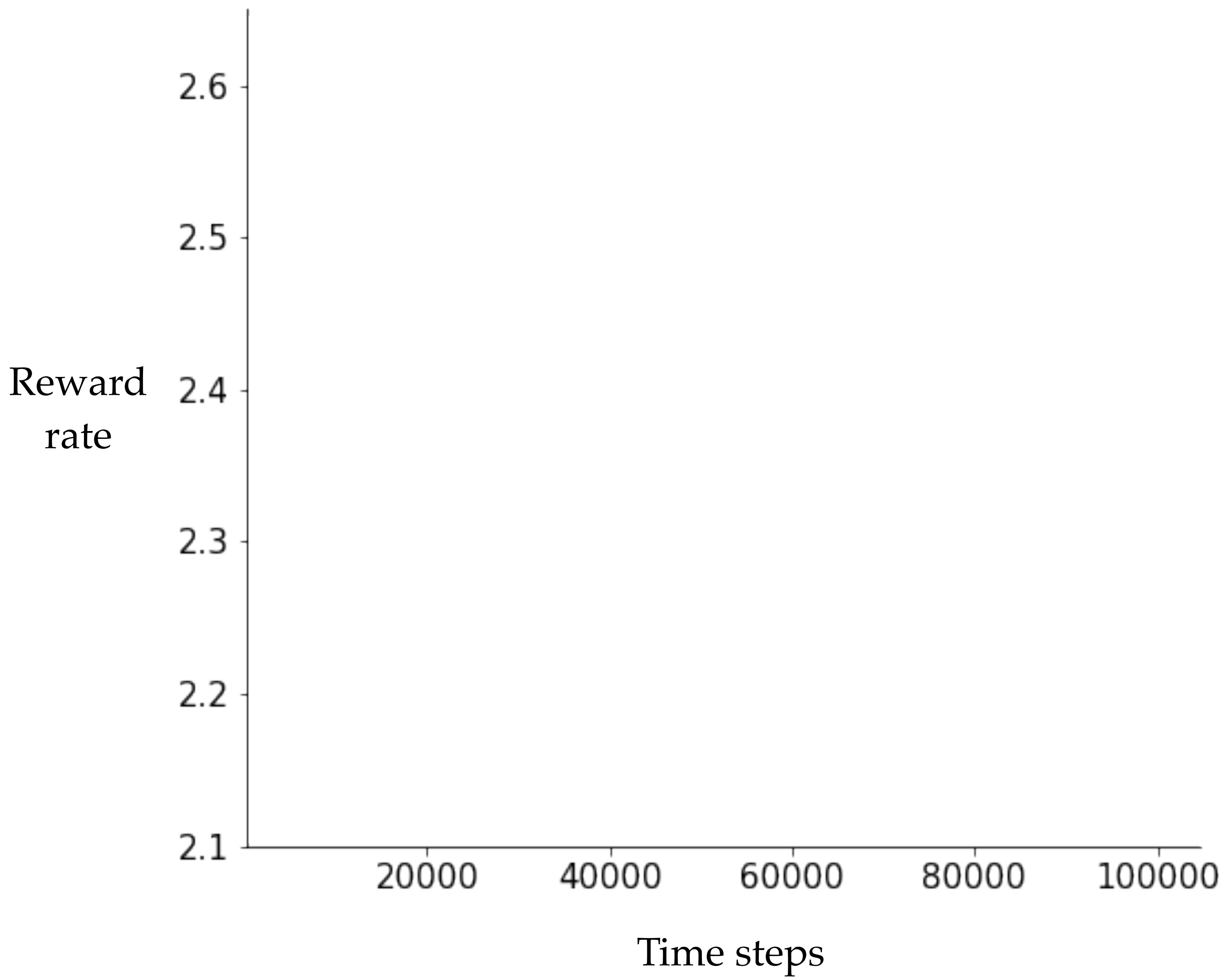
IMPROVING DISCOUNTING USING AVERAGE REWARD

Tea Time Talk
16 Aug 2023

Abhishek Naik

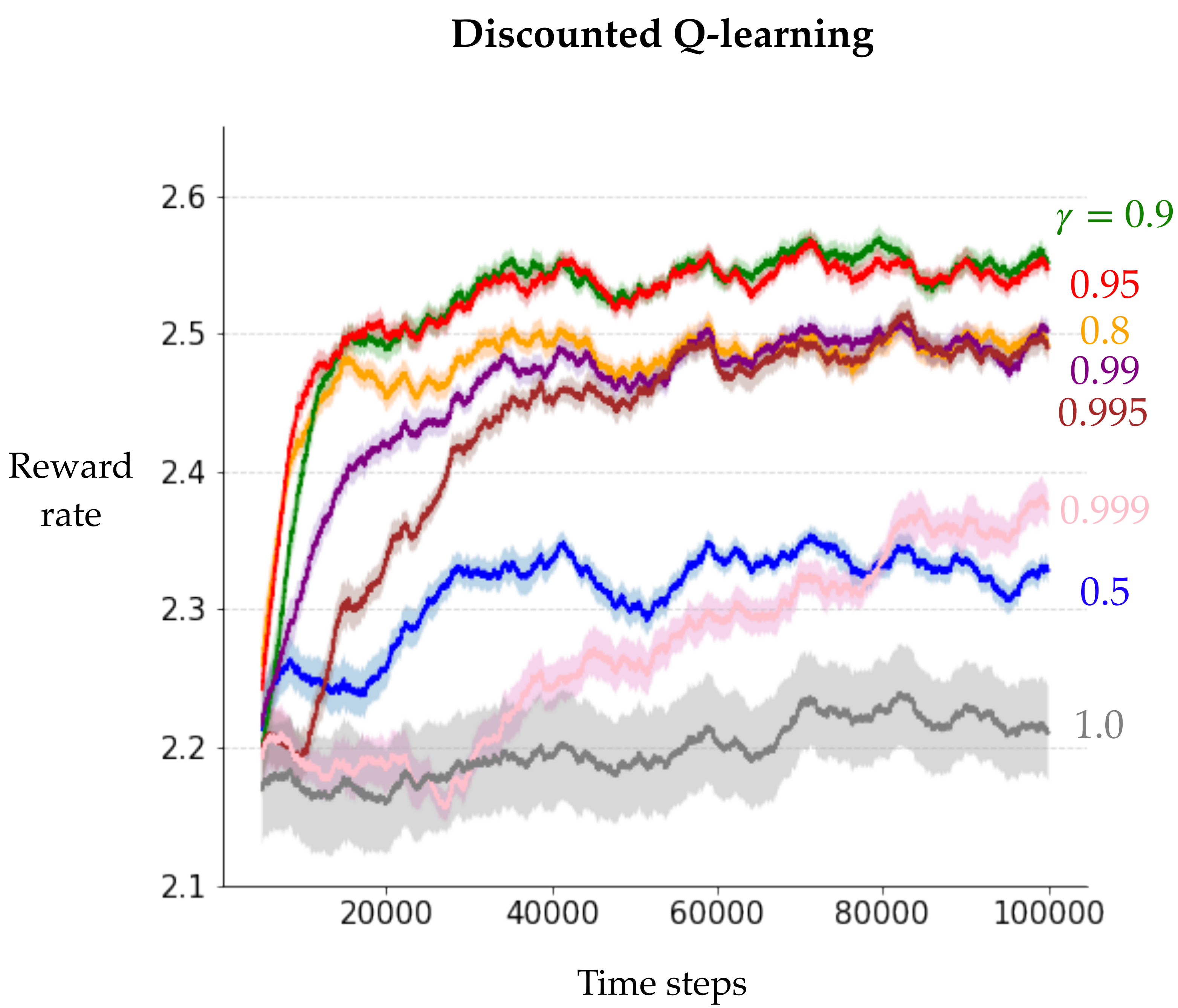
DISCOUNTING BECOMES UNSTABLE WHEN $\gamma \to 1$

DISCOUNTING BECOMES UNSTABLE WHEN $\gamma \to 1$



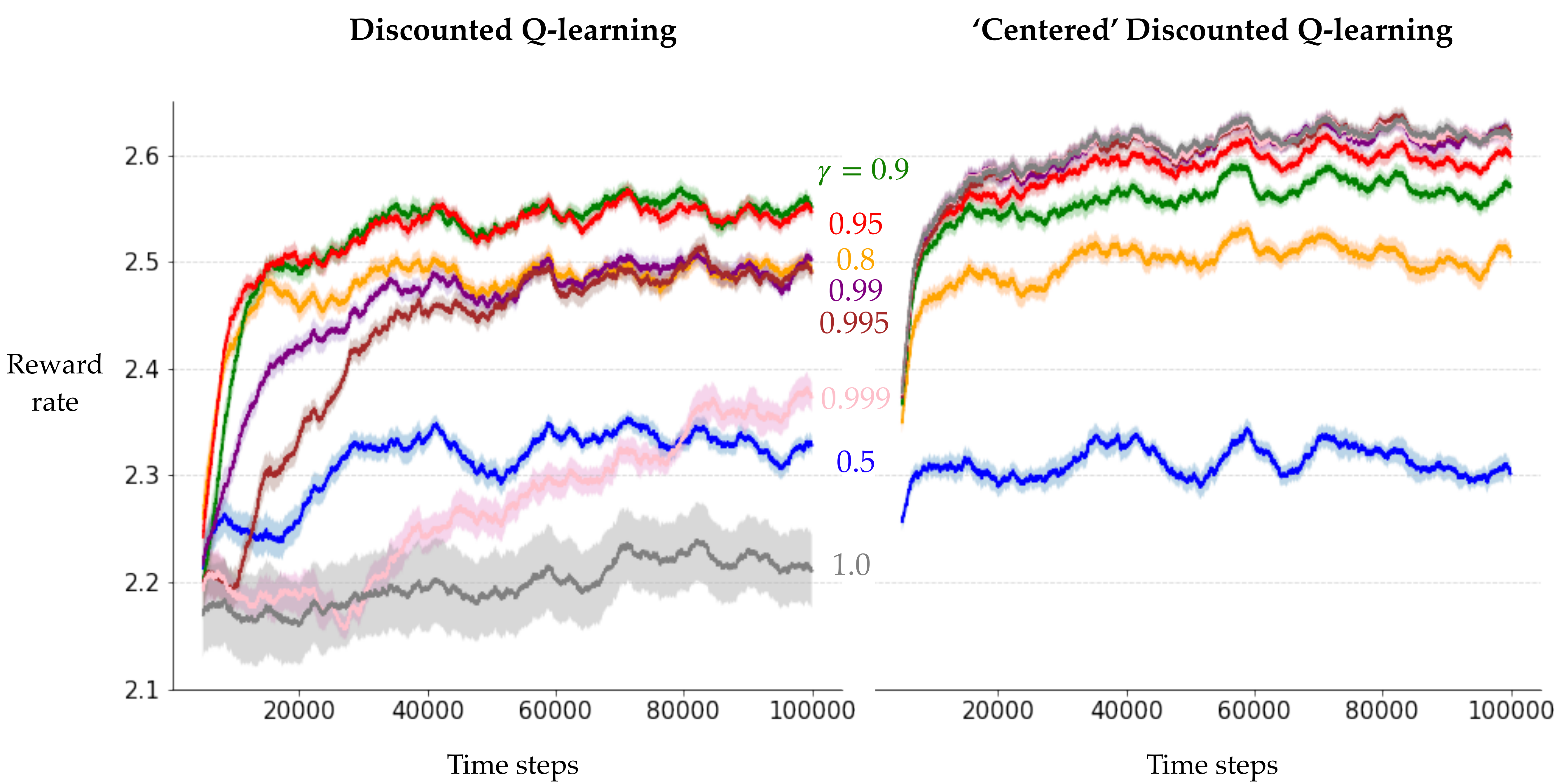
AccessControl (tabular)

