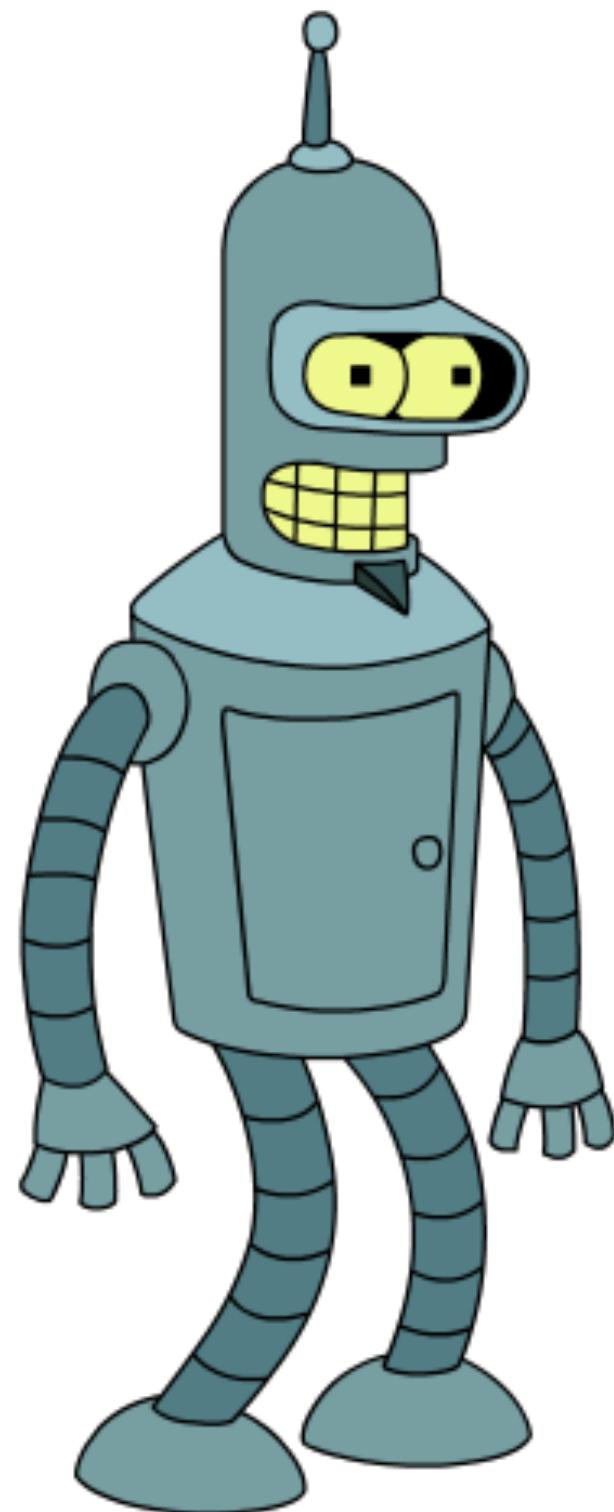


Reinforcement Learning

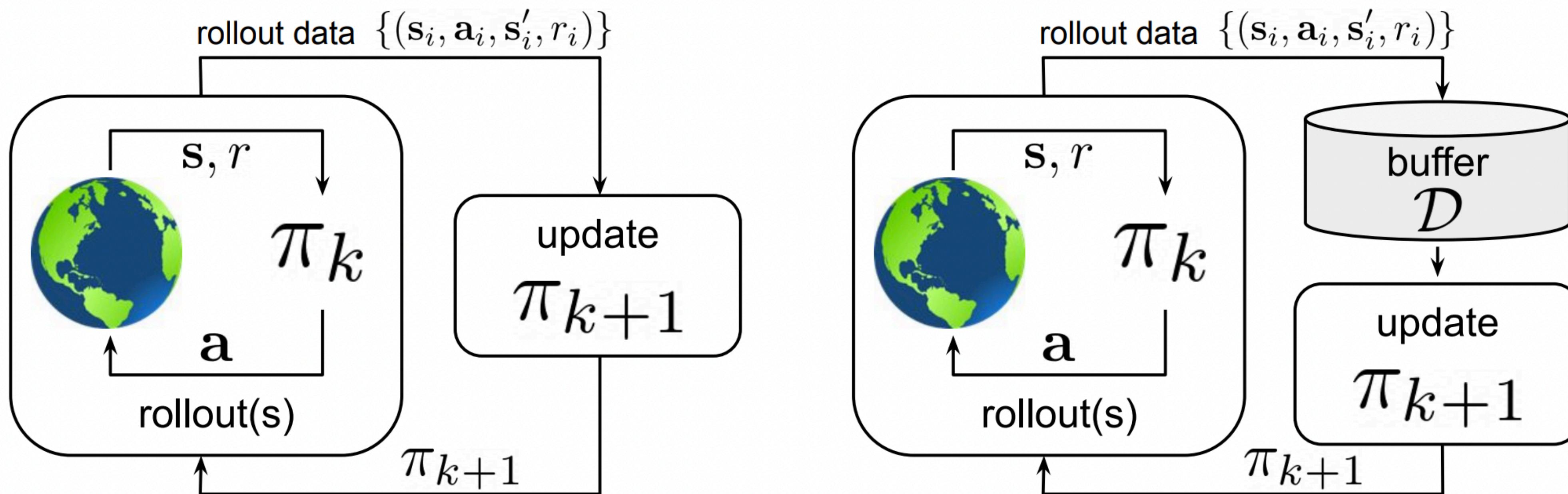
HSE, winter - spring 2024

Lecture 7: Offline RL

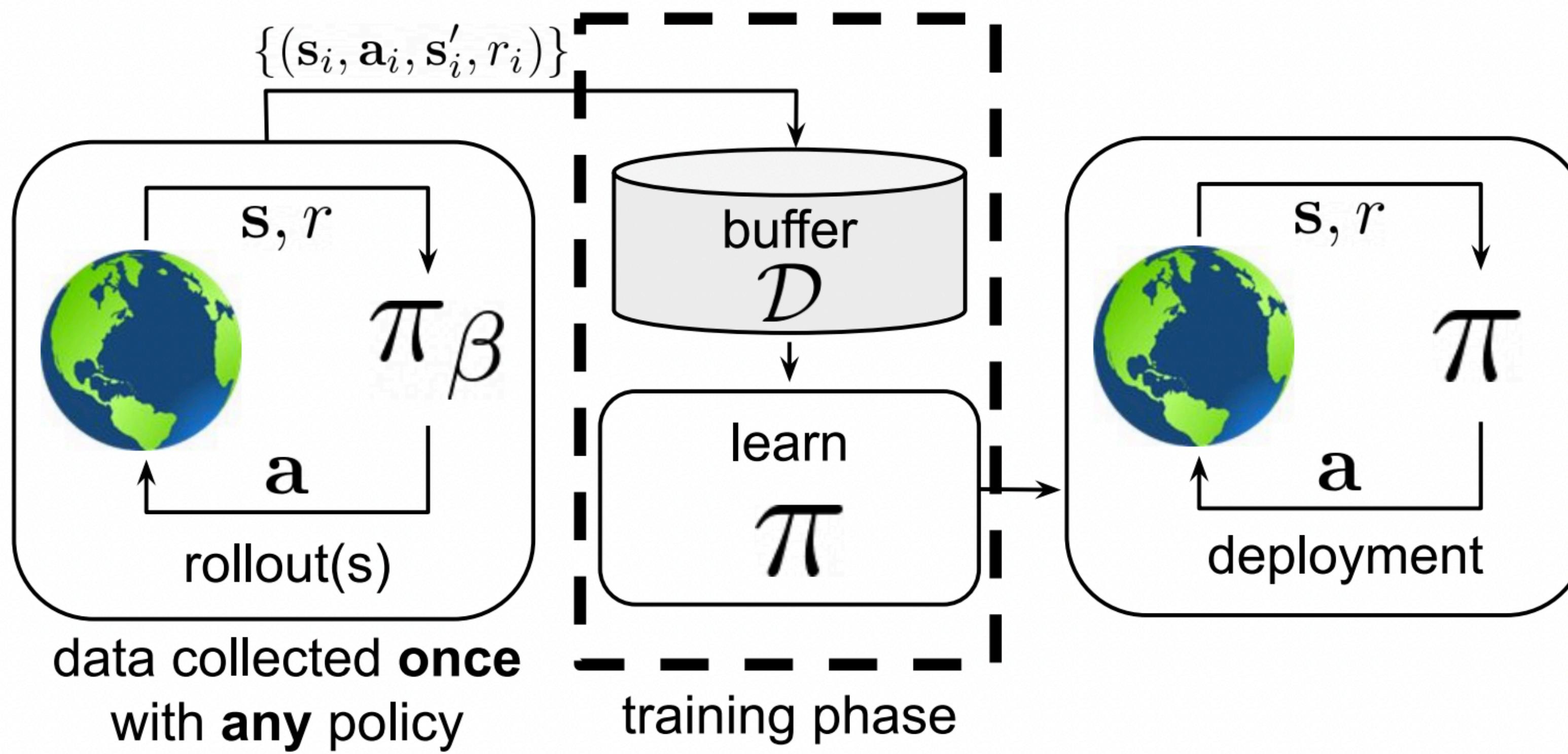


Sergei Laktionov
slaktionov@hse.ru
[LinkedIn](#)

Recap: On-policy vs Off-policy



Offline RL



$$D = \{(s_i, a_i, r_i, s'_i)\}$$

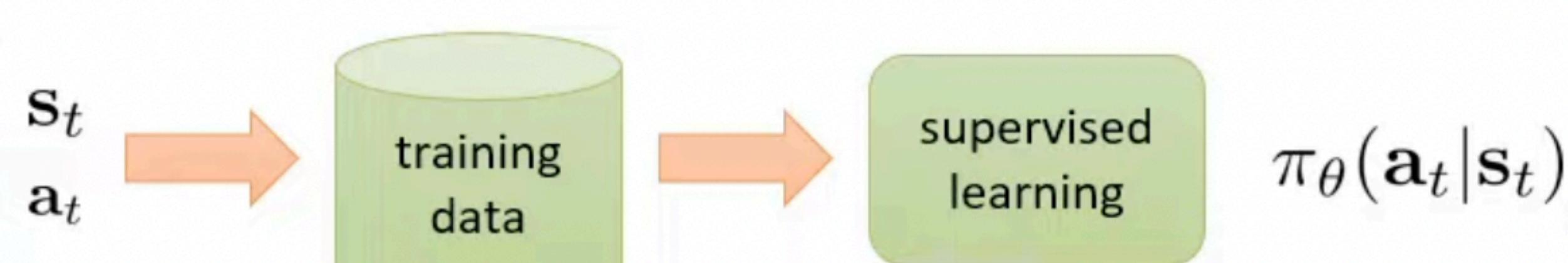
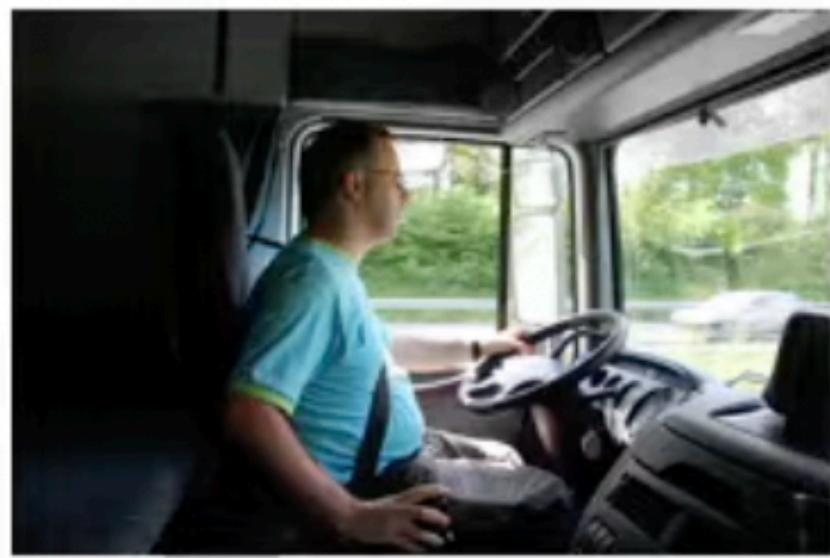
1. $a \sim \pi_\beta$
2. $s \sim d_{\pi_\beta}$
3. $s' \sim p(\cdot | s, a)$
4. $r = r(s, a)$

Offline RL

Offline RL algorithms must do two things:

