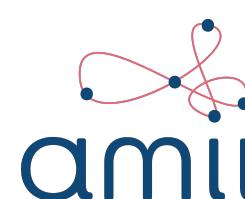


REINFORCEMENT LEARNING IN CONTINUING PROBLEMS USING AVERAGE REWARD

Defense Talk
28 March 2024

Abhishek Naik

with thanks to Rich, Yi, Janey, and many others



UNIVERSITY OF
ALBERTA



Additionally, problems of function approximation

- Remember, the policy improvement theorem does not hold in the function-approximation setting.
- In the tabular setting, we could compare two policies by a state-wise comparison of the value function.
- In the function-approximation setting, this cannot be done.



MY GOAL

To develop simple and practical learning algorithms
from first principles for long-lived agents

TOPICS I WORKED ON DURING MY PH.D.

- 0. The discounted formulation is not appropriate for AI
- 1. One-step average-reward methods
- 2. Multi-step average-reward methods
- 3. An idea to improve *discounted-reward* methods
- 4. A suite of continuing problems
- 5. Average-reward algorithms for the options framework
 - ▶ Planning with expectation models for control
 - ▶ Generalizing in the action space for large recommender systems

TOPICS I WORKED ON DURING MY PH.D.

0. The discounted formulation is not appropriate for AI

Naik et al. (2019). *Discounted Reinforcement Learning is Not an Optimization Problem*. Optimization Foundations of RL workshop at NeurIPS.

1. One-step average-reward methods

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

2. Multi-step average-reward methods

Naik & Sutton. (2022). *Multi-step Average-Reward Prediction via Differential TD(λ)*. RLDM.

Naik, Yu, Sutton. (2024). *Multi-Step Off-Policy Average-Reward Prediction with Eligibility Traces*. In preparation for a journal submission.

3. An idea to improve discounted-reward methods

Naik et al. (2024). *Reward Centering*. Under review.

4. A suite of continuing problems

Naik et al. (2021). *Towards Reinforcement Learning in the Continuing Setting*. Never-Ending RL workshop at ICLR. Also, Zhao et al. (2022). *CSuite*. github.com/google-deepmind/csuite

5. Average-reward algorithms for the options framework

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

▶ Planning with expectation models for control

Kudashkina et al. (2021). *Planning with Expectation Models for Control*. ArXiv:2104.08543

▶ Generalizing in the action space for large recommender systems

Naik et al. (2023). *Investigating Action-Space Generalization in Reinforcement Learning for Recommendation Systems*. RL4RecSys Workshop at WWW.

OUTLINE OF THIS TALK

Problem setting

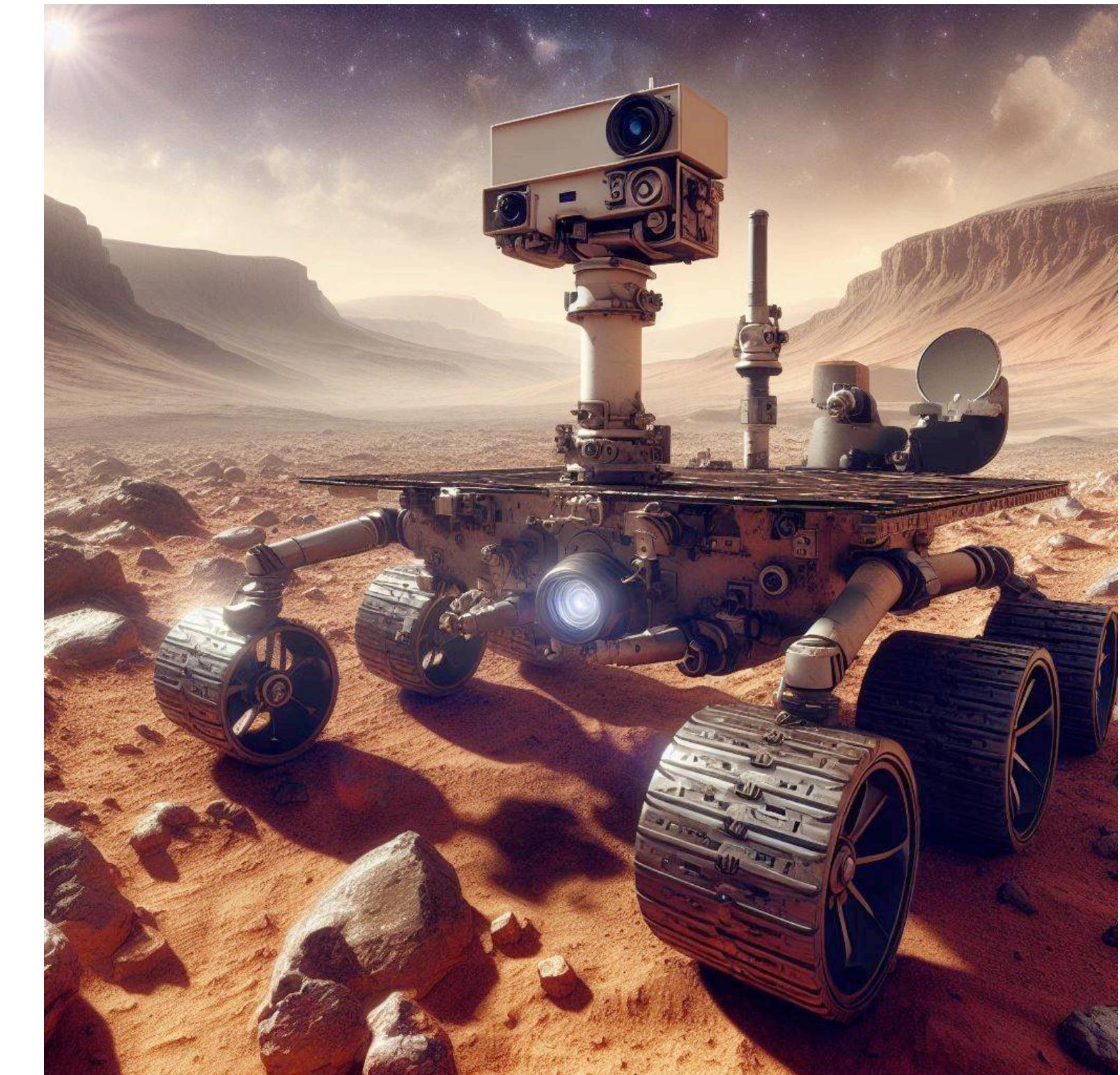
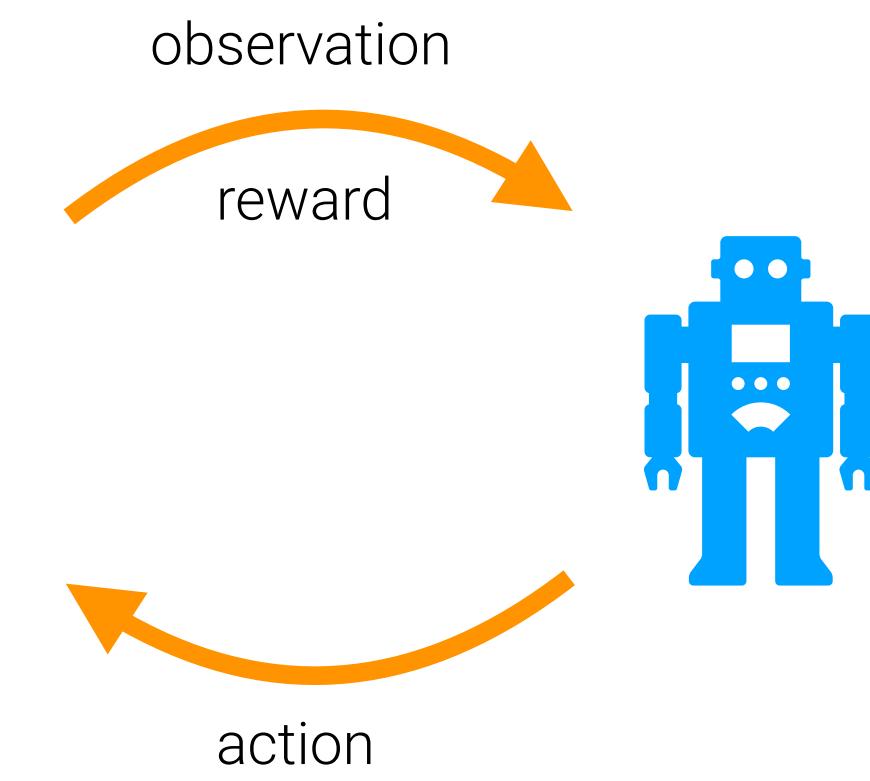
1. One-step average-reward methods
2. Multi-step average-reward methods
3. An idea to improve *discounted-reward* methods

Conclusions, limitations, and future work

Acknowledgments

PROBLEM SETTING

CONTINUING PROBLEMS



PROBLEM SETTING

CONTINUING PROBLEMS: FORMULATIONS

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

$$\max_{\pi} \sum_t^{\infty} R_t$$

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

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

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

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

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

PROBLEM SETTING

THE AVERAGE-REWARD FORMULATION

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

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

Differential value function $\longrightarrow \tilde{v}_\pi(s) \doteq \mathbb{E}_\pi[R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + \dots | S_t = s]$
 $v_\pi^\gamma(s) \doteq \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s]$

Estimate $r(\pi)$ and \tilde{v}_π
 while behaving according to b

Find π that maximizes $r(\pi)$
 while behaving according to b

off-policy

OUTLINE

Problem setting

1. One-step average-reward methods
2. Multi-step average-reward methods
3. An idea to improve *discounted-reward* methods

Conclusions, limitations, and future work

Acknowledgments

OUTLINE

Problem setting

1. *One-step average-reward methods*
2. *Multi-step average-reward methods*
3. An idea to improve *discounted-reward* methods

Conclusions, limitations, and future work

Acknowledgments

ONE-STEP AVERAGE-REWARD LEARNING METHODS

WITH PARTICULAR FOCUS ON THE OFF-POLICY CONTROL SETTING

- ▶ Schwartz (1993), Singh (1994), Tadepalli & Ok (1994), Bertsekas & Tsitsiklis (1996), Das et al. (1999), Ren & Krogh (2001), Gosavi (2004), Yang et al. (2016)
- ▶ Convergence results either not present, or, require special information about the problem.
- ▶ Abounadi, Bertsekas, & Borkar (2001): *a big step forward*

ESTIMATING THE AVERAGE REWARD FROM DATA

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

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

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

Off-policy?

