REINFORCEMENT LEARNING FOR CONTINUING TASKS

29 October 2026

Carleton-NTN Lab Seminar

Abhishek Naik

abhisheknaik22296@gmail.com



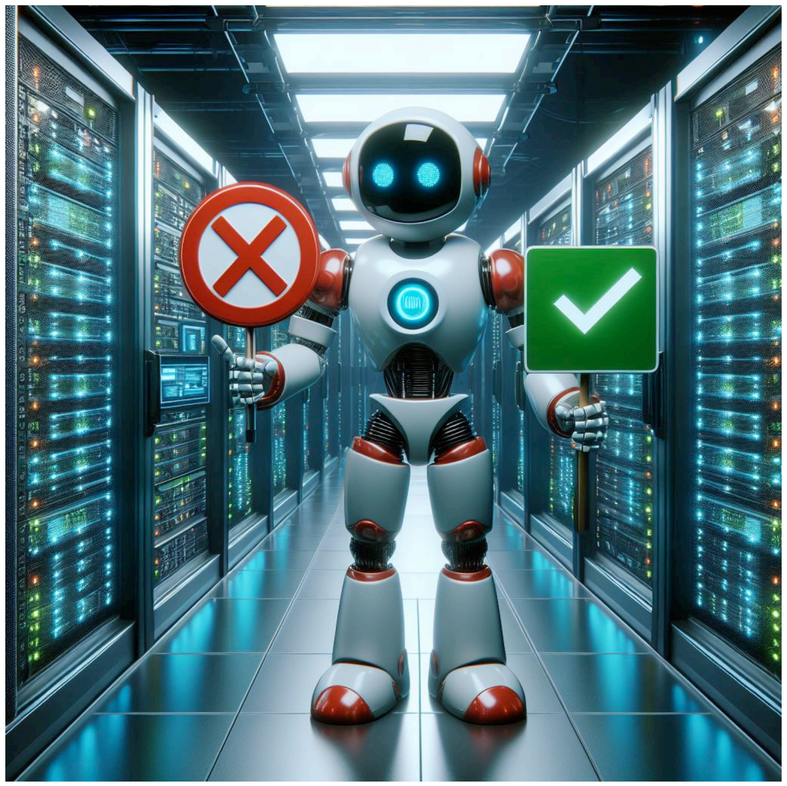
OUTLINE

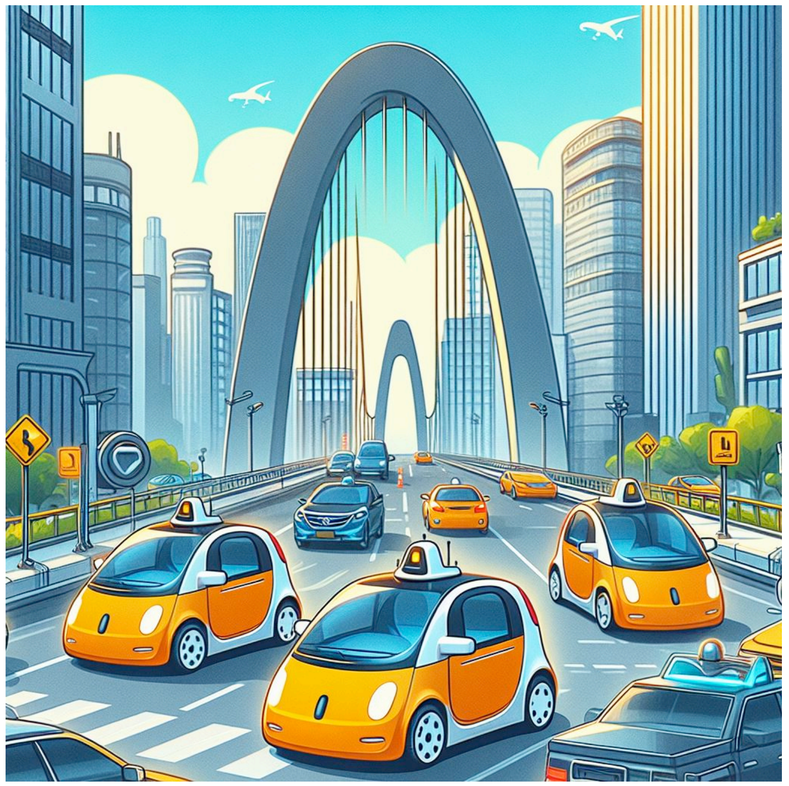
0. Problem setting: what are continuing tasks?
1. The discounted-reward formulation
2. The average-reward formulation
3. Improving discounted methods using average reward
   (if time permits)