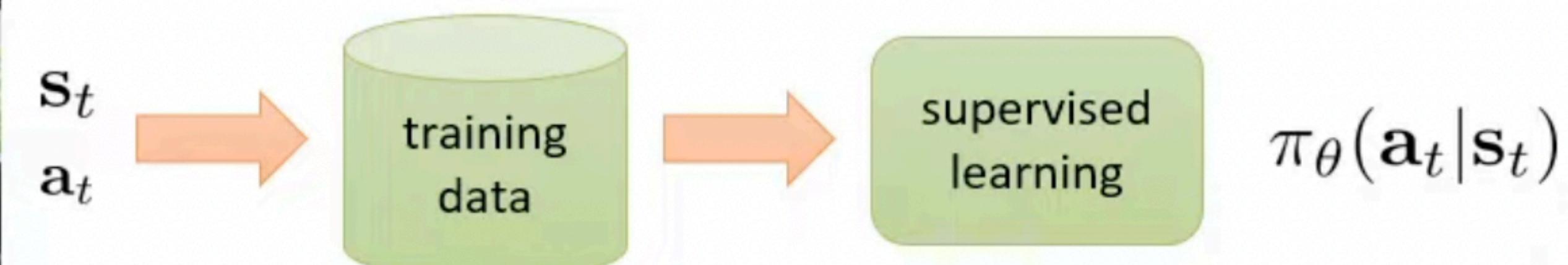
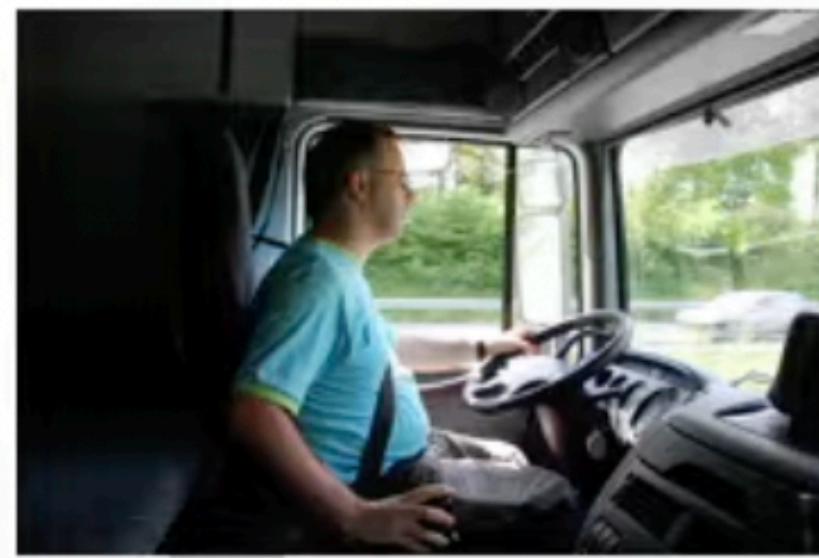
1. Maximise expected cumulative reward
2. Learn π_θ which is close to π_β

Imitation Learning



$$\sum_{(s,a) \in D} \log \pi_\theta(a | s) \rightarrow \max_{\theta}$$

Imitation Learning

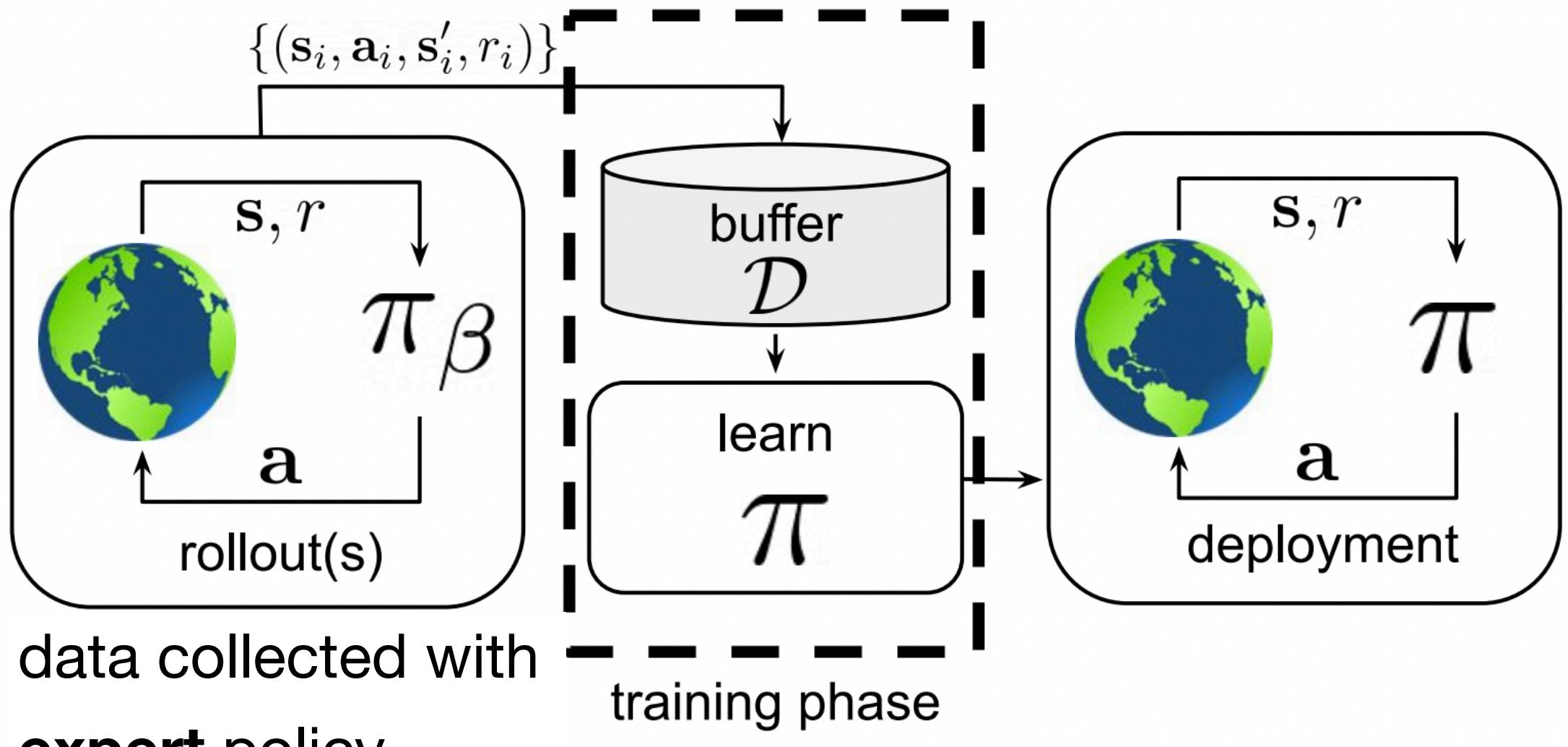


$$\sum_{(s,a) \in D} \log \pi_\theta(a | s) \rightarrow \max_\theta$$

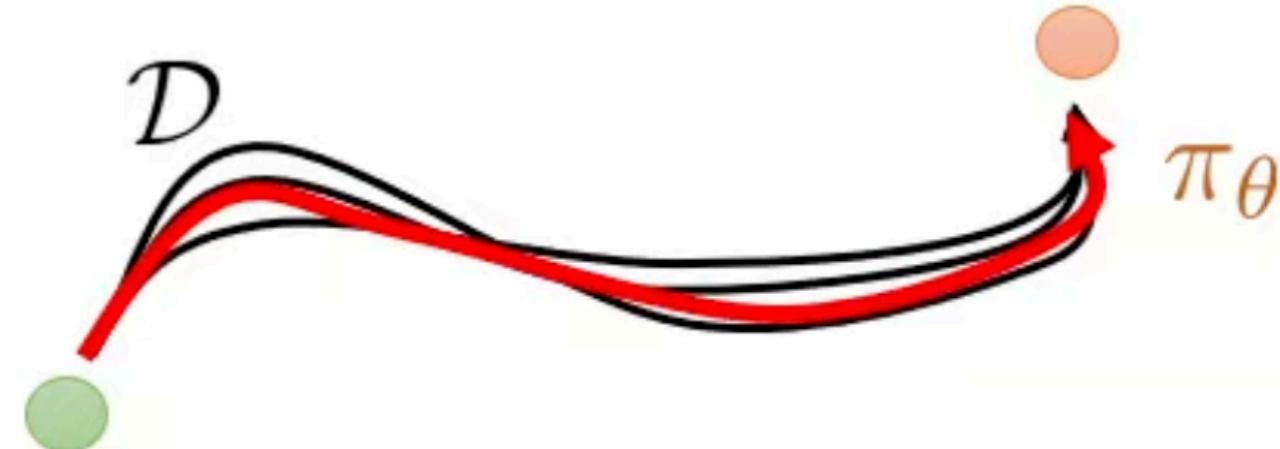
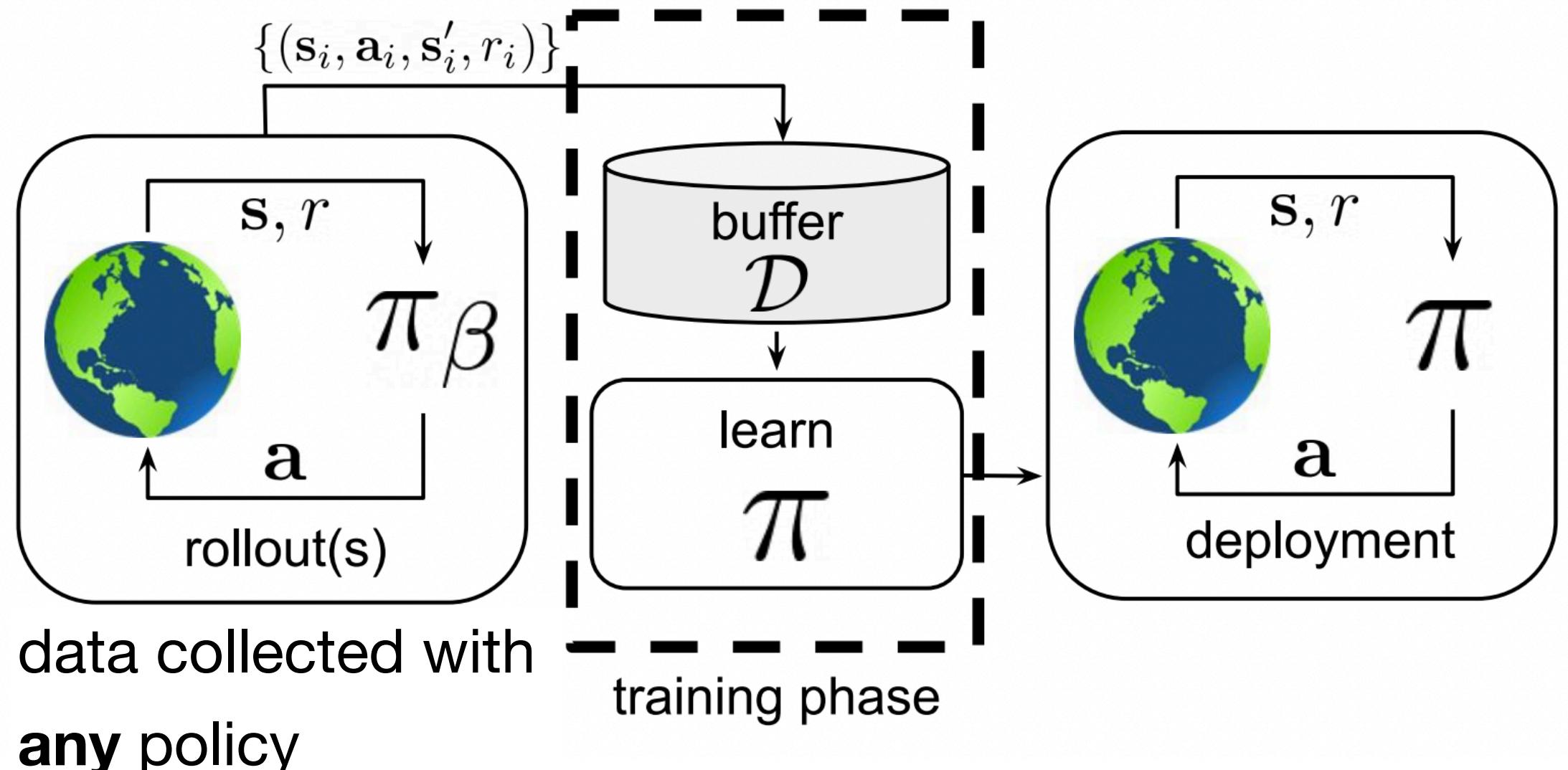
- + Very stable
- + Scalable
- + Easy to implement and debug
- Does not maximise the reward