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

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

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

```
new_estimate = old_estimate + stepsize*(new_target - old_estimate)
```

$$r(\pi) = \sum_s d_\pi(s) \sum_a \pi(a|s) \sum_r p(r|s,a) \ r$$

$$r(b) = \sum_s d_b(s) \sum_a b(a|s) \sum_r p(r|s,a) \ r$$

With $\rho_t \doteq \frac{\pi(A_t|S_t)}{b(A_t|S_t)}$

$$\bar{R}_\infty \not\rightarrow r(b)$$

$$\bar{R}_\infty \not\rightarrow r(\pi)$$

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

ESTIMATING THE VALUES FROM DATA

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

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

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

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

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

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

$\underbrace{\hspace{10em}}$

δ_t^γ

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

$\underbrace{\hspace{10em}}$

δ_t

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

ESTIMATING THE AVERAGE REWARD FROM DATA

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

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

δ_t

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

ESTIMATING THE AVERAGE REWARD FROM DATA

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

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

δ_t

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

ESTIMATING THE AVERAGE REWARD FROM DATA

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

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

δ_t

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

ESTIMATING THE AVERAGE REWARD FROM DATA

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

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

δ_t

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

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

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

$$\delta_t$$

(CONVERGENT) ALGORITHMS FOR OFF-POLICY CONTROL

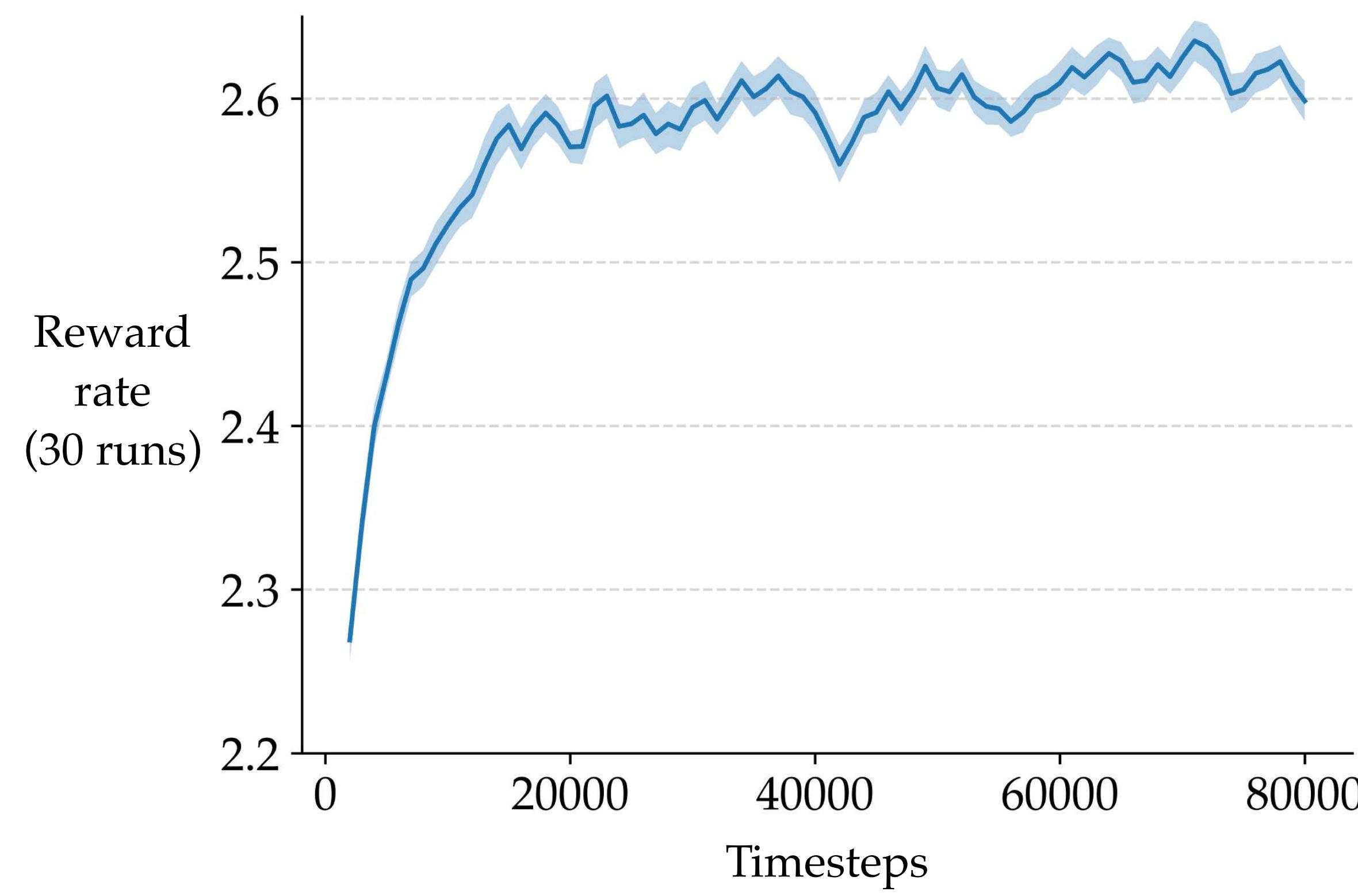
$$Q_{t+1}(S_t, A_t) \doteq Q_t(S_t, A_t) + \alpha_t [R_{t+1} - \bar{R}_t + \max_{a'} Q_t(S_{t+1}, a') - Q_t(S_t, A_t)]$$
$$\bar{R}_{t+1} \doteq \bar{R}_t + \frac{\eta \alpha_t \delta_t}{\beta_t}$$

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

Examples of f :

