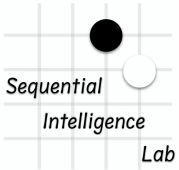
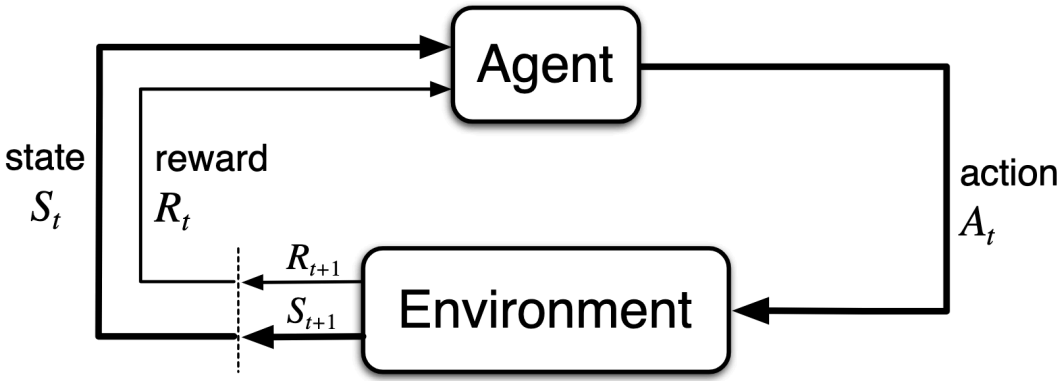
Towards Provable Emergence of In-Context Reinforcement Learning

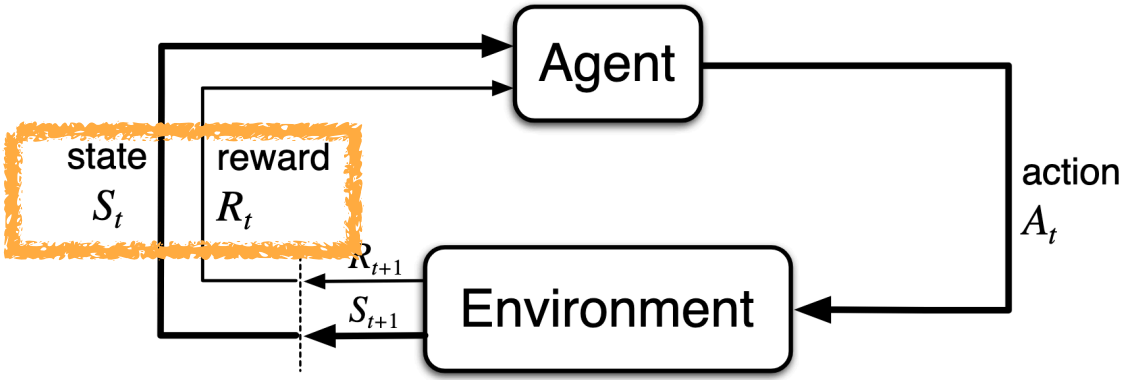
Jiuqi Wang, Rohan Chandra, Shangtong Zhang