Imitation Learning vs Offline RL

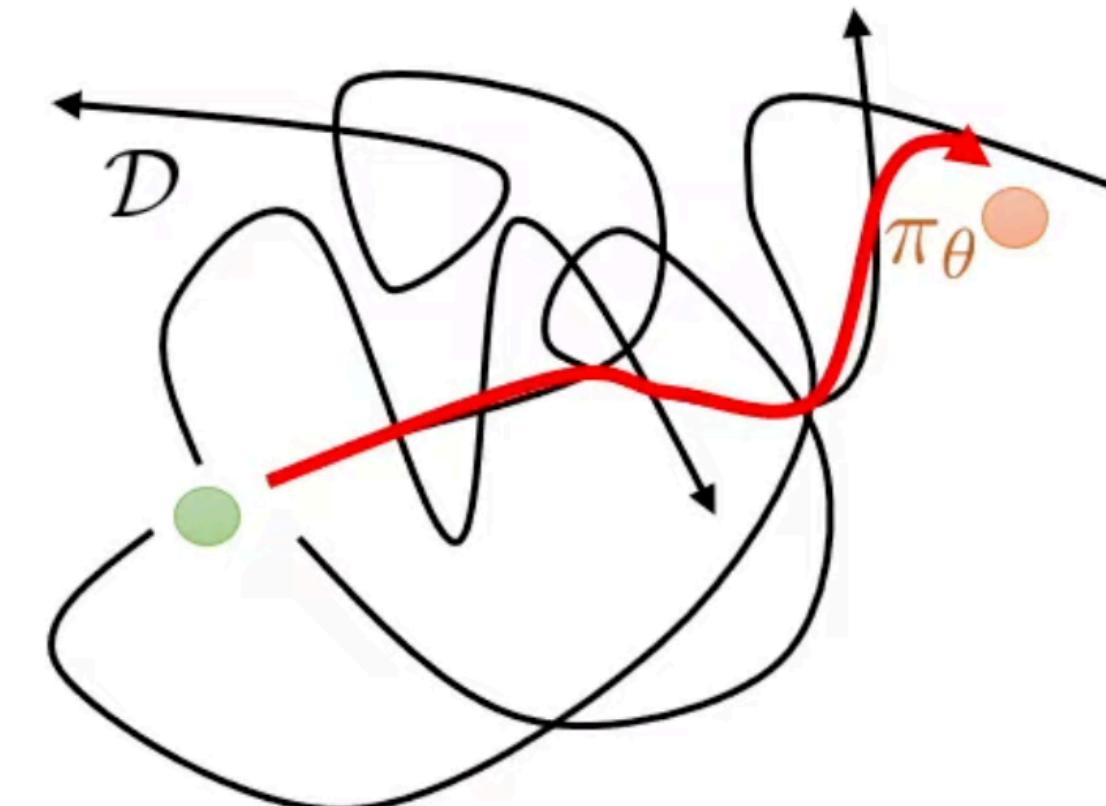
Imitation Learning



Offline RL



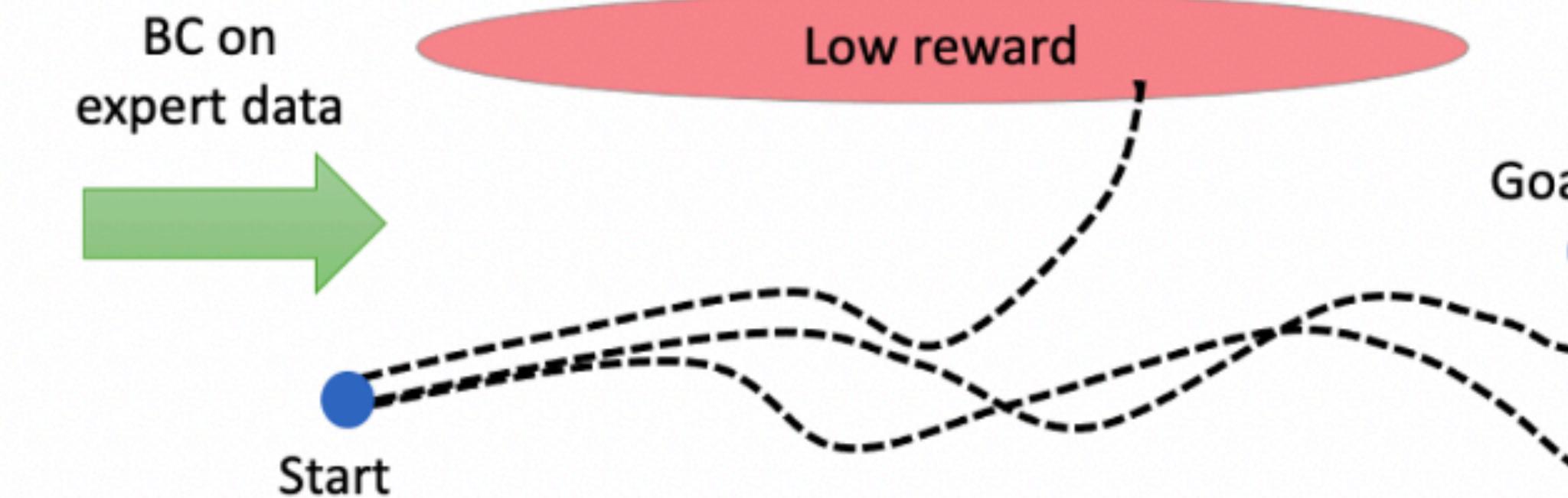
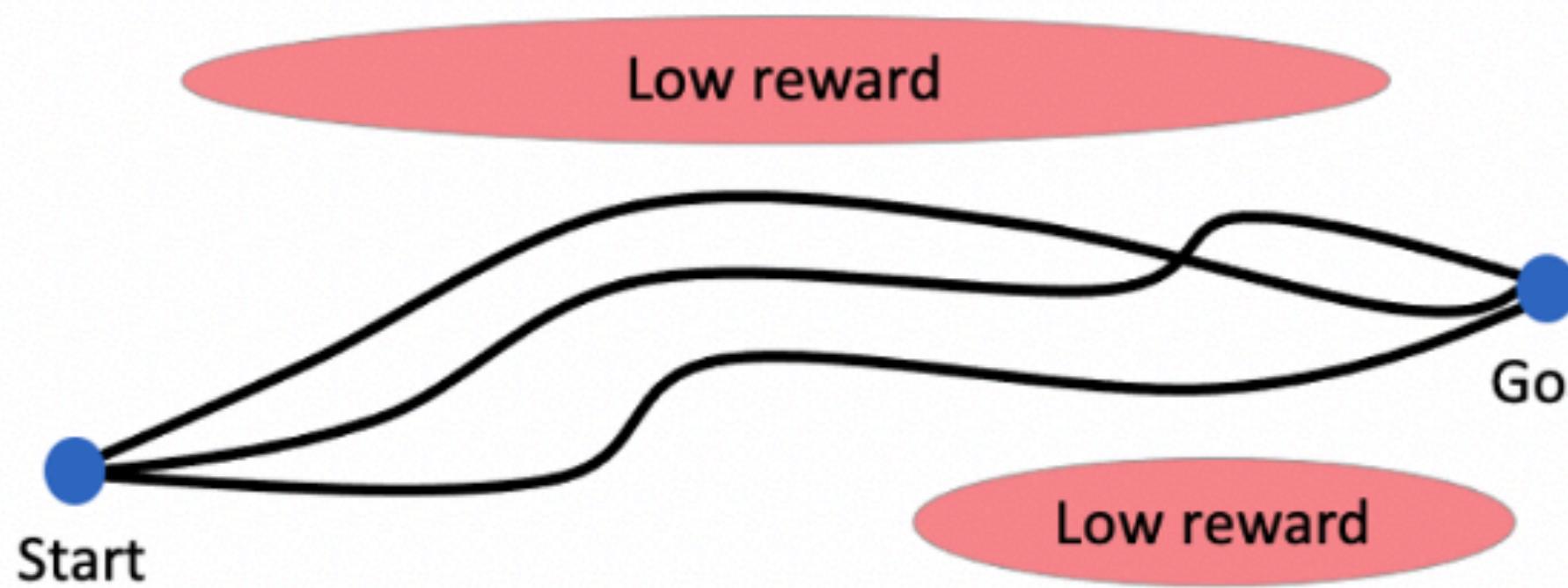
Imitate the data collected by experts



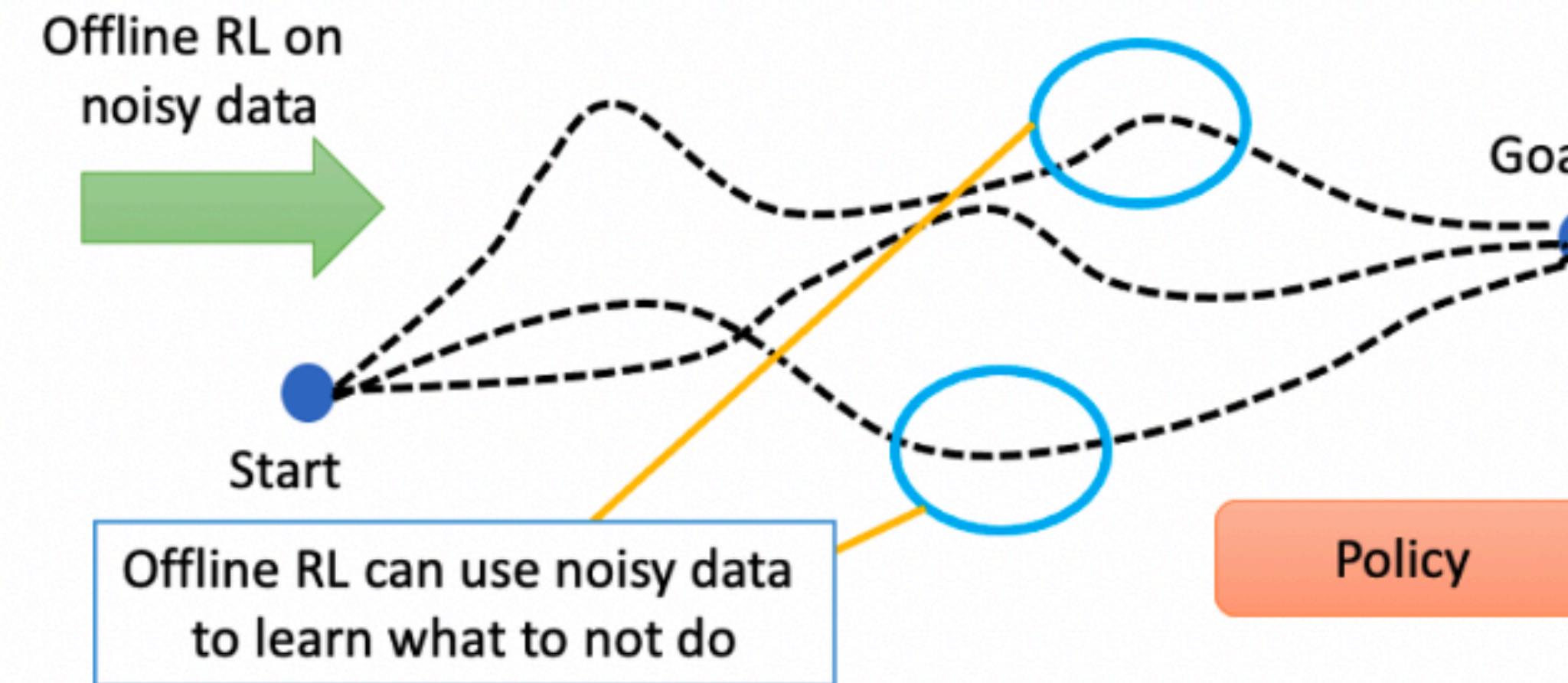
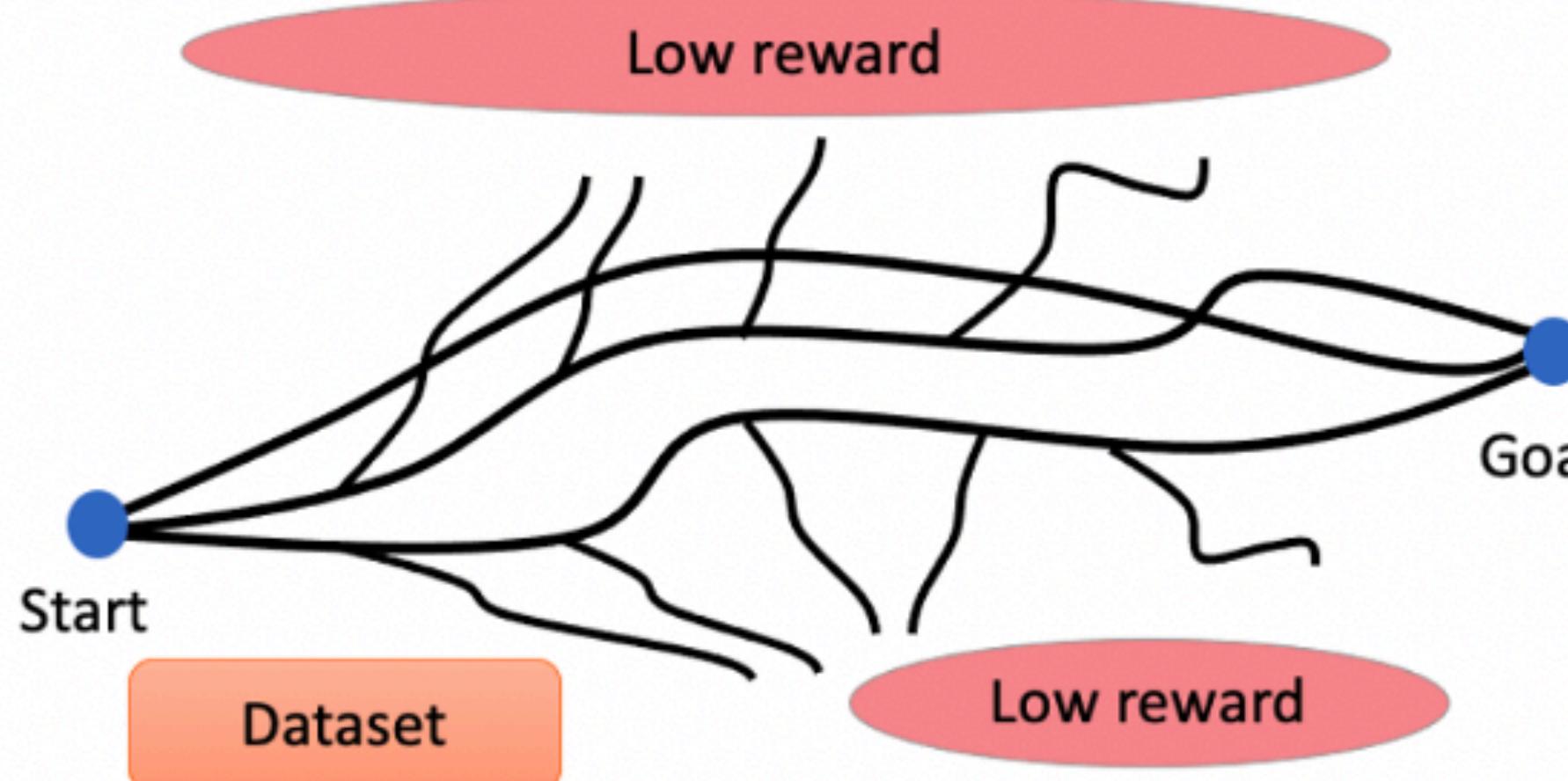
Extract the insights from noisy data