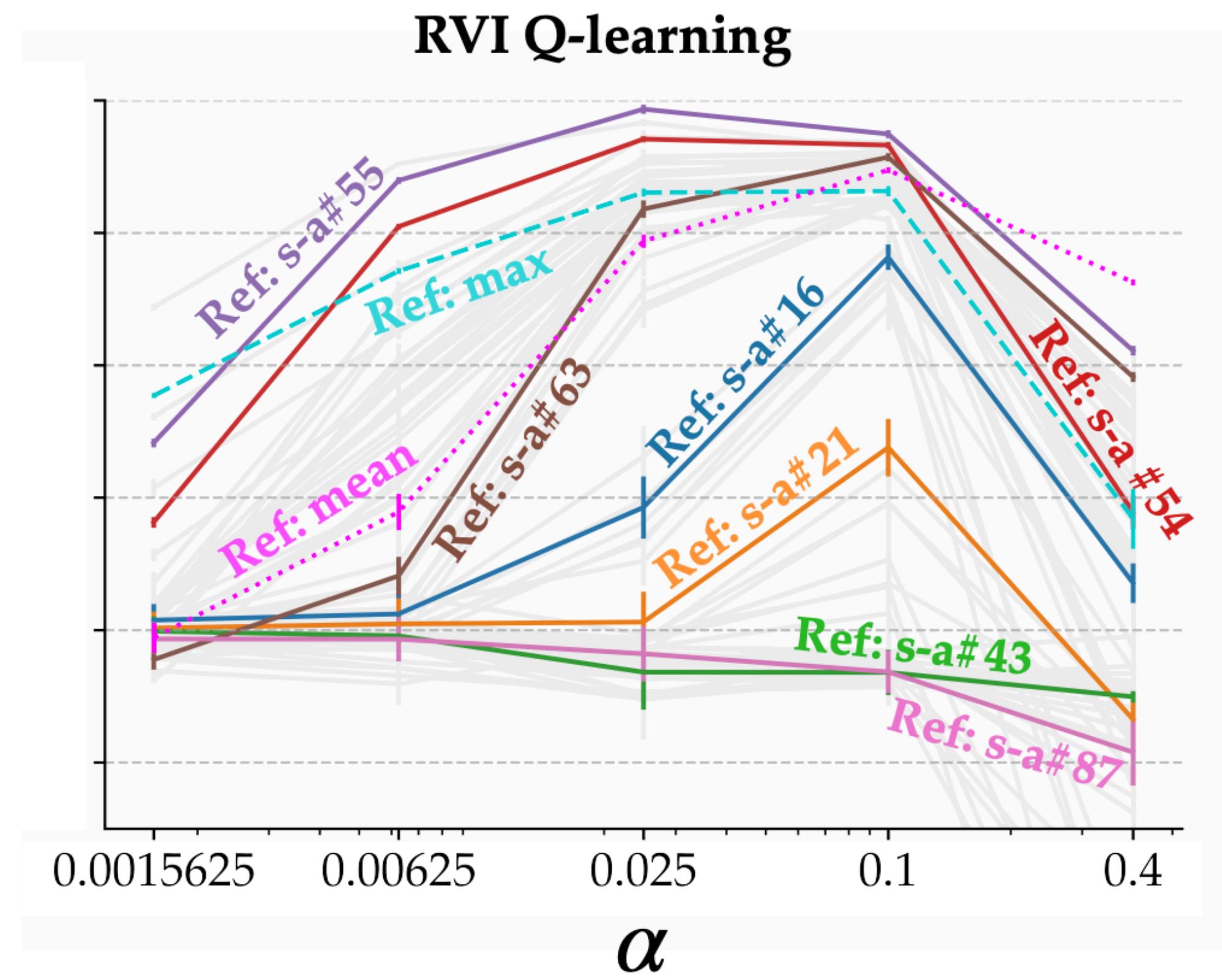
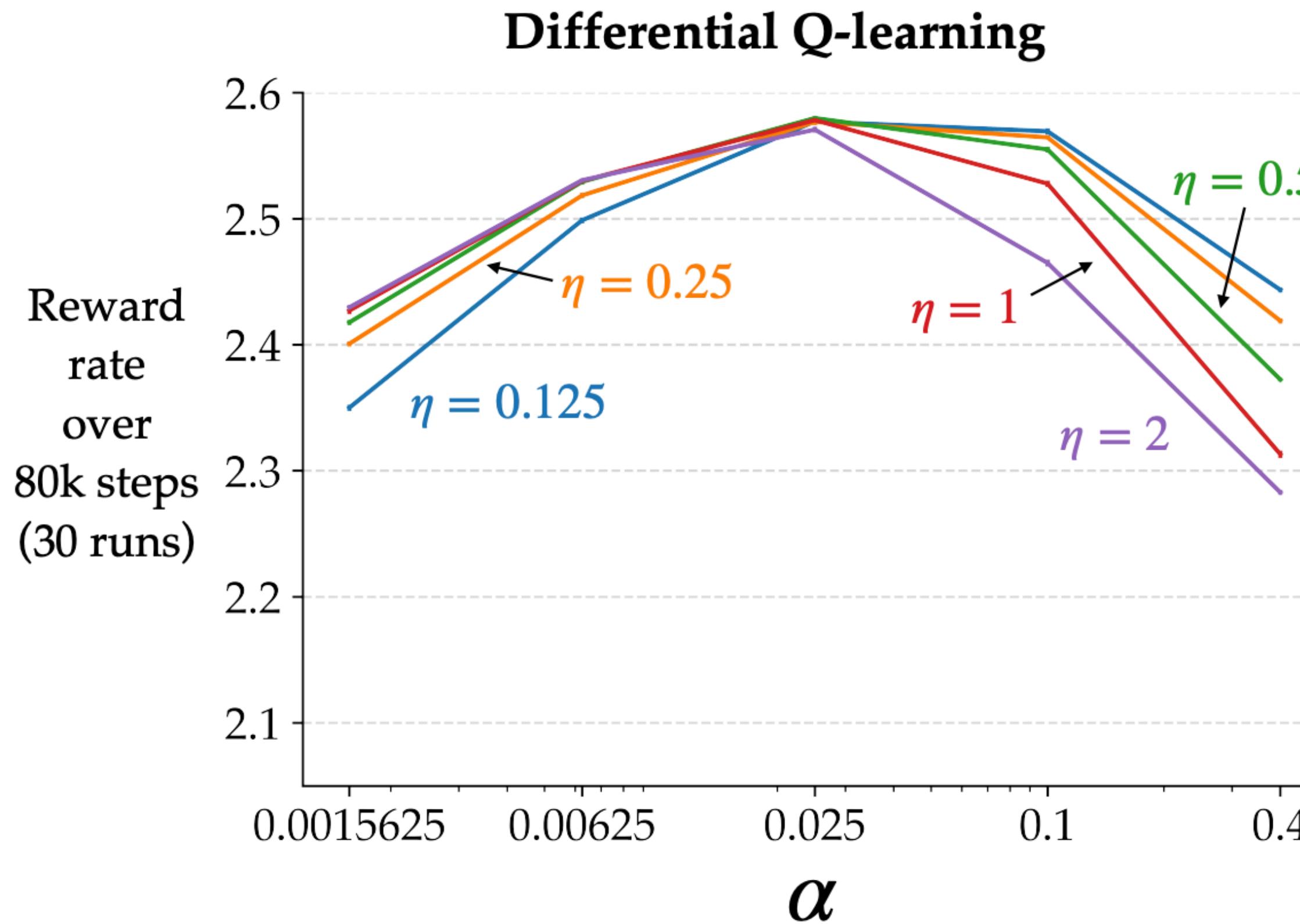
- ▶ value of a single state–action pair
- ▶ average of values of all state–action pairs

PERFORMANCE COMPARISON

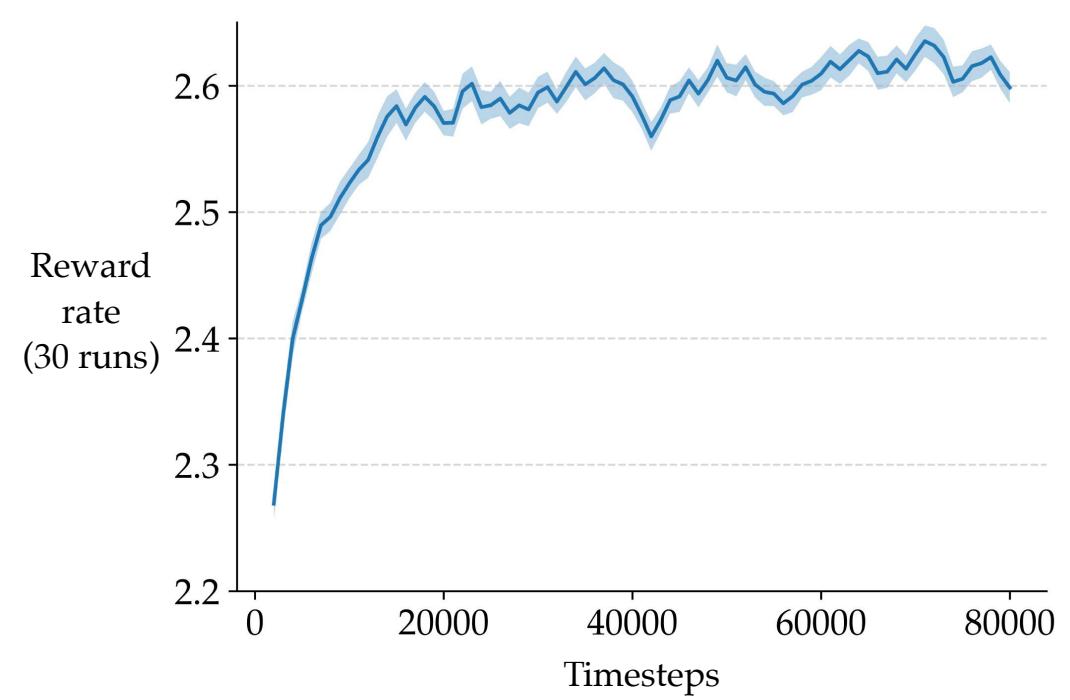


AccessControl

PERFORMANCE COMPARISON



AccessControl



TAKEAWAYS

- ▶ Differential Q-learning is a convergent algorithm for off-policy average-reward control.
 - ▶ does not require a reference function
 - ▶ is relatively easy to use
- ▶ Differential TD-learning is a convergent algorithm for off-policy average-reward prediction.
 - ▶ estimates both the average reward and the values accurately
 - ▶ is relatively easy to use
- ▶ Convergence results limited to the tabular case
- ▶ No temporal abstraction
- ▶ All algorithms are one-step methods

More experiments, planning variants of these learning algorithms, convergence proofs, etc.:

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

OUTLINE

Problem setting

1. *One-step average-reward methods*
2. *Multi-step average-reward methods*
3. An idea to improve *discounted-reward* methods