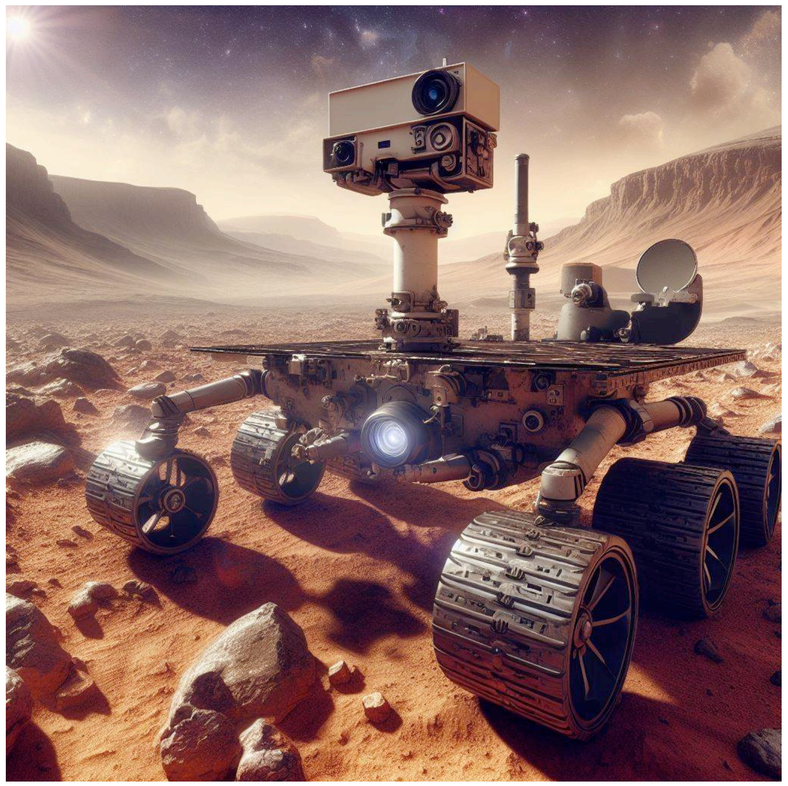
PROBLEM SETTING

CONTINUING PROBLEMS



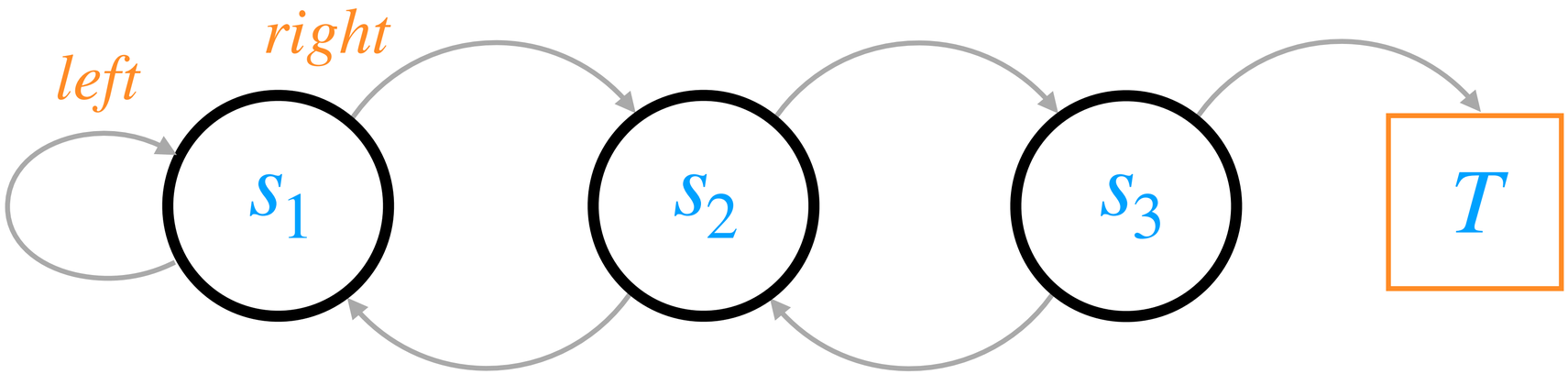


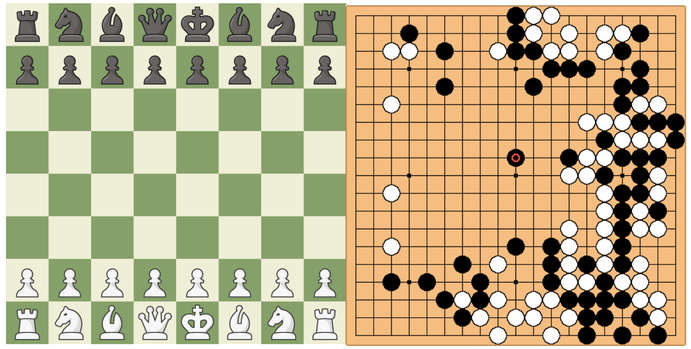




Images generated using DALL·E 3

RECALL, EPISODIC PROBLEMS





TIME SPANS OF DECISIONS' CONSEQUENCES ARE BOUNDED IN EPISODIC PROBLEMS

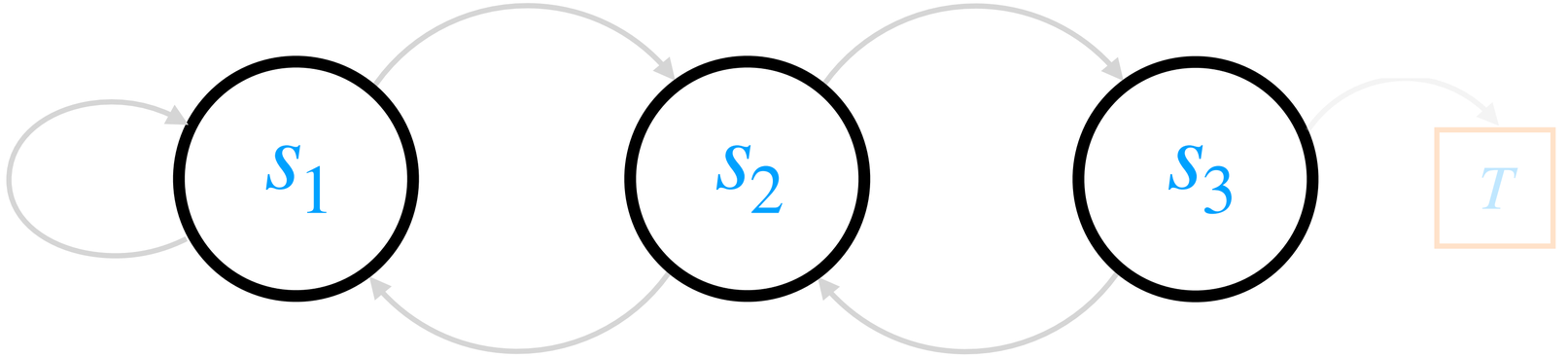
$$\ldots \quad S_{t-k} \quad \ldots \quad S_{t-1} \; A_{t-1} \; R_t \quad S_t \; A_t \; R_{t+1} \; S_{t+1} \; A_{t+1} \quad \ldots \quad S_{t+n} \quad \ldots$$

And no credit assignment occurs across episodic boundaries.