Imitation Learning vs Offline RL

Behavior cloning on expert:



Offline RL with noisy data:

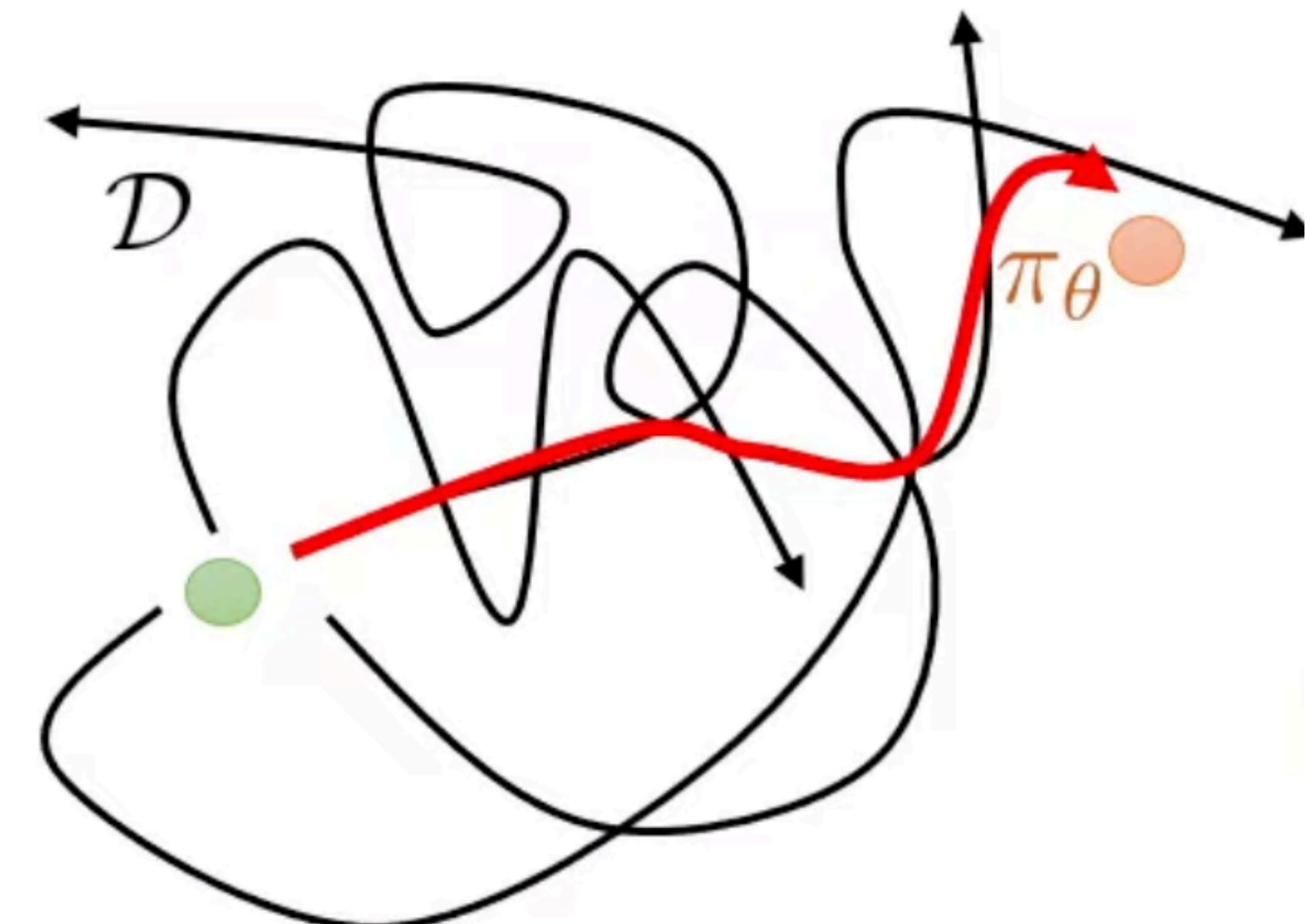


Offline RL can use noisy data
to learn what to not do

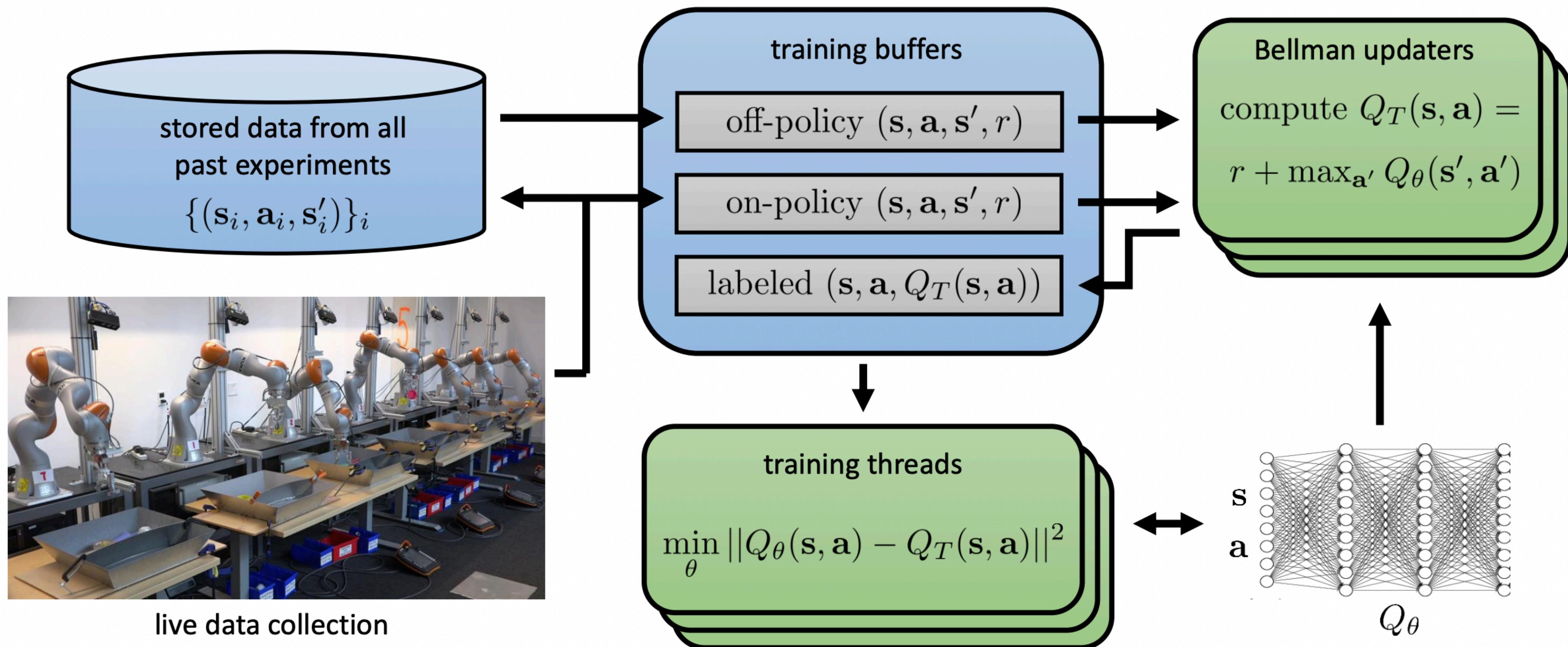
When should we prefer Offline RL over Behavioral Cloning?