Conclusions, limitations, and future work

Acknowledgments

OUTLINE

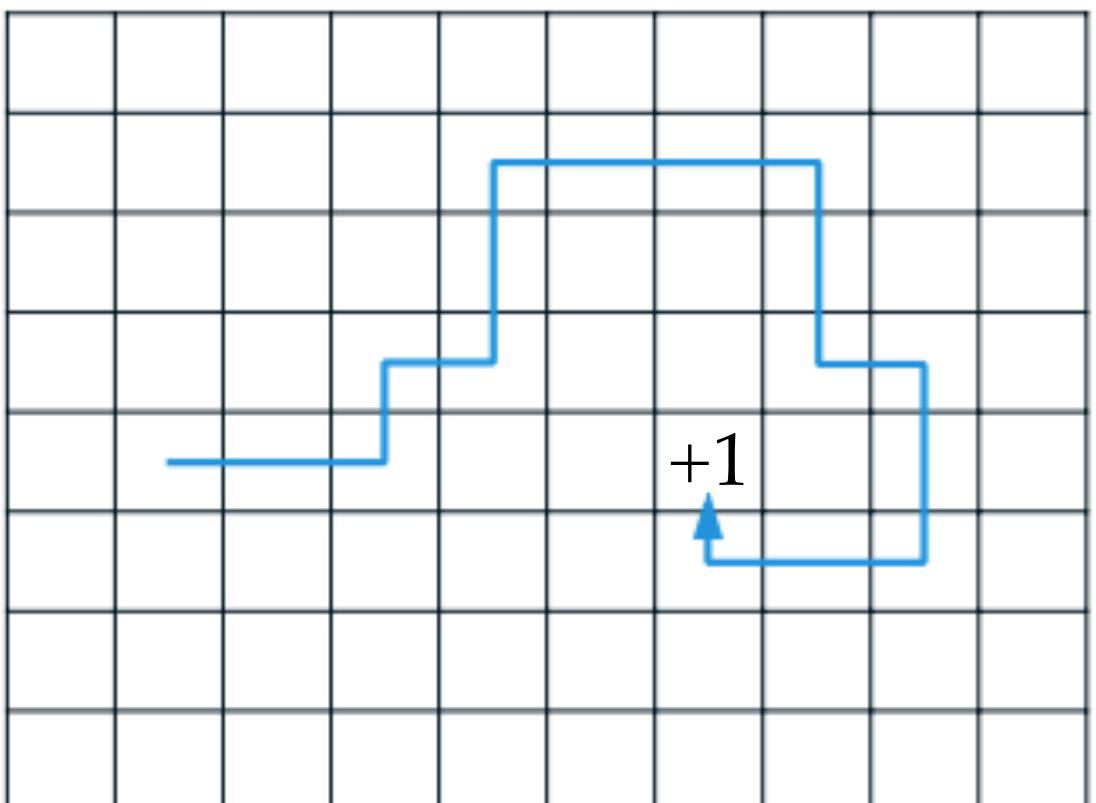
Problem setting

1. One-step average-reward methods
2. *Multi-step* average-reward methods
3. An idea to improve *discounted-reward* methods

Conclusions, limitations, and future work

Acknowledgments

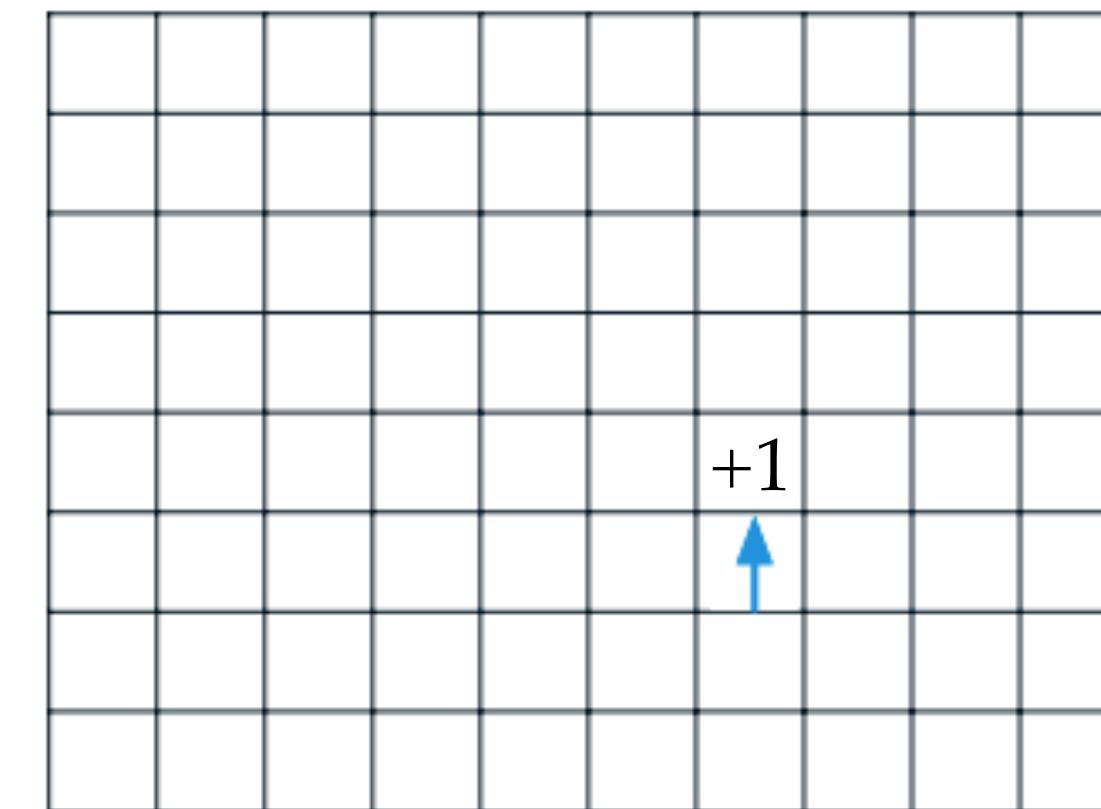
MULTI-STEP UPDATES CAN BE MORE EFFICIENT



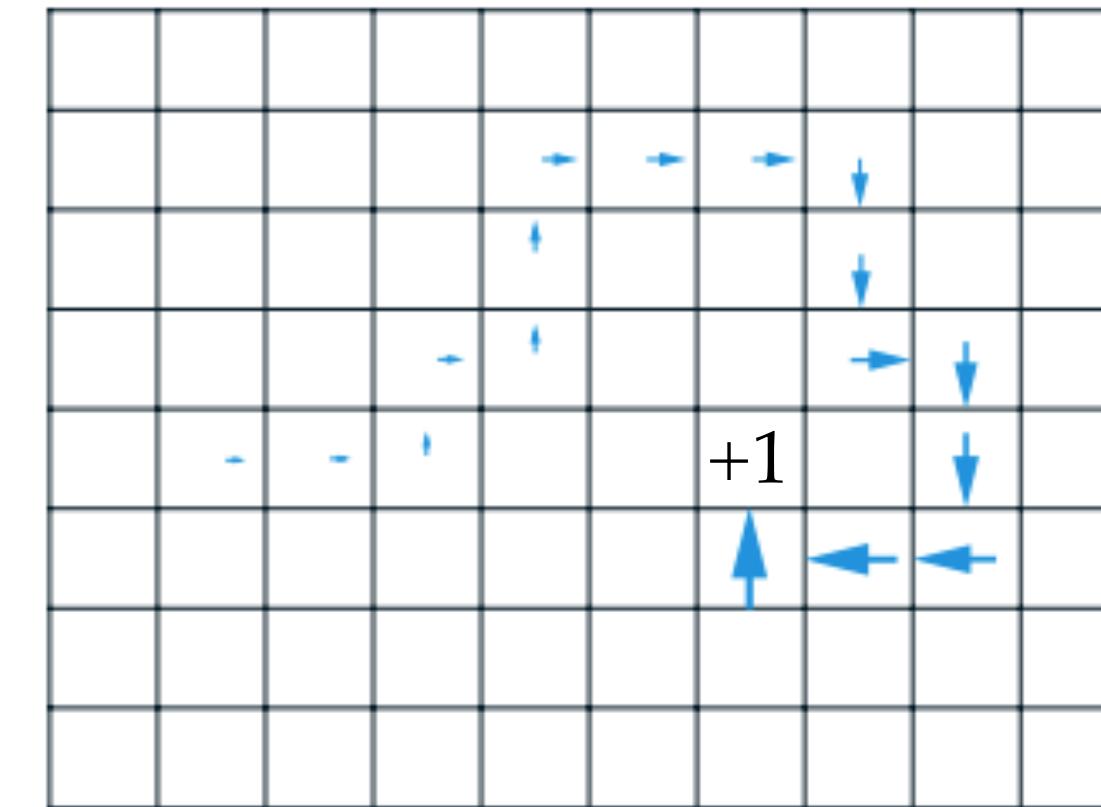
Trajectory

One-step
update

Multi-step
update



Learned values / policy



ON-POLICY PREDICTION

$$v_{\pi}(s) \approx \mathbf{w}^{\top} \mathbf{x}(s)$$

One-step Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{x}_t$$

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

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^{\top} \mathbf{x}_{t+1} + \mathbf{w}_t^{\top} \mathbf{x}_t$

Multi-step version

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

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

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

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^{\top} \mathbf{x}_{t+1} + \mathbf{w}_t^{\top} \mathbf{x}_t$

$$\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$$

Algorithm 1

Also guaranteed to converge,
under the same conditions

Average-Cost TD(λ)

Guaranteed to converge
(Tsitsiklis & Van Roy, 1999)

(WHAT I'VE LEARNED ABOUT)

PROVING CONVERGENCE OF SAMPLED-BASED ALGORITHMS USING THE ODE APPROACH

$$\underline{\mathbf{w}_0 \ \mathbf{w}_1 \ \dots \ \mathbf{w}_t \ \dots}$$

1. Show that the sequence of iterates is bounded and asymptotically converges to the solutions of an ODE.

2. Show the ODE has a globally stable equilibrium point.

Proving the convergence of Algorithm 1 was fairly straightforward.

