DEMYSTIFYING DISCOUNTING

Guest Lecture: CMPUT 655
27 Nov 2024

Abhishek Naik
abhisheknaik22296@gmail.com

Formerly:

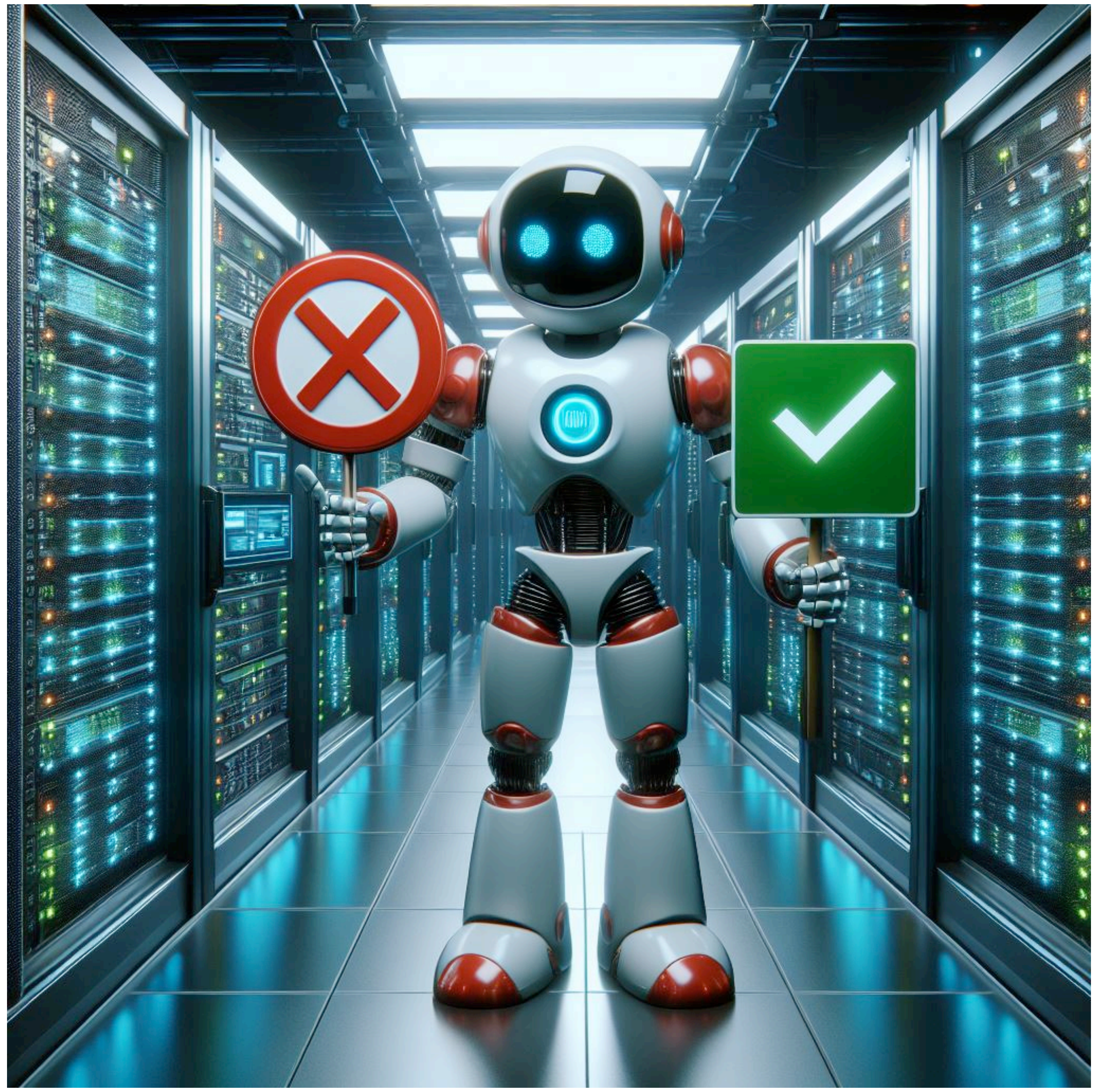
Now:

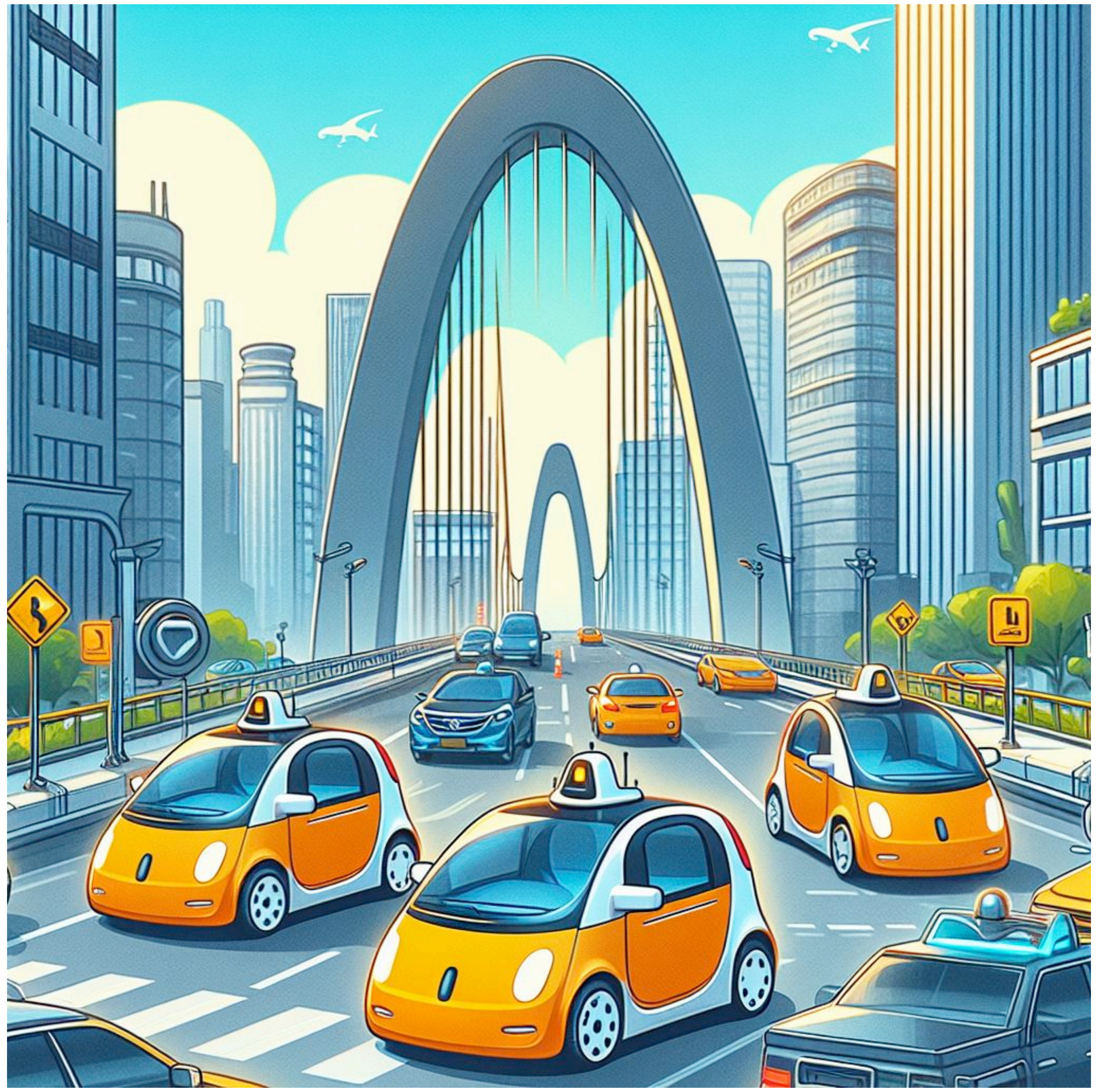
OUTLINE

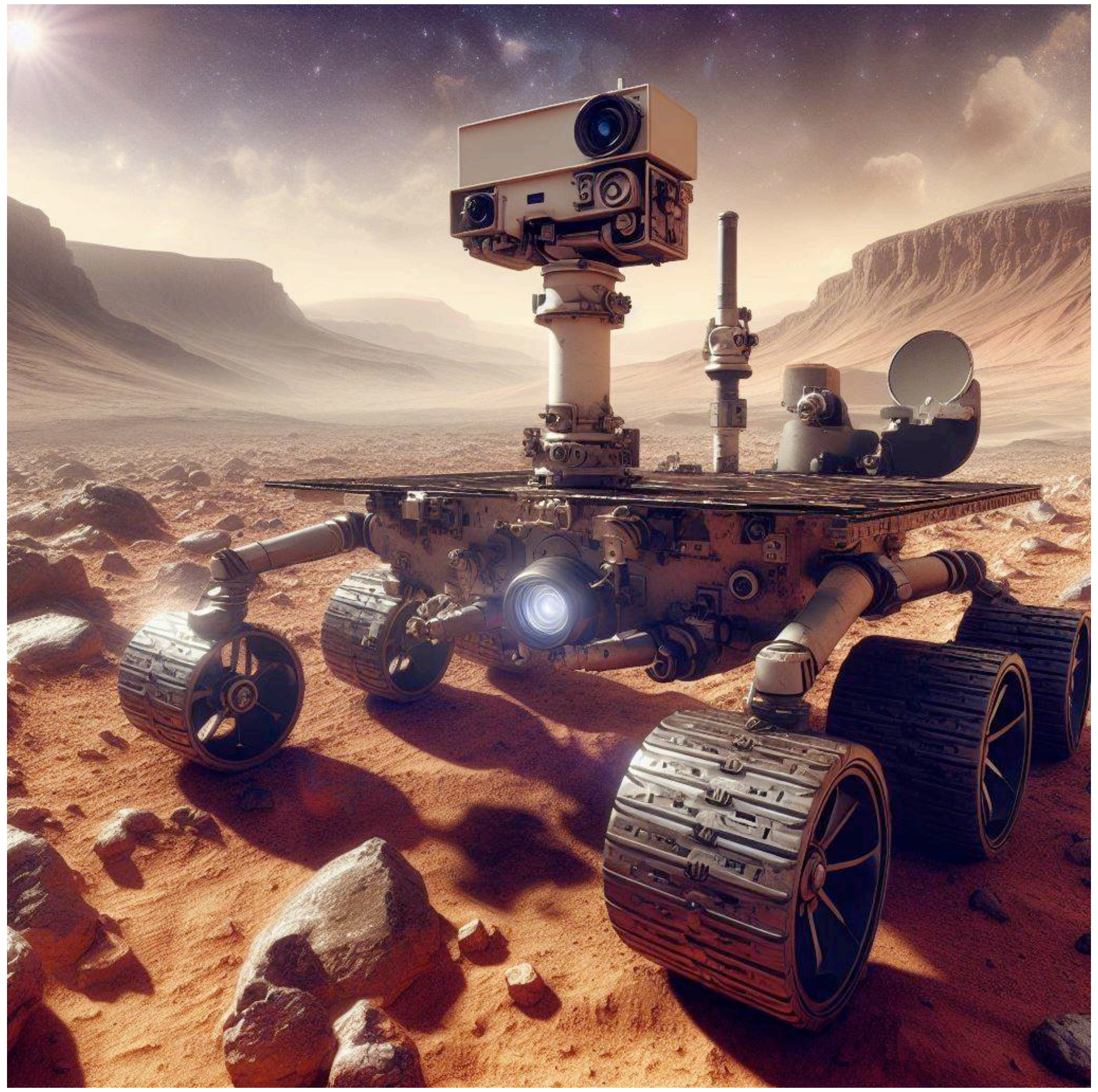
0. Problem setting
1. The discounted-reward formulation
2. The main issue with discounting
3. The average-reward formulation
4. Connections: improving discounted methods using average reward

