Offline2Online RL

TRPO/PPO/HPO

清華大學 智能产业研究院
Institute for AI Industry Research, Tsinghua University

TRPO: $\eta(\tilde{\pi}) = \eta(\pi) + \mathbb{E}_{s_0,a_0,s_1,a_1,\ldots}\left[\sum_{t=0}^{\infty}\gamma^t A_\pi(s_t,a_t)\right]$, where

$s_0 \sim \rho_0(s_0),\ a_t \sim \tilde{\pi}(a_t|s_t),\ s_{t+1} \sim P(s_{t+1}|s_t,a_t).$ (1)

$$\underset{\theta}{\text{minimize}}\ \mathbb{E}_{s\sim\rho_{\theta_{\text{old}}},a\sim q}\left[\frac{\pi_\theta(a|s)}{q(a|s)}Q_{\theta_{\text{old}}}(s,a)\right] \quad (15)$$

$$\text{subject to } \mathbb{E}_{s\sim\rho_{\theta_{\text{old}}}}\left[D_{\text{KL}}(\pi_{\theta_{\text{old}}}(\cdot|s)\,\|\,\pi_\theta(\cdot|s))\right] \le \delta.$$

PPO: $r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)},\ \text{so}\ r(\theta_{\text{old}}) = 1.$

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t\left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\right]$$

HPO: **Proposition 1.** *Given policies π_1 and π_2, π_1 improves upon π_2 if the following condition holds:*

$$\sum_{a\in\mathcal{A}} \pi_1(a|s)A^{\pi_2}(s,a) \ge 0,\ \forall s \in \mathcal{S}. \quad (5)$$

Proposition 2. *Given policies π_1 and π_2, π_1 improves upon π_2 if the following condition holds:*

$$(\pi_1(a|s) - \pi_2(a|s))A^{\pi_2}(s,a) \ge 0,\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}. \quad (6)$$

Online Policy improvement gurantee

SBAC

Lemma 1 (Performance difference (Kakade and Langford 2002)). *For any two policy π and μ,*

$$\Delta(\pi, \mu) = J(\pi) - J(\mu) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim d^{\pi} \\ a\sim\pi}} \left[A^{\mu}(s,a)\right]. \quad (3)$$

$$\Delta(\pi, \mu) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s\sim d^{\mu} \\ a\sim\pi}} \left[\frac{d^{\pi}(s)}{d^{\mu}(s)} A^{\mu}(s,a)\right]. \quad (5)$$

$$\max_{\pi} \mathop{\mathbb{E}}_{s\sim\mathcal{B}} \left[\mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)} \left[w^{\pi}(s) Q^{\mu}(s,a) + \alpha \log \mu(a|s)\right]\right]. \quad (8)$$