EXTENSION TO THE OFF-POLICY SETTING

One-step off-policy Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \rho_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

where $\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} + \mathbf{w}_t^\top \mathbf{x}_t$

$$\rho_t \doteq \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

Algorithm 1

$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t \\ \bar{R}_{t+1} &\doteq \bar{R}_t + \eta \alpha_t \delta_t\end{aligned}$$

where $\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$

Multi-step version?



$$\begin{aligned}\mathbf{w}_{t+1} &\doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t \\ \bar{R}_{t+1} &\doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t\end{aligned}$$

where $\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$

Algorithm 1_{off}

ANALYSIS OF (TABULAR) ALGORITHM 1OFF'S "A" MATRIX

$$\mathbf{A}^1 \doteq \begin{bmatrix} -\eta & \mathbf{0}^\top \\ \frac{-1}{1-\lambda} \mathbf{D}_\pi \mathbf{1} & \mathbf{D}_\pi (\mathbf{P}_\pi^\lambda - \mathbb{I}) \end{bmatrix}$$

is Hurwitz.

(Tsitsiklis & Van Roy's
(1999) Lemma 7)

$$\mathbf{A}^{1off} \doteq \begin{bmatrix} -\eta & \eta \mathbf{d}_b^\top (\mathbf{P}_\pi - \mathbb{I}) \\ \frac{-1}{1-\lambda} \mathbf{D}_b \mathbf{1} & \mathbf{D}_b (\mathbf{P}_\pi^\lambda - \mathbb{I}) \end{bmatrix}$$

is *not* Hurwitz.

(via a simulation analysis)

So Algorithm 1off can diverge... :(

EXTENSION TO THE OFF-POLICY SETTING

One-step *off-policy* Differential TD

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \rho_t \delta_t \mathbf{x}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \mathbf{w}_t^\top \mathbf{x}_{t+1} + \mathbf{w}_t^\top \mathbf{x}_t$$

$$\rho_t \doteq \frac{\pi(A_t | S_t)}{b(A_t | S_t)}$$

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t z_t^{\bar{R}}$$

where

$$\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$$

$$z_t^{\bar{R}} \doteq \rho_t (\lambda z_{t-1}^{\bar{R}} + 1)$$

Algorithm 2

Algorithm 1

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

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

$$\mathbf{z}_t \doteq \lambda \mathbf{z}_{t-1} + \mathbf{x}_t$$

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \rho_t \delta_t$$

where

$$\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$$

Algorithm 1 off

ANALYSIS OF (TABULAR) ALGORITHM 2

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha_t \delta_t \mathbf{z}_t$$

$$\bar{R}_{t+1} \doteq \bar{R}_t + \eta \alpha_t \delta_t z_t^{\bar{R}}$$

where $\mathbf{z}_t \doteq \rho_t (\lambda \mathbf{z}_{t-1} + \mathbf{x}_t)$

$$z_t^{\bar{R}} \doteq \rho_t (\lambda z_{t-1}^{\bar{R}} + 1)$$

$$\mathbf{A} = \mathbf{D}_b(\mathbf{P}_\pi^\lambda - \mathbb{I} - \frac{\eta}{1-\lambda} \mathbf{1}\mathbf{g}^\top) \text{ is Hurwitz!}$$

TAKEAWAYS

- ▶ **Algorithm 1** is a convergent multi-step algorithm for *on-policy* prediction with *linear* function approximation.
- ▶ **Algorithm 2** is a member of a *family* of multi-step *off-policy* prediction algorithms that are guaranteed to converge in the *tabular* case.
 - ▶ first convergence result for the multi-step off-policy setting!
- ▶ Convergence results can be further generalized
- ▶ Algorithms and analysis have to be fully extended to use function approximation
- ▶ Algorithm 2 may not be best among the family of algorithms

Both algorithms are extensions of the one-step Differential TD-learning algorithm.

Complete convergence analysis, experiments, etc.:

Naik, Huizhen, & Sutton. (2024). *Multi-Step Off-Policy Average-Reward Prediction with Eligibility Traces*. In preparation.

OUTLINE

Problem setting

1. One-step average-reward methods
2. *Multi-step* average-reward methods
3. An idea to improve *discounted-reward* methods

Conclusions, limitations, and future work

Acknowledgments

OUTLINE

Problem setting

1. One-step average-reward methods
2. Multi-step average-reward methods
3. An idea to improve *discounted-reward* methods

Conclusions, limitations, and future work

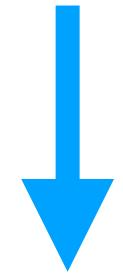
Acknowledgments

REWARD CENTERING

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

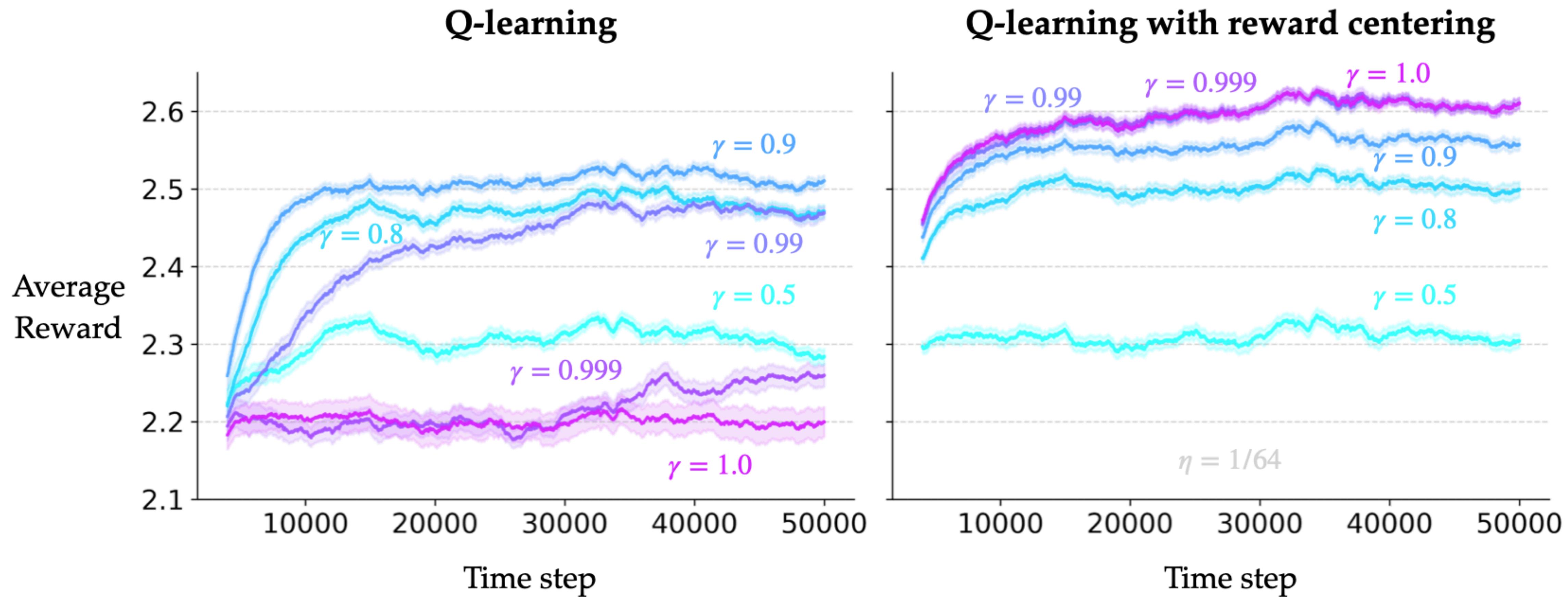
Estimate the average reward and subtract it from the observed rewards