DISCOUNTING BECOMES UNSTABLE WHEN $\gamma \to 1$



AccessControl (tabular)

DISCOUNTING BECOMES UNSTABLE WHEN $\gamma \rightarrow 1$

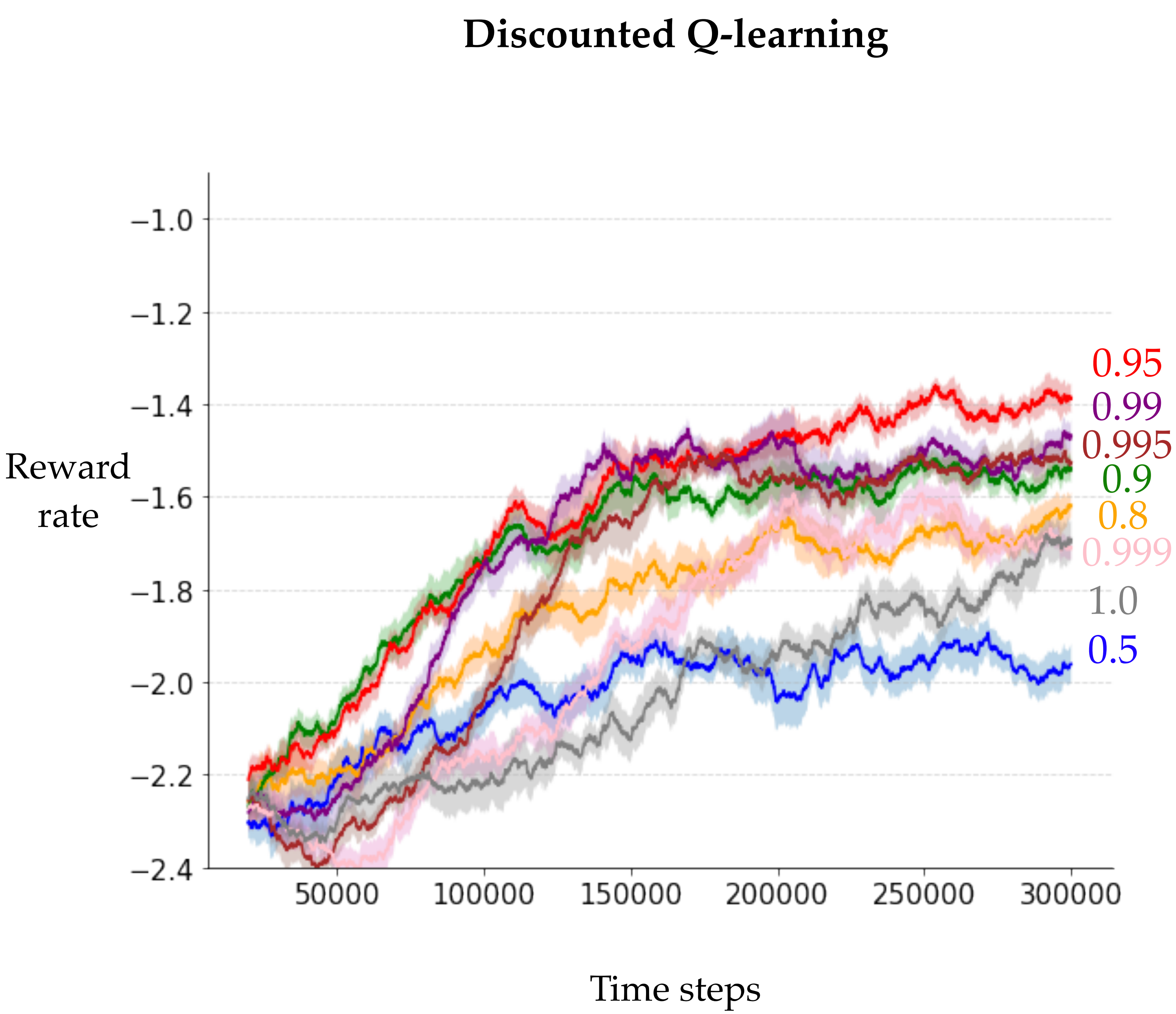


AccessControl (tabular)

SIMILAR TRENDS FOR LINEAR AND NON-LINEAR FA

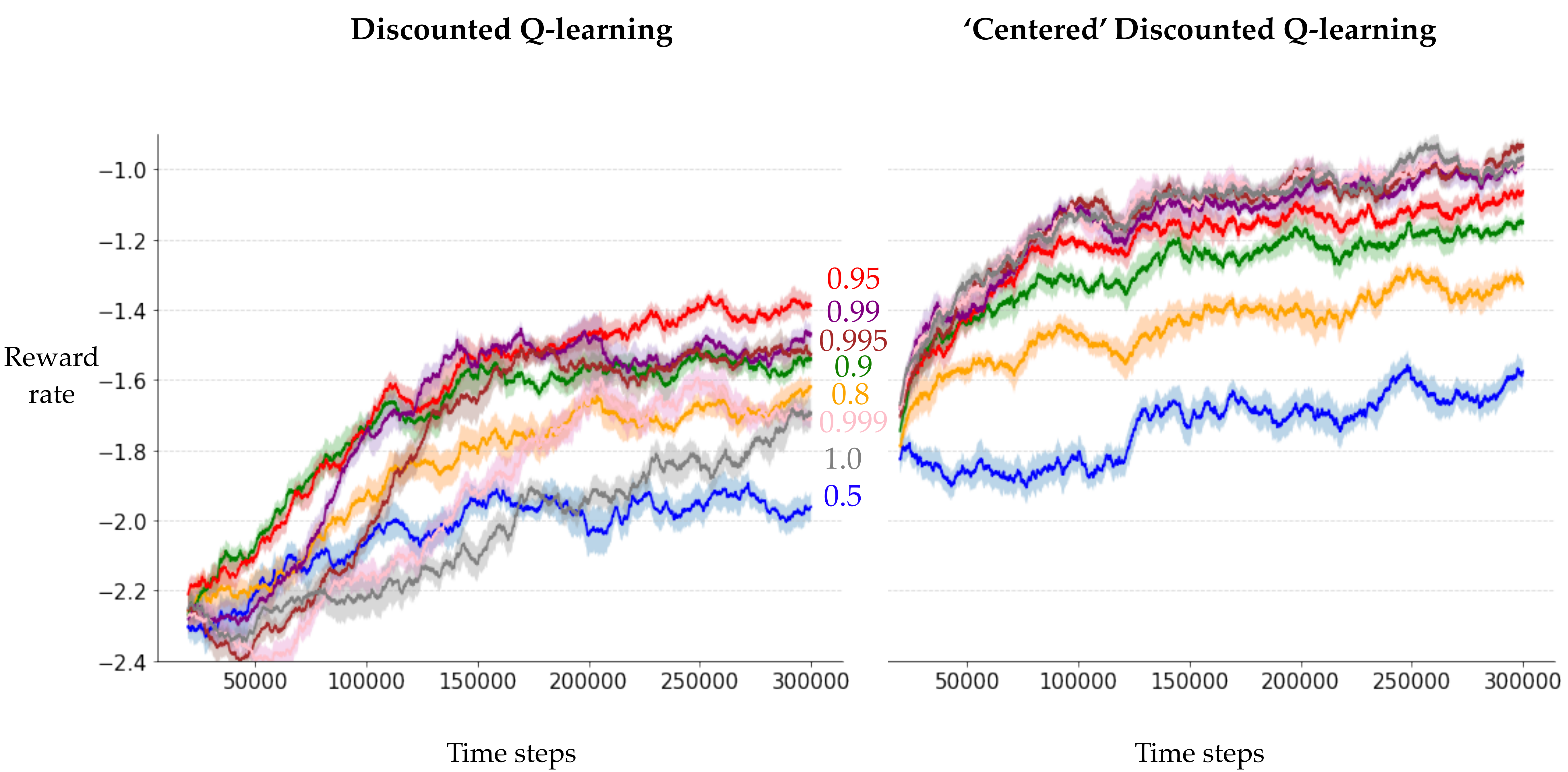
PuckWorld (linear FA)

SIMILAR TRENDS FOR LINEAR AND NON-LINEAR FA



PuckWorld (linear FA)