Order from Chaos

1. Extract good trajectories from both good and bad demonstrations
2. Generalise: good behaviour in one state may suggest good behaviour in another state
3. Recombine the parts of good behaviour

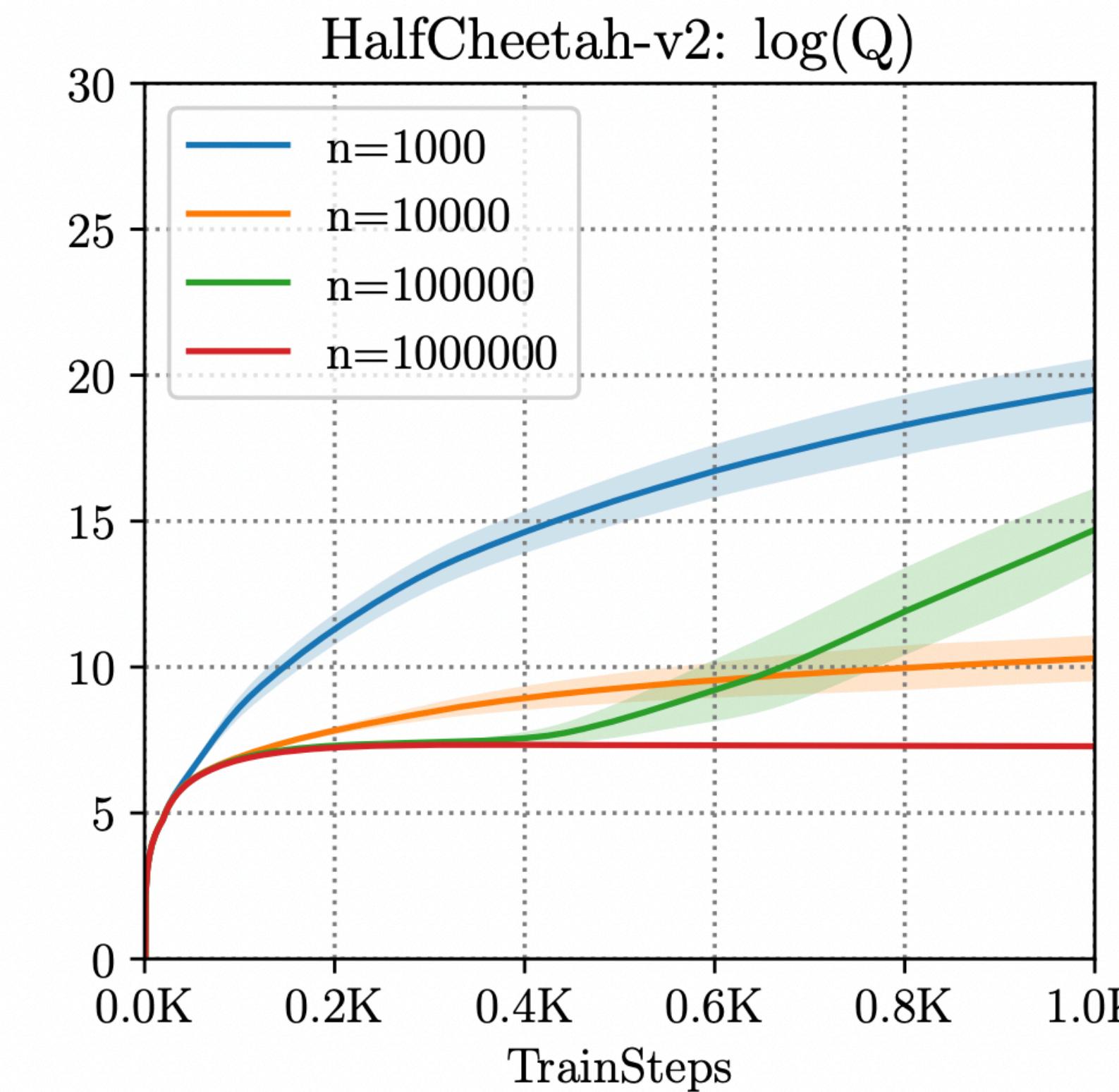
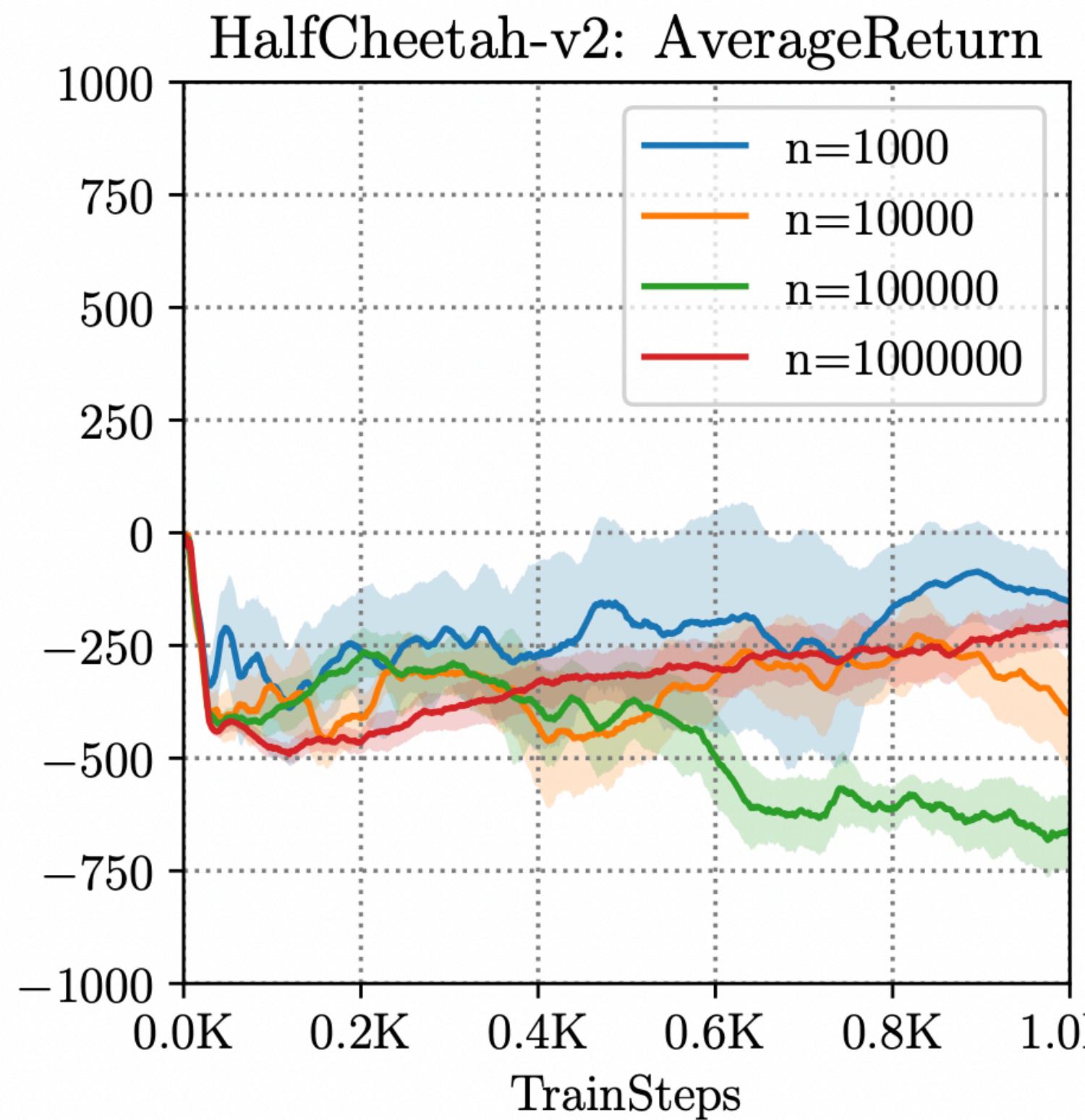


Full Pipeline



Bootstrapping Error

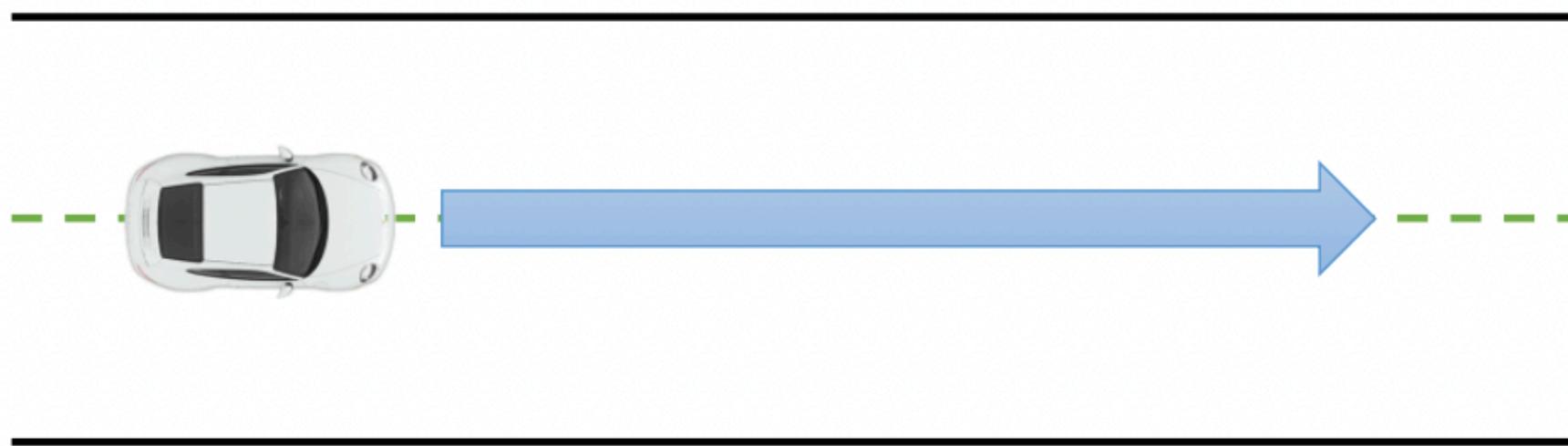
Performance of SAC



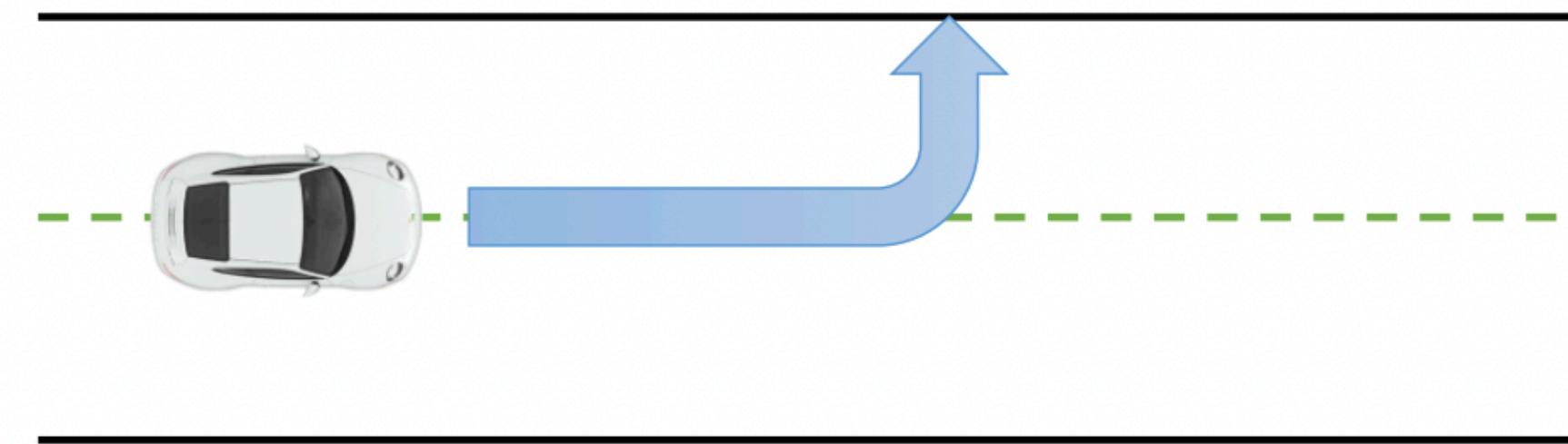
Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction