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



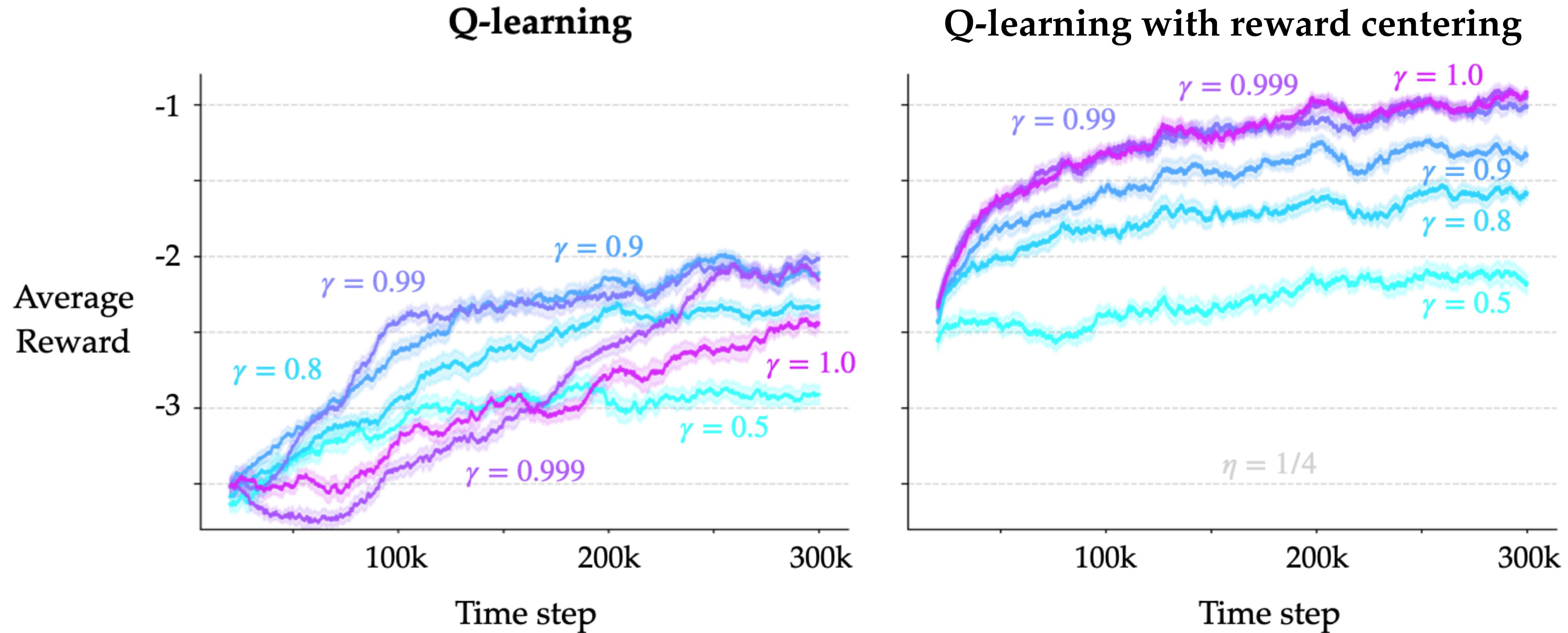
$$Q_{t+1}(S_t, A_t) \doteq Q_t(S_t, A_t) + \alpha_t [R_{t+1} - \bar{R}_t + \gamma \max_{a'} Q_t(S_{t+1}, a') - Q_t(S_t, A_t)]$$

NO INSTABILITY WITH LARGE DISCOUNT FACTORS



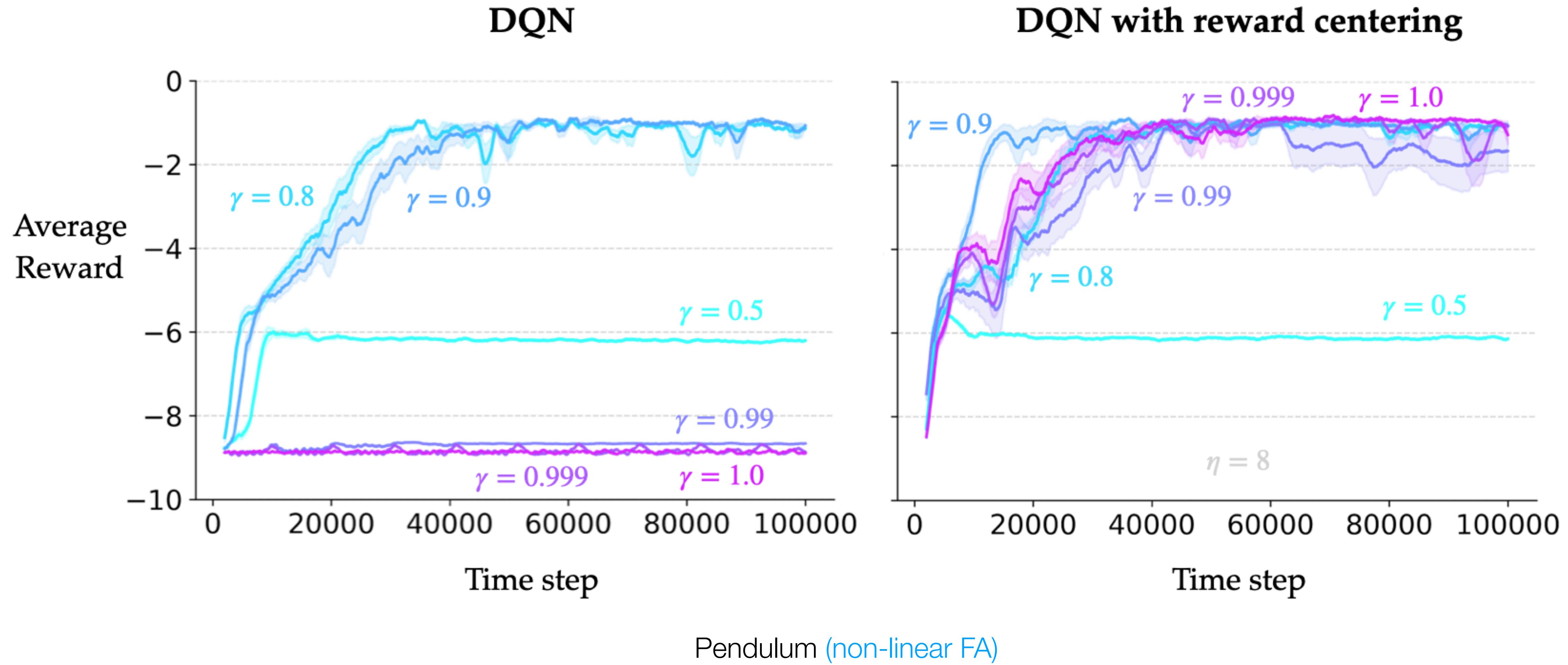
AccessControl (tabular)

NO INSTABILITY WITH LARGE DISCOUNT FACTORS



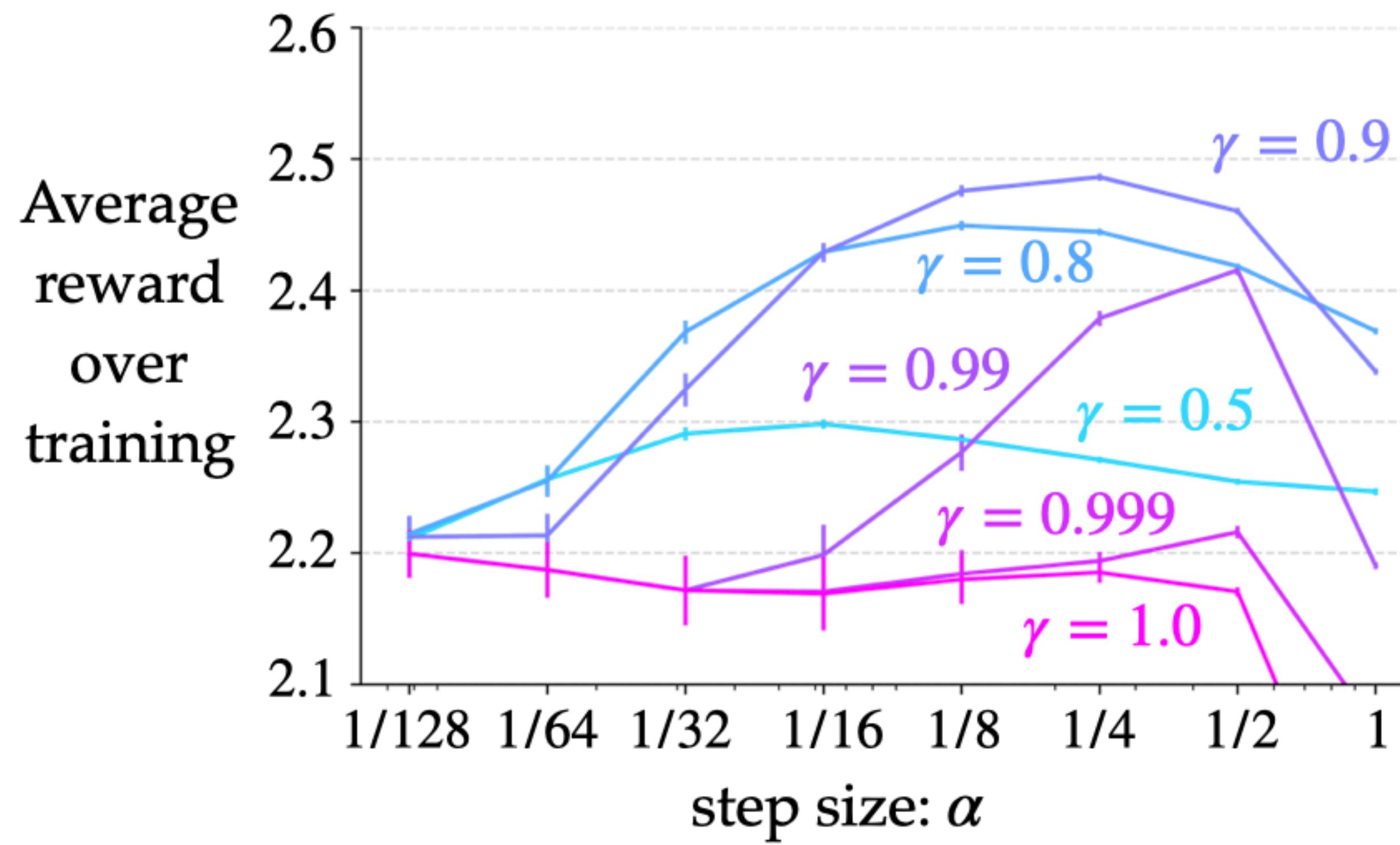
PuckWorld (linear FA)