SIMILAR TRENDS FOR LINEAR AND NON-LINEAR FA

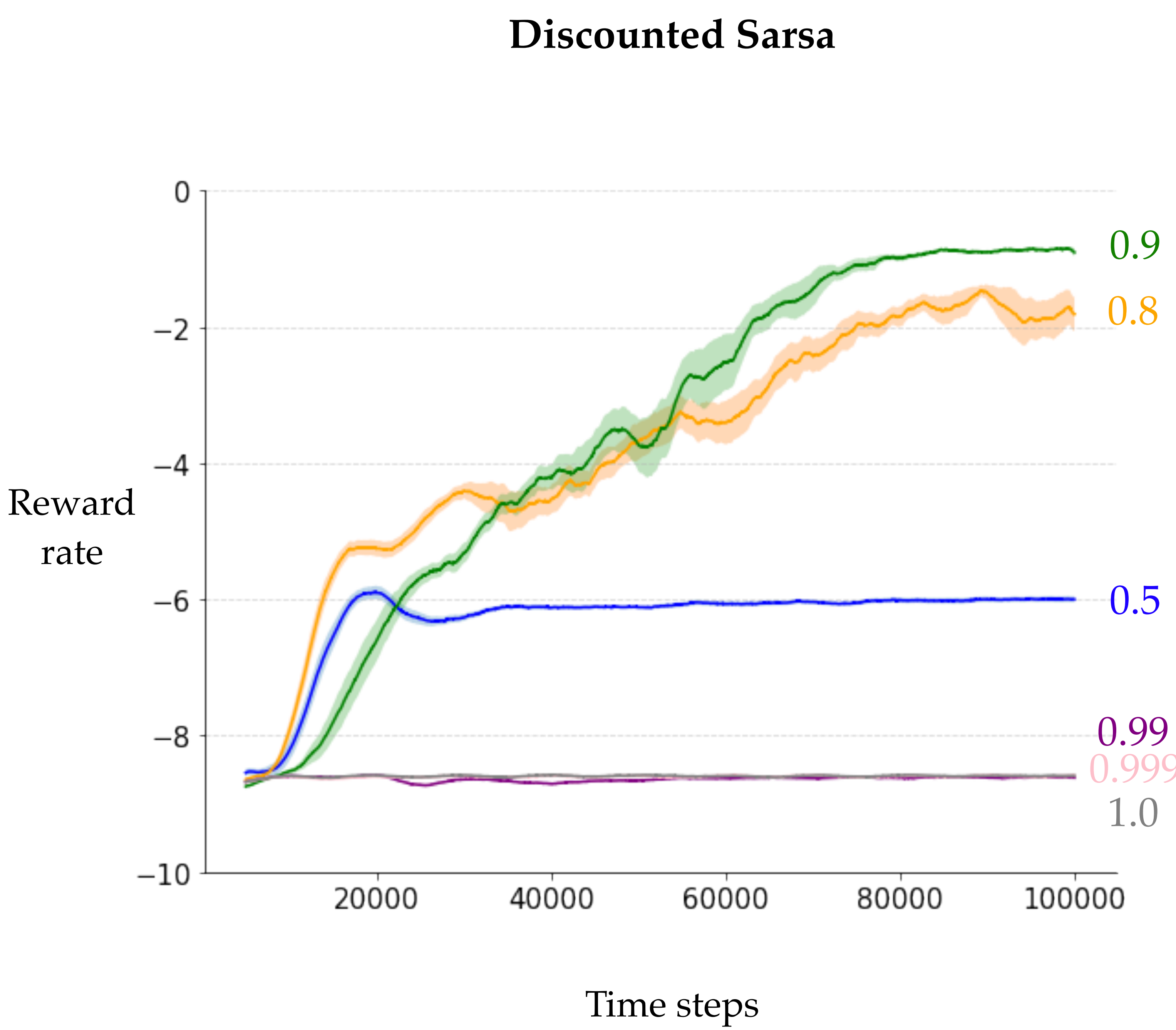


PuckWorld (linear FA)

SIMILAR TRENDS FOR LINEAR AND NON-LINEAR FA

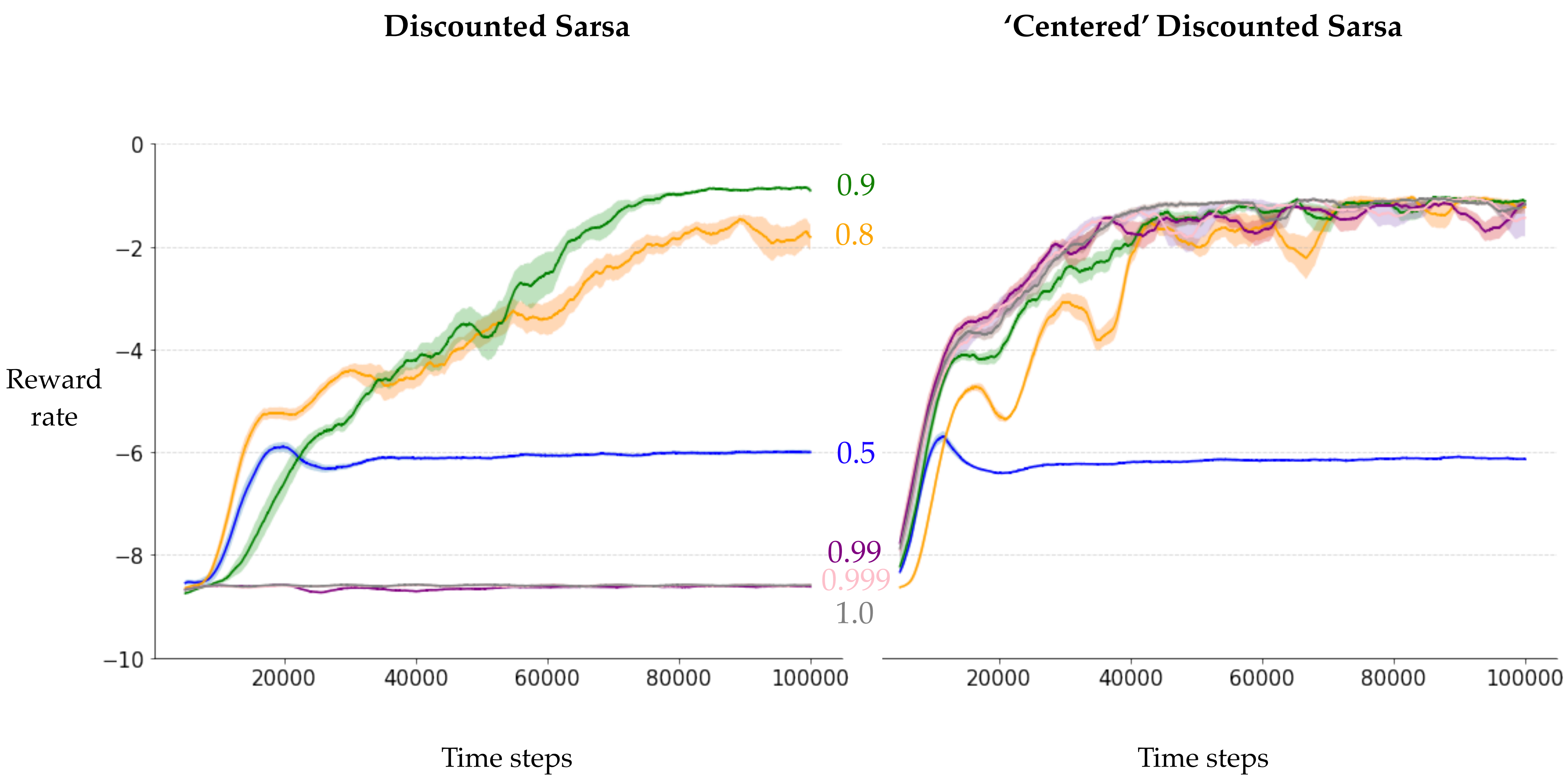
Pendulum (non-linear FA)

SIMILAR TRENDS FOR LINEAR AND NON-LINEAR FA



Pendulum (non-linear FA)

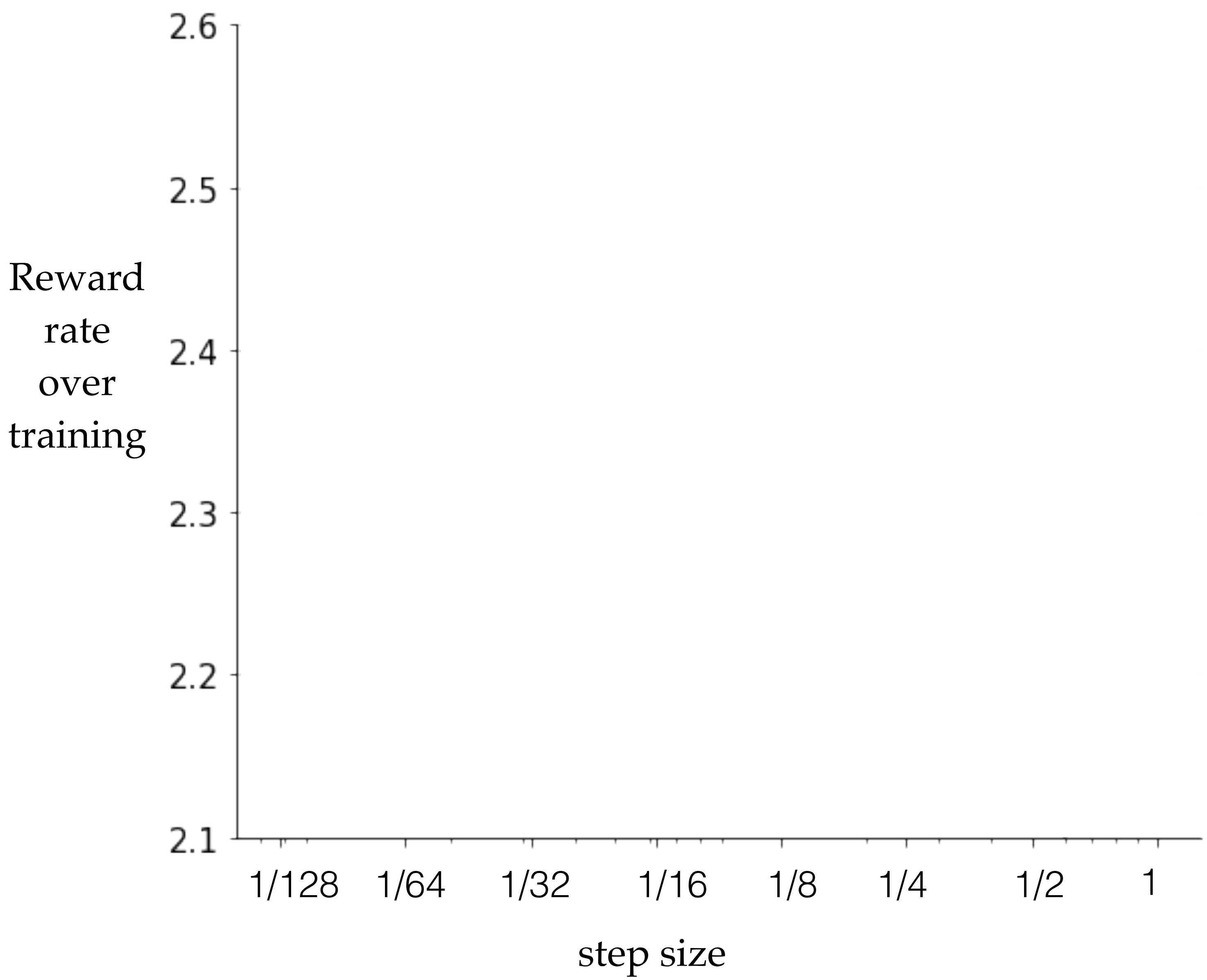
SIMILAR TRENDS FOR LINEAR AND NON-LINEAR FA