PROBLEM SETTING

CONTINUING PROBLEMS

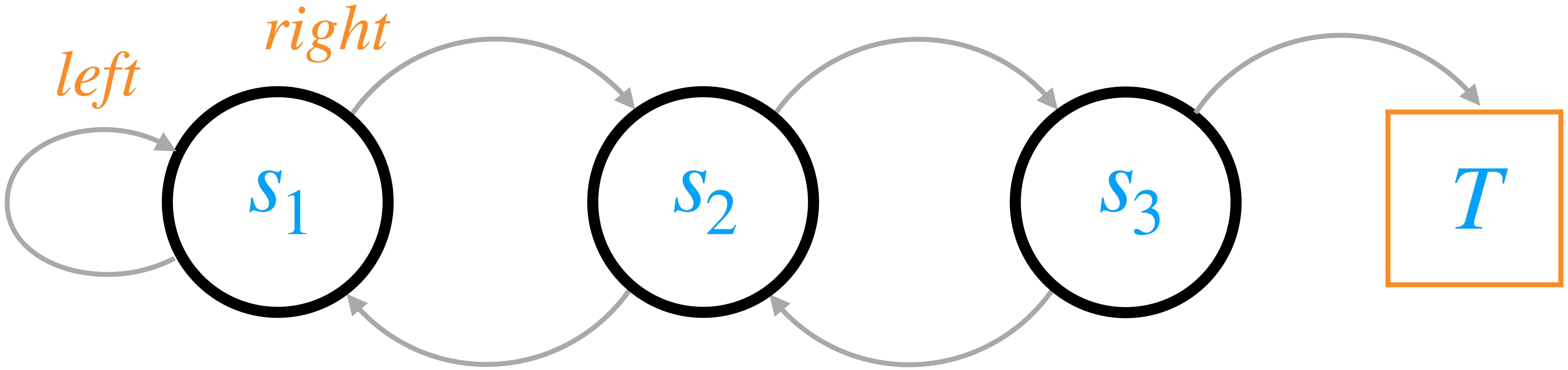


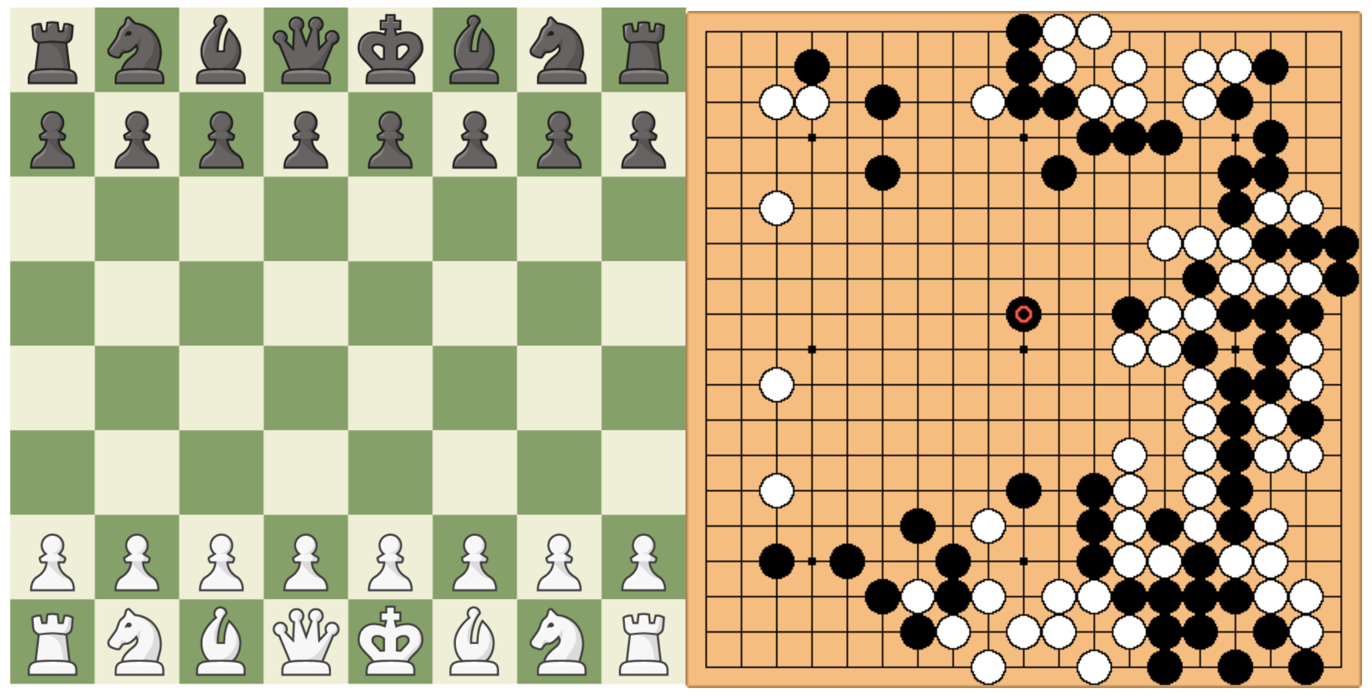




Images generated using DALL·E 3

RECAP: EPISODIC PROBLEMS





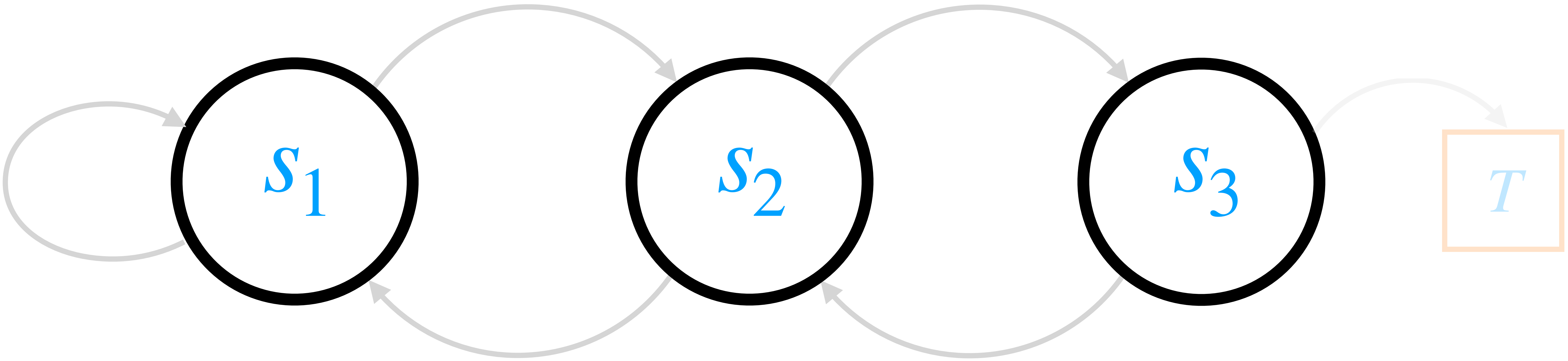
TIME SPANS OF DECISIONS' CONSEQUENCES ARE BOUNDED IN EPISODIC PROBLEMS

$$\ldots \quad S_{t-k} \;\ldots\; S_{t-1}\; A_{t-1}\; R_t \quad S_t\; A_t\; R_{t+1}\; S_{t+1}\; A_{t+1} \;\ldots\; S_{t+n} \quad \ldots$$

And no credit assignment occurs across episodic boundaries.

'Resets' don't really exist in life…

CONTINUING PROBLEMS



$$\dots\ S_{t-k}\ \dots\ S_{t-1}\ A_{t-1}\ R_t\ S_t\ A_t\ R_{t+1}\ S_{t+1}\ A_{t+1}\ \dots\ S_{t+n}\ \dots$$

ASIDE: IMPORTANT DISTINCTIONS WITH SIMILAR-SOUNDING TERMS

- Continual / never-ending / lifelong learning:
  emphasizes a learning agent's *continual* need to adapt to a non-stationary world.
  - Non-stationarity is orthogonal to the episodic or continuing nature of the agent-environment interaction.
  - Continuing problems can have non-stationary aspects.

- Continuous problems:

  have *continuous* state and/or action spaces
  - Continuing problems can have continuous state/action spaces.

CONTINUING PROBLEMS: FORMULATIONS

$$S_0\ A_0\ R_1\ S_1\ A_1, R_2 \ldots\ S_t\ A_t\ R_{t+1}\ S_{t+1}\ A_{t+1}\ R_{t+2}\ \ldots$$

$$\max_{\pi} \sum_{t}^{\infty} R_t$$

Discounted-Reward Formulation

$$\max_{\pi}\ v_{\pi}^{\gamma}(s), \forall s$$

Average-Reward Formulation

$$\max_{\pi}\ r(\pi)$$

$\gamma \in [0,1)$

$$R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$$

$$v_{\pi}^{\gamma}(s) \doteq \mathbb{E}_{\pi}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \mid S_t = s]$$

$$r(\pi) \doteq \lim_{n \to \infty} \frac{1}{n} \mathbb{E}_{\pi}\Big[\sum_{t=1}^{n} R_t\Big]$$

OUTLINE

0. Problem setting
1. The discounted-reward formulation
2. The main issue with discounting
3. The average-reward formulation
4. Connections: improving discounted methods using average reward