Typically, no 'resets' in life…

CONTINUING PROBLEMS



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

ASIDE: IMPORTANT DISTINCTIONS WITH SIMILAR-SOUNDING TERMS

- Continual / never-ending / lifelong learning:

  emphasizes a learning agent's *continual* need to adapt to a big or 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

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

$\gamma \in [0,1)$ $\quad R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots$

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

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

OUTLINE

0. Problem setting: what are continuing tasks?
1. The discounted-reward formulation
2. The average-reward formulation
3. Improving discounted methods using average reward
   (if time permits)

DISCOUNTED-REWARD FORMULATION

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

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

$$\pi^*_\gamma \rightarrow \max_{\pi}\ v^\gamma_\pi(s), \forall s \qquad \gamma \in [0,1)$$

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

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

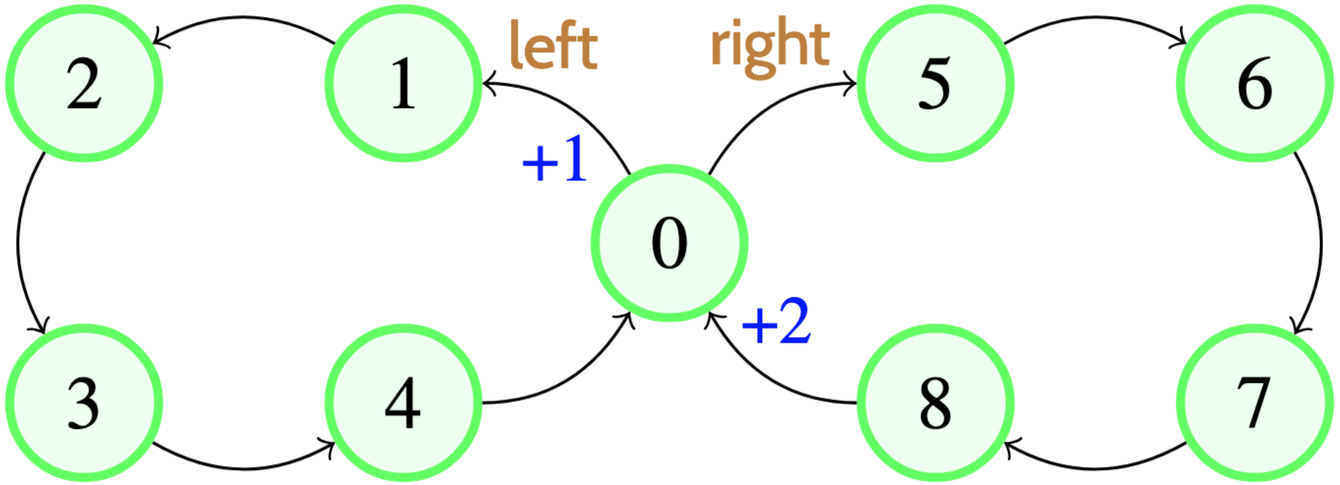
$$\pi^*_\gamma(s) = \arg\max_a q^\gamma_{\pi^*_\gamma}(s,a)$$

A PROTOTYPICAL ALGORITHM

Q-learning

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

THE BEST POLICY
DEPENDS ON THE DISCOUNT FACTOR



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

A USEFUL THEOREM

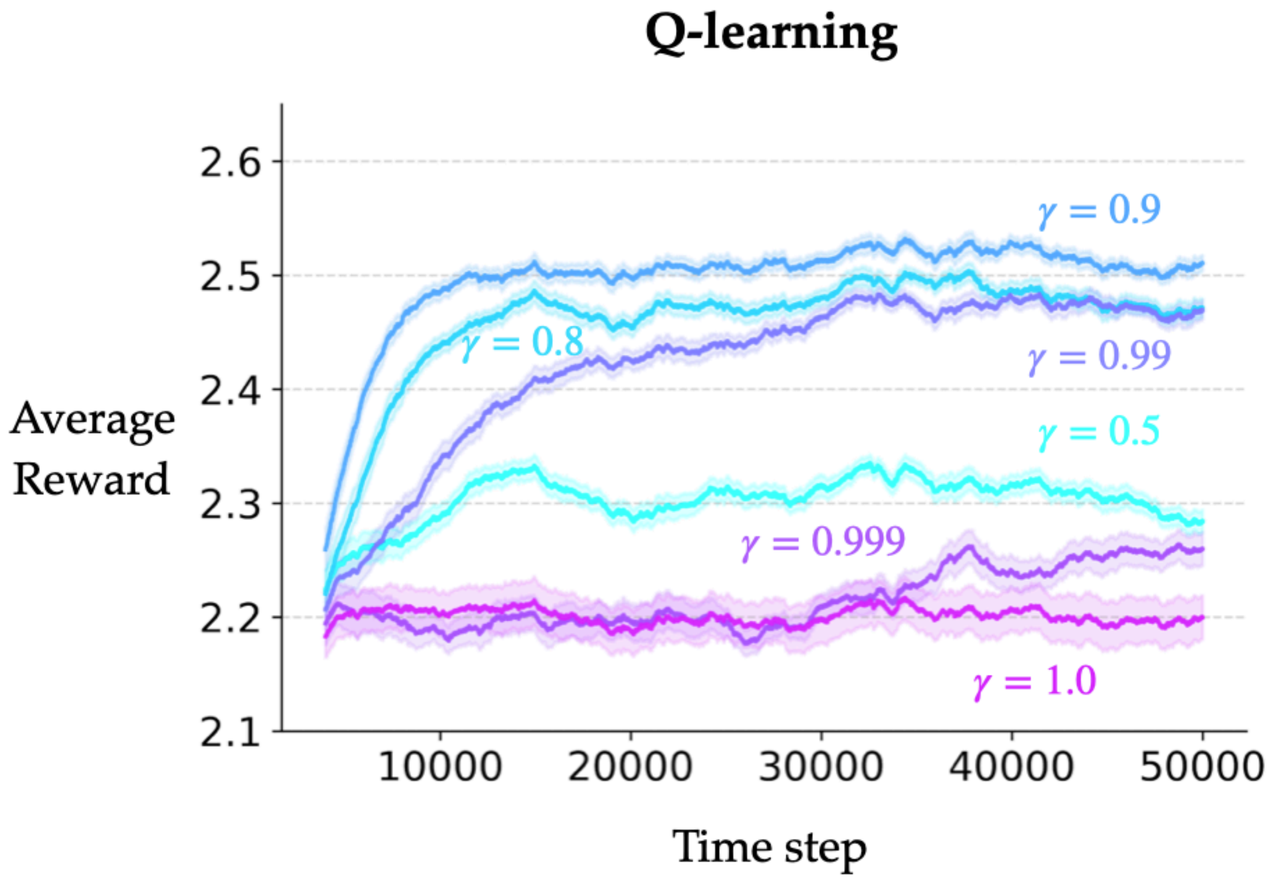
Blackwell, 1962; Grand-Clément & Petrik, 2023

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?