Pendulum (non-linear FA)

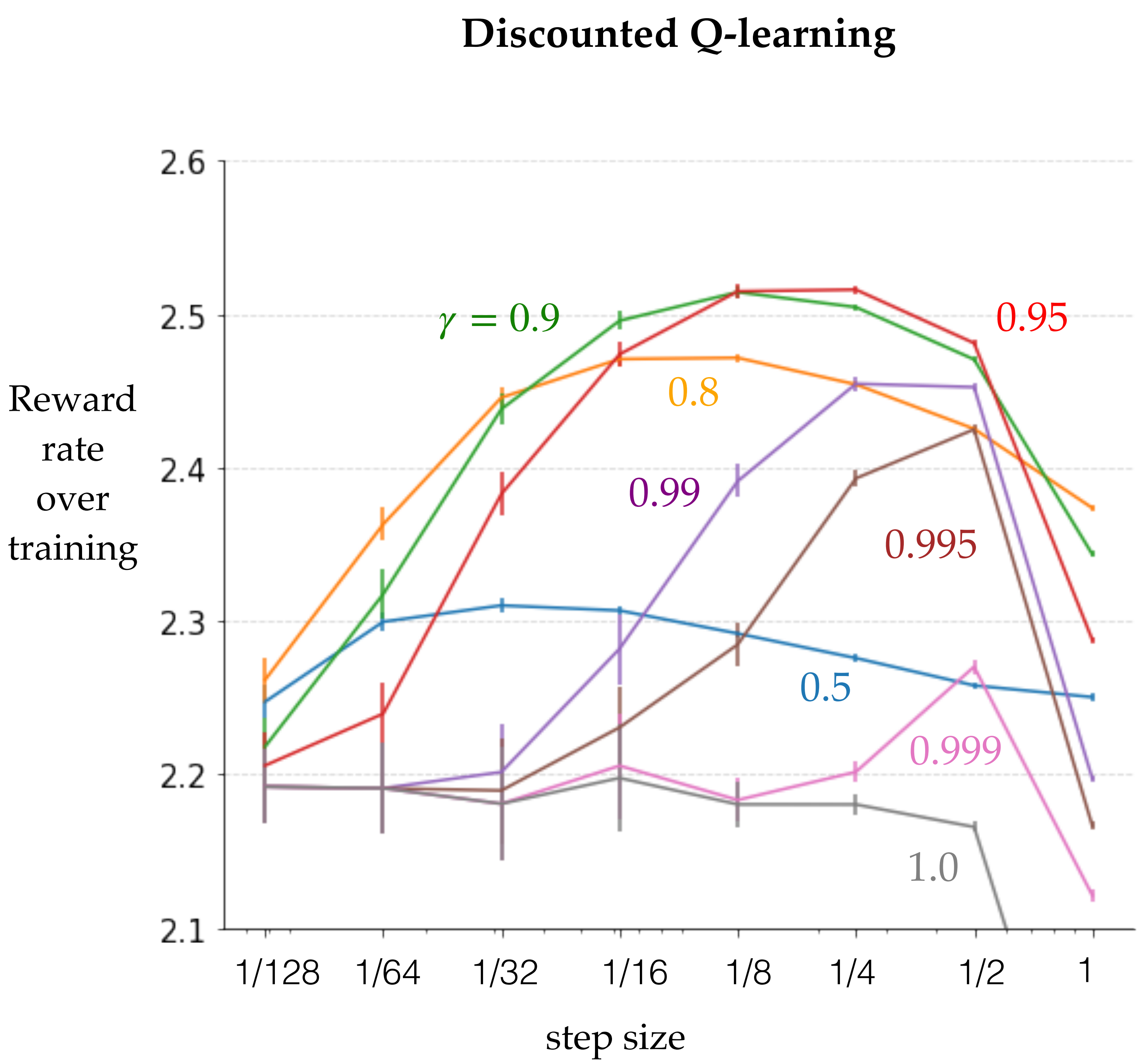
TRENDS ARE CONSISTENT ACROSS PARAMETERS

TRENDS ARE CONSISTENT ACROSS PARAMETERS



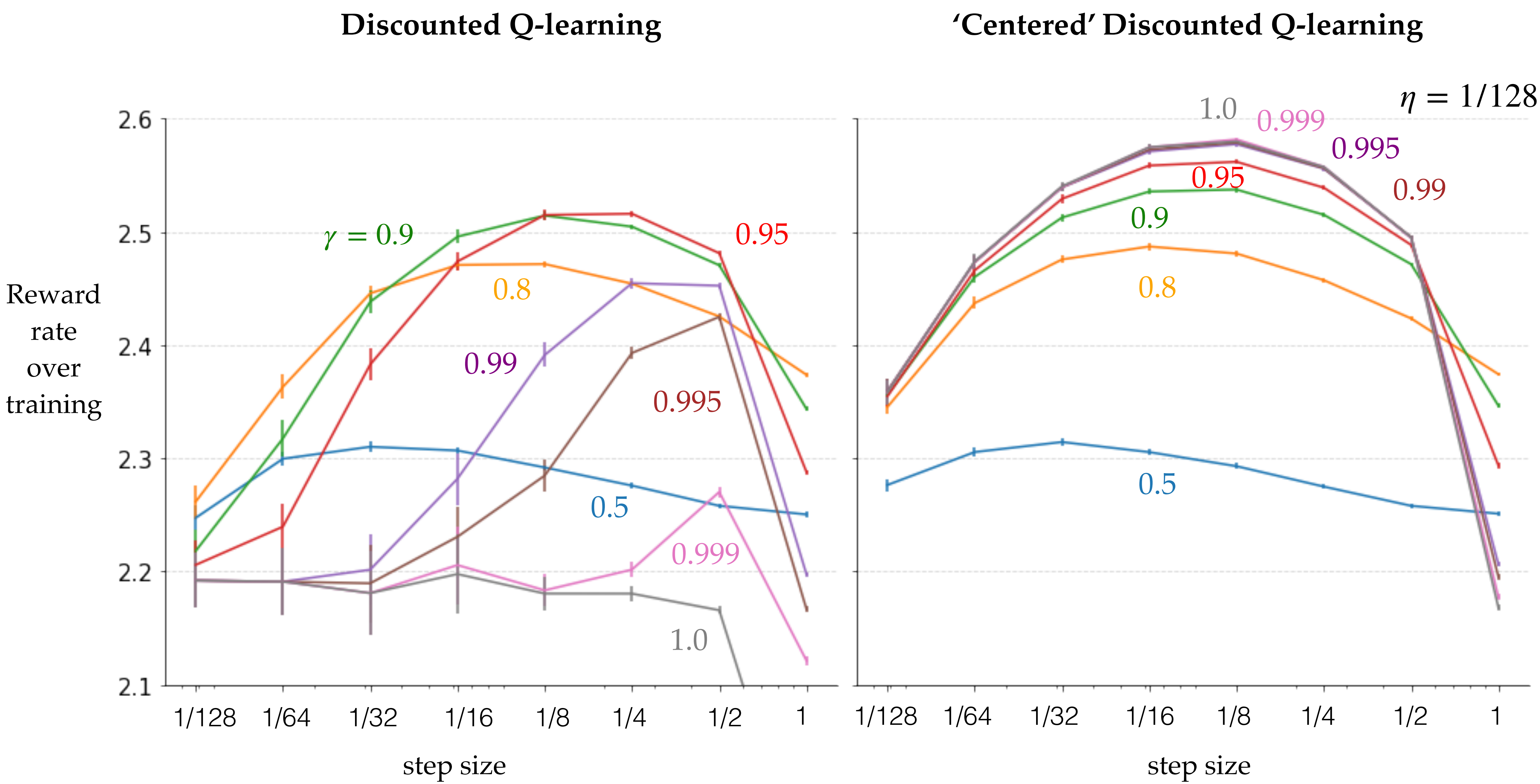
AccessControl (tabular)

TRENDS ARE CONSISTENT ACROSS PARAMETERS



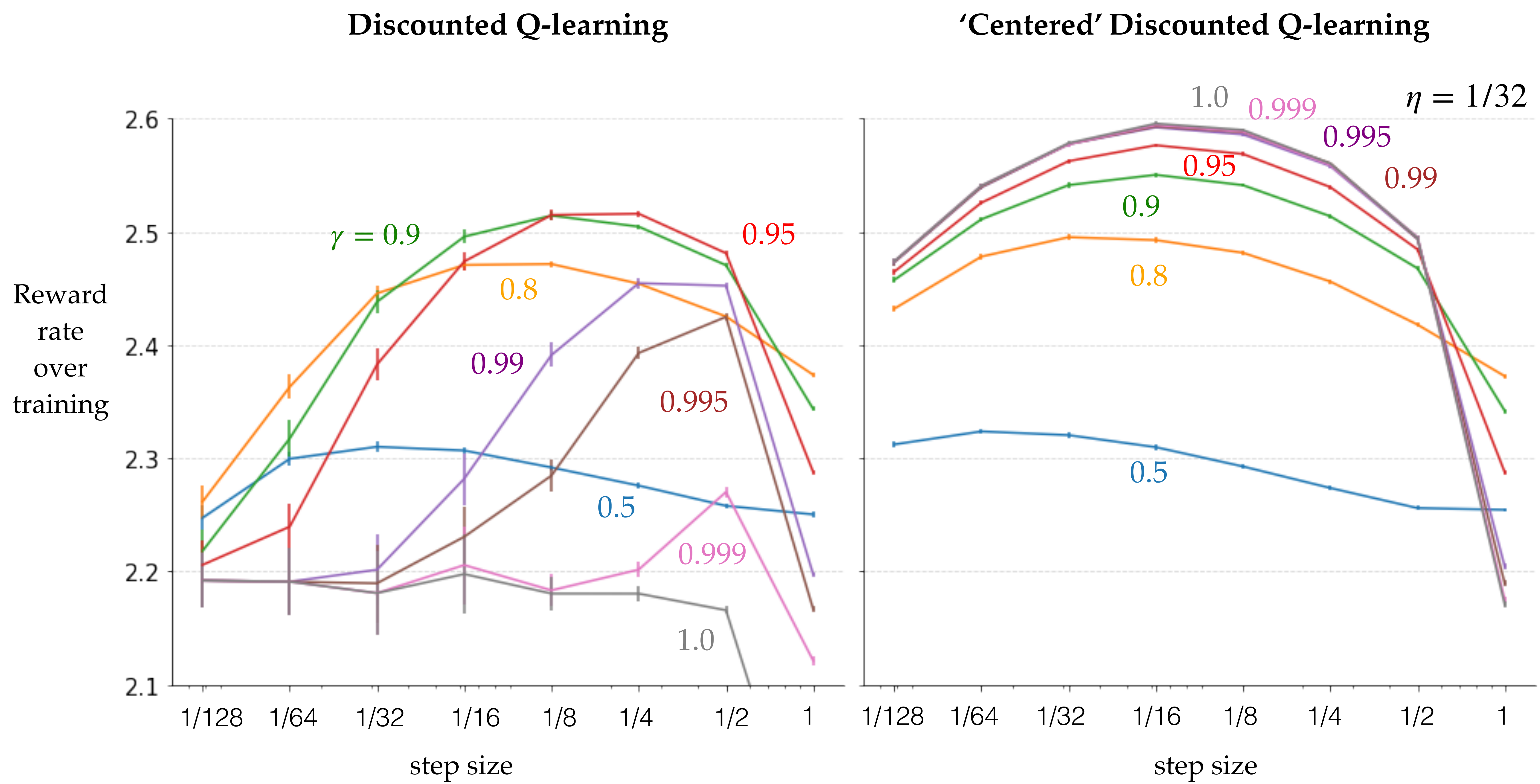
AccessControl (tabular)