DISCOUNTED-REWARD FORMULATION

$S_0\ A_0\ R_1\ S_1\ A_1, R_2 \dots\ S_t\ A_t\ R_{t+1}\ S_{t+1}\ A_{t+1}\ R_{t+2}\ \dots$

$$\max_{\pi} \sum_{t}^{\infty} R_t$$

$$\pi^*_\gamma \rightarrow \max_{\pi} v^\gamma_\pi(s), \forall s$$

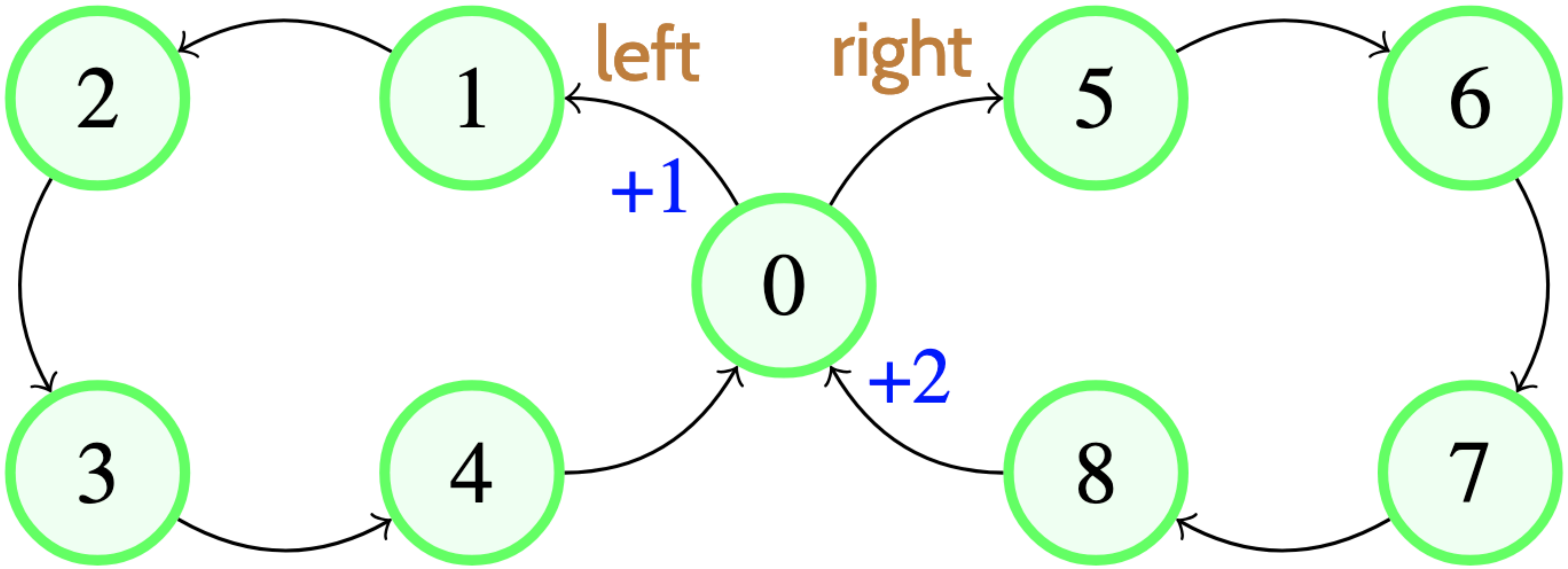
$\gamma \in [0,1)$

$$v^\gamma_\pi(s) \doteq \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s]$$

$$q^\gamma_\pi(s,a) \doteq \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s, A_t = a]$$

$$\pi^*_\gamma(s) = \arg\max_a q_{\pi^*_\gamma}(s,a)$$

THE BEST POLICY
DEPENDS ON THE DISCOUNT FACTOR



- $\pi^*_{\gamma=0}$: left
- $\pi^*_{\gamma=0.9}$: right

A USEFUL THEOREM

Blackwell, 1962; Grand-Clément & Petrik, 2023

In any *finite* MDP, there exists a discount factor $\gamma^* \in [0,1)$ such that $\forall \gamma \geq \gamma^*$, γ-optimal policies are also average-reward-optimal.

That is, π^*_γ maximizes the average reward for all $\gamma \geq \gamma^*$.

So just set a “high” value for γ?

OUTLINE

0. Problem setting
1. The discounted-reward formulation
2. The main issue with discounting
3. The average-reward formulation
4. Connections: improving discounted methods using average reward

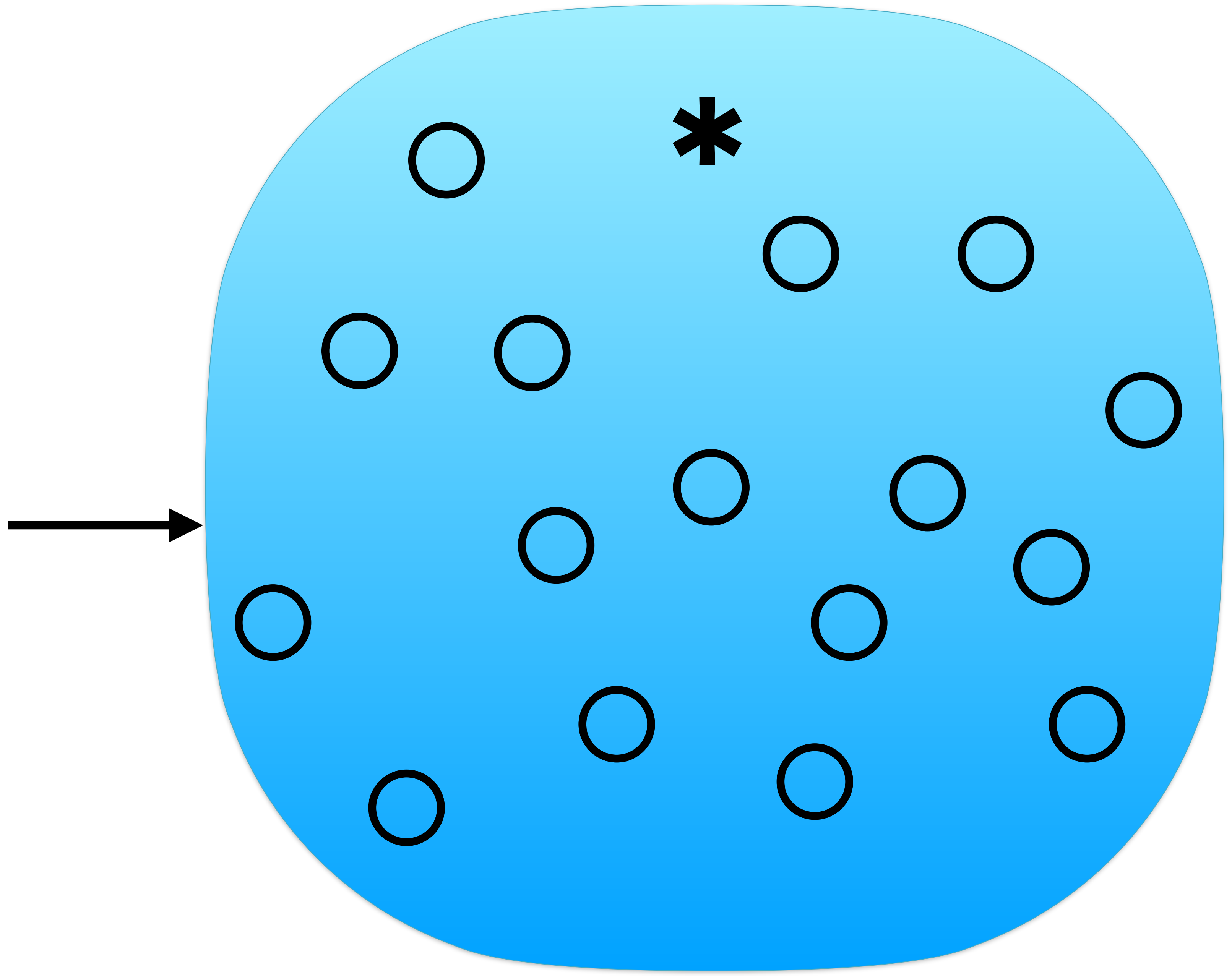
THE MAIN ISSUE

$$\max_{\pi} \; v_\pi^\gamma(s), \forall s$$

The discounted objective is not well-defined
for the problem setting of
continuing control with function approximation.

IN GENERAL, POLICIES ARE NOT COMPARABLE IN TERMS OF THE DISCOUNTED OBJECTIVE

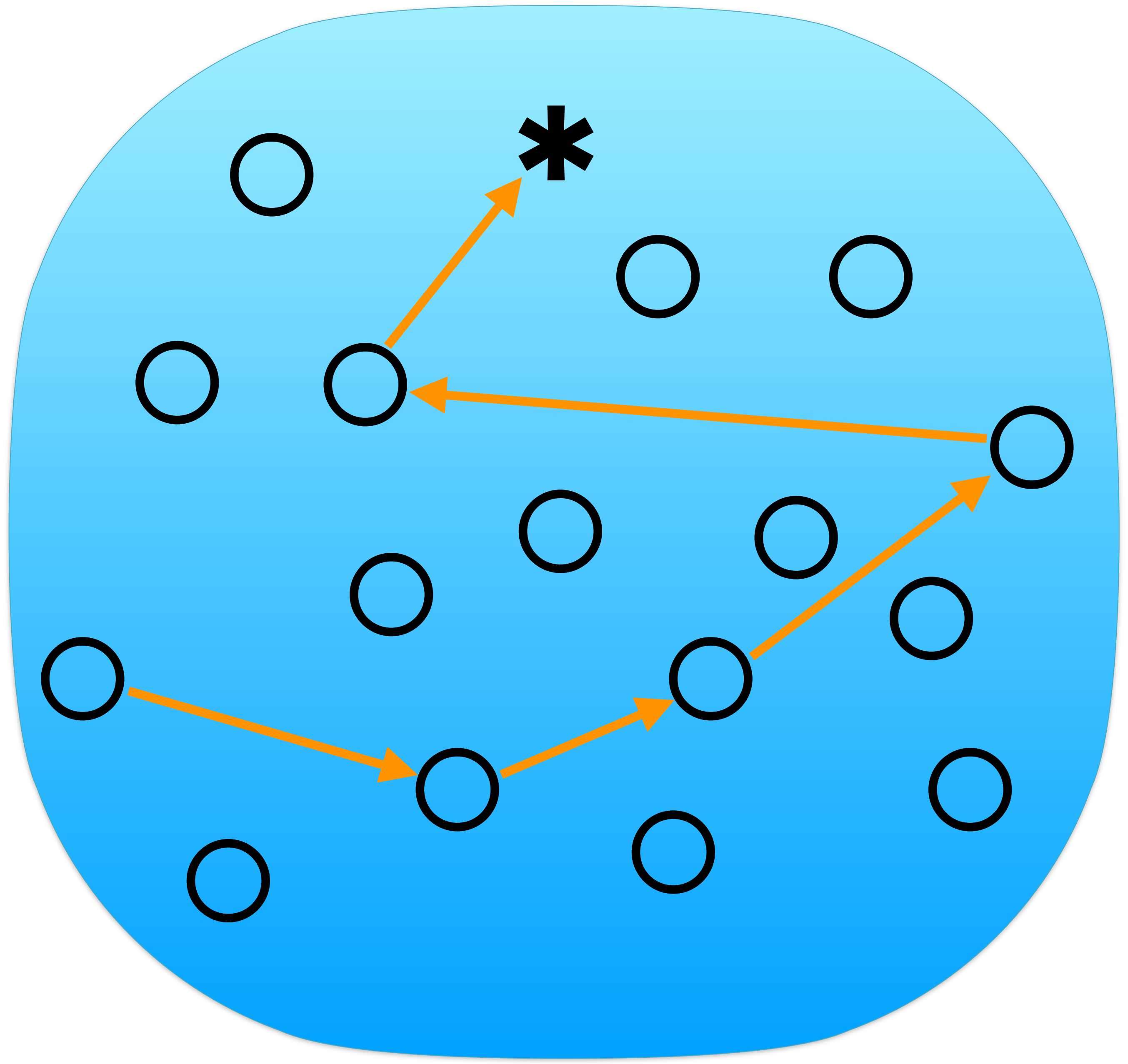
All policies



$$v_{\pi_a}(1) > v_{\pi_b}(1)$$
$$v_{\pi_a}(2) > v_{\pi_b}(2)$$
$$v_{\pi_a}(3) < v_{\pi_b}(3)$$
$$v_{\pi_a}(4) < v_{\pi_b}(4)$$

Which is better: π_a or π_b ?

