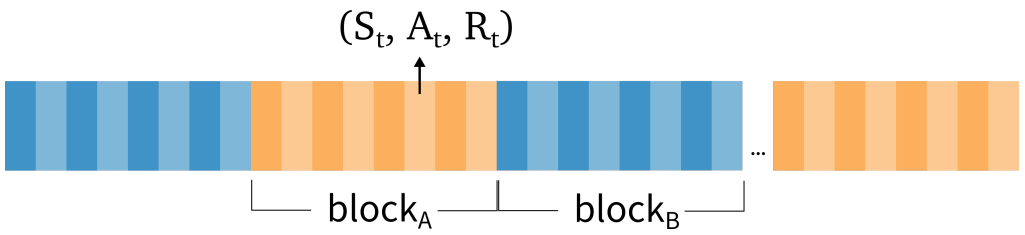


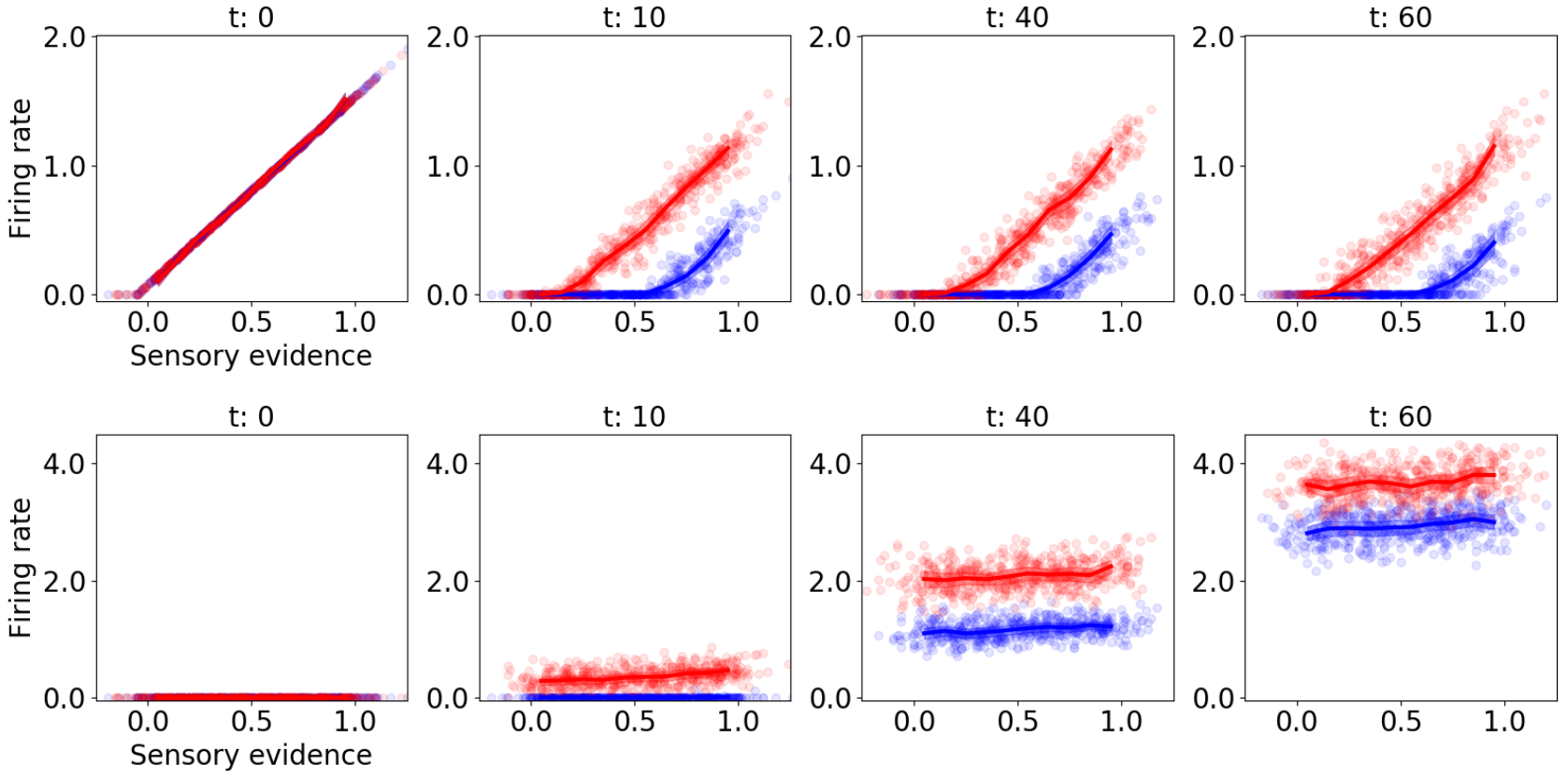






Tuning curve in time





Discussion

Thanks!