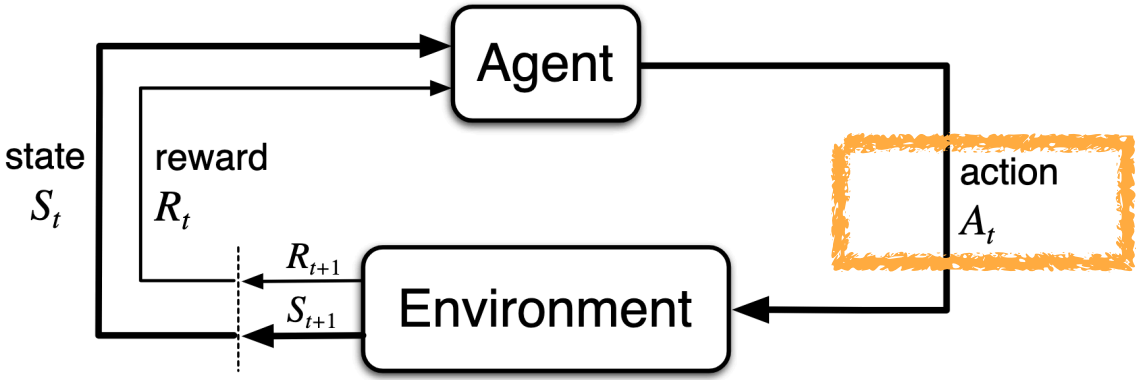
Reinforcement Learning Recap



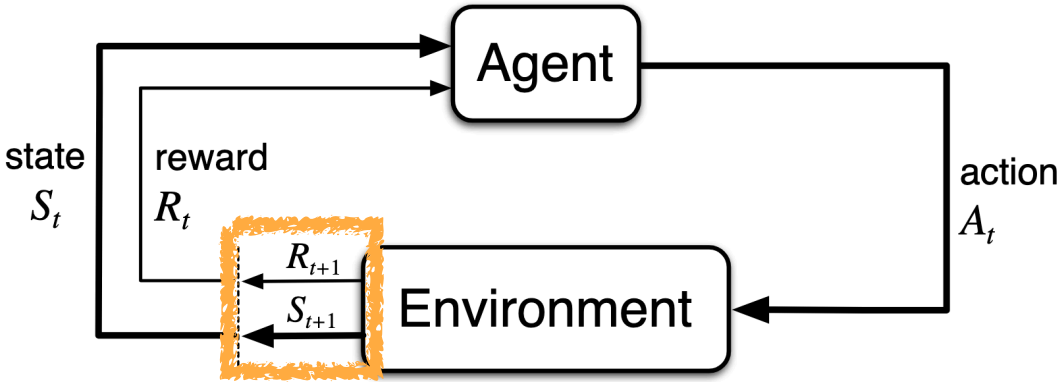
Reinforcement Learning Recap



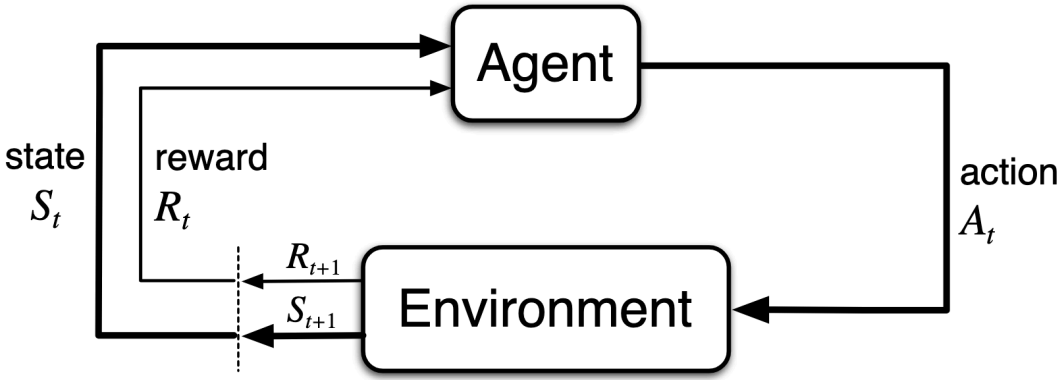
Reinforcement Learning Recap



Reinforcement Learning Recap



Policy Evaluation



The goal of the **policy evaluation** is to find the value function $v_\pi : \mathcal{S} \to \mathbb{R}$ for π, defined as

$$v_\pi(s) \doteq \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} \middle| S_0 = s \right]$$, where γ is a discount factor.