IN THE TABULAR SETTING, THE POLICY IMPROVEMENT THEOREM HELPS

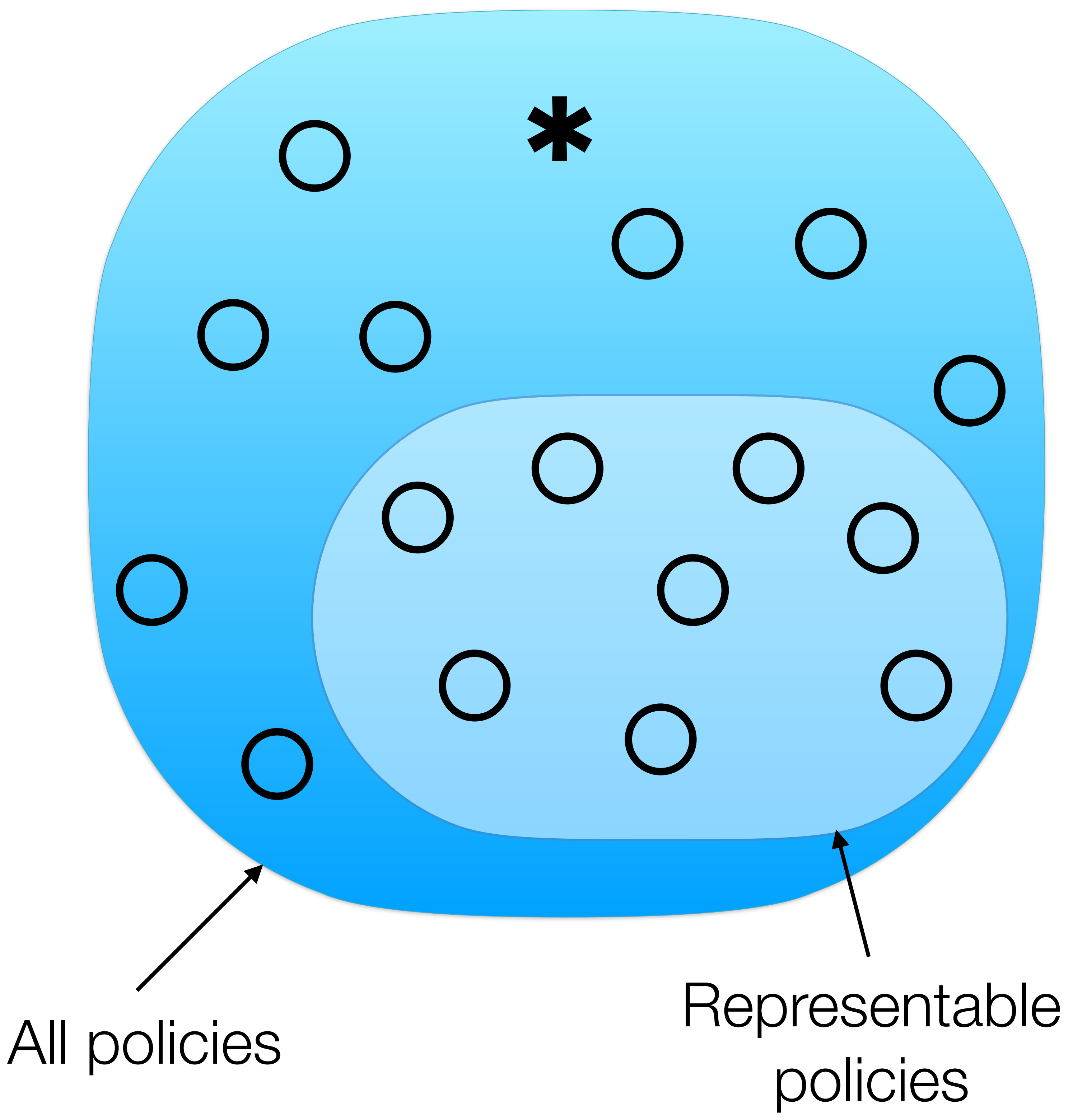


$\pi_0 \rightarrow \pi_1 \rightarrow \pi_2 \cdots\cdots\rightarrow \pi^*$

Start from any policy and eventually learn the optimal policy

The lack of comparability does not matter

WITH FUNCTION APPROXIMATION…



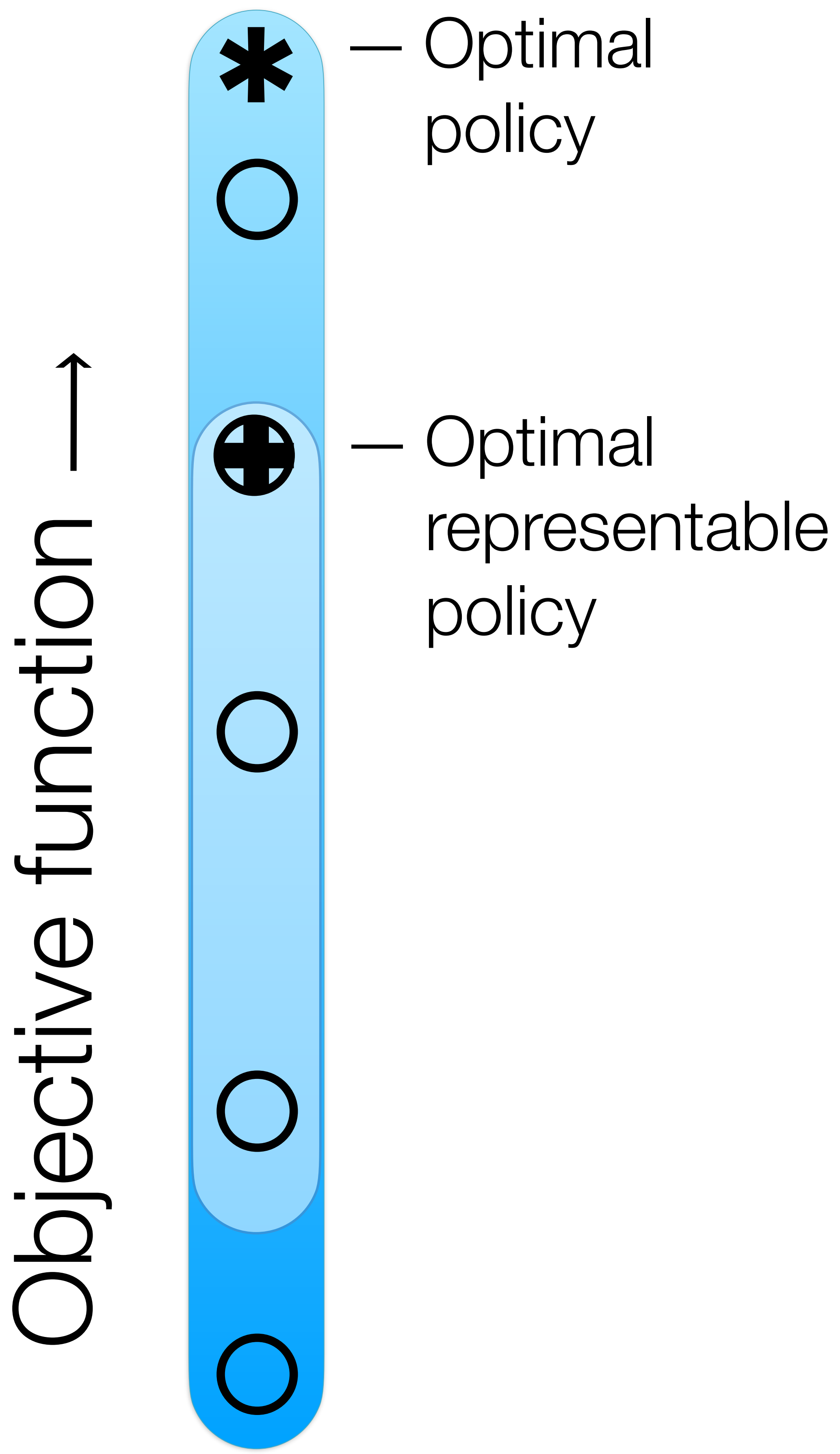
- The optimal/best policy is not representable under approximation.
- So we aim for the best representable policy.
- For that, we need to quantify the quality of a policy.

$v_{\pi_1}(1) > v_{\pi_2}(1)$
$v_{\pi_1}(2) > v_{\pi_2}(2)$
$v_{\pi_1}(3) < v_{\pi_2}(3)$
$v_{\pi_1}(4) < v_{\pi_2}(4)$

The standard optimality criterion in the discounted formulation does not rank-order policies.

Can we fix this issue?

RANKING POLICIES



- Can convert the vector to a scalar.

$$\left.\begin{matrix} v_\pi^\gamma(1) \\ v_\pi^\gamma(2) \\ v_\pi^\gamma(3) \\ v_\pi^\gamma(4) \end{matrix}\right\} \rightarrow J(\pi)$$

- What distributions can we use for averaging?
  - start-state distribution?
  - on-policy distribution?

ON-POLICY DISTRIBUTION OVER THE DISCOUNTED VALUE FUNCTION…

$$
\begin{aligned}
J(\pi) &= \sum_s \mu_\pi(s) v_\pi^\gamma(s) && \text{(where } v_\pi^\gamma \text{ is the discounted value function)} \\
&= \sum_s \mu_\pi(s) \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a) \left[r + \gamma v_\pi^\gamma(s')\right] && \text{(Bellman Eq.)} \\
&= r(\pi) + \sum_s \mu_\pi(s) \sum_a \pi(a|s) \sum_{s'} \sum_r p(s', r|s, a) \gamma v_\pi^\gamma(s') && \text{(from (10.7))} \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma(s') \sum_s \mu_\pi(s) \sum_a \pi(a|s) p(s'|s, a) && \text{(from (3.4))} \\
&= r(\pi) + \gamma \sum_{s'} v_\pi^\gamma(s') \mu_\pi(s') && \text{(from (10.8))} \\
&= r(\pi) + \gamma J(\pi) \\
&= r(\pi) + \gamma r(\pi) + \gamma^2 J(\pi) \\
&= r(\pi) + \gamma r(\pi) + \gamma^2 r(\pi) + \gamma^3 r(\pi) + \cdots \\
&= \frac{1}{1-\gamma} r(\pi).
\end{aligned}
$$

Section 10.4, Sutton & Barto (2018)

… is equivalent to the average-reward objective!