TRENDS ARE CONSISTENT ACROSS PARAMETERS



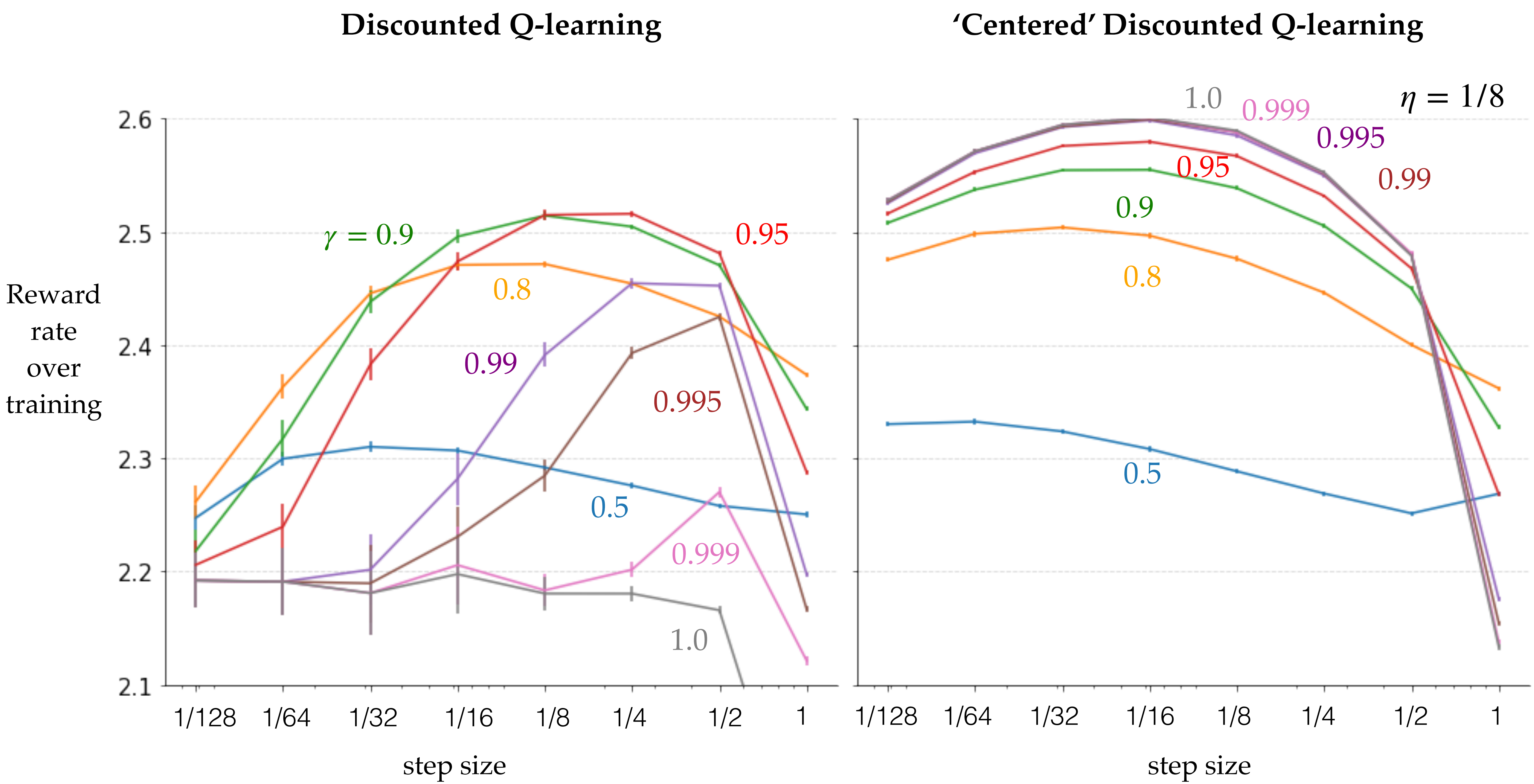
AccessControl (tabular)

TRENDS ARE CONSISTENT ACROSS PARAMETERS



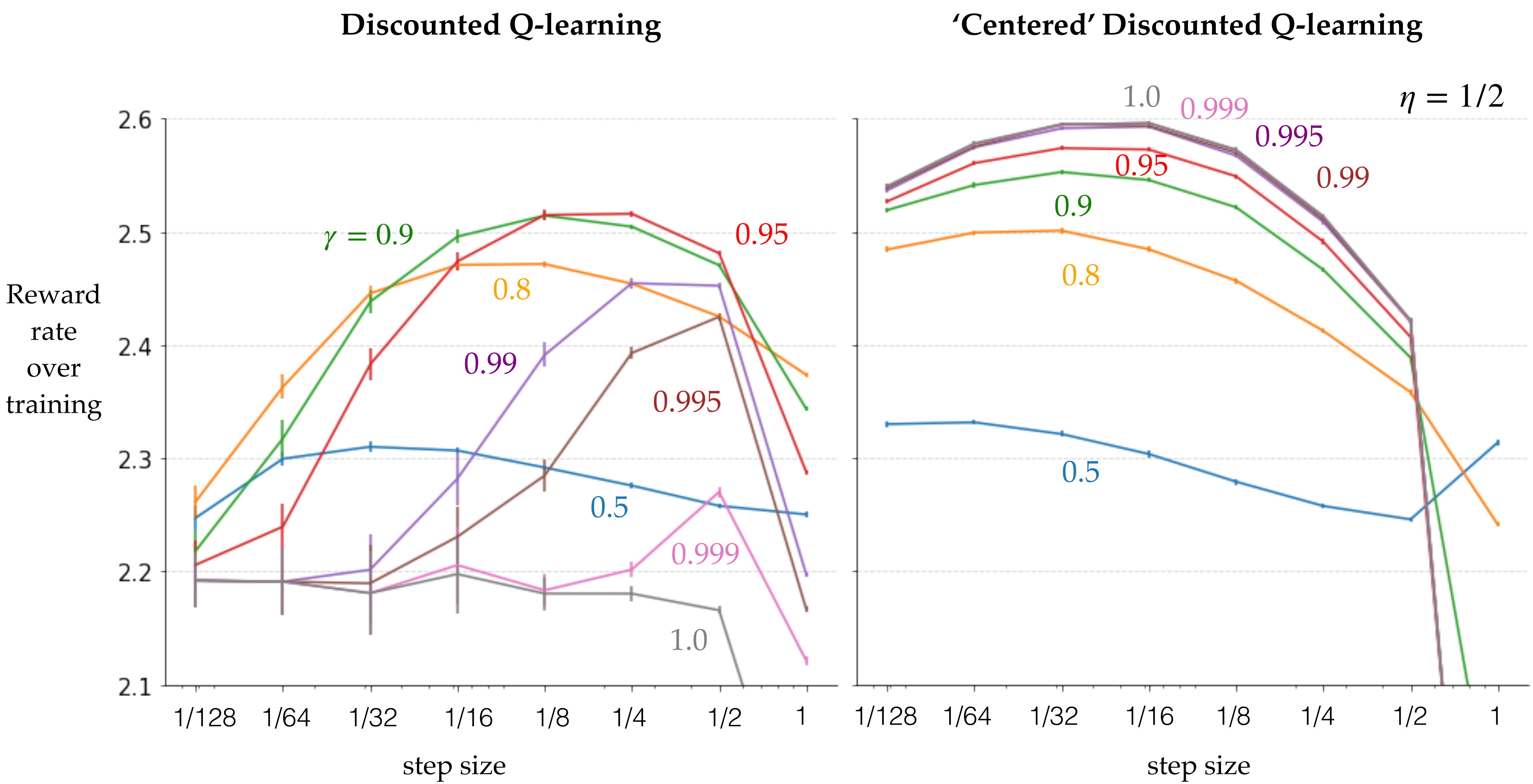
AccessControl (tabular)

TRENDS ARE CONSISTENT ACROSS PARAMETERS



AccessControl (tabular)

TRENDS ARE CONSISTENT ACROSS PARAMETERS



AccessControl (tabular)

KEY INSIGHT

KEY INSIGHT

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

KEY INSIGHT

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

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

Discounted value function

KEY INSIGHT

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

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

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

Discounted value function

KEY INSIGHT

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

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

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

Discounted value function

Laurent-series expansion

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

KEY INSIGHT

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

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

Discounted value function

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

Laurent-series expansion

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

KEY INSIGHT

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

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

Discounted value function

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

Laurent-series expansion

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

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

KEY INSIGHT

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

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

Discounted value function

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

Laurent-series expansion

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

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

KEY INSIGHT

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

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

← Discounted value function

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

Laurent-series expansion →

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

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

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

KEY INSIGHT

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

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

Discounted value function

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

Laurent-series expansion

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

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

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

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

KEY INSIGHT

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

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

Discounted value function

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

Laurent-series expansion

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

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

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

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

Centered discounted value function

ESTIMATING $r(\pi)$

ESTIMATING $r(\pi)$

- On-policy:

ESTIMATING $r(\pi)$

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

- On-policy:

ESTIMATING $r(\pi)$

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

- On-policy:
  - sample average of rewards

ESTIMATING $r(\pi)$

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

- On-policy:
  - sample average of rewards

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

ESTIMATING $r(\pi)$

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

- On-policy:
  - sample average of rewards

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

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

ESTIMATING $r(\pi)$

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

- On-policy:
  - sample average of rewards

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

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

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

ESTIMATING $r(\pi)$

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

- On-policy:
  - sample average of rewards

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

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

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

- Off-policy: ??