Distribution Shift

Training data

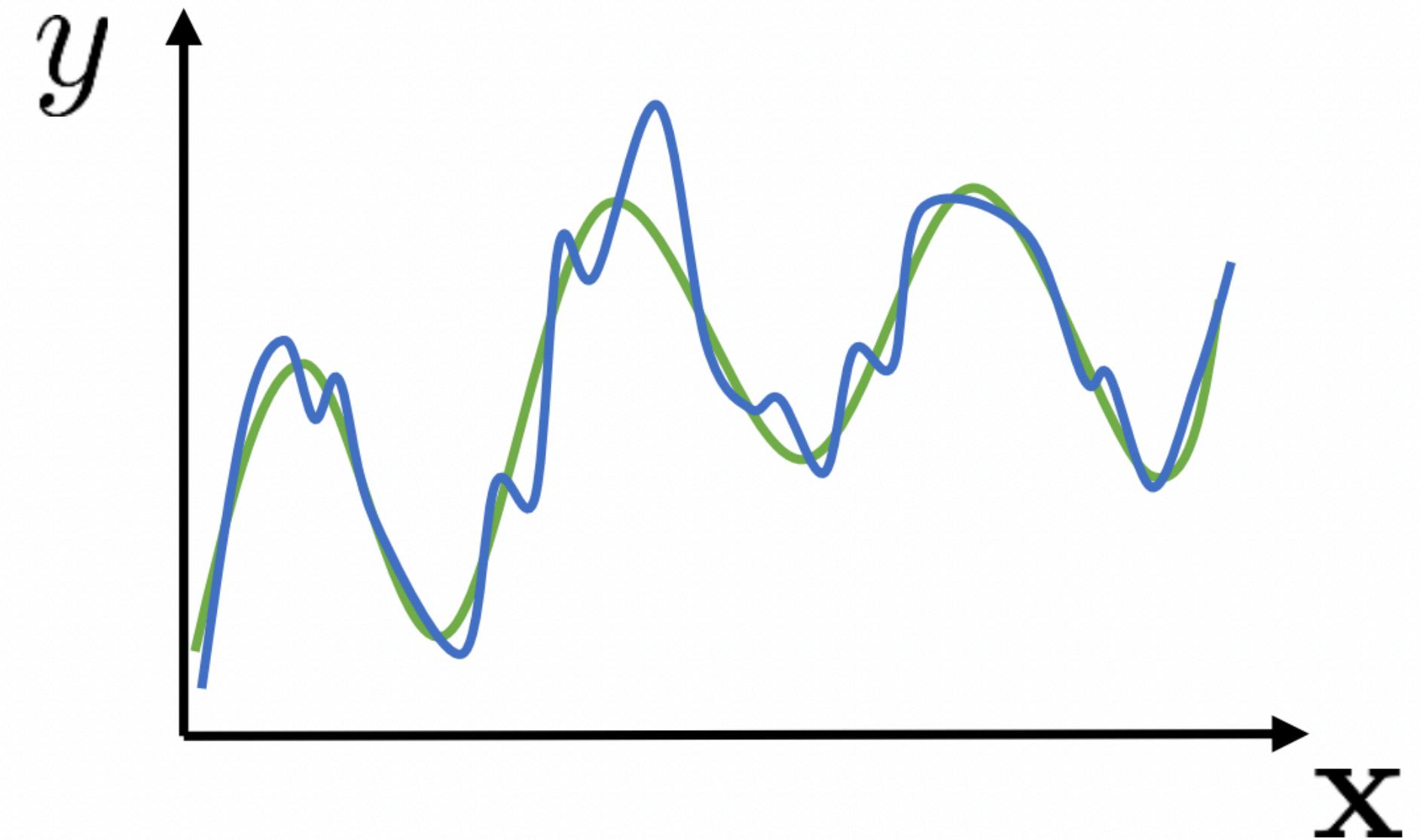


What the policy wants to do



$$y(s, a) = r + \gamma \mathbb{E}_{a' \sim \pi_{new}} [Q_{\phi^-}(s', a')]$$

$$\mathbb{E}_{s \sim d_{\pi_\beta}, a \sim \pi_\beta} [y(s, a) - Q_\phi(s, a)]^2 \rightarrow \min_{\phi}$$

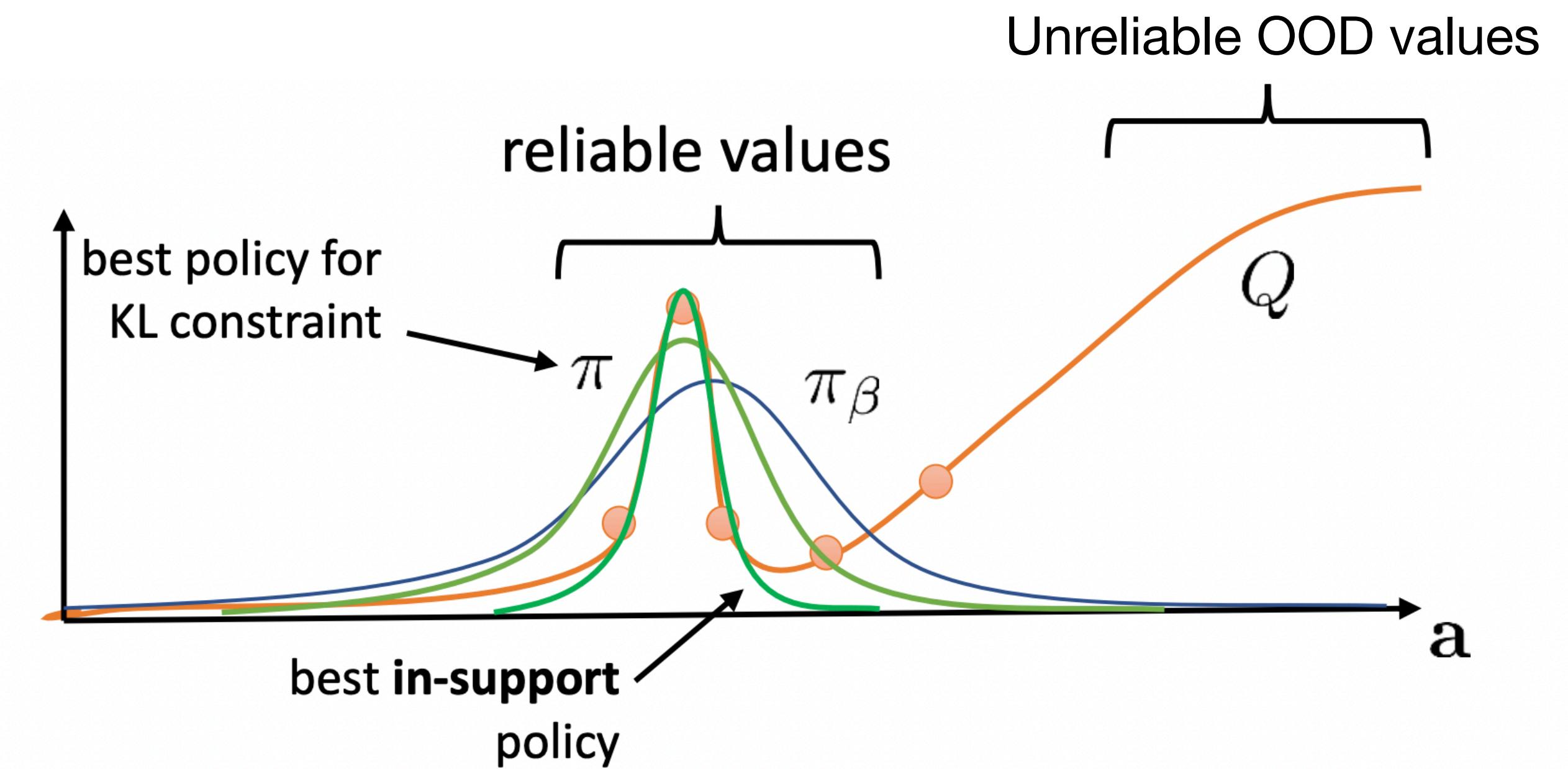


Policy Constraints

Actor objective:

$$\mathbb{E}_{s \sim D} \mathbb{E}_{a \sim \pi(\cdot | s)} [Q(s, a)] \rightarrow \max_{\pi}$$

$$\text{s.t. } D_{KL}(\pi(\cdot | s) || \pi_{\beta}(\cdot | s)) \leq \varepsilon$$



Implicit Policy Constraints

$$\mathbb{E}_{a \sim \pi(\cdot | s)} [A^\pi(s, a)] - \lambda D_{KL}(\pi(\cdot | s) || \pi_\beta(\cdot | s)) \rightarrow \max_{\pi}$$

$$\pi^*(a | s) = \frac{1}{Z(s)} \pi_\beta(a | s) \exp\left(\frac{A^\pi(s, a)}{\lambda}\right)$$

Weighted max loglikelihood:

$$\pi^{new}(a | s) = argmax_{\pi} \mathbb{E}_{(s,a) \sim \pi_\beta} \left[\log \pi(a | s) \frac{1}{Z(s)} \exp\left(\frac{A^{\pi_{old}}(s, a)}{\lambda}\right) \right]$$

Advantage Weighted Actor Critic

Actor: $\mathbb{E}_{(s,a) \sim \pi_\beta} \left[\log \pi_\theta(a | s) \frac{1}{Z(s)} \exp\left(\frac{A_\phi(s, a)}{\lambda}\right) \right] \rightarrow \max_{\theta}$,

where $A_\phi(s, a) = Q_\phi(s, a) - Q_\phi(s, a')$, $a' \sim \pi_\theta(\cdot | s)$.

Critic: $\mathbb{E}_{(s,a,r,s') \sim D} \left[r + \gamma \mathbb{E}_{a' \sim \pi_\theta(\cdot | s')} Q_{\phi^-}(s', a') - Q_\phi(s, a) \right]^2 \rightarrow \min_{\phi}$,

Advantage Weighted Actor Critic

Actor: $\mathbb{E}_{(s,a) \sim \pi_\beta} \left[\log \pi_\theta(a | s) \frac{1}{Z(s)} \exp\left(\frac{A_\phi(s, a)}{\lambda}\right) \right] \rightarrow \max_{\theta}$,

where $A_\phi(s, a) = Q_\phi(s, a) - Q_\phi(s, a')$, $a' \sim \pi_\theta(\cdot | s)$.

Critic: $\mathbb{E}_{(s,a,r,s') \sim D} \left[r + \gamma \mathbb{E}_{a' \sim \pi_\theta(\cdot | s')} Q_{\phi^-}(s', a') - Q_\phi(s, a) \right]^2 \rightarrow \min_{\phi}$,

Can we avoid OOD actions here?

Different Bootstrap Targets

$$\mathbb{E}_{(s,a,r,s') \sim D} [r + \gamma \mathbb{E}_{a' \sim \pi_\theta(\cdot | s')} Q_{\phi^-}(s', a') - Q_\phi(s, a)]^2$$

$$\mathbb{E}_{(s,a,r,s') \sim D} [r + \gamma Q_{\phi^-}(s', a') - Q_\phi(s, a)]^2, \text{ where } a' \sim \pi_\theta(\cdot | s')$$

$$\mathbb{E}_{(s,a,r,s') \sim D} [r + \gamma \max_{a'} Q_{\phi^-}(s', a') - Q_\phi(s, a)]^2$$

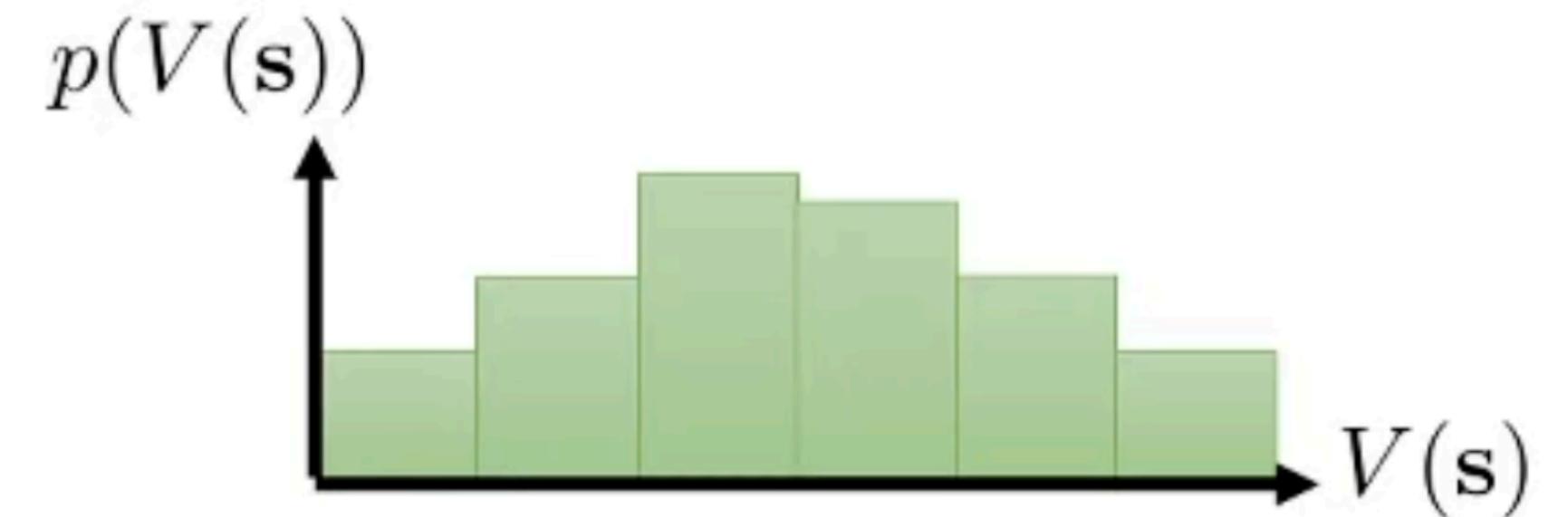
$$\mathbb{E}_{(s,a,r,s',a') \sim D} [r + \gamma Q_{\phi^-}(s', a') - Q_\phi(s, a)]^2$$

$$\mathbb{E}_{(s,a,r,s') \sim D} [r + \gamma \max_{a': \pi_\beta(a' | s') \geq \varepsilon} Q_{\phi^-}(s', a') - Q_\phi(s, a)]^2$$