Transformers Can Perform Policy Evaluation In-Context

Context

$S_0, R_1, S_1, R_2, \ldots, S_{m-1}, R_m, S_m$

Query

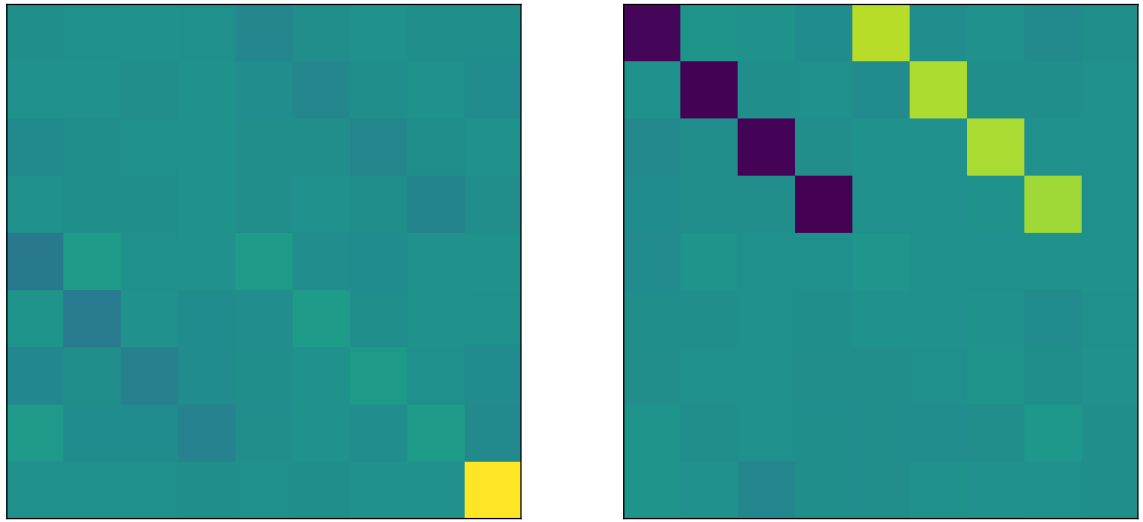
s_q

$\hat{v}(s_q) \approx v_\pi(s_q)$

Multi-Task Training Gives Rise to In-Context Policy Evaluation

- Let $TF_\theta\left(s_q; C\right)$ be the output of the Transformer parameterized by θ, given query state s_q, and conditioned on context C.
- **Multi-task TD training** updates θ as
  $\theta \leftarrow \left(r(s) + \gamma TF_\theta(s'; C) - TF_\theta(s; C)\right) \nabla TF_\theta(s; C)$.
- The update is simply the regular semi-gradient temporal difference update with an additional context in the input.

Multi-Task Training Gives Rise to In-Context Policy Evaluation



These Transformer parameters enable in-context policy evaluation!

Why do multi-task training give rise to the parameters that enable ICTD?

(Our Contribution) ICTD Parameters Minimize the NEU Loss!

- Let θ^{TD} denote the parameters that enable ICTD.
- The expected update of multi-task TD is
  $\Delta^{TD}(\theta) \doteq \mathbb{E}\left[(r(S) + \gamma TF_\theta(S'; C) - TF_\theta(S; C))\nabla TF_\theta(S; C)\right]$
- We proved that θ^{TD} is a **global minimizer** of the **norm of expected update (NEU)** loss, defined as $J(\theta) \doteq \left\| \Delta^{TD}(\theta) \right\|_1$.

(Our Contribution) Multi-Task Monte Carlo Also Gives Rise to ICTD Parameters!

- Keeping everything else unchanged, **multi-task MC** training updates θ as $\theta \leftarrow \left(G(s) - TF_\theta(s;C)\right) \nabla TF_\theta(s;C)$, where $G(s)$ is a return unrolled from s.
- θ^{TD} emerges from multi-task MC training as well!

(Our Contribution) Multi-Task Monte Carlo Also Gives Rise to ICTD Parameters!

Multi-task TD

Multi-task MC

Thank you!