ESTIMATING $r(\pi)$

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

- On-policy:
  - sample average of rewards

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

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

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

- Off-policy: ??

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

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$\bar{v}_\pi(s) = \sum_a \pi(a \,|\, s) \sum_{s',r} p(s', r \,|\, s, a)\big[r - r(\pi) + \bar{v}_\pi(s')\big]$$

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$\bar{v}_\pi(s) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s', r\,|\,s, a)\Big[r - r(\pi) + \bar{v}_\pi(s')\Big]$$

$$v_\pi^\gamma(s) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s', r\,|\,s, a)\Big[r + \gamma \bar{v}_\pi(s')\Big]$$

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$\bar{v}_\pi(s) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s', r\,|\,s, a)\Big[r - r(\pi) + \bar{v}_\pi(s')\Big]$$

$$v_\pi^\gamma(s) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s', r\,|\,s, a)\Big[r + \gamma \bar{v}_\pi(s')\Big]$$

$$\downarrow$$

$$V_{t+1}(S_t) \doteq V_t(S_t) + \alpha\Big(R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)\Big)$$

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$\bar{v}_\pi(s) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s', r\,|\,s, a)\big[r - r(\pi) + \bar{v}_\pi(s')\big]$$

$$v_\pi^\gamma(s) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s', r\,|\,s, a)\big[r + \gamma\bar{v}_\pi(s')\big]$$

$$\downarrow$$

$$V_{t+1}(S_t) \doteq V_t(S_t) + \alpha\big(R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)\big)$$

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$\bar{v}_\pi(s) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s', r\,|\,s, a)\big[r - r(\pi) + \bar{v}_\pi(s')\big]$$

$$v_\pi^\gamma(s) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s', r\,|\,s, a)\big[r + \gamma \bar{v}_\pi(s')\big]$$

$$\downarrow$$

$$V_{t+1}(S_t) \doteq V_t(S_t) + \alpha\Big(\underbrace{R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)}_{\delta_t}\Big)$$

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$\bar{v}_\pi(s) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s', r\,|\,s, a)\big[r - r(\pi) + \bar{v}_\pi(s')\big]$$

$$\bar{V}_{t+1}(S_t) \doteq \bar{V}_t(S_t) + \alpha\,\delta_t$$

$$v^\gamma_\pi(s) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s', r\,|\,s, a)\big[r + \gamma\bar{v}_\pi(s')\big]$$

$$\downarrow$$

$$V_{t+1}(S_t) \doteq V_t(S_t) + \alpha\Big(\underbrace{R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)}_{\delta_t}\Big)$$

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$\bar{v}_\pi(s) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s',r\,|\,s,a)\big[r - r(\pi) + \bar{v}_\pi(s')\big]$$

$$\bar{V}_{t+1}(S_t) \doteq \bar{V}_t(S_t) + \alpha\,\delta_t$$

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

$$v_\pi^\gamma(s) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s',r\,|\,s,a)\big[r + \gamma\bar{v}_\pi(s')\big]$$

$$\downarrow$$

$$V_{t+1}(S_t) \doteq V_t(S_t) + \alpha\big(\underbrace{R_{t+1} + \gamma V_t(S_{t+1}) - V_t(S_t)}_{\delta_t}\big)$$

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$\bar{v}_\pi(s) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s', r\,|\,s, a)\Big[r - r(\pi) + \bar{v}_\pi(s')\Big]$$

$$\bar{V}_{t+1}(S_t) \doteq \bar{V}_t(S_t) + \alpha\,\delta_t$$

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

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$\bar{v}_\pi(s) = \sum_a \pi(a\,|\,s)\sum_{s',r} p(s', r\,|\,s, a)\Big[r + \bar{v}_\pi(s')\Big] - r(\pi)$$

$$\bar{V}_{t+1}(S_t) \doteq \bar{V}_t(S_t) + \alpha\,\delta_t$$

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

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$r(\pi) + \bar{v}_\pi(s) = \sum_a \pi(a \,|\, s) \sum_{s',r} p(s', r \,|\, s, a)\big[r + \bar{v}_\pi(s')\big] - r(\pi) + r(\pi)$$

$$\bar{V}_{t+1}(S_t) \doteq \bar{V}_t(S_t) + \alpha\,\delta_t$$

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

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$r(\pi) + \bar{v}_\pi(s) = \sum_a \pi(a\,|\,s)\sum_{s',r} p(s', r\,|\,s, a)\Big[r + \bar{v}_\pi(s')\Big]$$

$$\bar{V}_{t+1}(S_t) \doteq \bar{V}_t(S_t) + \alpha\,\delta_t$$

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

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$r(\pi) + \bar{v}_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)\big[r + \bar{v}_\pi(s')\big] - \bar{v}_\pi(s)$$

$$- \bar{v}_\pi(s)$$

$$\bar{V}_{t+1}(S_t) \doteq \bar{V}_t(S_t) + \alpha\,\delta_t$$

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

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$r(\pi) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s', r\,|\,s, a)\Big[r + \bar{v}_\pi(s') - \bar{v}_\pi(s)\Big]$$

$$\bar{V}_{t+1}(S_t) \doteq \bar{V}_t(S_t) + \alpha\,\delta_t$$

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

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$r(\pi) = \sum_a \pi(a\,|\,s) \sum_{s',r} p(s', r\,|\,s, a)\Big[r + \bar{v}_\pi(s') - \bar{v}_\pi(s)\Big]$$

$$\bar{V}_{t+1}(S_t) \doteq \bar{V}_t(S_t) + \alpha\,\delta_t$$

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

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$r(\pi) = \sum_a \pi(a \,|\, s) \sum_{s',r} p(s', r \,|\, s, a)\big[r + \bar{v}_\pi(s') - \bar{v}_\pi(s)\big]$$

$$\bar{V}_{t+1}(S_t) \doteq \bar{V}_t(S_t) + \alpha\,\delta_t$$

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

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$r(\pi) = \sum_a \pi(a|s) \sum_{s',r} p(s', r | s, a)\Big[r + \bar{v}_\pi(s') - \bar{v}_\pi(s)\Big]$$

$$\bar{V}_{t+1}(S_t) \doteq \bar{V}_t(S_t) + \alpha\,\delta_t$$

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

new_estimate = old_estimate + stepsize * (new_target - old_estimate)

$$\bar{R}_{t+1} \doteq \bar{R}_t + \beta\big(R_{t+1} + \bar{V}_t(S_{t+1}) - \bar{V}_t(S_t) - \bar{R}_t\big)$$

ESTIMATING $r(\pi)$ IN THE AVG-REWARD FORMULATION

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

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

$$r(\pi) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)\Big[r + \bar{v}_\pi(s') - \bar{v}_\pi(s)\Big]$$

$$\bar{V}_{t+1}(S_t) \doteq \bar{V}_t(S_t) + \alpha\,\delta_t$$

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

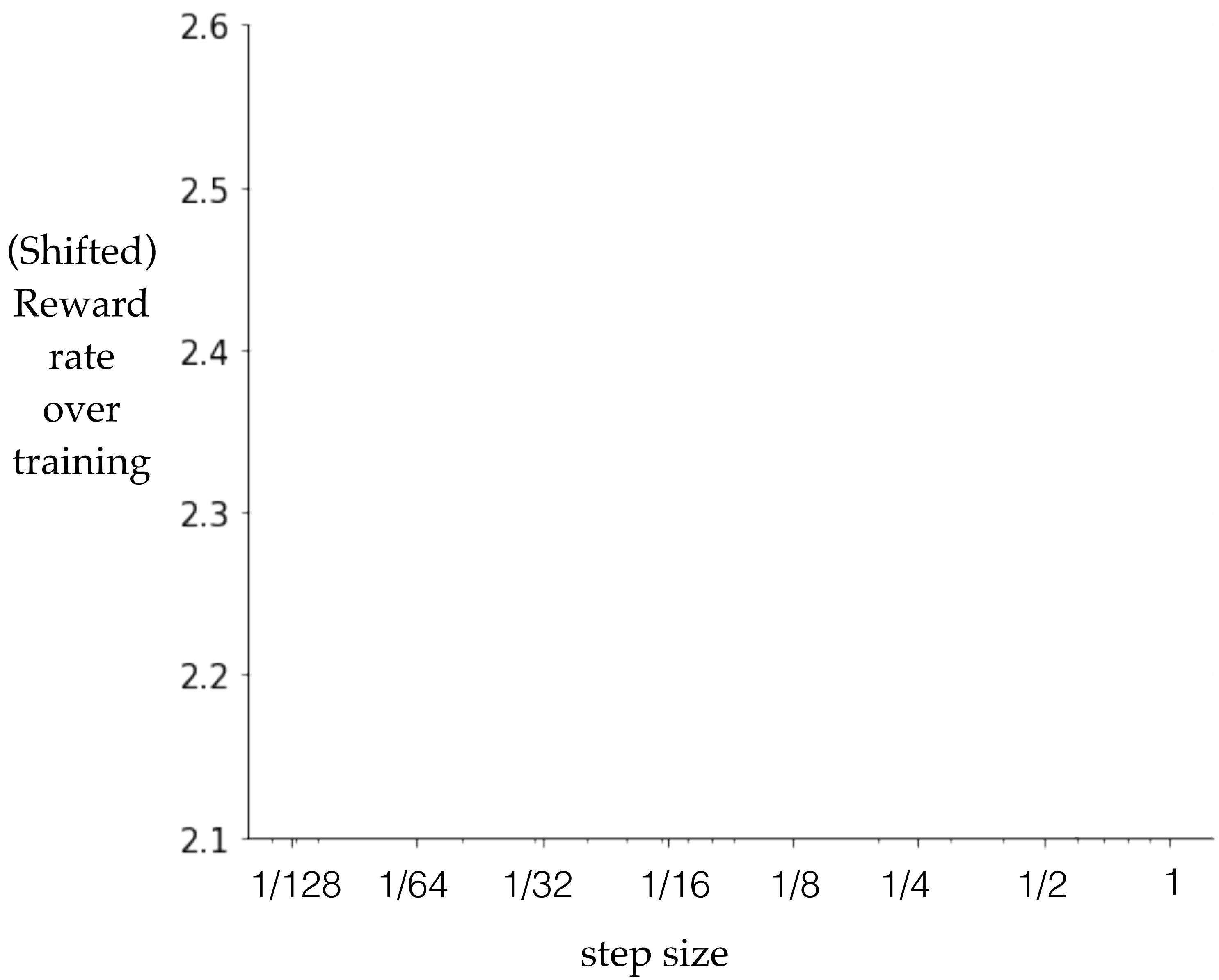
new_estimate = old_estimate + stepsize * (new_target - old_estimate)

$$\bar{R}_{t+1} \doteq \bar{R}_t + \beta\big(R_{t+1} + \bar{V}_t(S_{t+1}) - \bar{V}_t(S_t) - \bar{R}_t\big)$$