Implicit Q-Learning

$$\mathbb{E}_{a \sim \pi(\cdot|s)} Q^\pi(s, a) = V^\pi(s)$$

Approximate it with a neural network: $V^\pi \approx V_\psi$

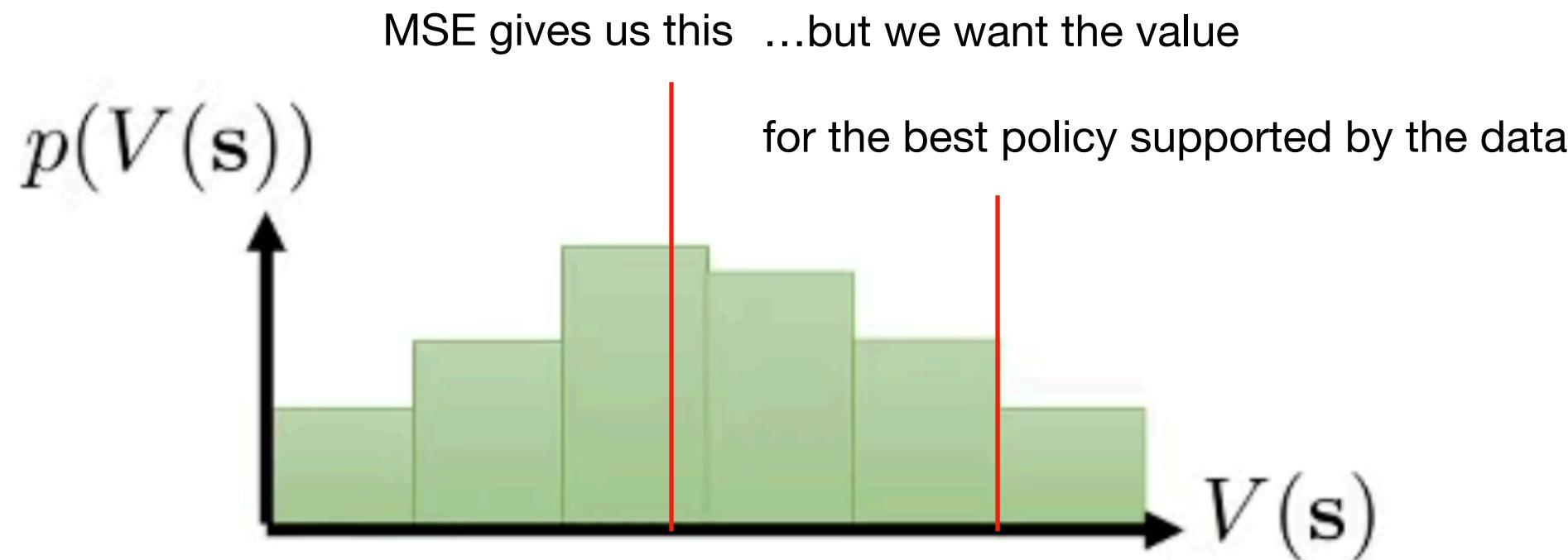


Similar states can give the information about state-value distribution

Implicit Q-Learning

$$\mathbb{E}_{a \sim \pi(\cdot|s)} Q^\pi(s, a) = V^\pi(s)$$

Approximate it with a neural network: $V^\pi \approx V_\psi$



Similar states can give the information about
state-value distribution

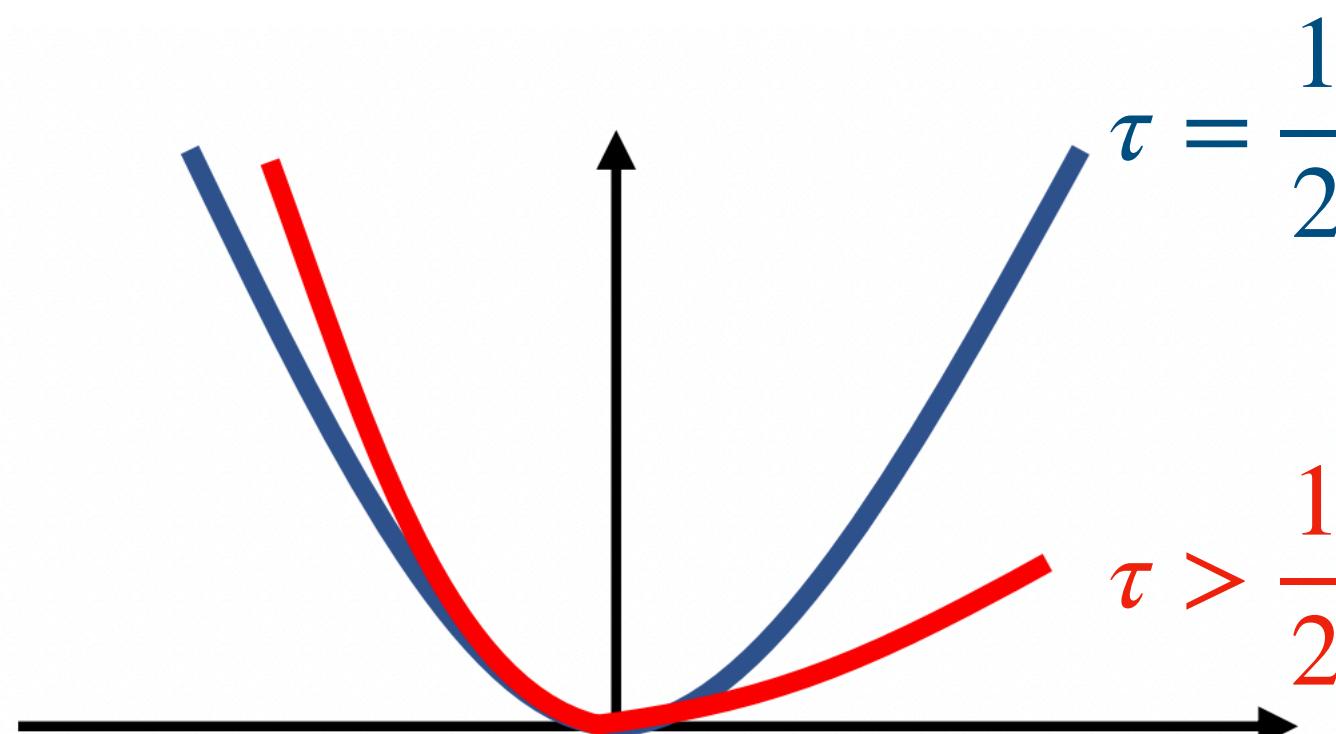
Implicit Q-Learning

$$\mathbb{E}_{a \sim \pi(\cdot|s)} Q^\pi(s, a) = V^\pi(s)$$

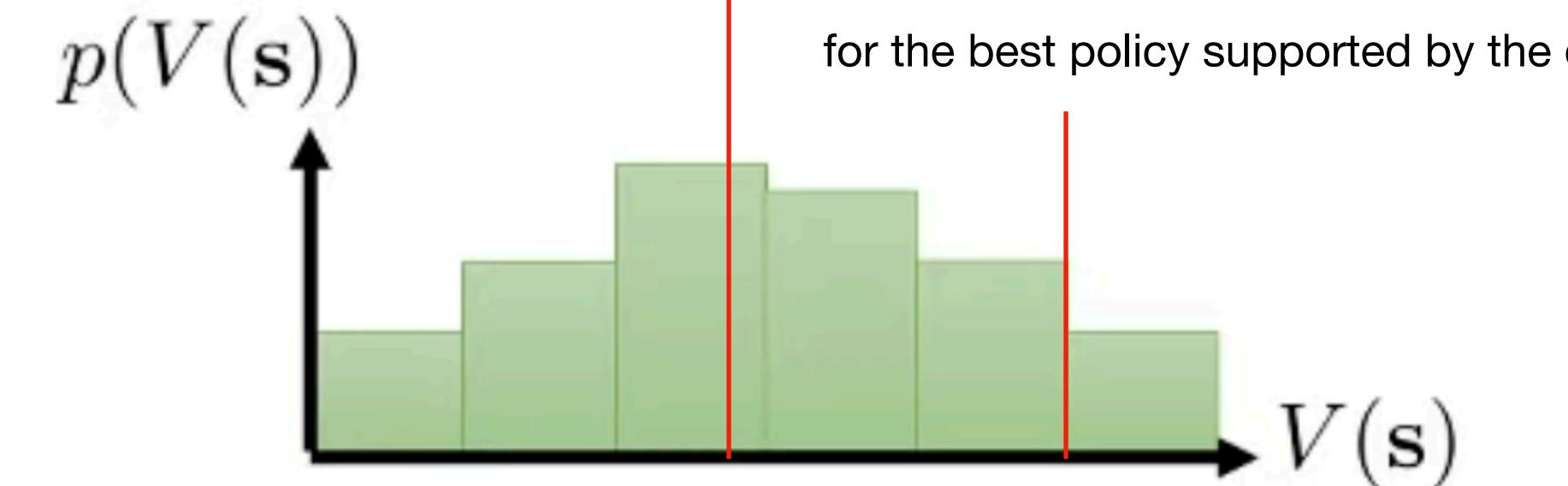
Approximate it with a neural network: $V^\pi \approx V_\psi$

Expectile loss:

$$L_2^\tau(u) = |\tau - \mathbb{I}(u < 0)| u^2$$



MSE gives us this ...but we want the value

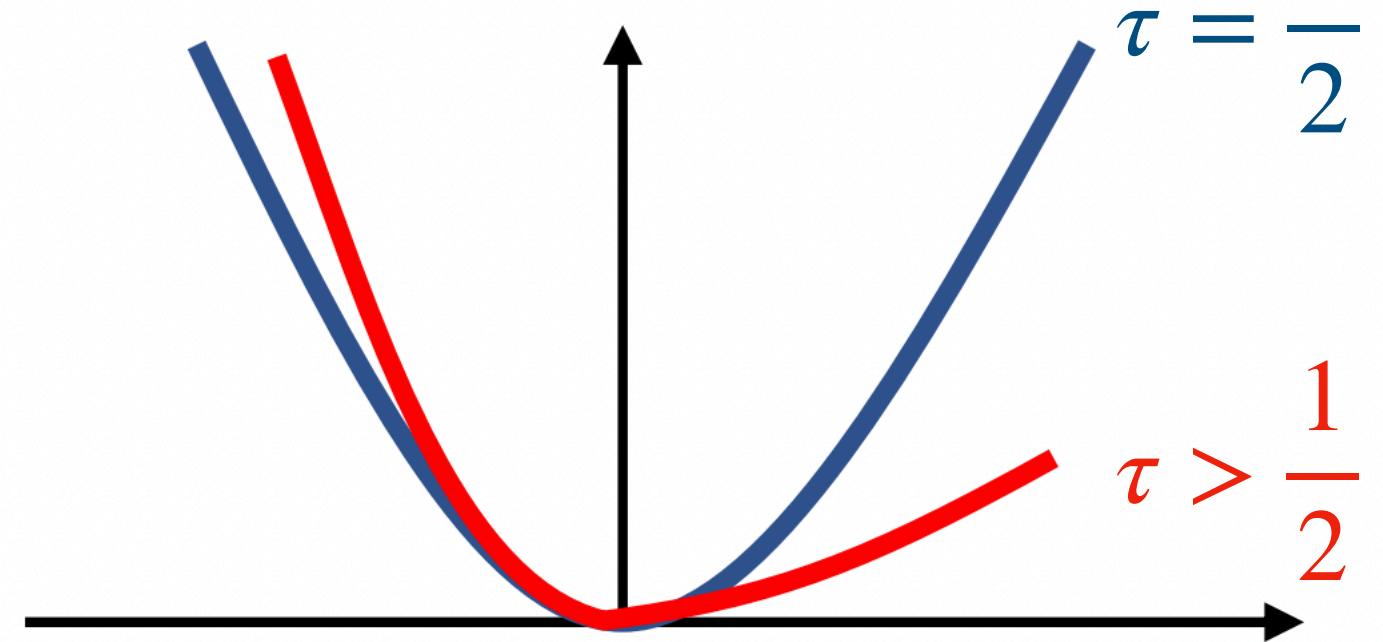


Similar states can give the information about state-value distribution

Implicit Q-Learning

Expectile loss:

$$L_2^\tau(u) = |\tau - \mathbb{I}(u < 0)| u^2$$



IQL:

$$L_V = \mathbb{E}_{(s,a) \sim D} [L_2^\tau(Q_{\phi^-}(s, a) - V_\psi(s))]$$

$$L_Q = \mathbb{E}_{(s,a,r,s') \sim D} [r + \gamma V_\psi(s') - Q_\phi(s, a)]^2$$

$$L_\pi = -\mathbb{E}_{(s,a) \sim D} \left[\log \pi_\theta(a | s) \frac{1}{Z(s)} \exp\left(\frac{A(s, a)}{\lambda}\right) \right],$$

$$\text{where } A(s, a) = Q_\phi(s, a) - V_\psi(s)$$

Conservative Q-Learning

$$L_{CQL} = \max_{\mu} [\alpha(\mathbb{E}_{s \sim D, a \sim \mu(\cdot|s)} Q_{\phi}(s, a)) + R(\mu)] - \alpha \mathbb{E}_{(s, a) \sim D} Q_{\phi}(s, a) + \\ + \frac{1}{2} \mathbb{E}_{(s, a, r, s') \sim D} [r + \gamma \mathbb{E}_{a' \sim \pi} Q_{\phi^-}(s', a') - Q_{\phi}(s, a)]^2 \rightarrow \min_{\phi}$$