THE PROBLEM SPECIFICATION DOES NOT INVOLVE GAMMA

$$J(\pi) = \sum_s \mu_\pi(s)\, v_\pi^\gamma(s) = \frac{r(\pi)}{1-\gamma}$$

$$r(\pi_1) > r(\pi_2) \implies J(\pi_1) > J(\pi_2) \quad \forall \gamma$$

that is, γ does not play a role in the problem definition.

RECALL:
DIFFERENCE BETWEEN PROBLEM AND SOLUTION METHODS

Problem

Find a policy that maximizes total reward

$$\max_{\pi} \sum_{t}^{\infty} R_t$$

Maximize the discounted sum of rewards *from each state*

$$\max_{\pi} \; v_{\pi}^{\gamma}(s), \forall s$$

Maximize the discounted sum of rewards *averaged over each state* ≡ Maximize the average reward

$$\sum_{s} \mu_{\pi}(s) \, v_{\pi}^{\gamma}(s) = \frac{r(\pi)}{1-\gamma}$$

$$r(\pi)$$

Solution methods

Q-learning, Sarsa, …

Differential Q-learning, Differential Sarsa, …

TAKEAWAYS SO FAR

- “Continuing control with function approximation” is an important problem setting for AI.
- The policy-improvement theorem does not hold with function approximation.
- As a result, the standard discounted objective is not well-defined in this problem setting.

- The on-policy average of the discounted value function is sensible way to rank-order policies. It is equivalent to the average-reward objective.

Naik, Shariff, Yasui, Yao, Sutton (2019). *Discounted RL is not an Optimization Problem.* OptRL workshop at NeurIPS.

OUTLINE

0. Problem setting
1. The discounted-reward formulation
2. The main issue with discounting
3. The average-reward formulation
4. Connections: improving discounted methods using average reward

THE AVERAGE-REWARD FORMULATION

$$\max_{\pi} \sum_{t}^{\infty} R_t$$

$$S_0 \; A_0 \; R_1 \; S_1 \; A_1, R_2 \ldots \; S_t \; A_t \; R_{t+1} \; S_{t+1} \; A_{t+1} \; R_{t+2} \; \ldots$$

Average Reward →

$$r(\pi) \doteq \lim_{n\to\infty} \frac{1}{n} \mathbb{E}_{\pi}\Big[\sum_{t=1}^{n} R_t\Big]$$

Differential value function →

$$\tilde{v}_{\pi}(s) \doteq \mathbb{E}_{\pi}[R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + \ldots \mid S_t = s]$$

How is this finite?

$$v_{\pi}^{\gamma}(s) \doteq \mathbb{E}_{\pi}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \mid S_t = s]$$

IF THE REWARDS ARE BOUNDED, THE AVERAGE REWARD IS FINITE

$$S_0 \; A_0 \; R_1 \; S_1 \; A_1, R_2 \dots \; S_t \; A_t \; R_{t+1} \; S_{t+1} \; A_{t+1} \; R_{t+2} \; \dots$$

$$|R_i| < k \in \mathbb{R}^+$$

$$\mathbb{E}[R_i] < k$$

$$\mathbb{E}[\sum_{i=1}^{n} R_i] < nk$$

$$\mathbb{E}[A + B] = \mathbb{E}[A] + \mathbb{E}[B]$$

$$\lim_{n \to \infty} \mathbb{E}[\sum_{i=1}^{n} R_i] \to \infty$$

$$\implies \frac{1}{n} \mathbb{E}[\sum_{i=1}^{n} R_i] < k$$

i.e., the average reward is finite

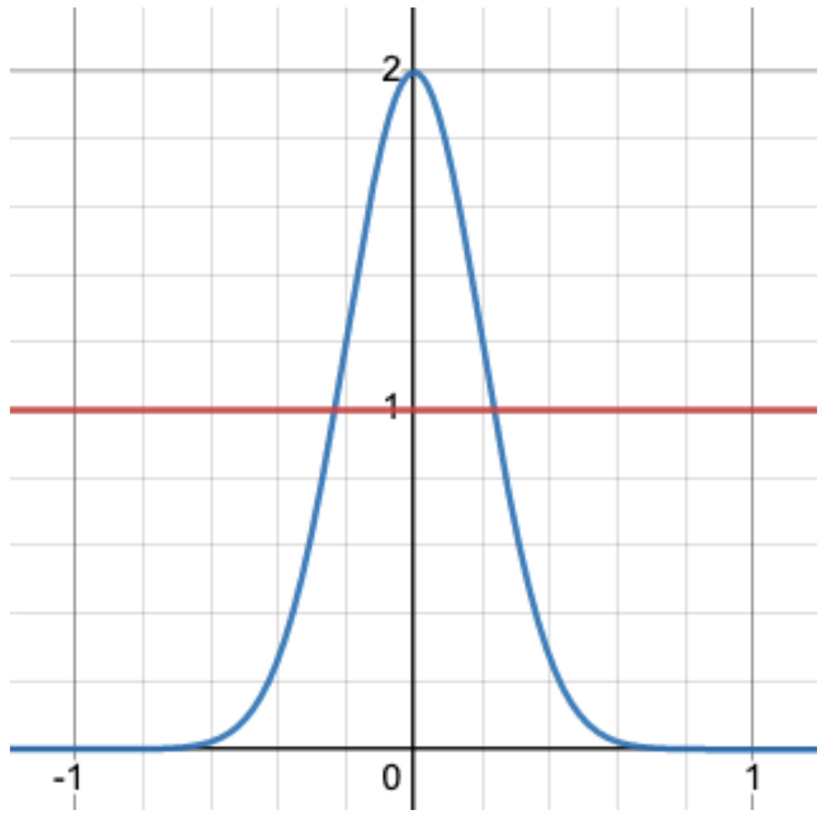
$S_0\ A_0\ R_1\ S_1\ A_1, R_2 \ldots\ S_t\ A_t\ R_{t+1}\ S_{t+1}\ A_{t+1}\ R_{t+2}\ \ldots$

$|R_i| < k \in \mathbb{R}^+$

If $R_i \sim U(-k, k)$

$$\mathbb{E}[R_i] = 0$$

$$\mathbb{E}\left[\sum_{i=1}^{n} R_i\right] = 0$$



If $R_i \sim N(0, \sigma^2)$

$$\mathbb{E}[R_i] = 0$$

$$\mathbb{E}\left[\sum_{i=1}^{n} R_i\right] = 0$$

If all the random variables have zero mean,
then the sum of the random variables also has zero mean.

THE DIFFERENTIAL VALUE FUNCTION IS FINITE

$$S_0\ A_0\ R_1\ S_1\ A_1, R_2 \dots\ S_t\ A_t\ R_{t+1}\ S_{t+1}\ A_{t+1}\ R_{t+2}\ \dots$$

$$|R_i| < k \in \mathbb{R}^+$$

$$\mathbb{E}[R_i] = \bar{r}_i$$

$$\mathbb{E}[R_i] - \bar{r}_i = 0$$

$$\mathbb{E}[R_i - \bar{r}_i] = 0$$

$$\mathbb{E}\left[\sum_i (R_i - \bar{r}_i)\right] = 0$$

$$\bar{r}_i = \bar{r} \quad \forall\, i$$

under the assumption of ergodicity

$$\mu(s) \doteq \lim_{t\to\infty} Pr(S_t = s \,|\, A_{0:t-1} \sim \pi)$$ exists

$$\sum_s \mu(s) \sum_a \pi(a\,|\,s) \sum_{s'} p(s'\,|\,s, a) = \mu(s')$$

ESTIMATING THE AVERAGE REWARD FROM DATA

$R_1 \quad R_2 \quad R_3 \quad \dots \quad R_{t-1} \quad R_t \quad R_{t+1} \quad \dots$

$$\bar{R}_t \doteq \frac{1}{t}\sum_{i=1}^{t} R_i$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \frac{1}{t+1}\left(R_{t+1} - \bar{R}_t\right)$$

Off-policy?

$\bar{R}_\infty \to r(\pi)$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \beta_t (R_{t+1} - \bar{R}_t)$$

$\bar{R}_\infty \to r(b)$

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

$$r(\pi) = \sum_s \mu_\pi(s) \sum_a \pi(a \mid s) \sum_r p(r \mid s, a)\ r$$

$$r(b) = \sum_s \mu_b(s) \sum_a b(a \mid s) \sum_r p(r \mid s, a)\ r$$

With $\rho_t \doteq \dfrac{\pi(A_t \mid S_t)}{b(A_t \mid S_t)}$

$\bar{R}_\infty \nrightarrow r(b)$

$\bar{R}_\infty \nrightarrow r(\pi)$

If $\bar{R}_{t+1} \doteq \bar{R}_t + \beta_t \delta_t$ then $\bar{R}_\infty \to r(\pi)$

ESTIMATING THE VALUES FROM DATA

$$S_0\ A_0\ R_1\ S_1\ A_1, R_2 \ldots\ S_t\ A_t\ R_{t+1}\ S_{t+1}\ A_{t+1}\ R_{t+2}\ \ldots$$

$$q^\gamma_\pi(s,a) \doteq \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \mid S_t = s, A_t = a]$$

$$q^\gamma_*(s,a) = \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma \max_{a'} q^\gamma_*(s',a')\right]$$

$$\tilde{q}_\pi(s,a) \doteq \mathbb{E}_\pi[R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + \ldots \mid S_t = s, A_t = a]$$

$$\tilde{q}_*(s,a) = \sum_{s',r} p(s',r \mid s,a)\left[r - \bar{r} + \max_{a'} \tilde{q}_*(s',a')\right]$$

Discounted Q-learning

$$Q_{t+1}(S_t,A_t) \doteq Q_t(S_t,A_t) + \alpha_t \underbrace{\left[R_{t+1} + \gamma \max_{a'} Q_t(S_{t+1},a') - Q_t(S_t,A_t)\right]}_{\delta^\gamma_t}$$

Differential Q-learning

$$Q_{t+1}(S_t,A_t) \doteq Q_t(S_t,A_t) + \alpha_t \underbrace{\left[R_{t+1} - \bar{R}_t + \max_{a'} Q_t(S_{t+1},a') - Q_t(S_t,A_t)\right]}_{\delta_t}$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \beta_t \delta_t$$

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

ESTIMATING THE AVERAGE REWARD FROM DATA

$S_0\ A_0\ R_1\ S_1\ A_1, R_2 \dots\ S_t\ A_t\ R_{t+1}\ S_{t+1}\ A_{t+1}\ R_{t+2}\ \dots$

Differential Q-learning

$$Q_{t+1}(S_t, A_t) \doteq Q_t(S_t, A_t) + \alpha_t \underbrace{\left[R_{t+1} - \bar{R}_t + \max_{a'} Q_t(S_{t+1}, a') - Q_t(S_t, A_t)\right]}_{\delta_t}$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \beta_t \delta_t$$

$$\tilde{q}_*(s, a) = \sum_{s', r} p(s', r \mid s, a)\left[r - \bar{r} + \max_{a'} \tilde{q}_*(s', a')\right]$$