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

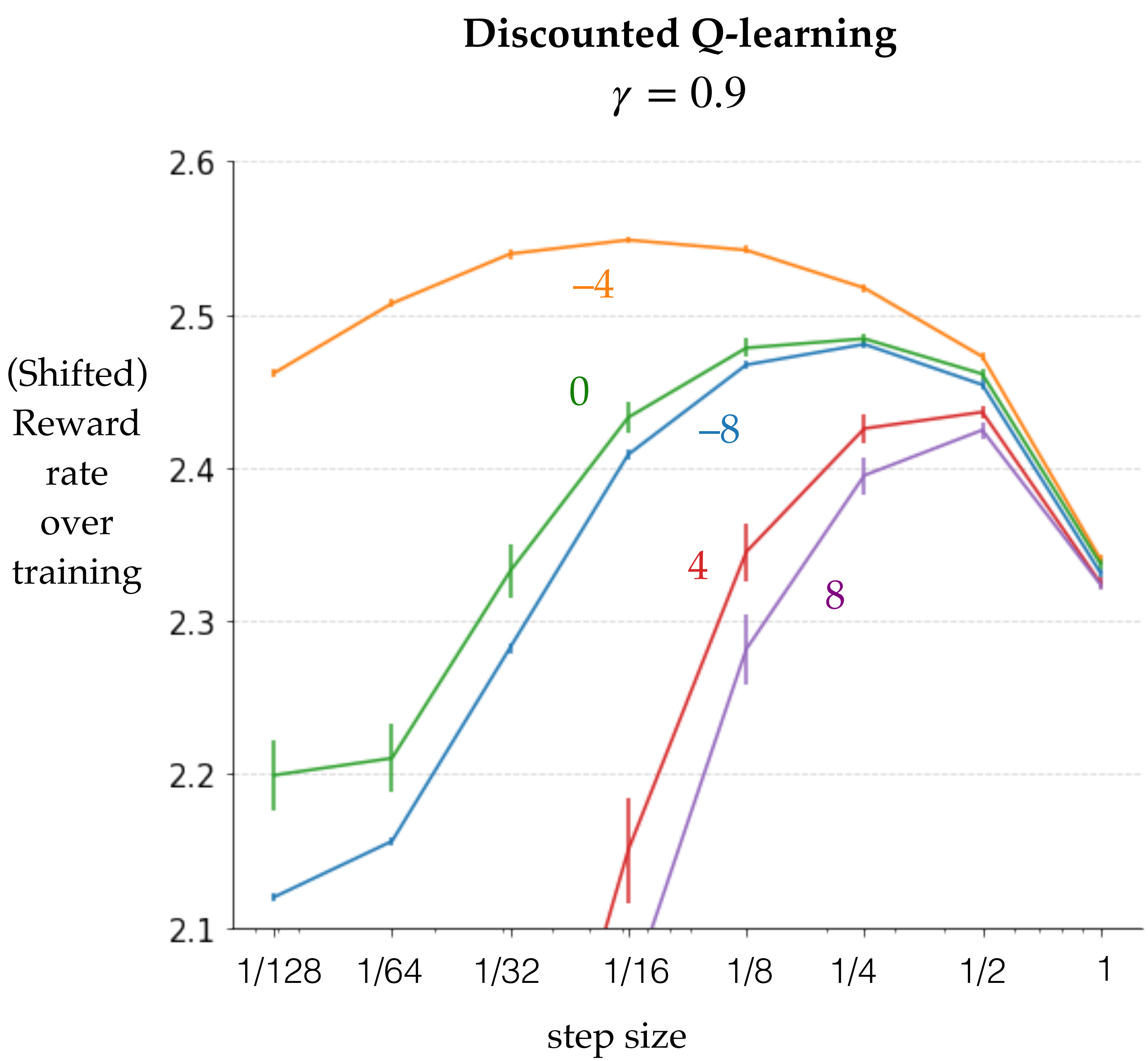
SIMILAR EFFECT WITH SHIFTED REWARDS

SIMILAR EFFECT WITH SHIFTED REWARDS



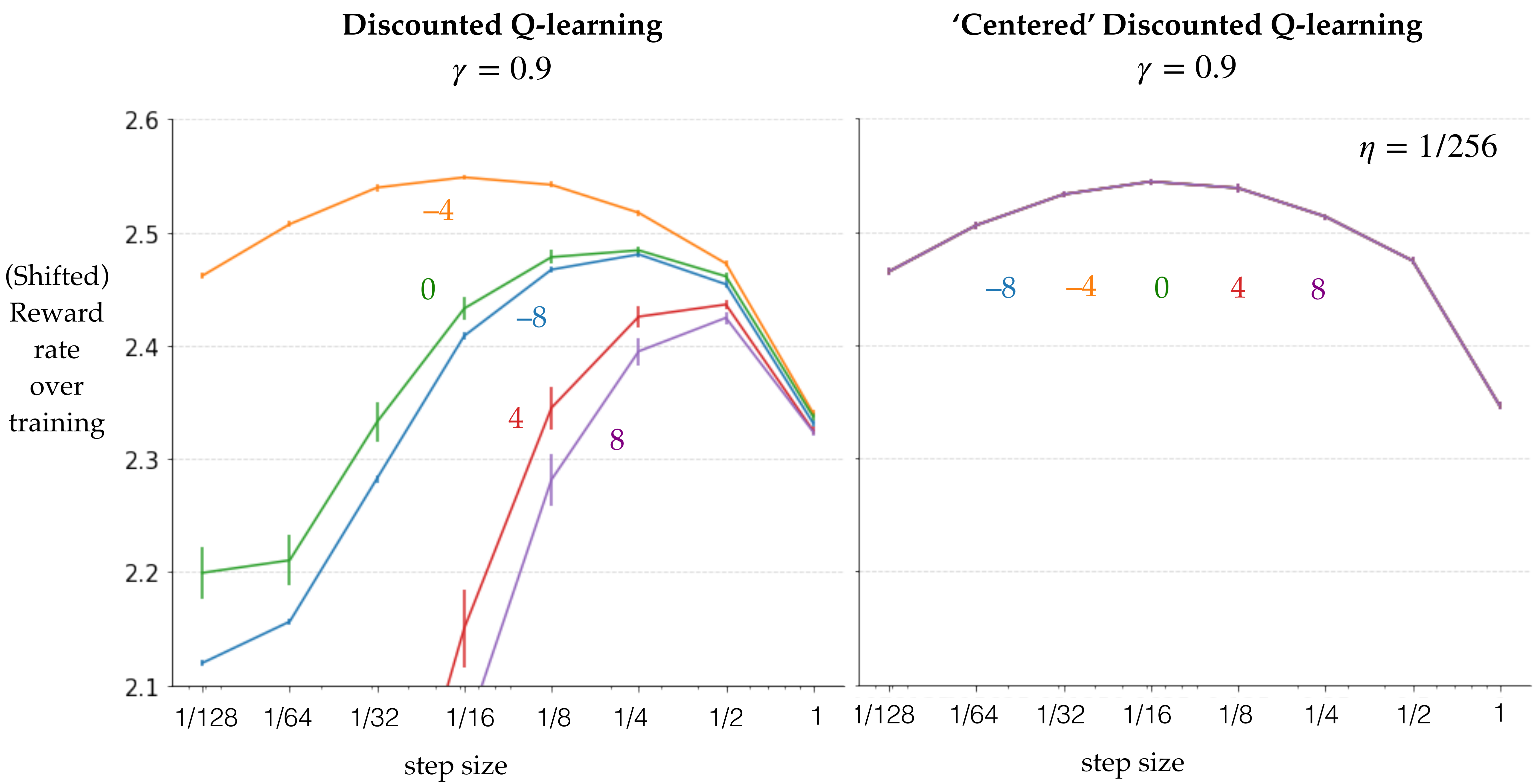
AccessControl (tabular)

SIMILAR EFFECT WITH SHIFTED REWARDS



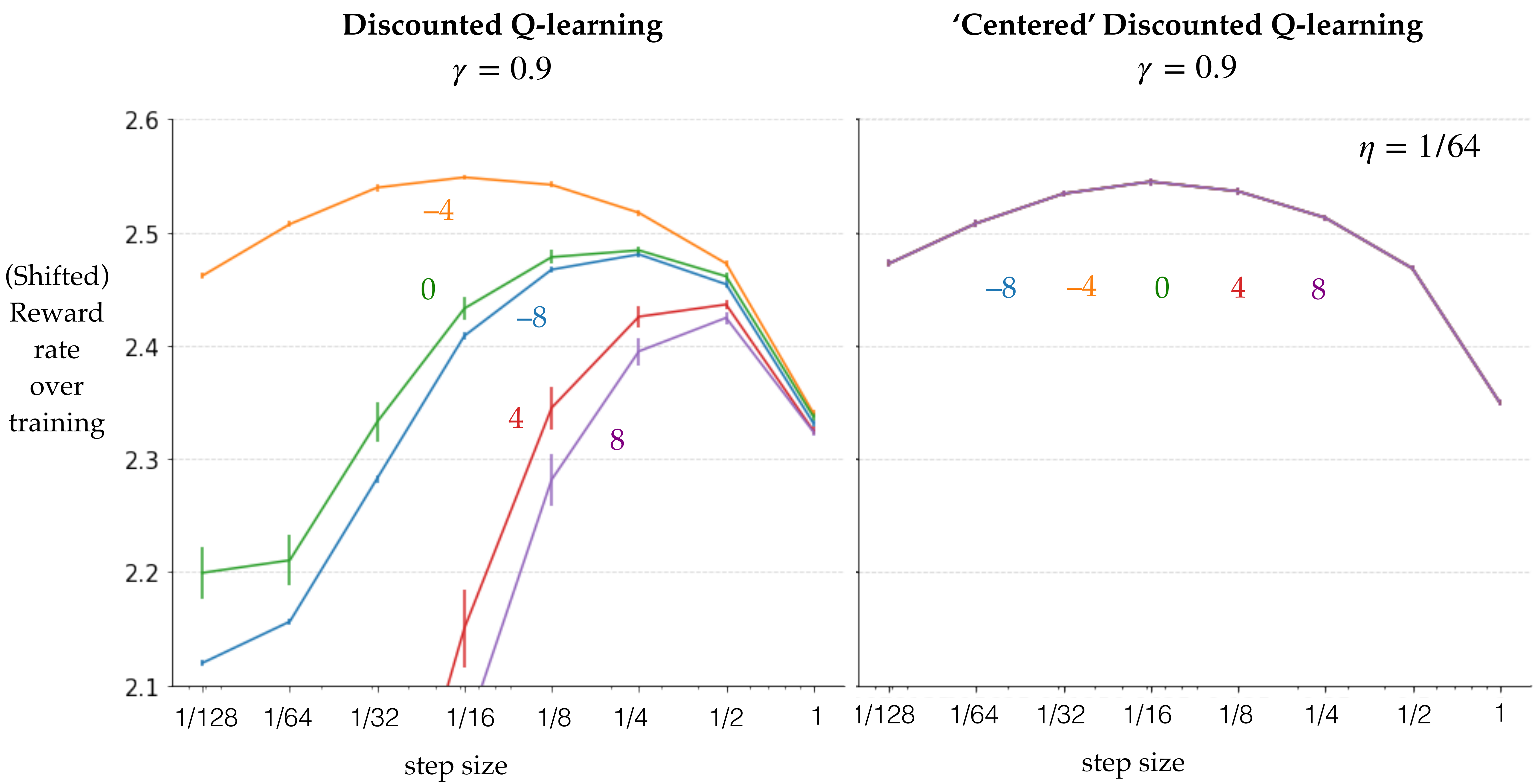
AccessControl (tabular)