Conservative Q-Learning

Common choice $R(\mu) = \mathbb{E}_{s \sim D} H(\mu(\cdot | s))$

$$\mathbb{E}_{s \sim D, a \sim \mu(\cdot | s)} [Q_\phi(s, a) + H(\mu(\cdot | s))] \rightarrow \max_{\mu} \implies \mu(a | s) \propto \exp(Q_\phi(s, a))$$

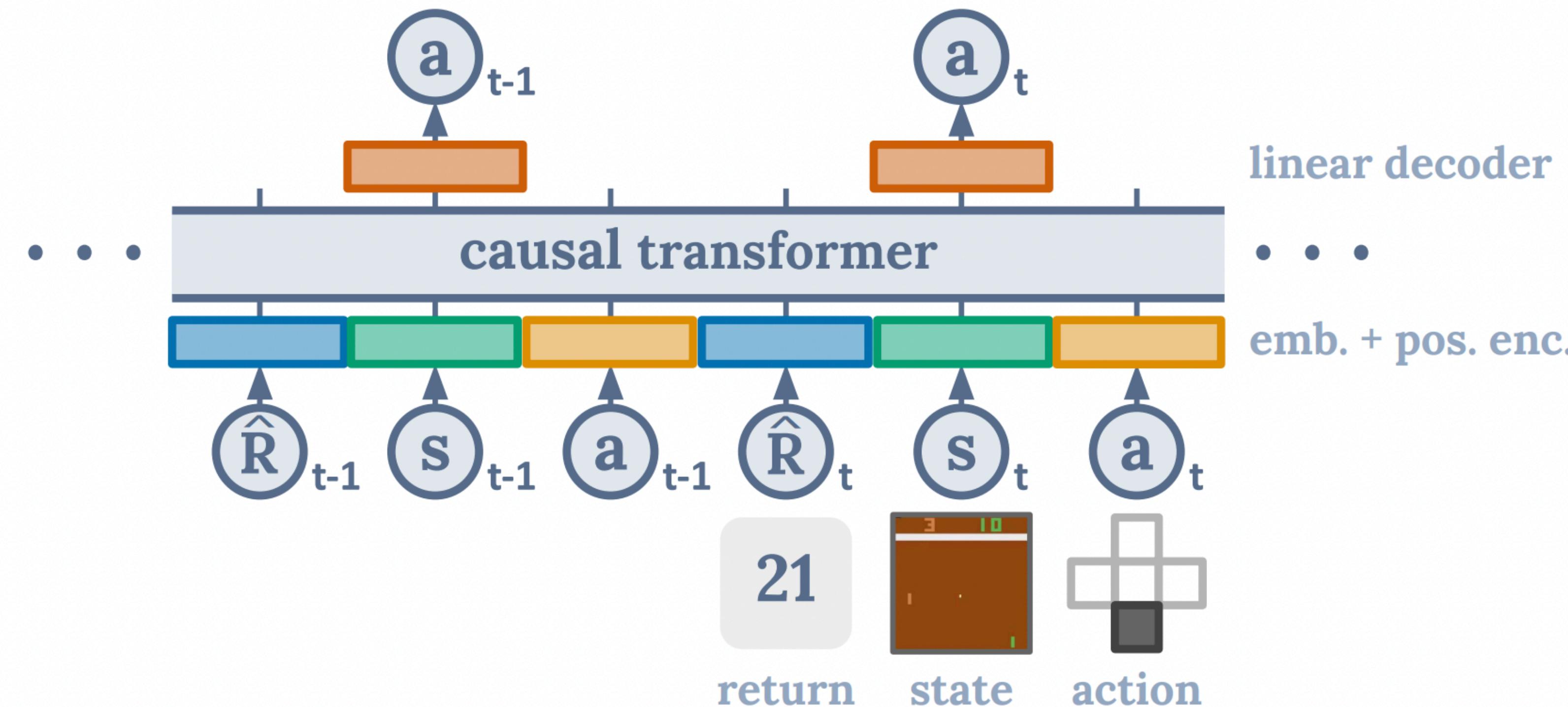
$$L_{CQL} = \alpha (\mathbb{E}_{s \sim D} \log \sum_a \exp(Q(s, a))) - \alpha \mathbb{E}_{(s, a) \sim D} Q_\phi(s, a)$$

$$+ \frac{1}{2} \mathbb{E}_{(s, a, r, s') \sim D} [r + \gamma \mathbb{E}_{a' \sim \pi} Q_{\phi^-}(s', a') - Q_\phi(s, a)]^2 \rightarrow \min_{\phi}$$

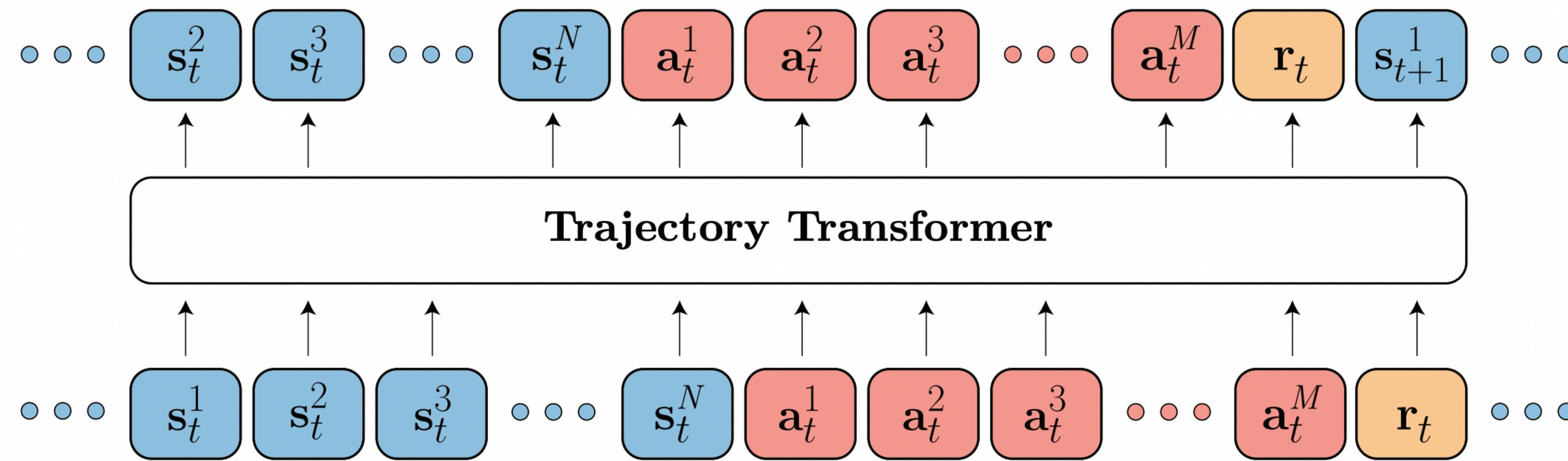
1. Learn Q_ϕ optimising L_{CQL} on offline data
2. Update policy with SAC-style entropy regularization:

$$\mathbb{E}_{s \sim D, a \sim \pi_\theta} [Q_\phi(s, a) - \log \pi_\theta(a | s)] \rightarrow \max_{\theta}$$

Decision Transformer



Trajectory Transformer



$$\mathcal{L}(\tau) = \sum_{t=1}^T \left(\sum_{i=1}^N \log P_\theta(s_t^i | \mathbf{s}_t^{<i}, \tau_{<t}) + \sum_{j=1}^M \log P_\theta(a_t^j | \mathbf{a}_t^{<j}, \mathbf{s}_t, \tau_{<t}) + \log P_\theta(r_t | \mathbf{a}_t, \mathbf{s}_t, \tau_{<t}) \right),$$

Planning with Beam Search using sum of rewards

Recap: TD3

1. Deterministic Policy Gradient:

$$\mathbb{E}_{s \sim D}[Q_\phi(s, \pi_\theta(s))] \rightarrow \max_{\theta}$$

2. Clipped Double-Q Learning:

$$y = r + \gamma \min_{i=1,2} Q_{\phi_i^-}(s', \mu_{\theta^-}(s'))$$

3. Delayed Policy Updates:

Update the policy (and target networks) less frequently than the Q-function.

$$y = r + \gamma \min_{i=1,2} Q_{\phi_i^-}(s', \mu_{\theta^-}(s'))$$

4. Target Policy Smoothing:

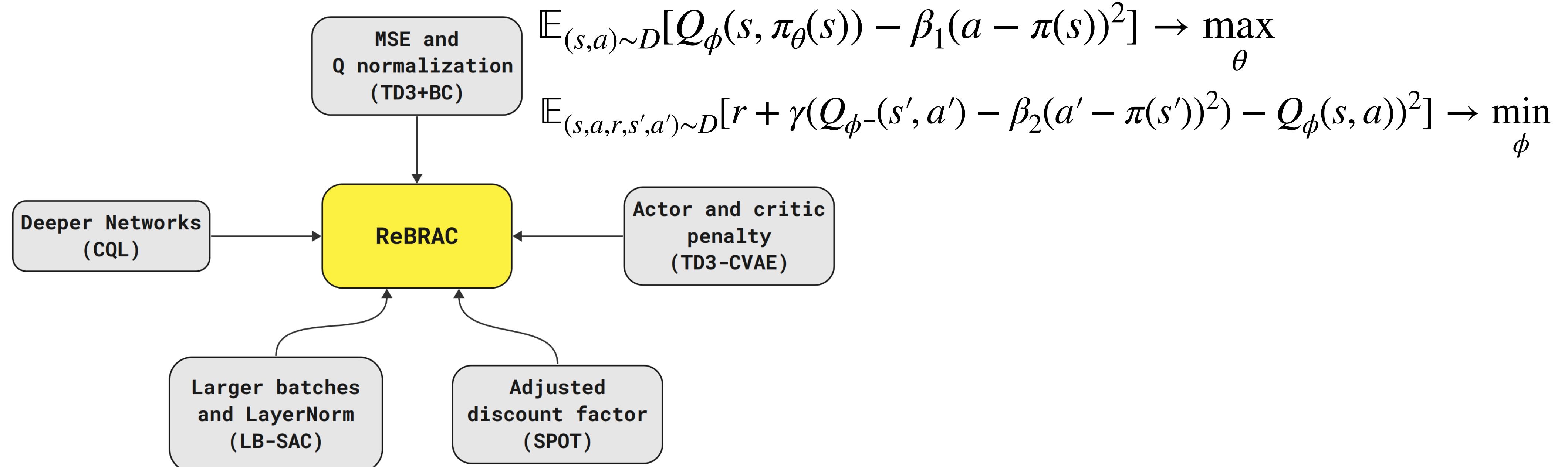
$$y = r + \gamma \min_{i=1,2} Q_{\phi_i^-}(s', a'), a' = \mu_{\theta^-}(s) + \varepsilon', \\ \varepsilon' \sim clip(\mathcal{N}(0, \sigma' I), -c, c)$$

The Minimalist Approach to Offline RL

TD3+BC:

$$\mathbb{E}_{(s,a) \sim D} [\lambda Q_\phi(s, \pi_\theta(s)) - (a - \pi_\theta(s))^2] \rightarrow \max_{\theta}$$

ReBRAC:



Not covered yet

1. Uncertainty-Based Offline Reinforcement Learning with Diversified Q-Ensemble
2. Anti-Exploration by Random Network Distillation

Background

1. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems
2. RL course by Sergey Levine, Lectures 15-16, Playlist
3. CORL: Research-oriented Deep Offline Reinforcement Learning Library

Thank you for your attention!