Reward Centering

Abhishek Naik, Yi Wan, Manan Tomar, Richard S. Sutton

{abhishek.naik,mtomar,rsutton}@ualberta.ca, yiwan@meta.com

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

THE IDEA

Estimate the average reward and subtract it from the observed rewards.

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

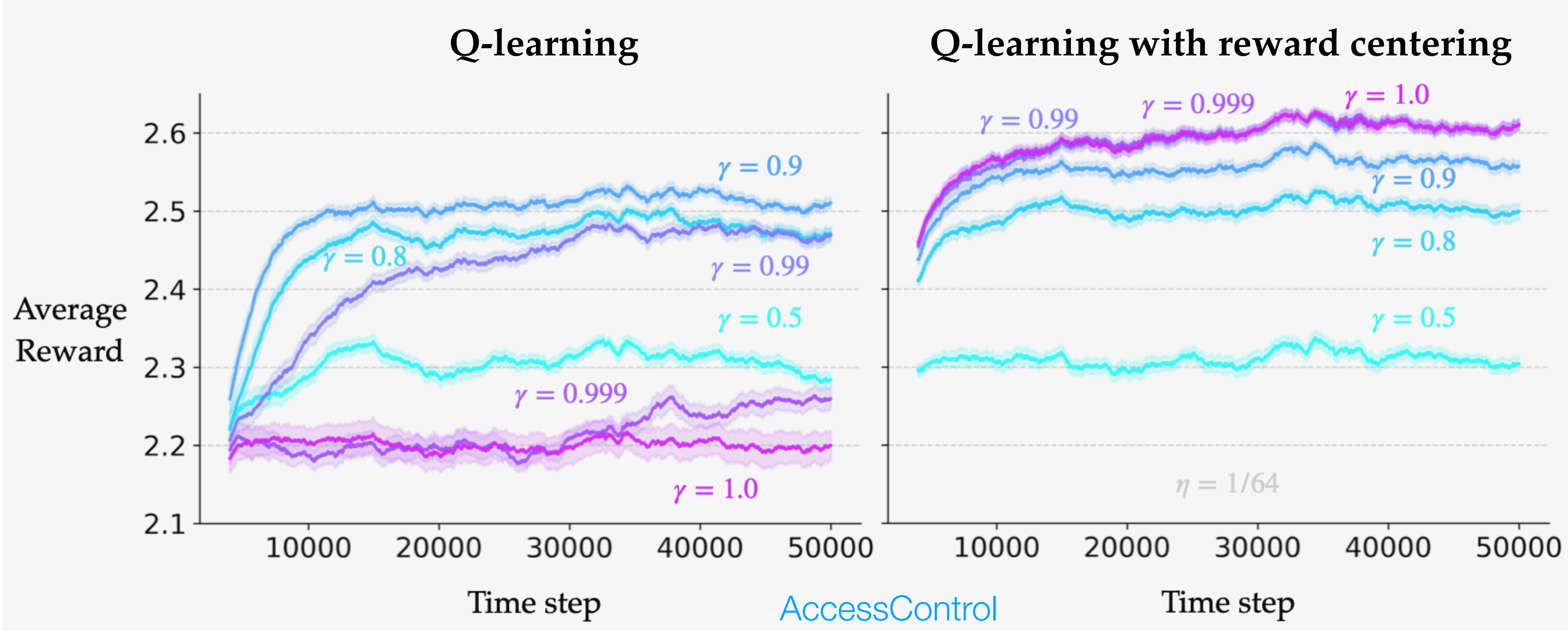
$$\downarrow$$

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

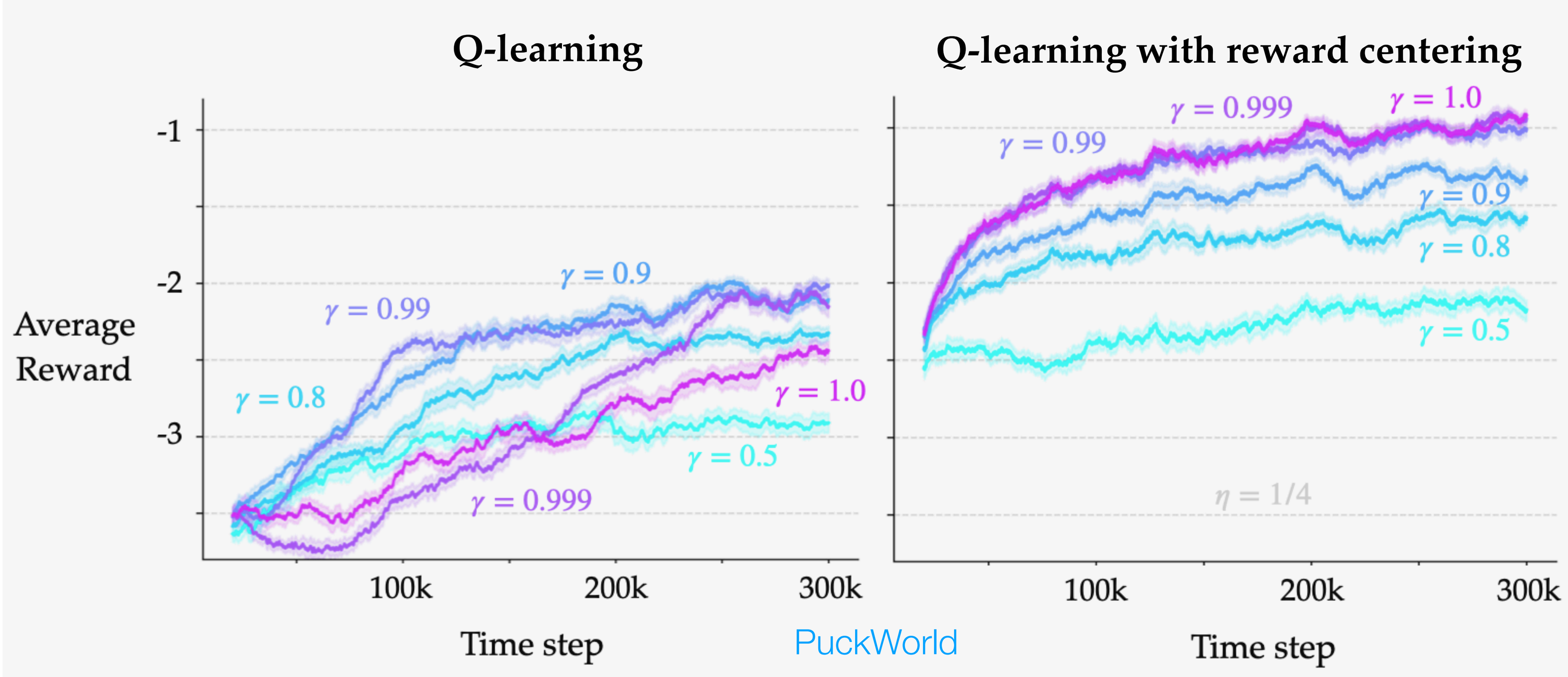
$$S_0\ A_0\ R_1\ S_1\ A_1, R_2 \ldots\ S_t\ A_t\ R_{t+1}\ S_{t+1}\ A_{t+1}\ R_{t+2}\ \ldots$$