NO INSTABILITY WITH LARGE DISCOUNT FACTORS

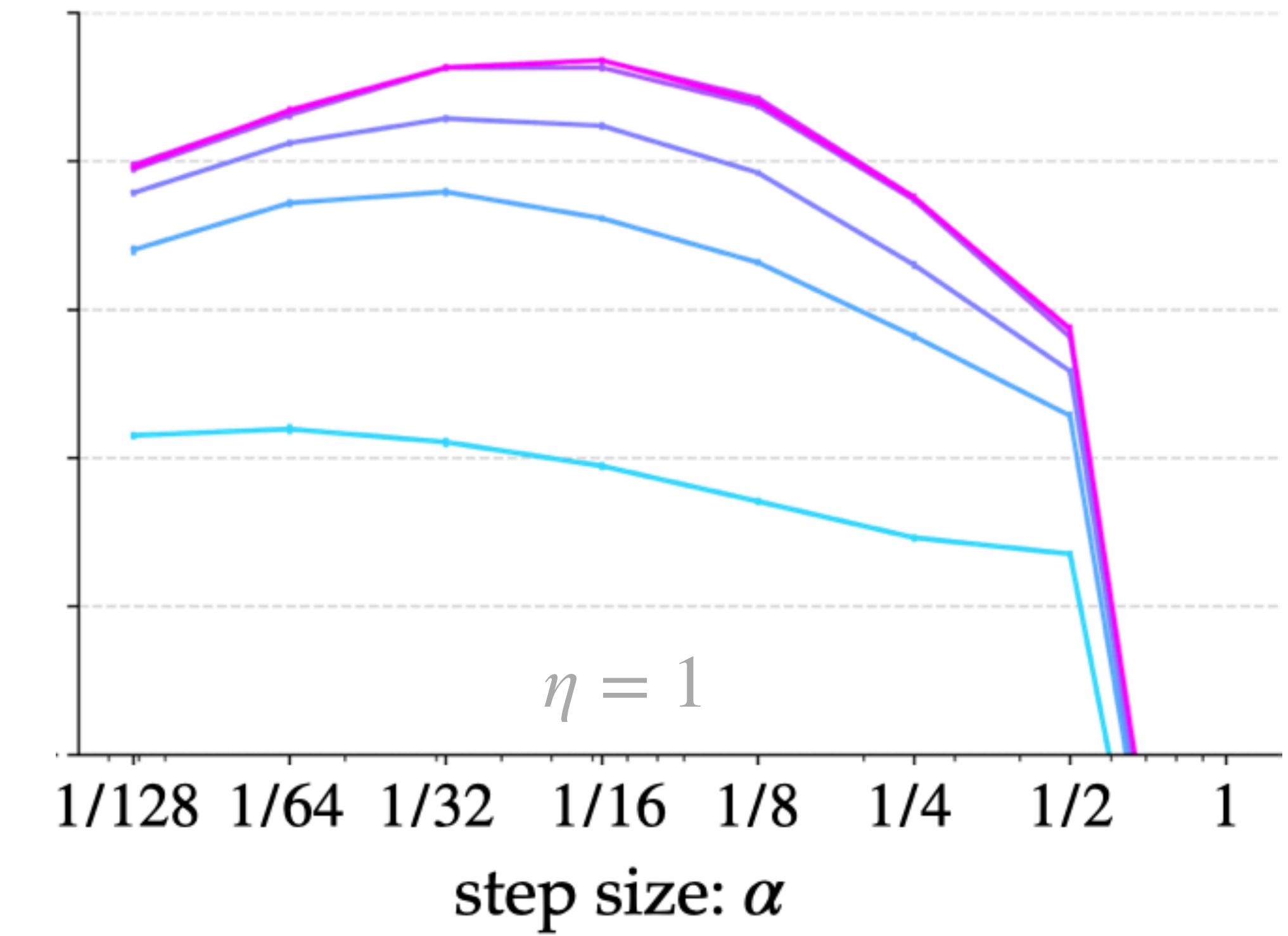


TRENDS ARE CONSISTENT ACROSS PARAMETERS

Q-learning



Q-learning with reward centering



AccessControl (tabular)

KEY INSIGHT

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

$$v_\pi^\gamma(s) \doteq \mathbb{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots | S_t = s] \quad \xleftarrow{\text{Standard discounted value function}}$$

$$= \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | S_t = s \right]$$

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

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

Centered
discounted
value function

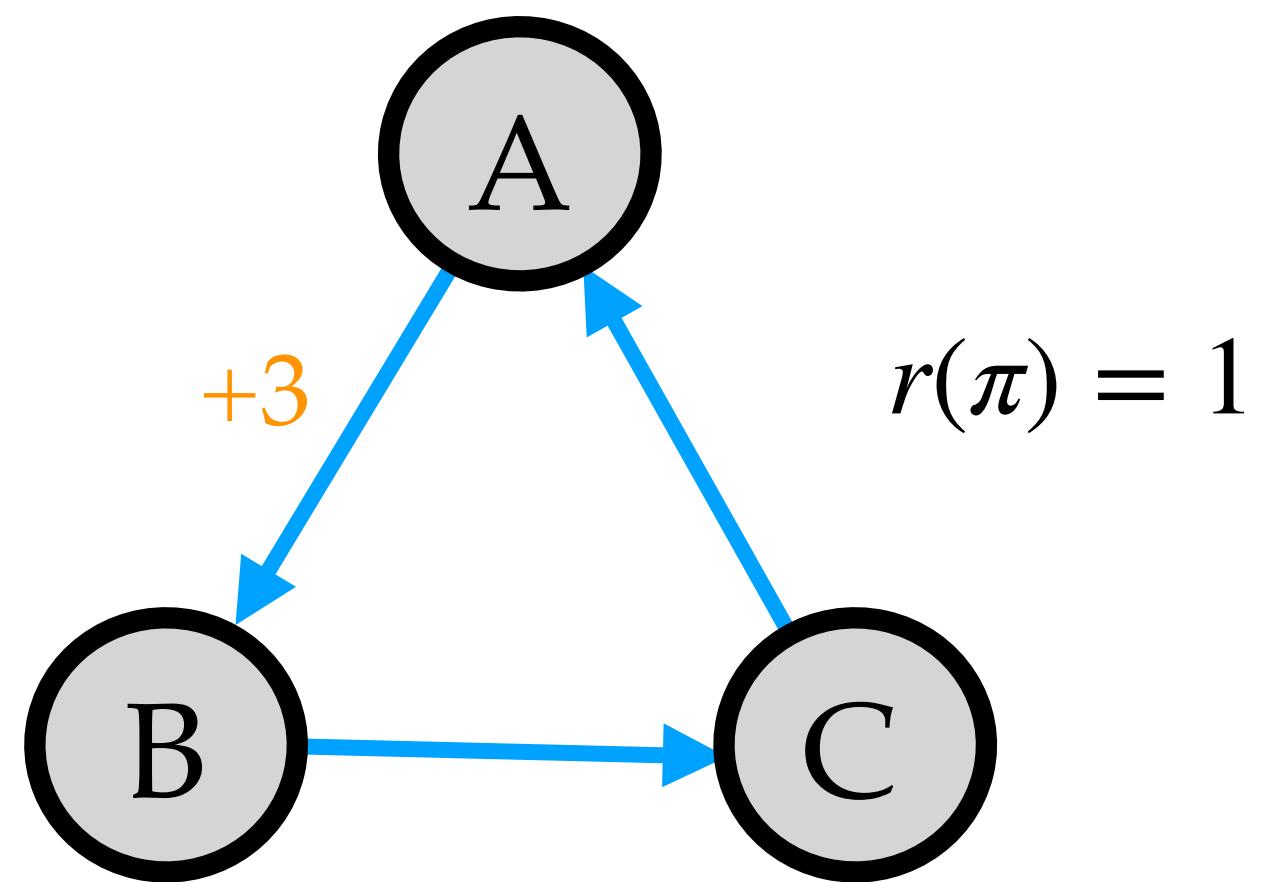
MORE INTUITION

$$\frac{\frac{r(\pi)}{1 - \gamma}}{\gamma}$$

$\gamma = 0.8$	5
$\gamma = 0.9$	10
$\gamma = 0.99$	100

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

$\tilde{v}_{\pi}^{\gamma}(s)$



	s_A	s_B	s_C	
Standard discounted values	$\gamma = 0.8$	6.15	3.93	4.92
	$\gamma = 0.9$	11.07	8.97	9.96
	$\gamma = 0.99$	101.01	98.99	99.99
Centered discounted values	$\gamma = 0.8$	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	

