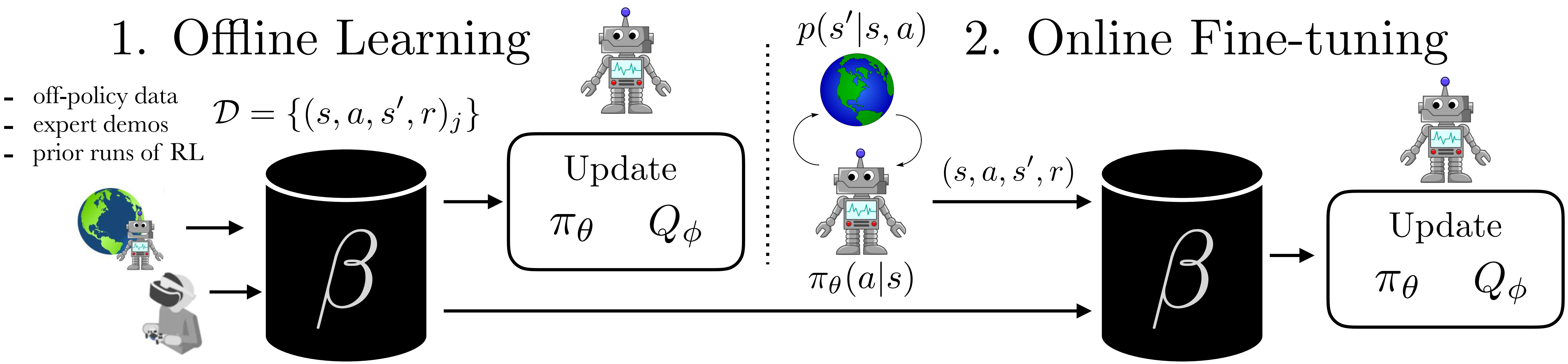
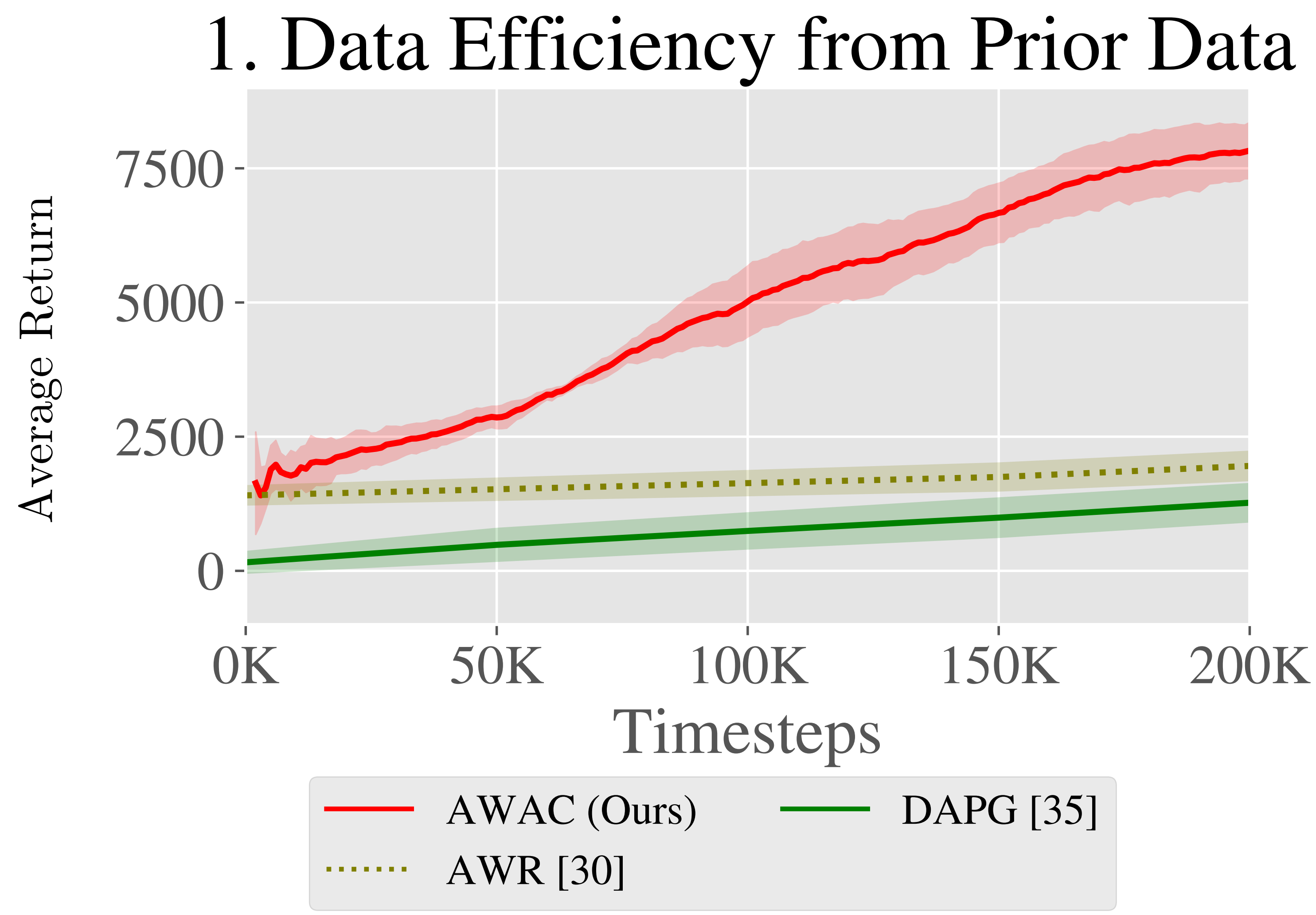
Accelerating Online Reinforcement Learning with Offline Datasets