INSTABILITY WITH LARGE DISCOUNT FACTORS



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

BEYOND EMPIRICAL ISSUES...

The discounted-reward formulation is not well-suited for continuing problems beyond the tabular setting.

There are deeper theoretical issues.

10.4 Deprecating the Discounted Setting

The continuing, discounted problem formulation has been very useful in the tabular case, in which the returns from each state can be separately identified and averaged. But in the approximate case it is questionable whether one should ever use this problem formulation.

To see why, consider an infinite sequence of returns with no beginning or end, and no clearly identified states. The states might be represented only by feature vectors, which may do little to distinguish the states from each other. As a special case, all of the feature vectors may be the same. Thus one really has only the reward sequence (and the actions), and performance has to be assessed purely from these. How could it be done? One way is by averaging the rewards over a long interval—this is the idea of the average-reward setting. How could discounting be used? Well, for each time step we could measure the discounted return. Some returns would be small and some big, so again we would

Discounted Reinforcement Learning Is Not an Optimization Problem

Abhishek Naik[1,2*] **Roshan Shariff**[1] **Niko Yasui**[1] **Hengshuai Yao**[2] **Richard S. Sutton**[1]
{anaik1,rshariff,yasui,rsutton}@ualberta.ca hengshuai.yao@huawei.com

[1]Department of Computing Science
University of Alberta

[2]Huawei Technologies
Edmonton, Canada

Abstract

Discounted reinforcement learning is fundamentally incompatible with function approximation for control in continuing tasks. It is not an optimization problem in its usual formulation, so when using function approximation there is no optimal policy. We substantiate these claims, then go on to address some misconceptions about discounting and its connection to the average reward formulation. We encourage researchers to adopt rigorous optimization approaches, such as maximizing average reward, for reinforcement learning in continuing tasks.

Sutton & Barto (2018: Section 10.4). *Reinforcement Learning: An Introduction.* MIT Press.
Naik, Shariff, Yasui, Yao, Sutton (2019). *Discounted RL is not an Optimization Problem.* OptRL workshop at NeurIPS.

OUTLINE

0. Problem setting: what are continuing tasks?
1. The discounted-reward formulation
2. The average-reward formulation
3. Improving discounted methods using average reward
   (if time permits)

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 ⟶

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

This is finite.

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

THE ALGORITHMS FOR THE TWO FORMULATIONS LOOK QUITE SIMILAR IMPLEMENTATION-WISE

Discounted Q-learning

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

Differential Q-learning

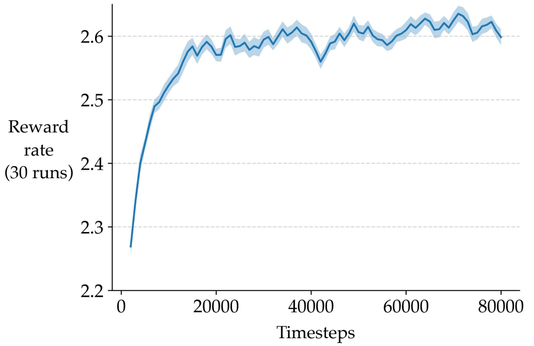
$$Q_{t+1}(S_t, A_t) \doteq Q_t(S_t, A_t) + \alpha_t \underbrace{\left[R_{t+1} - \bar{R}_t + \max_{a'} Q_t(S_{t+1}, a') - Q_t(S_t, A_t)\right]}_{\delta_t}$$

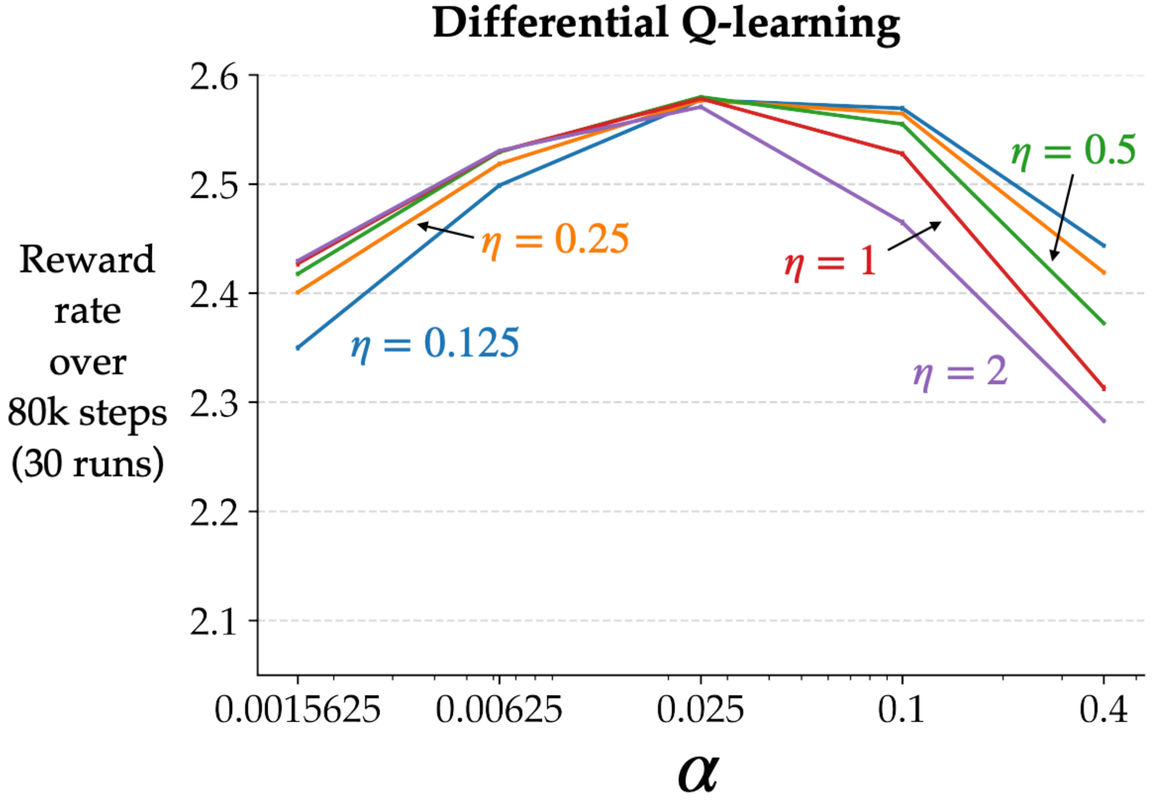
$$\bar{R}_{t+1} \doteq \bar{R}_t + \beta_t \, \delta_t$$