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

ESTIMATING THE AVERAGE REWARD FROM DATA

$$S_0\ A_0\ R_1\ S_1\ A_1, R_2 \ldots\ S_t\ A_t\ R_{t+1}\ S_{t+1}\ A_{t+1}\ R_{t+2}\ \ldots$$

Differential Q-learning

$$Q_{t+1}(S_t, A_t) \doteq Q_t(S_t, A_t) + \alpha_t \big[\underbrace{R_{t+1} - \bar{R}_t + \max_{a'} Q_t(S_{t+1}, a') - Q_t(S_t, A_t)}_{\delta_t}\big]$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \beta_t \delta_t$$

$$\tilde{q}_*(s, a) = \sum_{s', r} p(s', r \mid s, a)\big[r + \max_{a'} \tilde{q}_*(s', a')\big] - \bar{r}$$

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

ESTIMATING THE AVERAGE REWARD FROM DATA

$$S_0 \; A_0 \; R_1 \; S_1 \; A_1, R_2 \ldots \; S_t \; A_t \; R_{t+1} \; S_{t+1} \; A_{t+1} \; R_{t+2} \; \ldots$$

Differential Q-learning

$$Q_{t+1}(S_t, A_t) \doteq Q_t(S_t, A_t) + \alpha_t \underbrace{\left[R_{t+1} - \bar{R}_t + \max_{a'} Q_t(S_{t+1}, a') - Q_t(S_t, A_t)\right]}_{\delta_t}$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \beta_t \delta_t$$

$$\bar{r} = \sum_{s',r} p(s', r \mid s, a)\left[r + \max_{a'} \tilde{q}_*(s', a')\right] - \tilde{q}_*(s, a)$$

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

ESTIMATING THE AVERAGE REWARD FROM DATA

$$S_0\ A_0\ R_1\ S_1\ A_1, R_2 \ldots\ S_t\ A_t\ R_{t+1}\ S_{t+1}\ A_{t+1}\ R_{t+2}\ \ldots$$

Differential Q-learning

$$Q_{t+1}(S_t, A_t) \doteq Q_t(S_t, A_t) + \alpha_t \underbrace{\left[R_{t+1} - \bar{R}_t + \max_{a'} Q_t(S_{t+1}, a') - Q_t(S_t, A_t)\right]}_{\delta_t}$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \beta_t \delta_t$$

$$\bar{r} = \sum_{s',r} p(s', r \mid s, a)\left[r + \max_{a'} \tilde{q}_*(s', a') - \tilde{q}_*(s, a)\right]$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \beta_t \underbrace{(R_{t+1} + \max_{a'} Q_t(S_{t+1}, a') - Q_t(S_t, A_t) - \bar{R}_t)}_{\delta_t}$$

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

THE TWO ALGORITHMS LOOK QUITE SIMILAR

$$S_0\ A_0\ R_1\ S_1\ A_1, R_2 \dots\ S_t\ A_t\ R_{t+1}\ S_{t+1}\ A_{t+1}\ R_{t+2}\ \dots$$

Differential Q-learning

$$Q_{t+1}(S_t, A_t) \doteq Q_t(S_t, A_t) + \alpha_t \underbrace{\left[R_{t+1} - \bar{R}_t + \max_{a'} Q_t(S_{t+1}, a') - Q_t(S_t, A_t)\right]}_{\delta_t}$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \beta_t \delta_t$$

Discounted Q-learning

$$Q_{t+1}(S_t, A_t) \doteq Q_t(S_t, A_t) + \alpha_t \underbrace{\left[R_{t+1} + \gamma \max_{a'} Q_t(S_{t+1}, a') - Q_t(S_t, A_t)\right]}_{\delta_t^\gamma}$$

The algorithms are very similar implementation-wise;
the theoretical analysis is significantly different

Wan*, Naik*, & Sutton. (2021). *Learning and Planning in Average-Reward Markov Decision Processes*. ICML.

ADVANCED ALGORITHMS

- Hierarchical learning via options
  - Differential intra-option, inter-option, interruption algorithms.
  - Proved to converge in the tabular setting.

Wan, Naik, Sutton (2021). *Average-Reward Learning and Planning with Options*. NeurIPS.

- More efficient learning algorithms
  - Multi-step TD(λ)-style algorithms with eligibility traces.
  - Proved to converge with linear function approximation.

Naik & Sutton (2022). *Multi-Step Average-Reward Prediction via Differential TD(λ).* RLDM.
Naik (2024). *Reinforcement Learning in Continuing Problems using Average Reward.* Ph.D. dissertation.

OUTLINE

0. Problem setting
1. The discounted-reward formulation
2. The main issue with discounting
3. The average-reward formulation
4. Connections: improving discounted methods using average reward

THE MAIN MESSAGE

The performance of standard discounted-reward methods
such as TD-learning or Q-learning
can be significantly improved
by estimating the average reward
and subtracting it from the observed rewards.

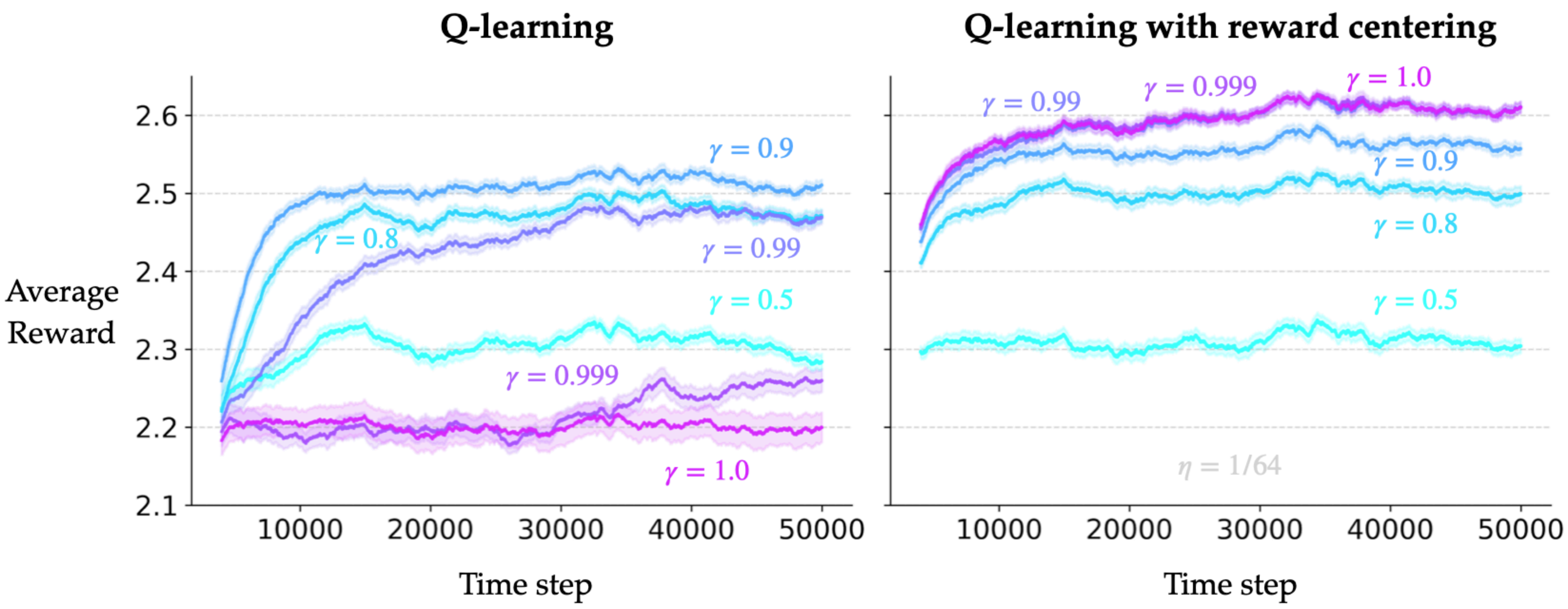
$$S_0 \; A_0 \; R_1 \; S_1 \; A_1, R_2 \ldots \; S_t \; A_t \; R_{t+1} \; S_{t+1} \; A_{t+1} \; R_{t+2} \; \ldots$$

$$Q_{t+1}(S_t, A_t) \doteq Q_t(S_t, A_t) + \alpha_t \big[R_{t+1} + \gamma \max_{a'} Q_t(S_{t+1}, a') - Q_t(S_t, A_t) \big]$$

↓

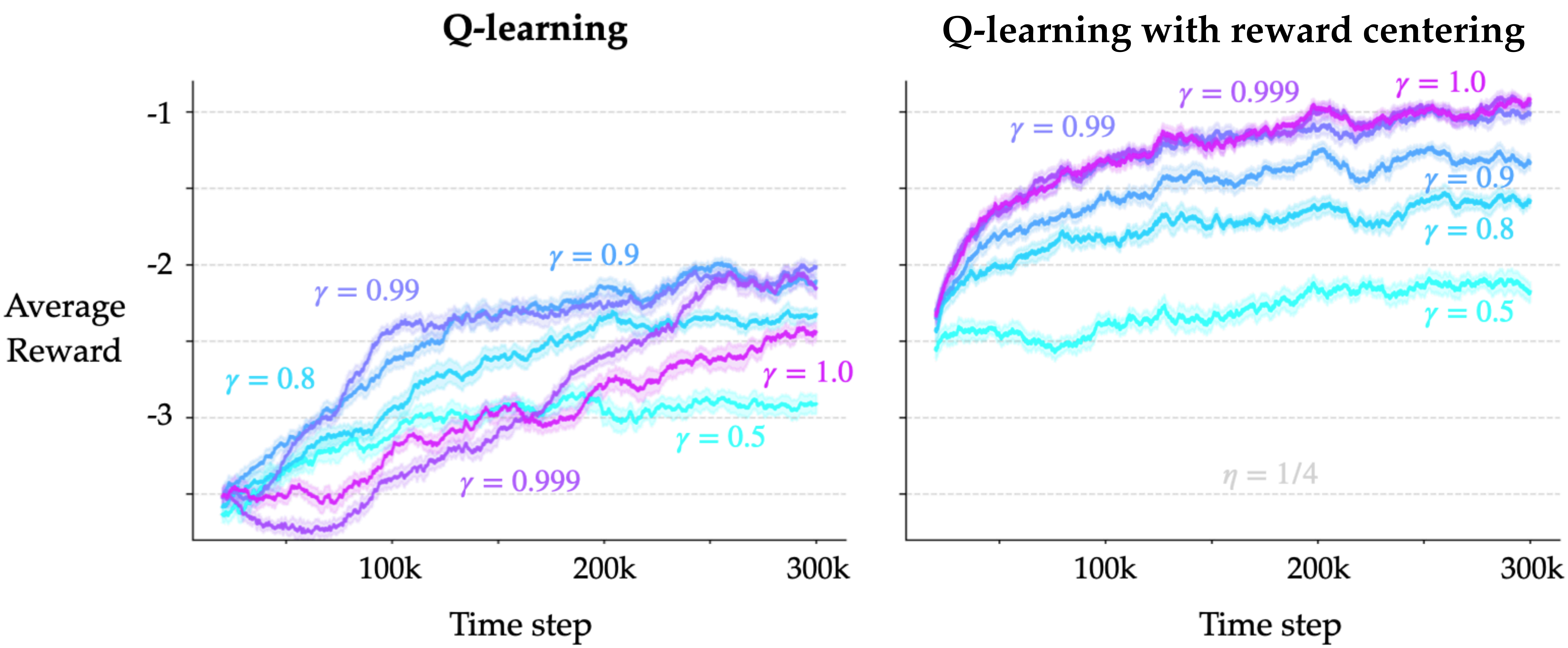
$$Q_{t+1}(S_t, A_t) \doteq Q_t(S_t, A_t) + \alpha_t \big[R_{t+1} - \bar{R}_t + \gamma \max_{a'} Q_t(S_{t+1}, a') - Q_t(S_t, A_t) \big]$$