Ashvin Nair, Murtaza Dalal, Abhishek Gupta, Sergey Levine
UC Berkeley

How can we learn difficult tasks with little online fine-tuning by using prior datasets?



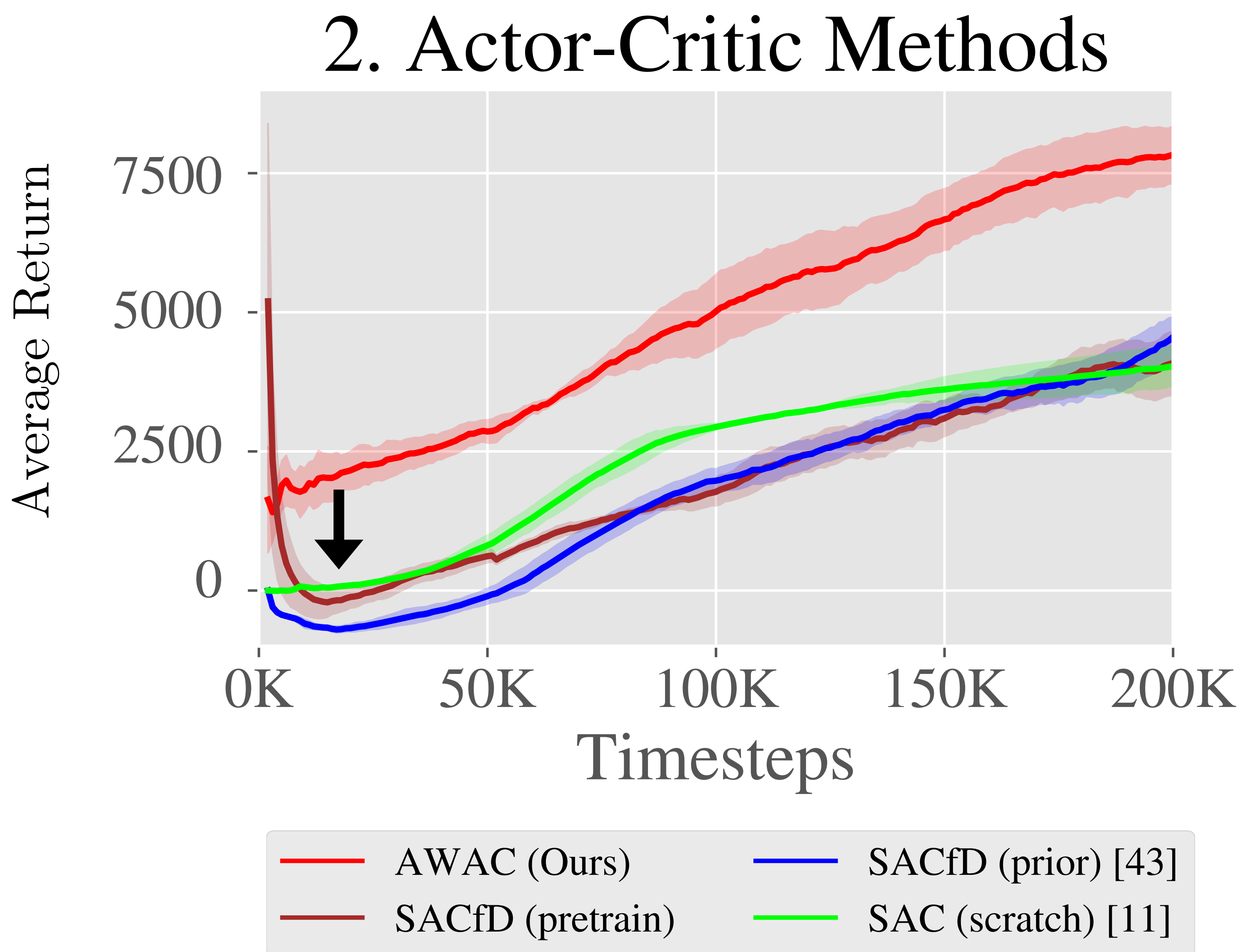
Challenges Fine-tuning with Existing Methods



1. On-policy fine-tuning methods exhibit slow online improvement.

[30] Advantage Weighted Regression: Simple and Scalable Off-Policy Reinforcement Learning. Peng et al. 2019.
[35] Learning Complex Dextrous Manipulation with Deep Reinforcement Learning and Demonstrations. Rajeswaran et al. 2017.

Challenges Fine-tuning with Existing Methods



2. Standard actor-critic methods do not take advantage of offline training, even if the policy is pretrained with behavioral cloning.

[11] Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. Haarnoja et al. 2019.
[43] Leveraging Demonstrations for Deep Reinforcement Learning on Robotics Problems with Sparse Rewarsd. Vecerik et al. 2017.

Challenges Fine-tuning with Existing Methods

Policy Improvement Step

$$\pi_{k+1} = \arg\max_{\pi \in \Pi} \mathbb{E}_{\mathbf{a} \sim \pi(\cdot|\mathbf{s})} [Q^{\pi_k}(\mathbf{s}, \mathbf{a})]$$

$$\text{s.t. } D(\pi(\cdot|\mathbf{s}) || \pi_\beta(\cdot|\mathbf{s})) \leq \epsilon$$