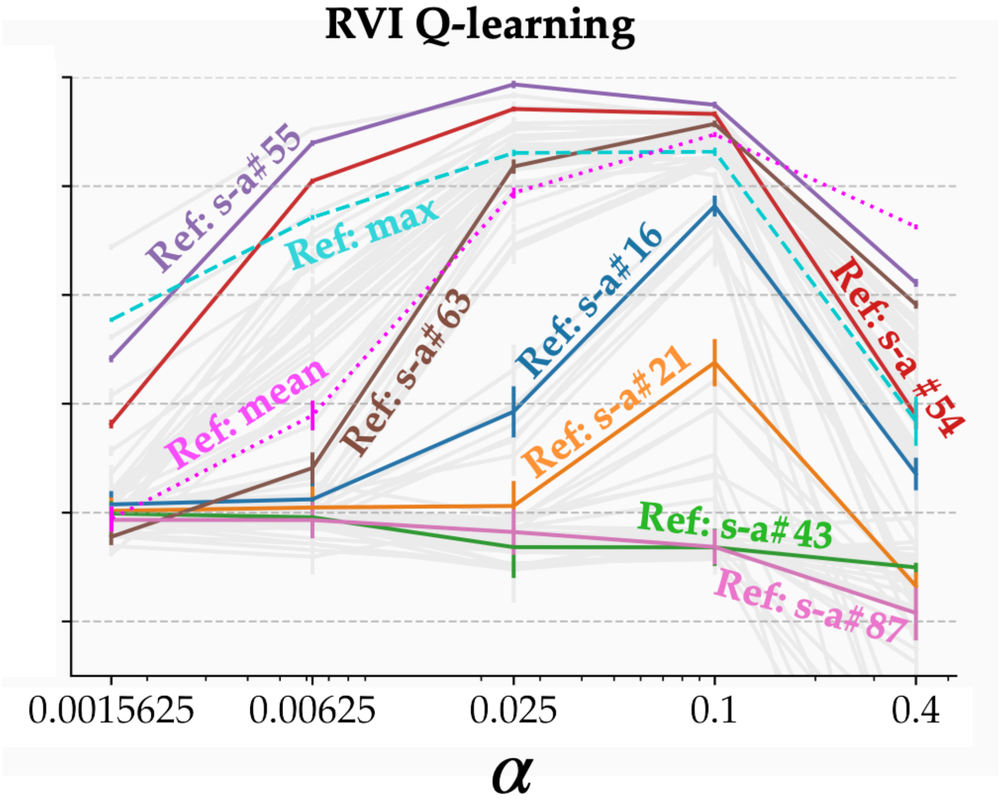
The theoretical analysis is significantly different, though.

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

THE ALGORITHM WORKS WELL IN PRACTICE AND IS ROBUST







AccessControl

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

THERE ARE NOW AVERAGE-REWARD EQUIVALENTS OF MANY COMMON DISCOUNTED-REWARD ALGORITHMS

- Average-reward TRPO

  Zhang, Ross (2021). *On-Policy Deep Reinforcement Learning for the Average-Reward Criterion*. ICML.

- Average-reward PPO

  Ma, Tang, Xia, Yang, Zhao (2021). *Average-reward reinforcement learning with trust region methods. IJCAI.*

- Average-reward DDPG

  Saxena, Khastagir, Kolathaya, Bhatnagar (2023). *Off-policy average reward actor-critic with deterministic policy search*. ICML.

- Average-reward SAC

  Adamczyk, Makarenko, Tiomkin, Kulkarni (2025). *Average-reward soft actor-critic.* RLC.

- Average-reward distributional algorithms

  Rojas, Lee (2026). *A differential perspective on distributional reinforcement learning*. AAAI.

⋮

OUTLINE

0. Problem setting: what are continuing tasks?
1. The discounted-reward formulation
2. The average-reward formulation
3. 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.

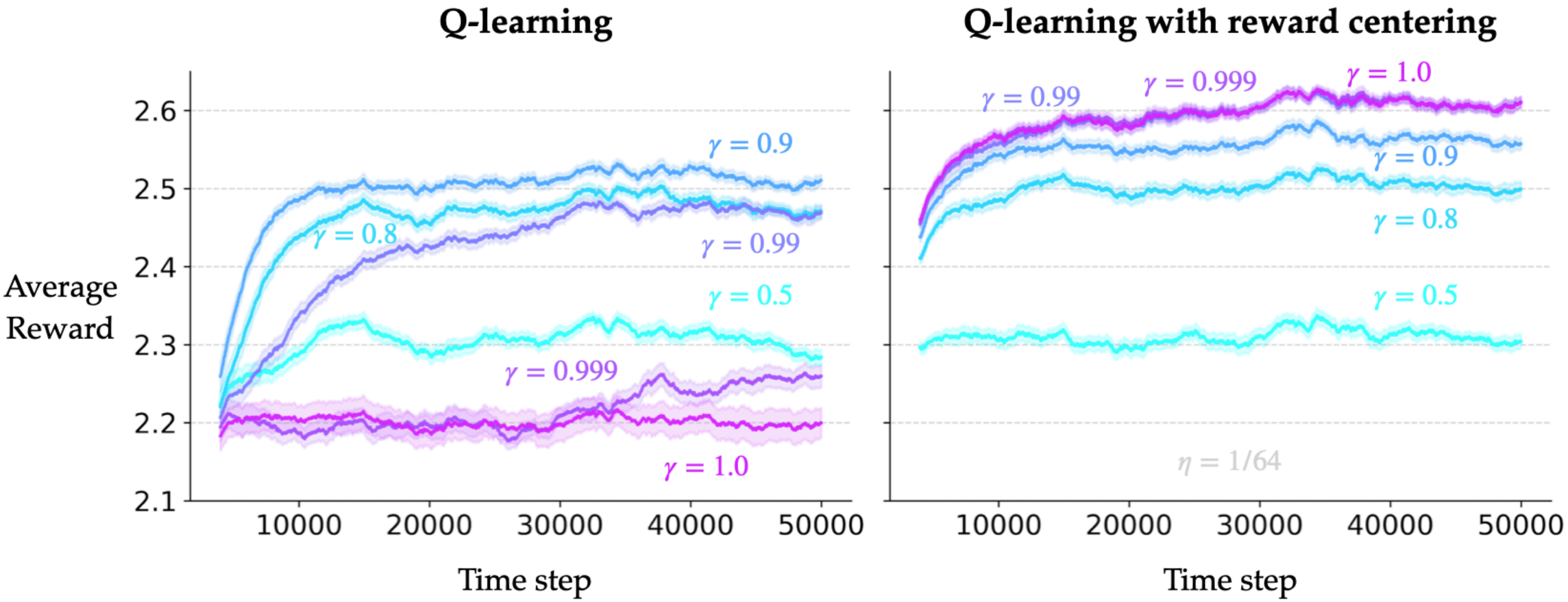
$$S_0\ A_0\ R_1\ S_1\ A_1, R_2 \ldots\ S_t\ A_t\ R_{t+1}\ S_{t+1}\ A_{t+1}\ R_{t+2}\ \ldots$$