ESTIMATING $r(\pi)$

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

On-policy

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

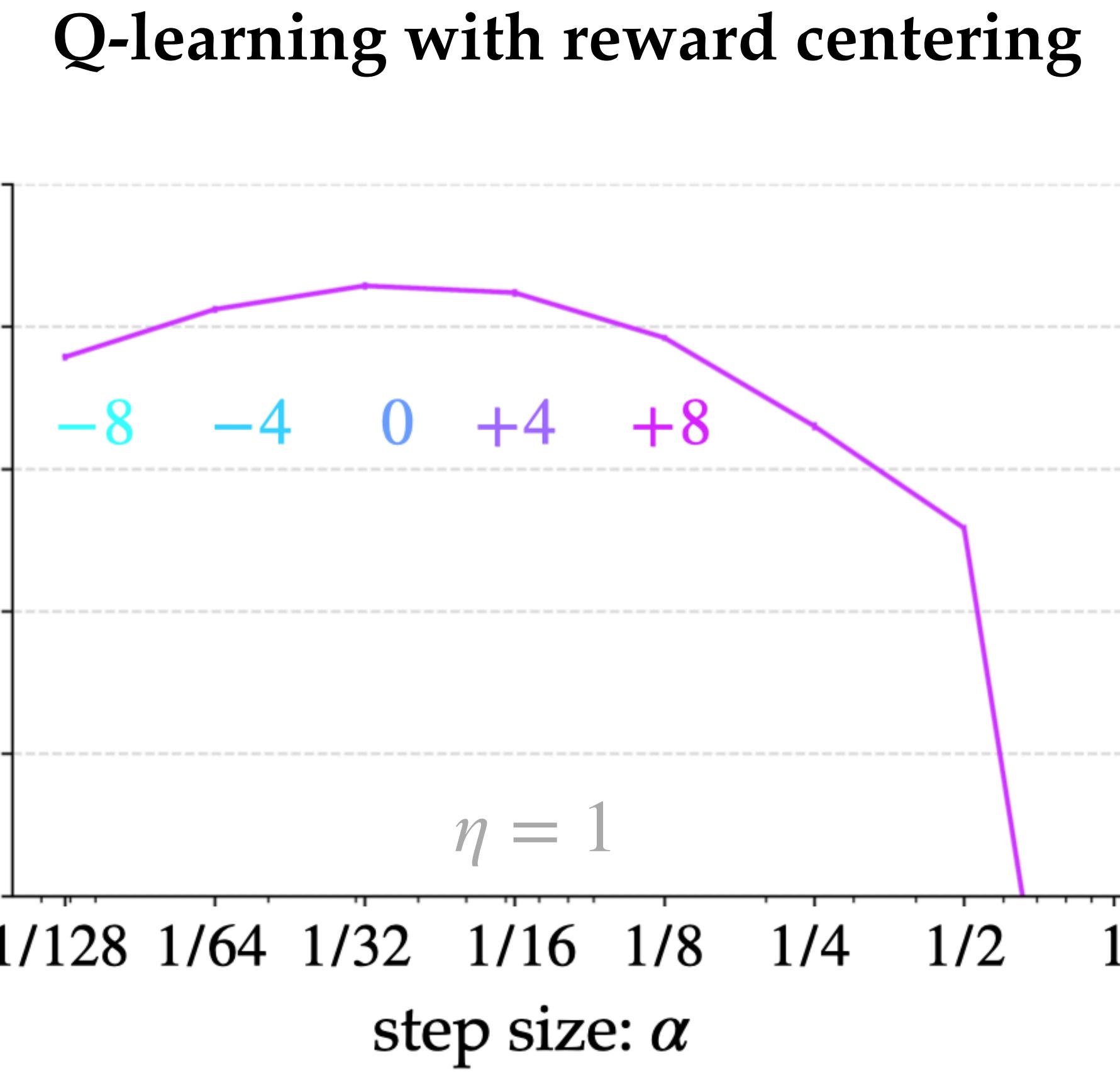
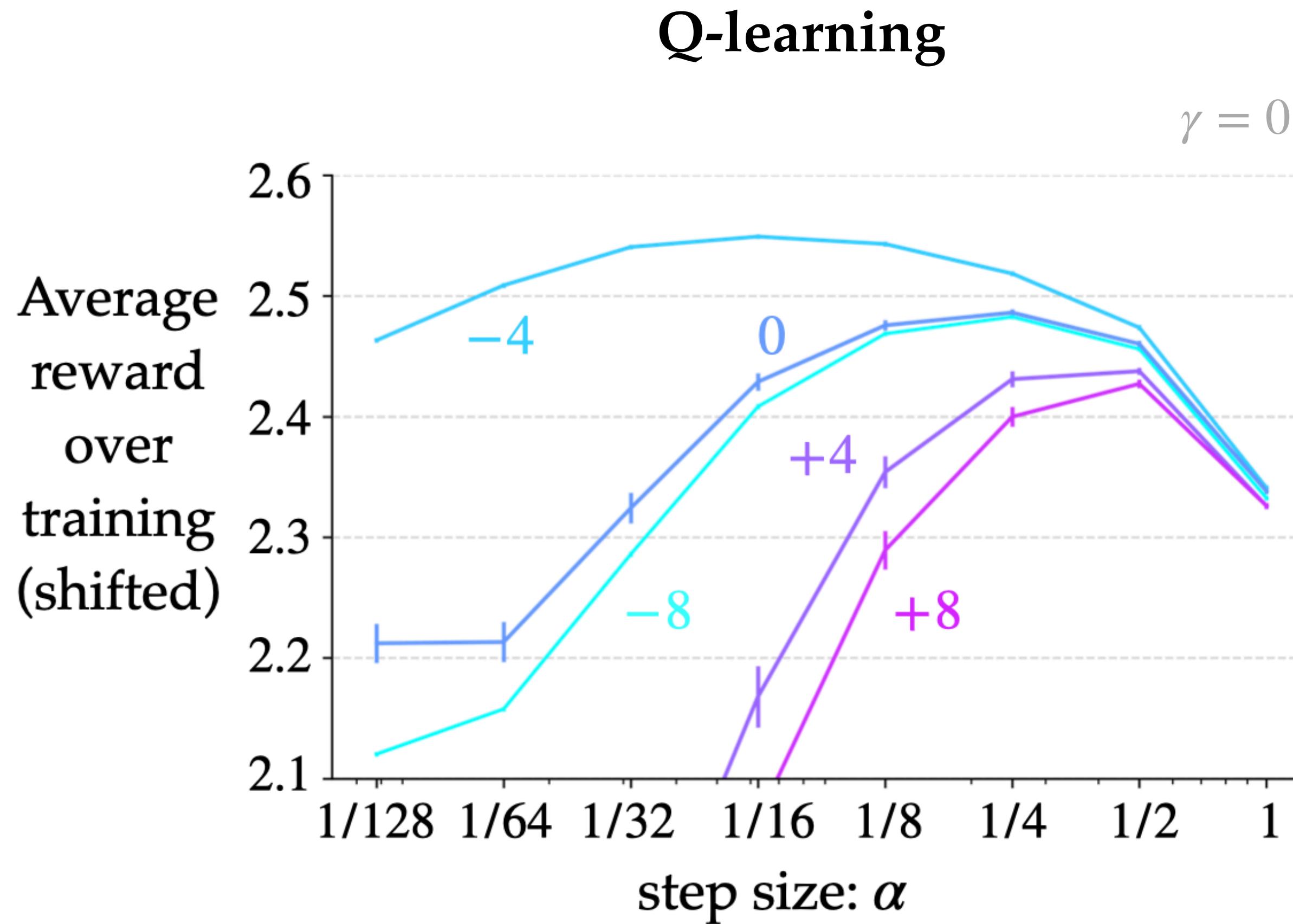
Off-policy

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

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

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



AccessControl (tabular)

TAKEAWAYS

- ▶ Reward centering can improve the performance of discounted methods for all discount factors, especially as $\gamma \rightarrow 1$.
- ▶ 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*. Under review.

OUTLINE

Problem setting

1. One-step average-reward methods
2. Multi-step average-reward methods
3. An idea to improve *discounted-reward* methods

Conclusions, limitations, and future work

Acknowledgments

OUTLINE

Problem setting

1. One-step average-reward methods
2. Multi-step average-reward methods
3. An idea to improve *discounted-reward* methods

Conclusions, limitations, and future work

Acknowledgments

OVERALL TAKEAWAYS

Contributions

- ▶ one-step tabular average-reward learning algorithms for on- and off-policy prediction and control
- ▶ multi-step average-reward learning algorithms for on- and off-policy prediction using eligibility traces
- ▶ the reward-centering idea to improve discounted-reward algorithms for continuing problems

Future Work

- ▶ Extensive empirical evaluation of average-reward and discounted-reward methods
- ▶ A suite of continuing problems
- ▶ Policy-based variants
- ▶ Model-based variants
- ▶ Exploration techniques for continuing problems

“To develop simple and practical learning algorithms from first principles for long-lived agents”

ACKNOWLEDGMENTS



THANK YOU

Questions?