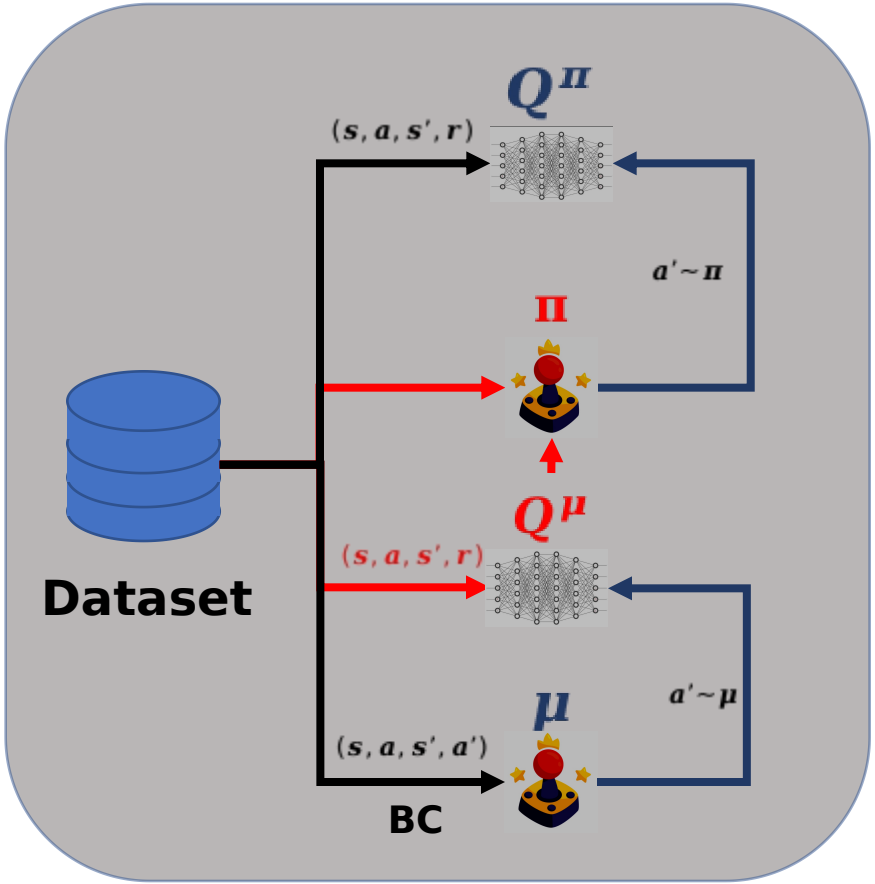
Offline Policy improvement gurantee

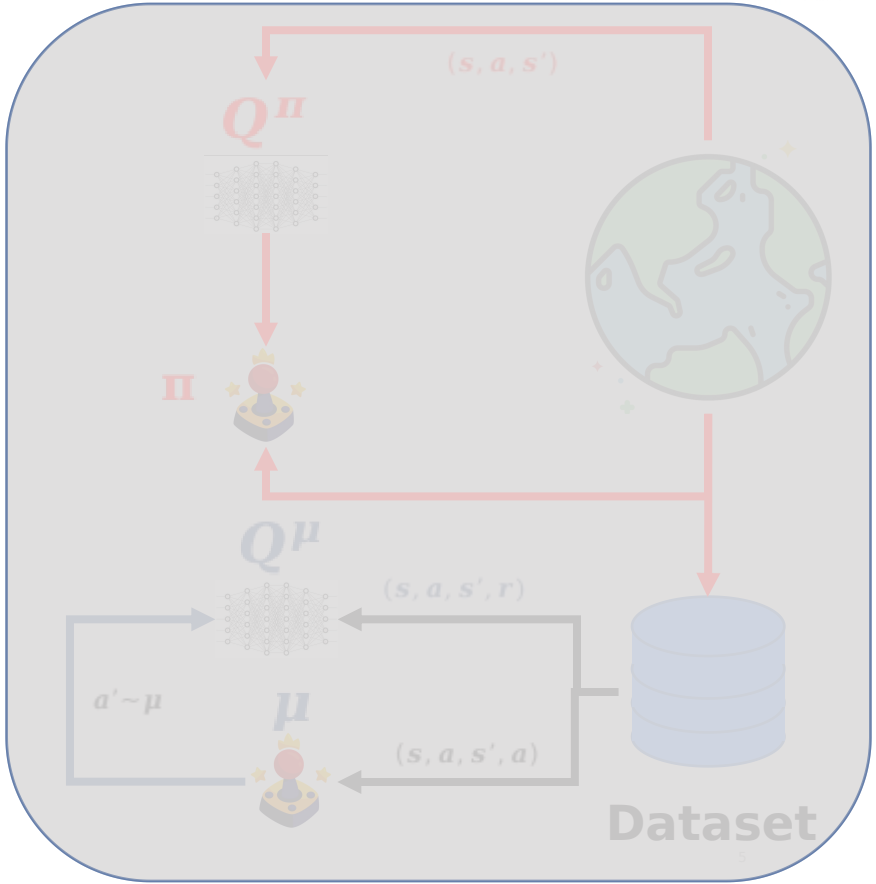
Offline2Online

清華大學 智能产业研究院
Institute for AI Industry Research, Tsinghua University

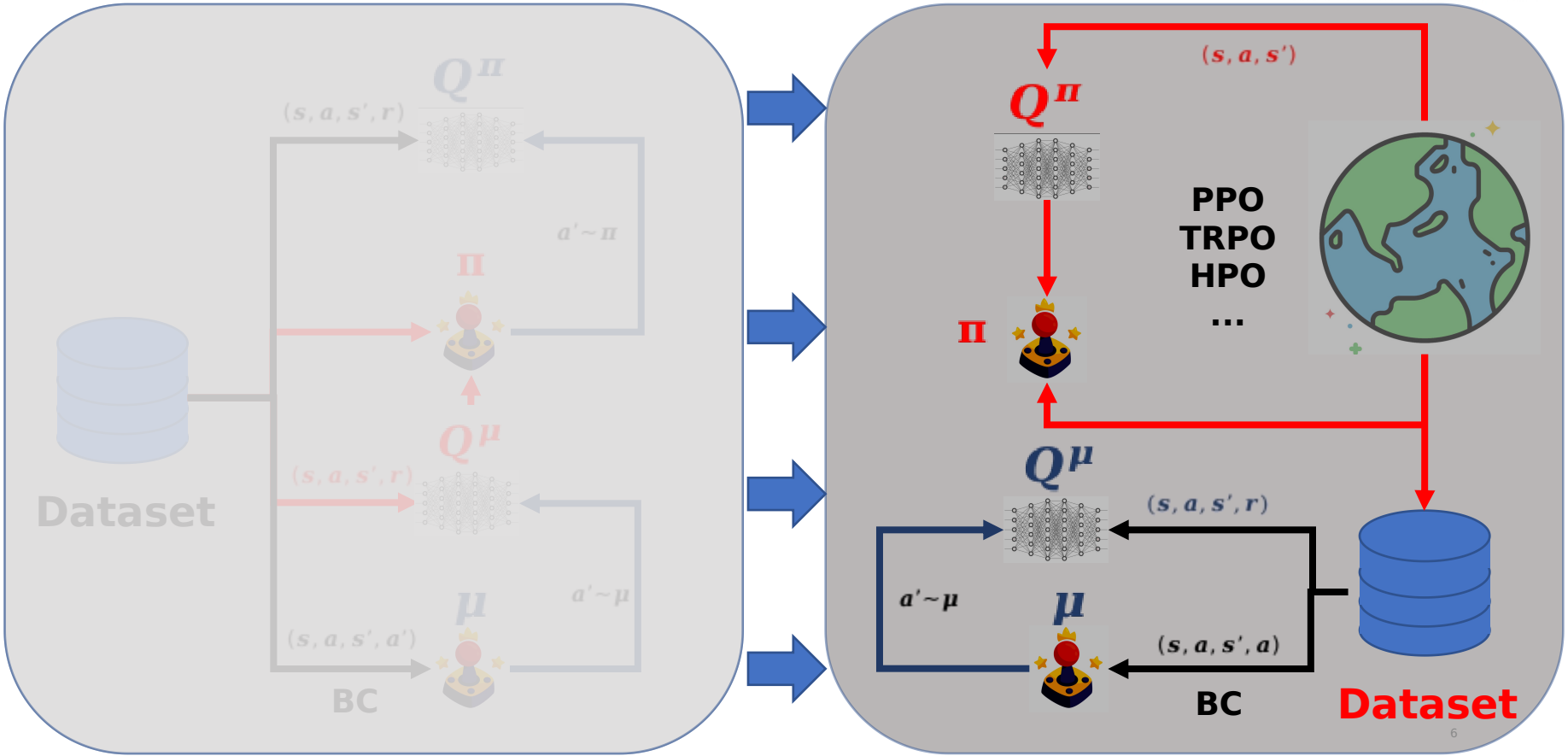
Offline2Online: Offline

清華大學 智能产业研究院
Institute for AI Industry Research, Tsinghua University



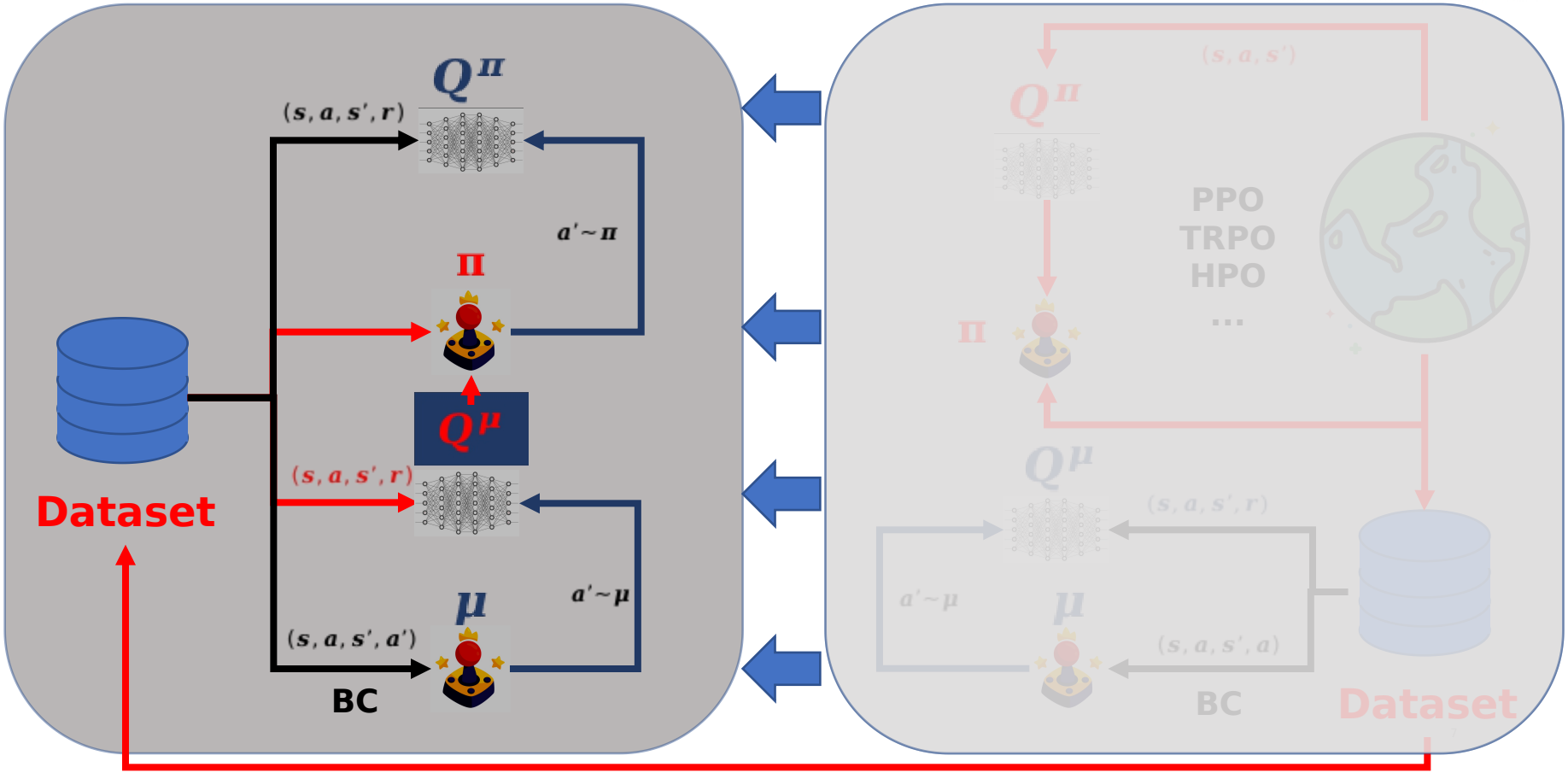