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

$$\downarrow$$

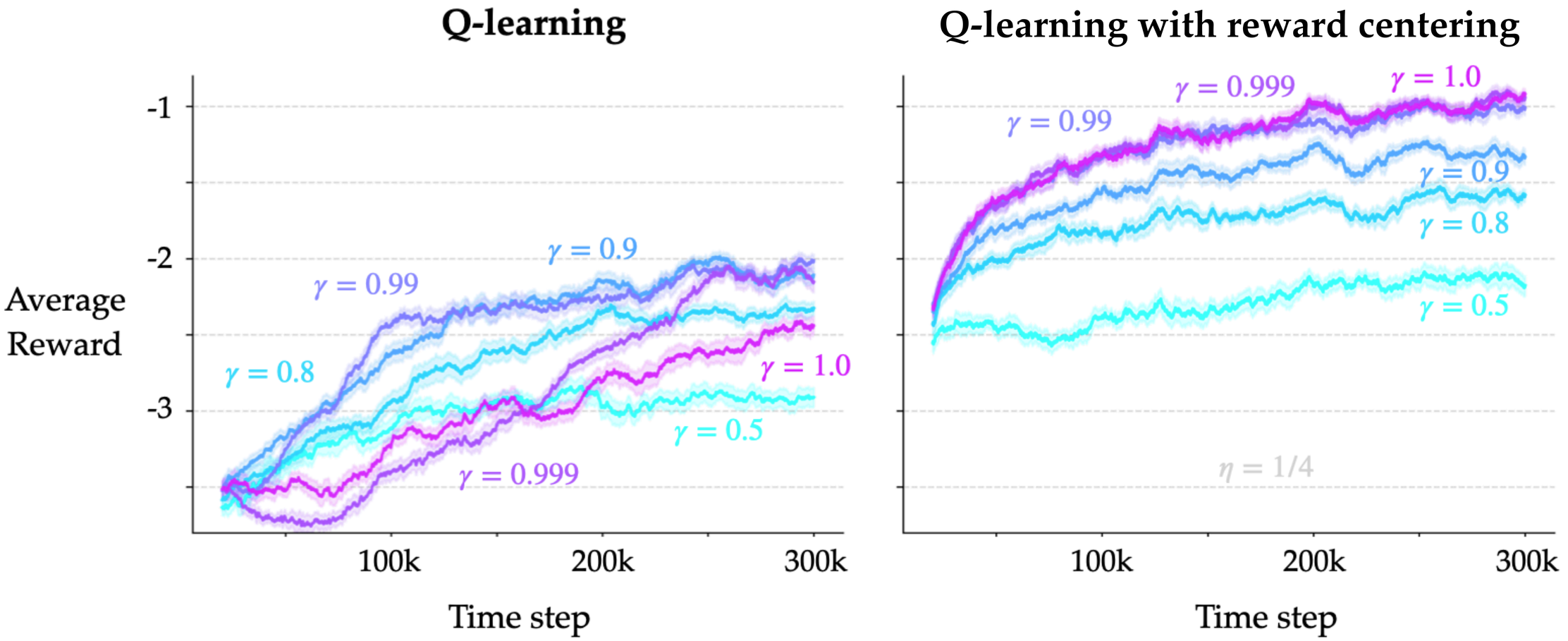
$$Q_{t+1}(S_t, A_t) \doteq Q_t(S_t, A_t) + \alpha_t \left[R_{t+1} - \bar{R}_t + \gamma \max_{a'} Q_t(S_{t+1}, a') - Q_t(S_t, A_t) \right]$$