NO INSTABILITY WITH LARGE DISCOUNT FACTORS



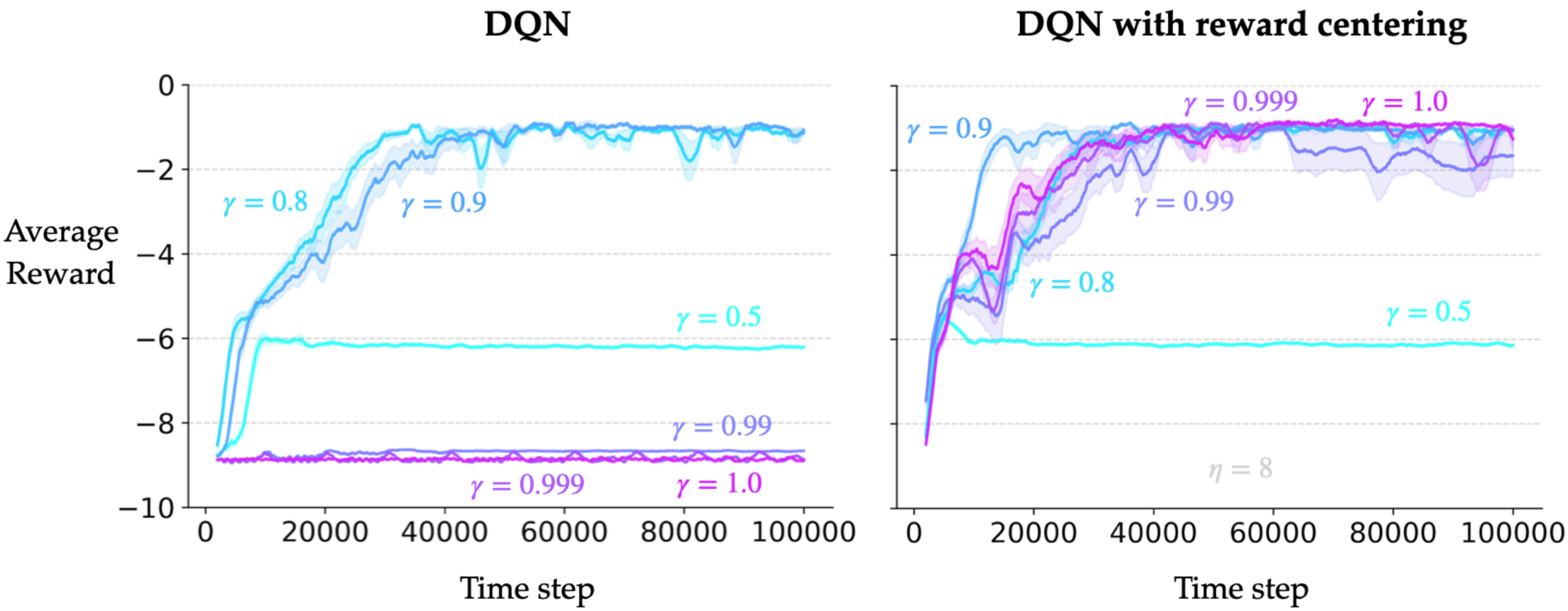
AccessControl (tabular)

NO INSTABILITY WITH LARGE DISCOUNT FACTORS



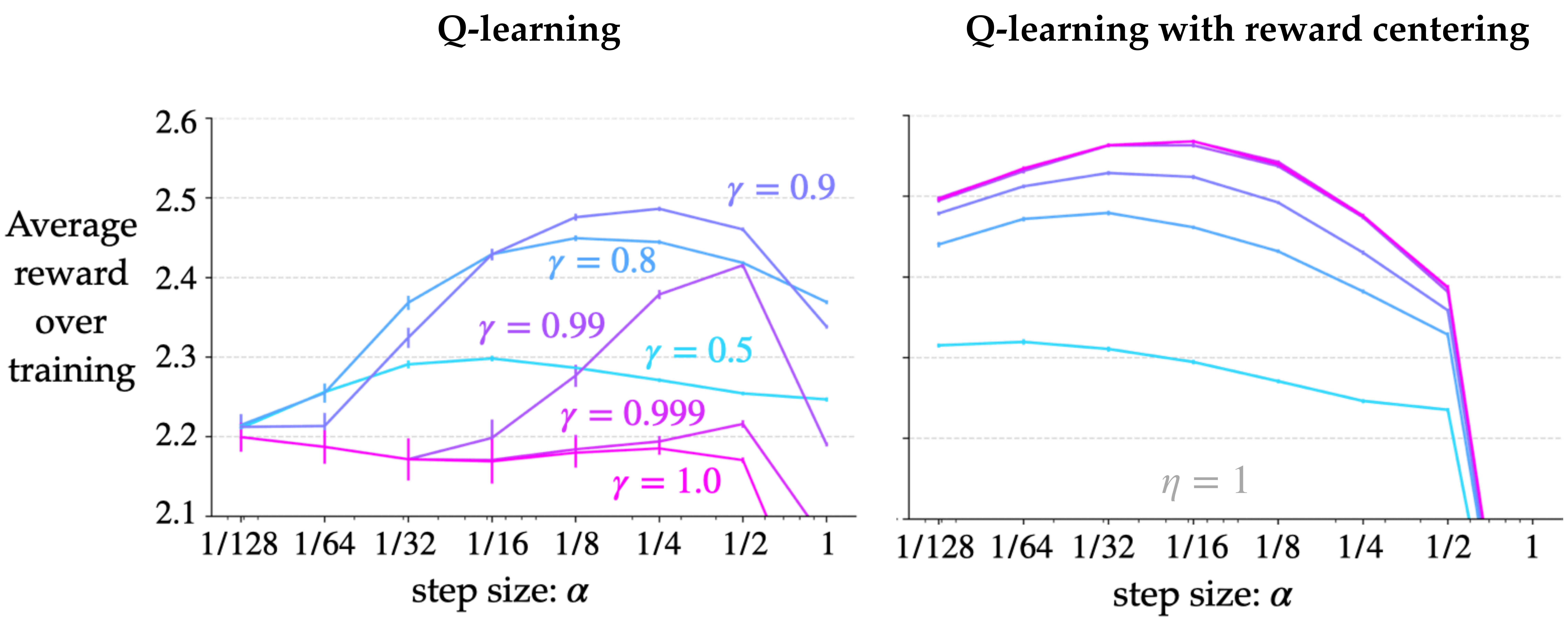
PuckWorld (linear FA)

NO INSTABILITY WITH LARGE DISCOUNT FACTORS



Pendulum (non-linear FA)

TRENDS ARE CONSISTENT ACROSS PARAMETERS



AccessControl (tabular)

UNDERLYING THEORY

$$v_\pi^\gamma(s) = \frac{r(\pi)}{1-\gamma} + \tilde{v}_\pi(s) + e_\pi^\gamma(s)$$

$$R_{t+1} \quad R_{t+2} \quad R_{t+3} \quad \dots \quad R_{t+n} \quad \dots$$

Standard discounted value function

$$v_\pi^\gamma(s) \doteq \mathbb{E}_\pi\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \mid S_t = s\right] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\right]$$

Average reward

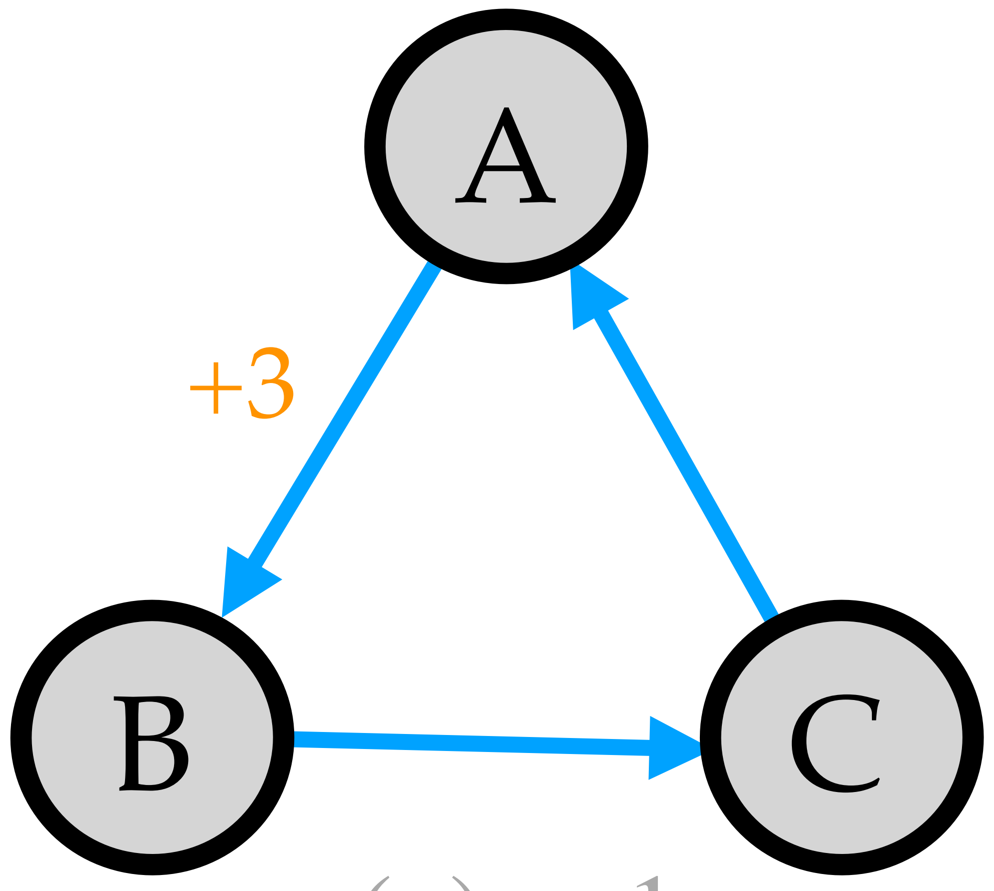
$$r(\pi) \doteq \lim_{n\to\infty} \frac{1}{n} \mathbb{E}_\pi\left[\sum_{t=1}^{n} R_t\right]$$