Offline2Online: Offline2Online



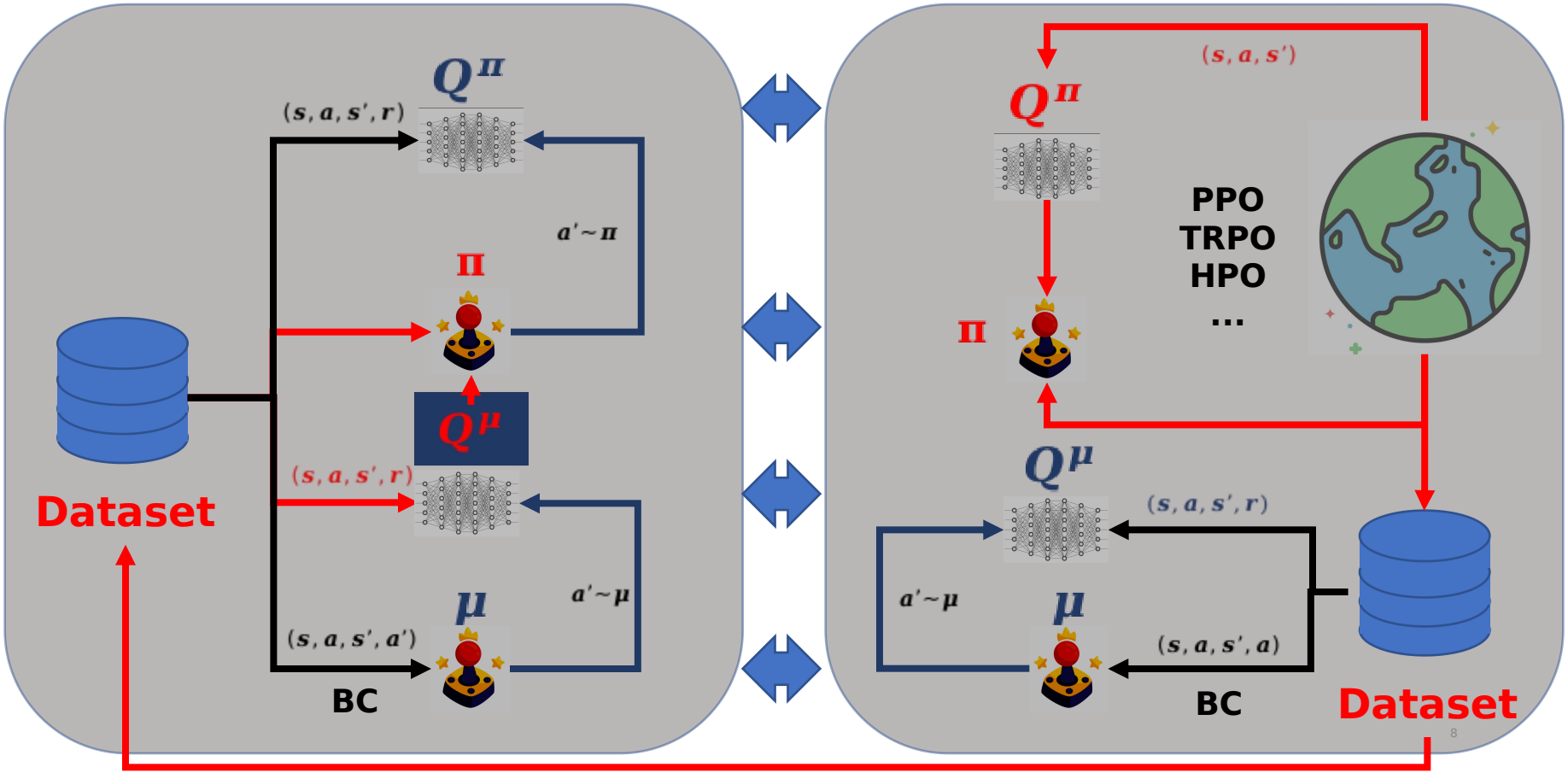
Offline2Online: Online2Offline

清華大學 智能产业研究院
Institute for AI Industry Research, Tsinghua University

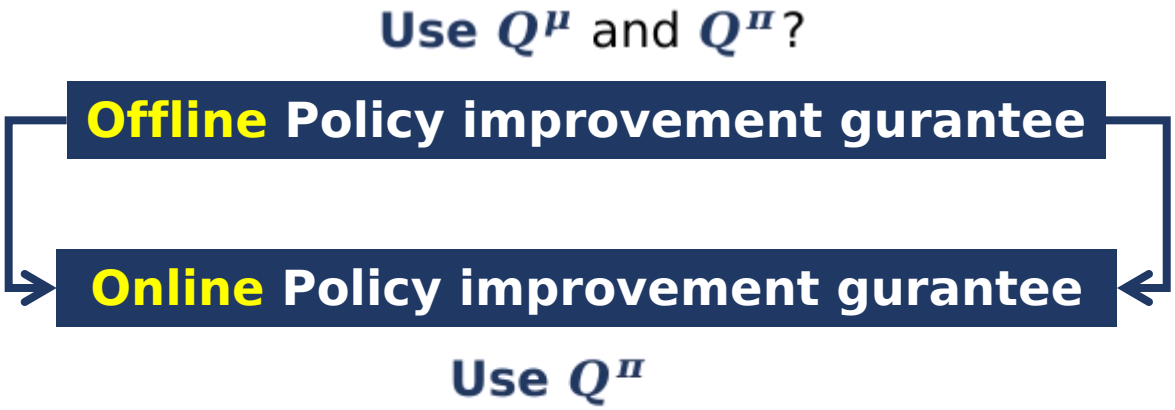


Offline2Online: Online2Offline

清華大學 智能产业研究院
Institute for AI Industry Research, Tsinghua University



Insight



- Q^{π} is optimistic, encourages agent to explore (Online)
- Q^{μ} is pessimistic, prevents agent from explore (Offline)
- $Q^{\mu} < Q^{\pi}$ in most of cases
- When the number of deployments tends to infinite, $Q^{\pi} \approx Q^{\mu}, \pi \approx \mu$

Offline2Online: Online2Offline

清華大學 智能产业研究院
Institute for AI Industry Research, Tsinghua University

HPO: **Proposition 2.** *Given policies π_1 and π_2, π_1 improves upon π_2 if the following condition holds:*

$$(\pi_1(a|s) - \pi_2(a|s))A^{\pi_2}(s,a) \geq 0,\ \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{6}$$

Full-life policy improvement gurantee:

$$\sum_{i=1}^{k} (\pi_{k+1}(a|s) - \pi_i(a|s))A^{\pi_i}(s,a)$$

Best policy improvement gurantee

$$(\pi_{k+1}(a|s) - \pi_{max}(a|s))A^{\pi_{max}}(s,a)$$