NO INSTABILITY WITH LARGE DISCOUNT FACTORS

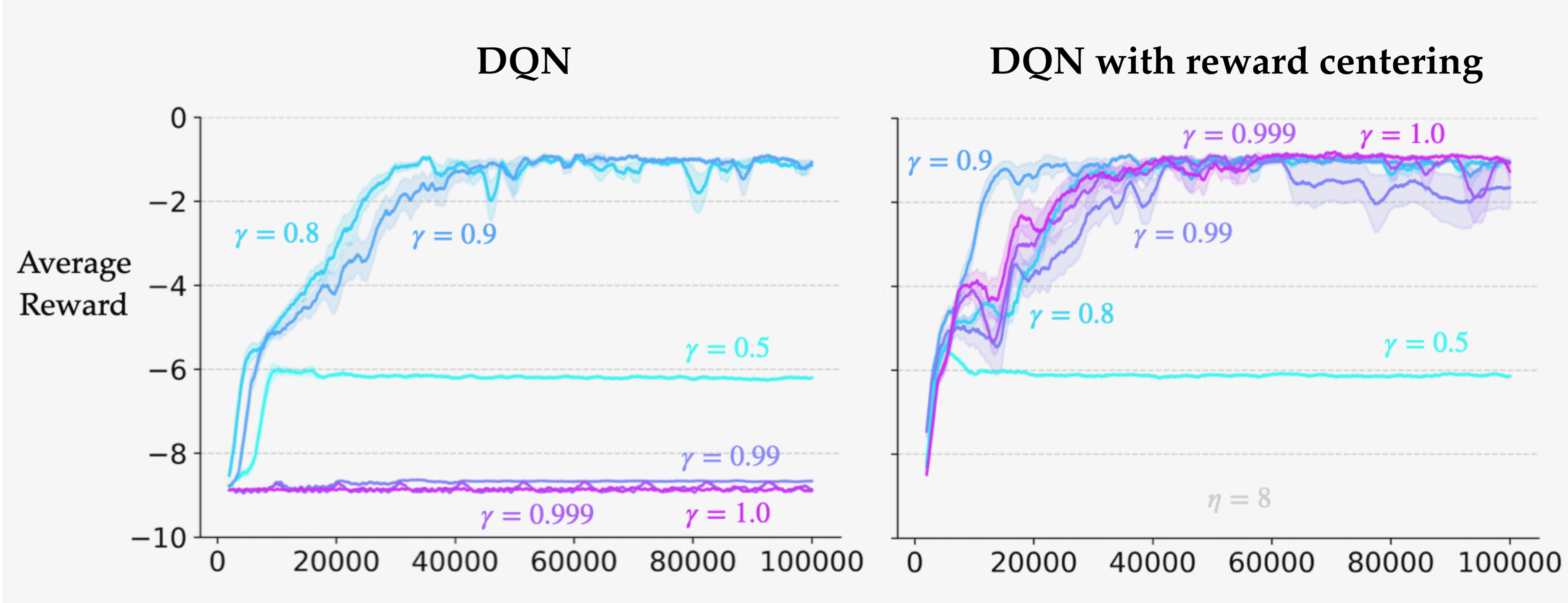
Implication #1



AccessControl

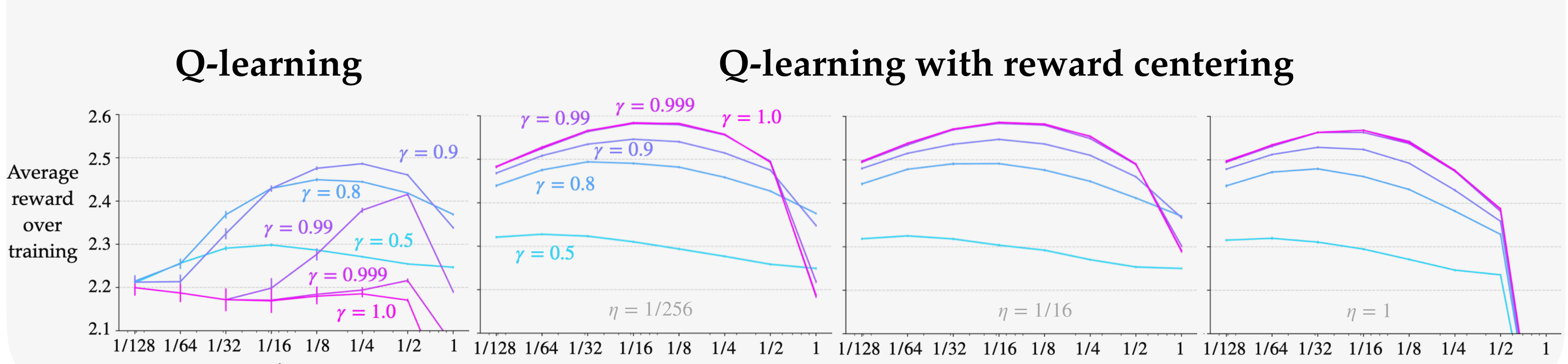


PuckWorld



Pendulum

TRENDS ARE CONSISTENT ACROSS PARAMETERS



THEORY

Full Paper

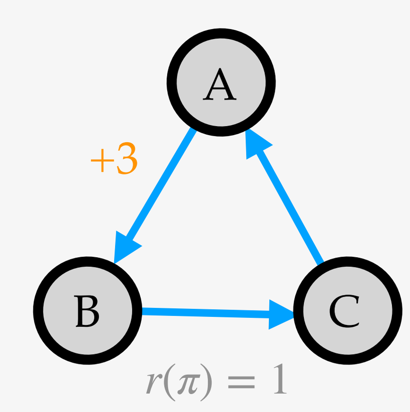
$$v^\gamma_\pi(s) = \frac{r(\pi)}{1-\gamma} + \underbrace{\tilde{v}_\pi(s) + e^\gamma_\pi(s)}_{\tilde{v}^\gamma_\pi(s)}, \qquad \forall s$$

Standard discounted value function: $v^\gamma_\pi(s) \doteq \mathbb{E}_\pi\big[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \mid S_t = s\big]$

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