SIMILAR EFFECT WITH SHIFTED REWARDS



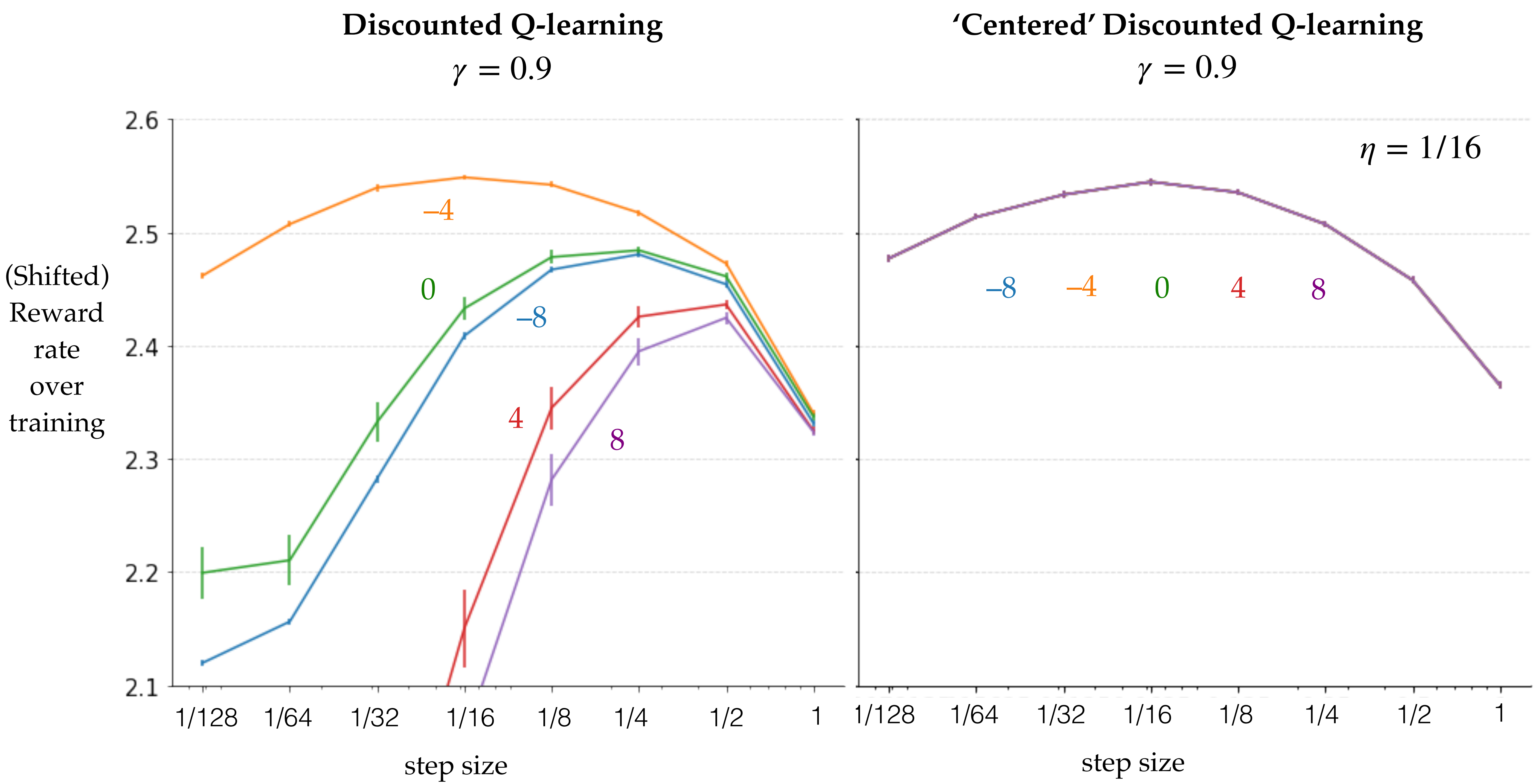
AccessControl (tabular)

SIMILAR EFFECT WITH SHIFTED REWARDS



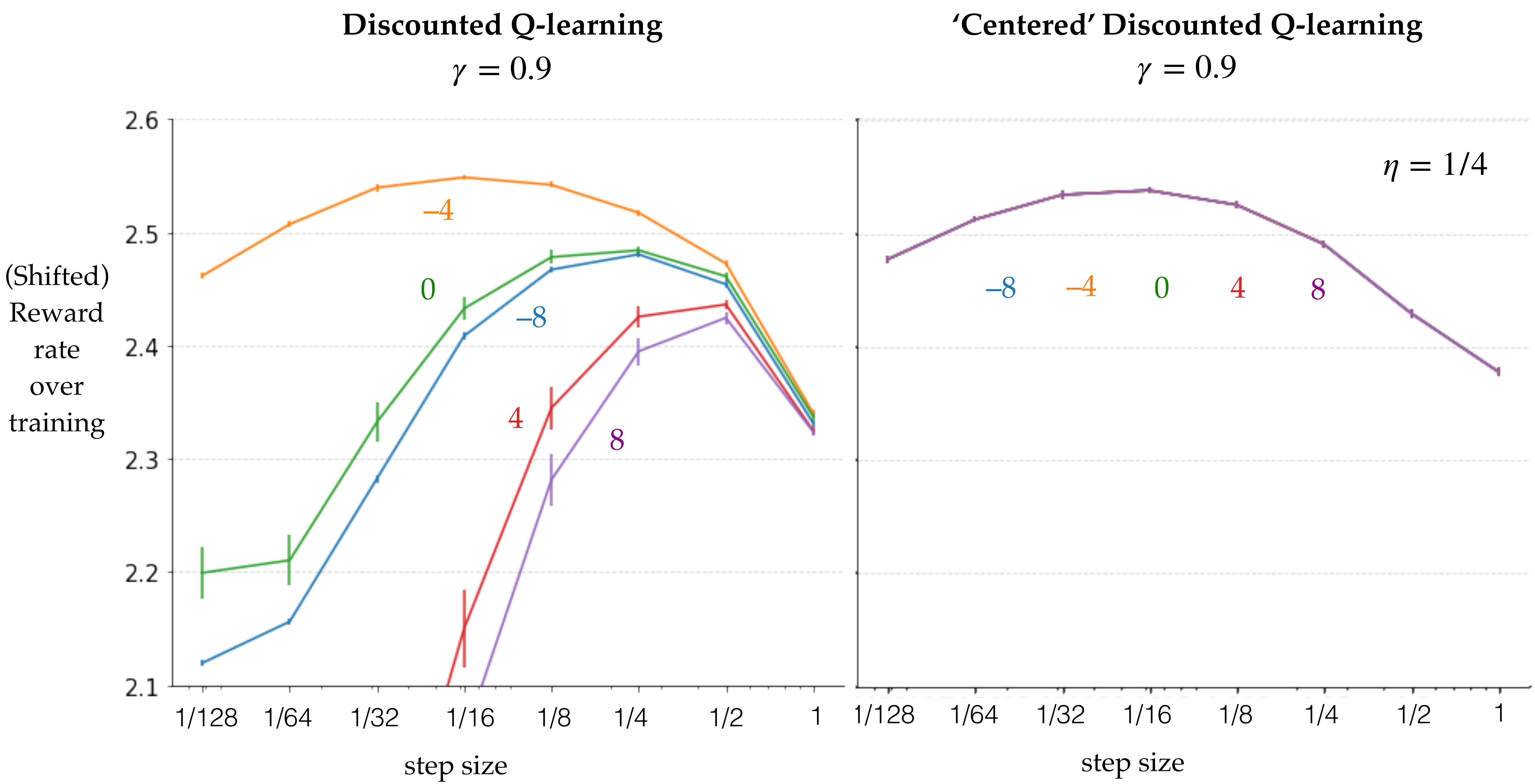
AccessControl (tabular)

SIMILAR EFFECT WITH SHIFTED REWARDS



AccessControl (tabular)

SIMILAR EFFECT WITH SHIFTED REWARDS



AccessControl (tabular)

TAKEAWAYS

TAKEAWAYS

- Standard discounted solution methods are unstable when the discount factor approaches 1.

TAKEAWAYS

- Standard discounted solution methods are unstable when the discount factor approaches 1.

- Centered discounted solution methods are not: the discount factor can be raised to 1 without degradation in performance.

TAKEAWAYS

- Standard discounted solution methods are unstable when the discount factor approaches 1.
- Centered discounted solution methods are not: the discount factor can be raised to 1 without degradation in performance.
- Centered methods are also more robust to any shifting in rewards (or initial values).

TAKEAWAYS

- Standard discounted solution methods are unstable when the discount factor approaches 1.
- Centered discounted solution methods are not: the discount factor can be raised to 1 without degradation in performance.
- Centered methods are also more robust to any shifting in rewards (or initial values).

Centered methods should be the new status quo.

TAKEAWAYS

- Standard discounted solution methods are unstable when the discount factor approaches 1.

- Centered discounted solution methods are not: the discount factor can be raised to 1 without degradation in performance.

- Centered methods are also more robust to any shifting in rewards (or initial values).

Centered methods should be the new status quo.

maybe with $\gamma = 1$;)

THANK YOU

Questions?