Differential value function

$$\tilde{v}_\pi(s) \doteq \mathbb{E}_\pi[R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + \dots \mid S_t = s]$$

INTUITION THROUGH AN EXAMPLE

$$v_\pi^\gamma(s) = \frac{r(\pi)}{1-\gamma} + \tilde{v}_\pi(s) + e_\pi^\gamma(s)$$



		s_A	s_B	s_C	$\frac{r(\pi)}{1-\gamma}$
Standard discounted values	$\gamma = 0.8$	6.15	3.93	4.92	5
	$\gamma = 0.9$	11.07	8.97	9.96	10
	$\gamma = 0.99$	101.01	98.99	99.99	100
	$\gamma = 0.8$	1.15	-1.07	-0.08	
	$\gamma = 0.9$	1.07	-1.03	-0.04	
	$\gamma = 0.99$	1.01	-1.01	-0.01	
Differential values		1	-1	0	

Centered discounted value function

$$\tilde{v}_\pi^\gamma(s) \doteq \mathbb{E}_\pi\Big[\sum_{k=0}^{\infty} \gamma^k \big(R_{t+k+1} - r(\pi)\big) \mid S_t = s\Big] = v_\pi^\gamma(s) - \frac{r(\pi)}{1-\gamma}$$

ESTIMATING $r(\pi)$

$$S_0\ A_0\ R_1\ S_1\ A_1, R_2 \ldots\ S_t\ A_t\ R_{t+1}\ S_{t+1}\ A_{t+1}\ R_{t+2}\ \ldots$$

On-policy $$\bar{R}_{t+1} \doteq \bar{R}_t + \beta_t (R_{t+1} - \bar{R}_t)$$

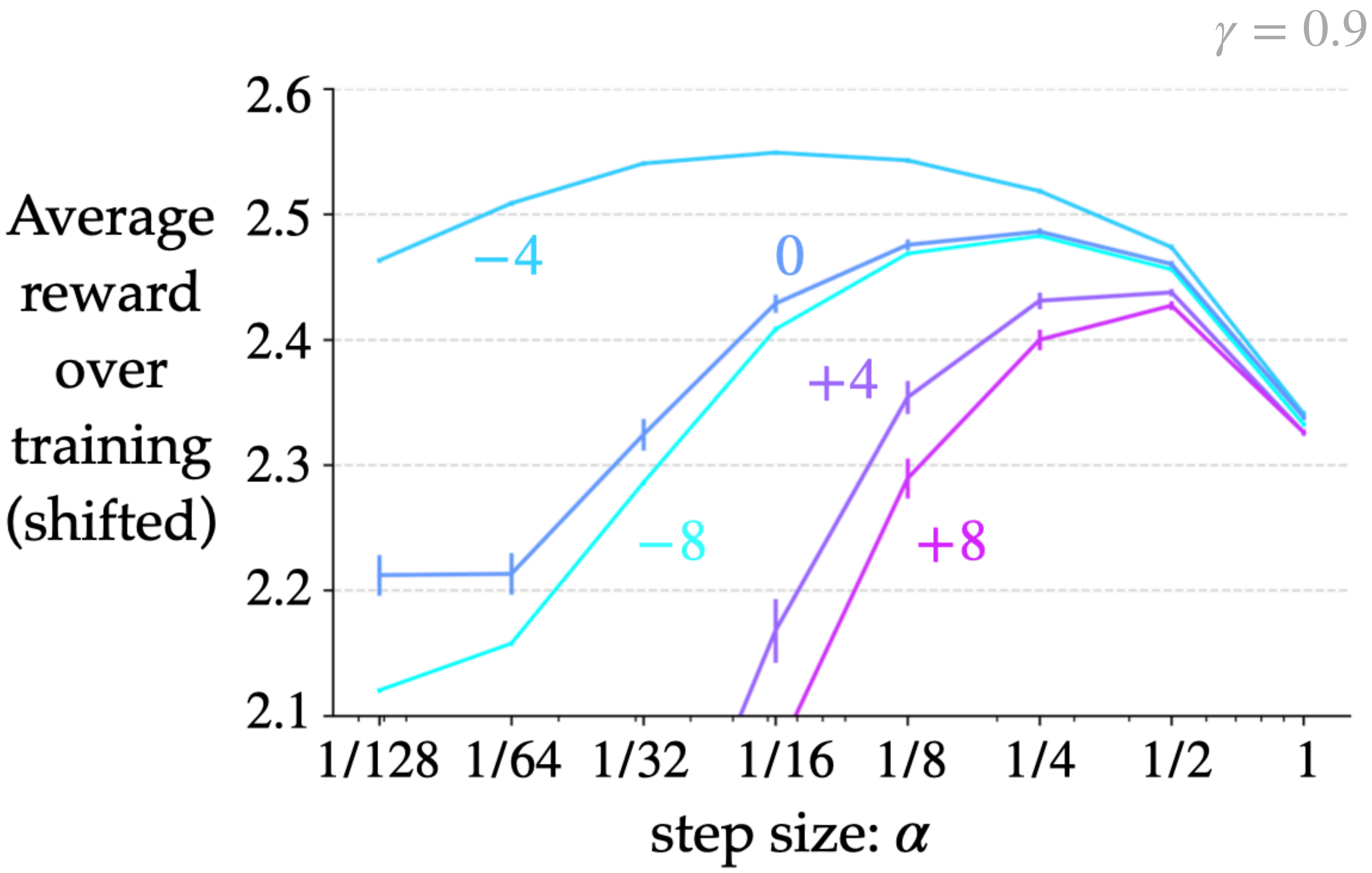
Off-policy $$\bar{R}_{t+1} \doteq \bar{R}_t + \beta_t \delta_t$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \gamma V_t(S_{t+1}) - V_t(S_t)$

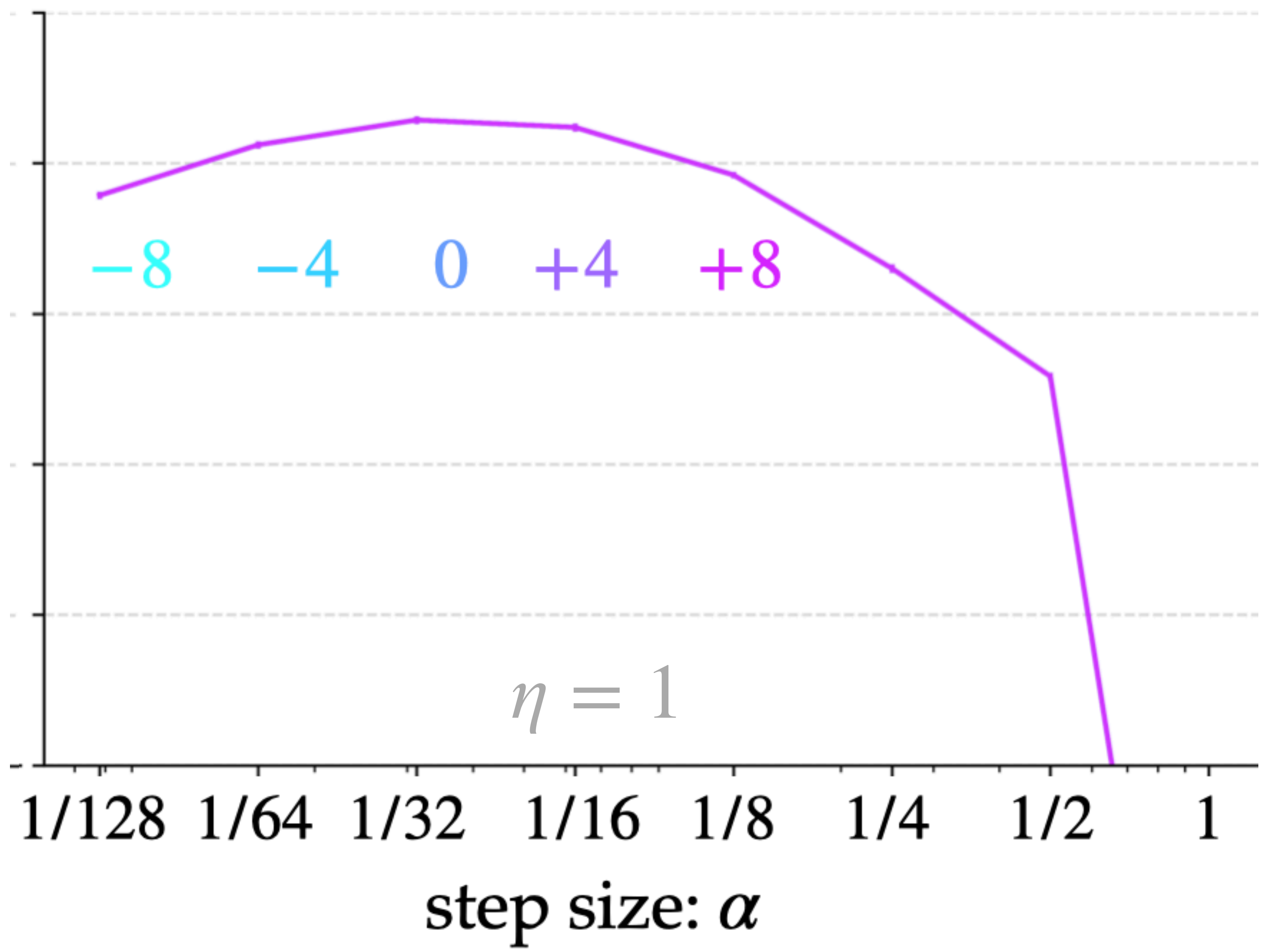
MORE ROBUST TO SHIFTED REWARDS

$$v_\pi^\gamma(s) = \frac{r(\pi)}{1-\gamma} + \tilde{v}_\pi(s) + e_\pi^\gamma(s)$$

Q-learning



Q-learning with reward centering



AccessControl (tabular)

TAKEAWAYS

- Reward centering can improve the performance of discounted methods for all discount factors, especially as $\gamma \to 1$.
- Reward centering can also make discounted methods robust to shifts in the problems' rewards.
- Both techniques of centering are quite effective; using the TD error is more appropriate for the off-policy setting.

- Additional non-stationarity; step-size adaptation would help!
- Should be combined with techniques for reward *scaling*
- Unlocks algorithms in which the discount factor can be efficiently adapted over time

Every RL algorithm will benefit with reward centering!

Analysis, more experiments, etc.:

Naik, Wan, Tomar, & Sutton. (2024). *Reward Centering*. Reinforcement Learning Conference.

https://arxiv.org/abs/2405.09999

OUTLINE

0. Continuing problems
1. The discounted-reward formulation
2. The main issue with discounting
3. The average-reward formulation
4. Connections: improving discounted methods using average reward

THANK YOU

Questions?

abhisheknaik96.github.io
abhisheknaik22296@gmail.com