NO INSTABILITY WITH LARGE DISCOUNT FACTORS



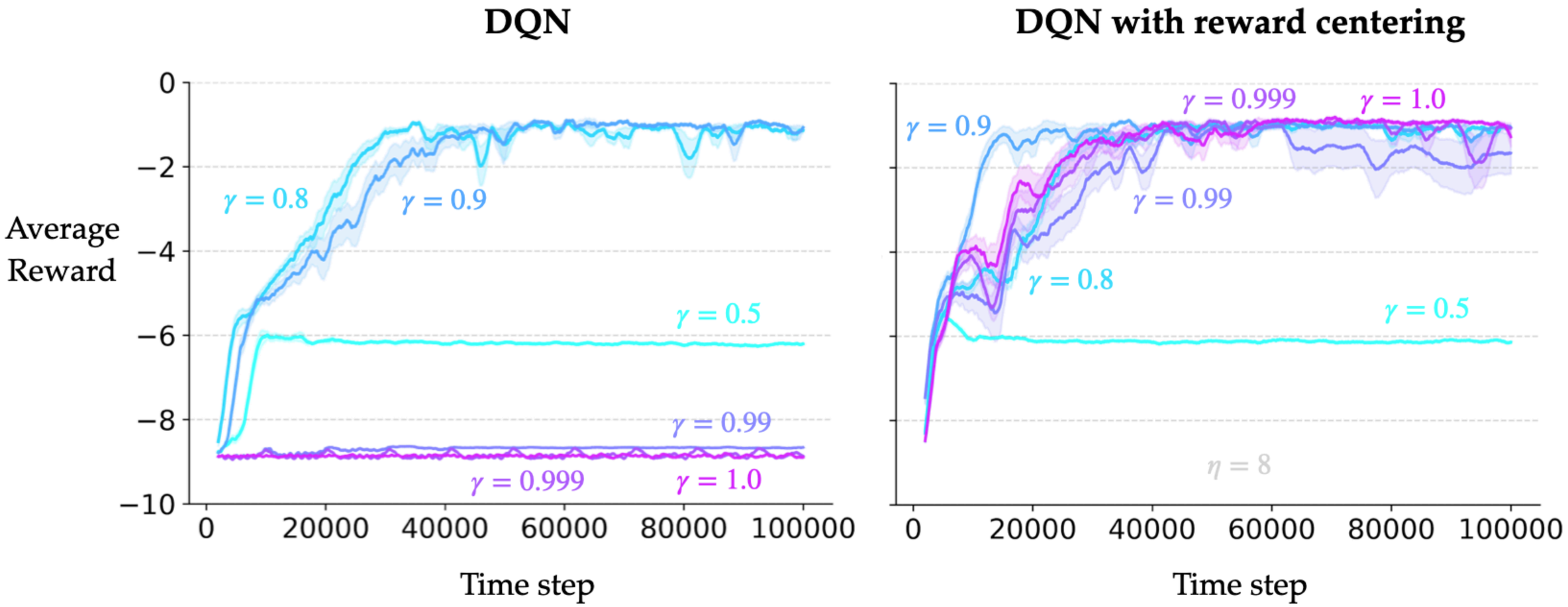
AccessControl (tabular)

NO INSTABILITY WITH LARGE DISCOUNT FACTORS



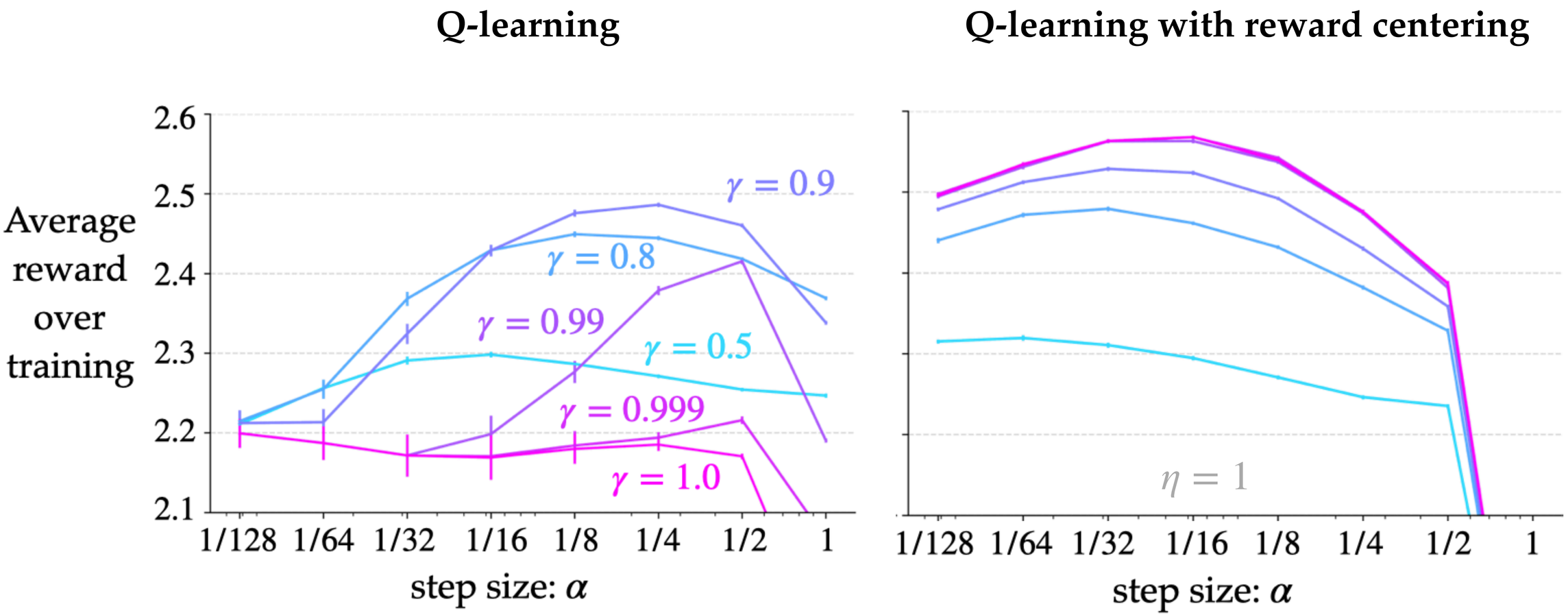
PuckWorld (linear FA)

NO INSTABILITY WITH LARGE DISCOUNT FACTORS



Pendulum (non-linear FA)

TRENDS ARE CONSISTENT ACROSS PARAMETERS



AccessControl (tabular)

UNDERLYING THEORY

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

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

Standard discounted value function

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

Average reward

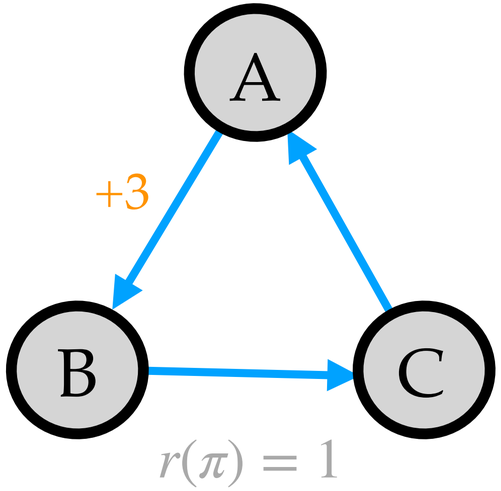
$$r(\pi) \doteq \lim_{n\to\infty} \frac{1}{n} \mathbb{E}_\pi\left[\sum_{t=1}^{n} R_t\right]$$