Centered discounted value function: $\tilde{v}^\gamma_\pi(s) \doteq \mathbb{E}_\pi\big[\sum_{k=0}^{\infty} \gamma^k \big(R_{t+k+1} - r(\pi)\big) \mid S_t = s\big]$



		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
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	

TWO WAYS TO ESTIMATE THE AVERAGE REWARD

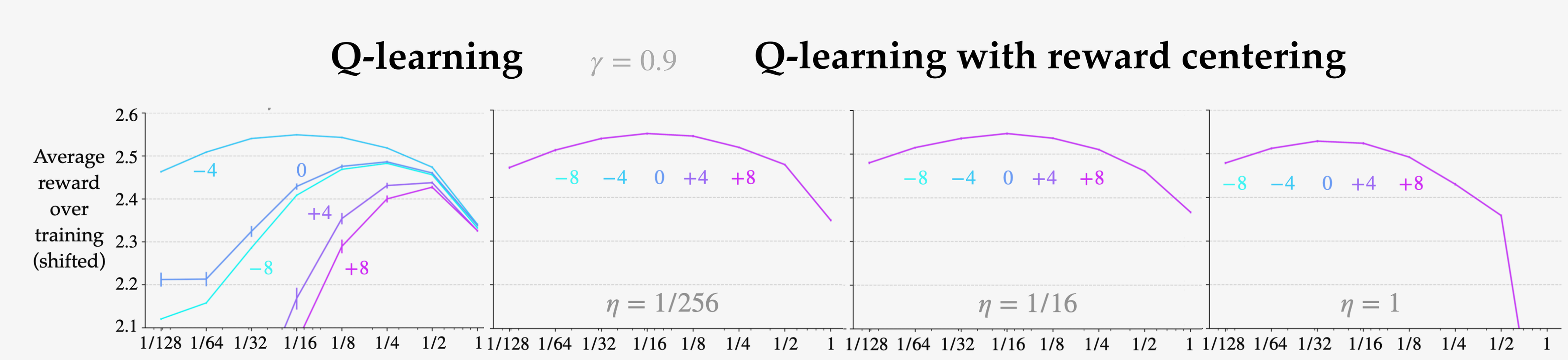
On-policy: $\bar{R}_{t+1} \doteq \bar{R}_t + \underbrace{\beta_t}_{\doteq \eta\alpha_t}(R_{t+1} - \bar{R}_t)$

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

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

ALSO MORE ROBUST TO SHIFTED REWARDS

Implication #2



AccessControl

TAKEAWAY

Reward Centering can improve the performance of *every* discounted algorithm for continuing problems, especially as $\gamma \to 1$.