Differential value function

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

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

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

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

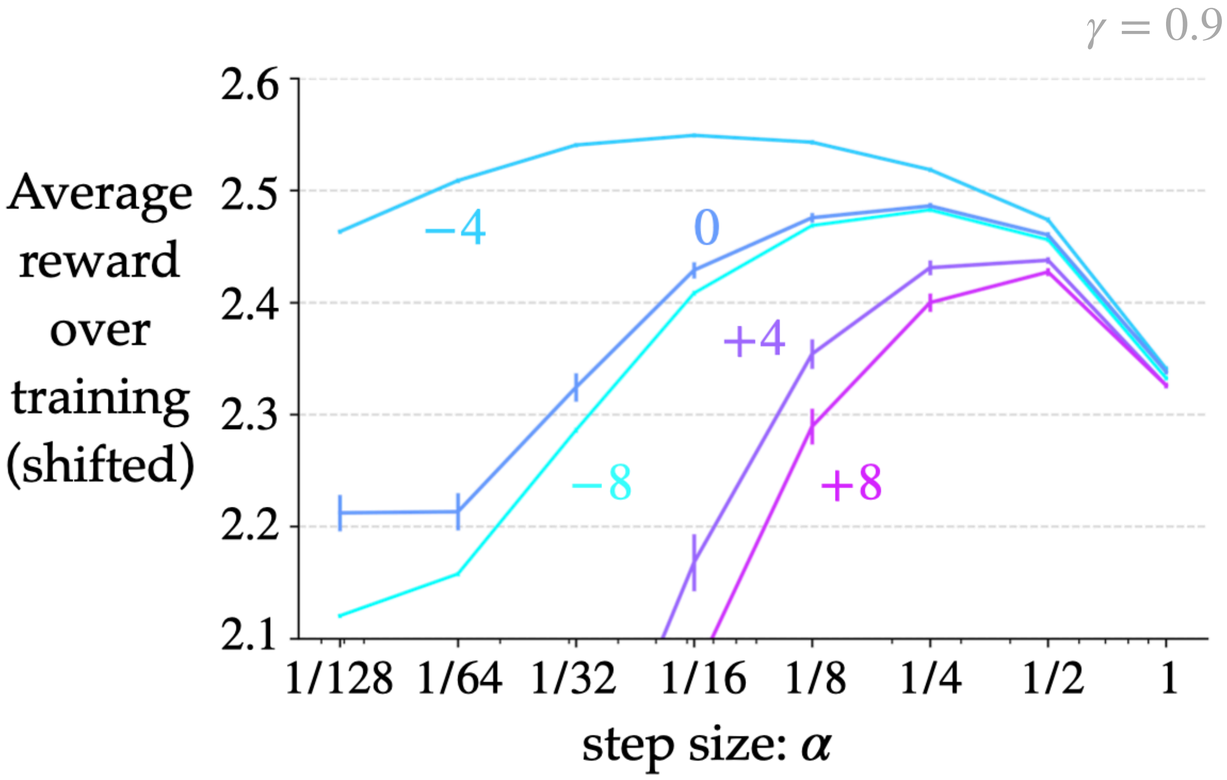
Off-policy $$\bar{R}_{t+1} \doteq \bar{R}_t + \beta_t\, \delta_t$$

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

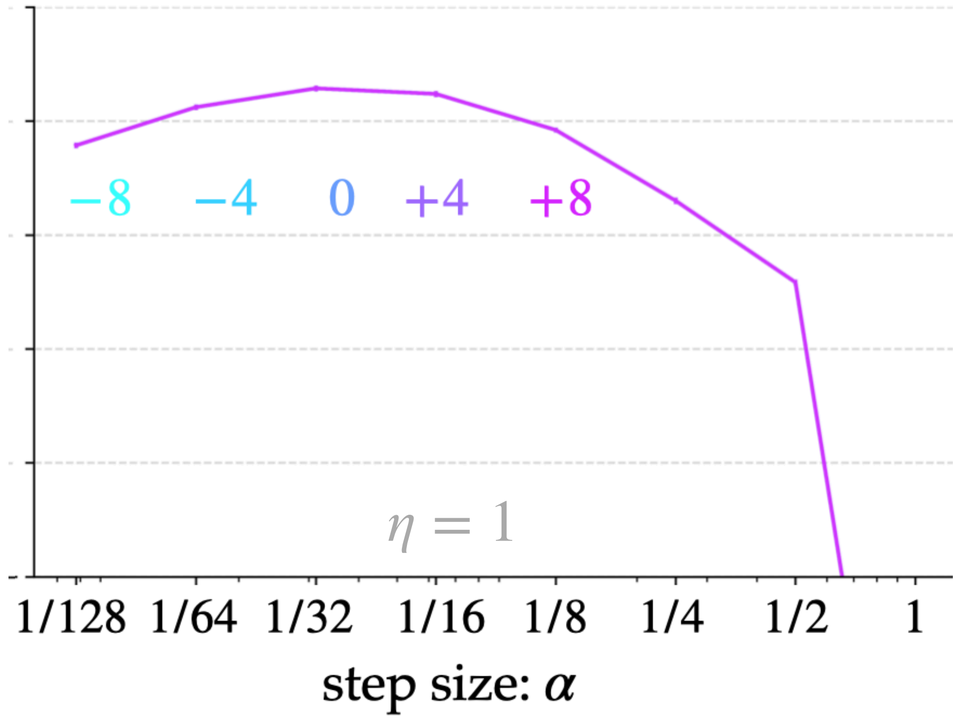
MORE ROBUST TO SHIFTED REWARDS

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

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.

Every RL algorithm will benefit with reward centering!

- 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

Reward centering has now been extended to episodic tasks as well!

De Asis, Elsayed, He (2026). *Extending Differential Temporal Difference Methods for Episodic Problems*. RLC.

Analysis, more experiments, etc.:

Naik, Wan, Tomar, & Sutton. (2024). *Reward Centering*. RLC.

OUTLINE

0. Problem setting: what are continuing tasks?
1. The discounted-reward formulation
2. The average-reward formulation
3. Improving discounted methods using average reward

My dissertation has more details on all these topics:

Naik (2024), *Reinforcement Learning in Continuing Problems using Average Reward,* PhD dissertation, University of Alberta.

THANK YOU

Questions?

abhisheknaik96.github.io
abhisheknaik22296@gmail.com