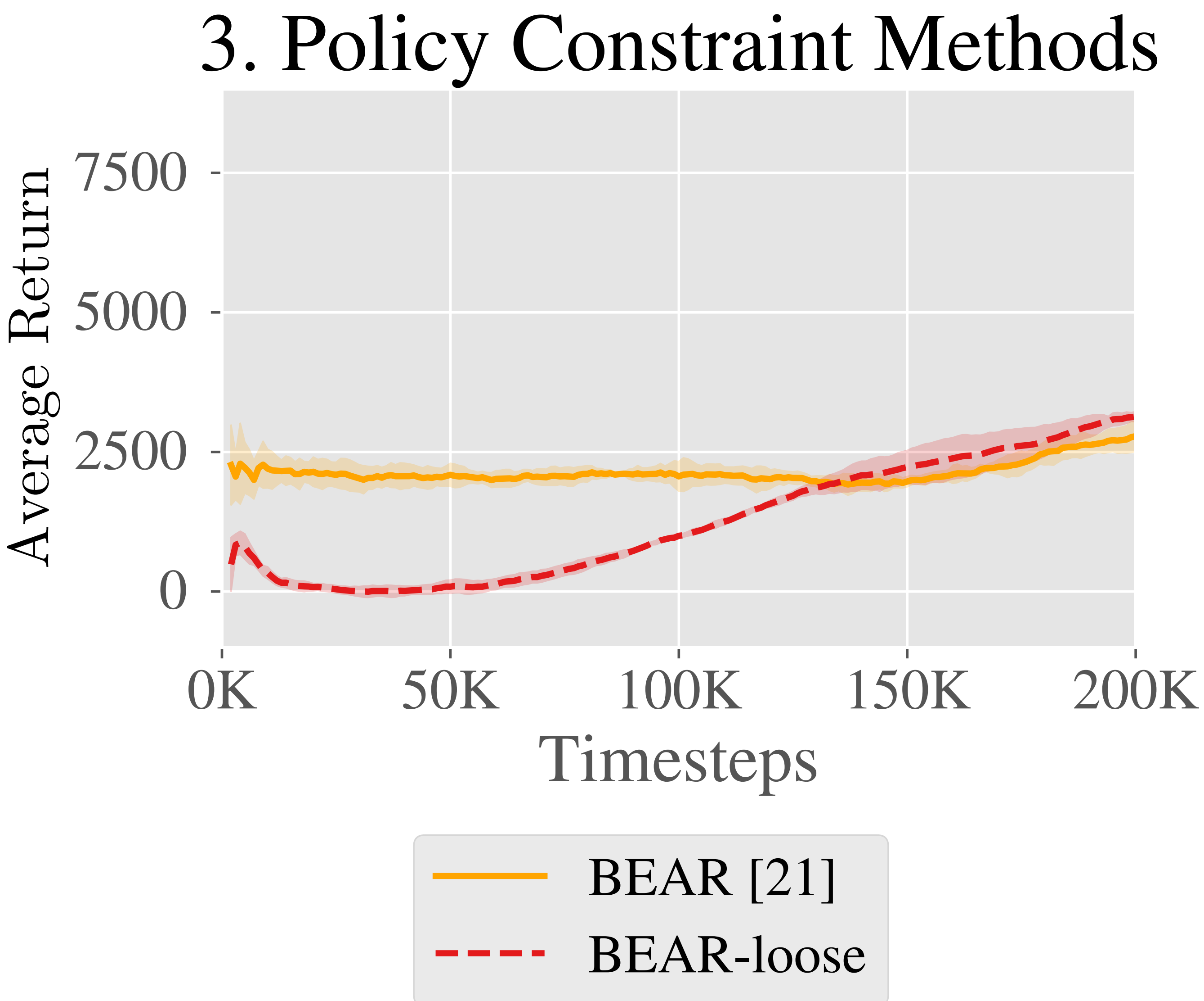
Maximize estimated returns

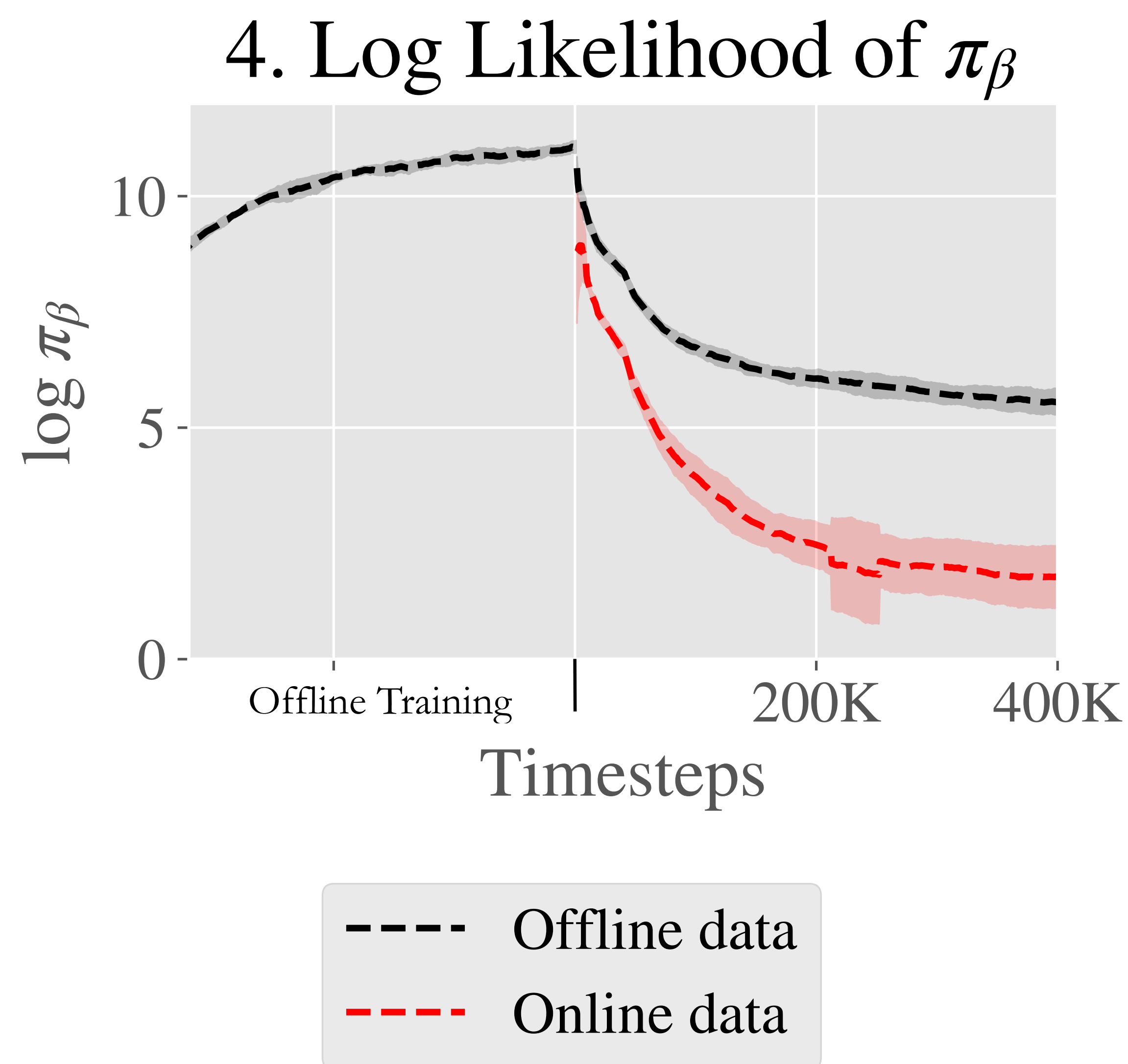
while staying close to the data distribution

Trained by supervised learning

Actor-critic methods can be stabilized for offline training by incorporating a policy constraint in the policy improvement step.

Challenges Fine-tuning with Existing Methods





3. Existing policy constraint methods (BEAR [21], ABM [38], BRAC [46]) rely on behavior models of prior data, which are difficult to train online.

[21] Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction. Kumar et al. 2019.
[38] Keep Doing What Worked: Behavior Modelling Priors for Offline Reinforcement Learning. Siegel et al. 2019.
[46] Behavior Regularized Offline Reinforcement Learning. Yifan Wu et al. 2019.

Advantage Weighted Actor Critic (AWAC)

Policy Improvement Step

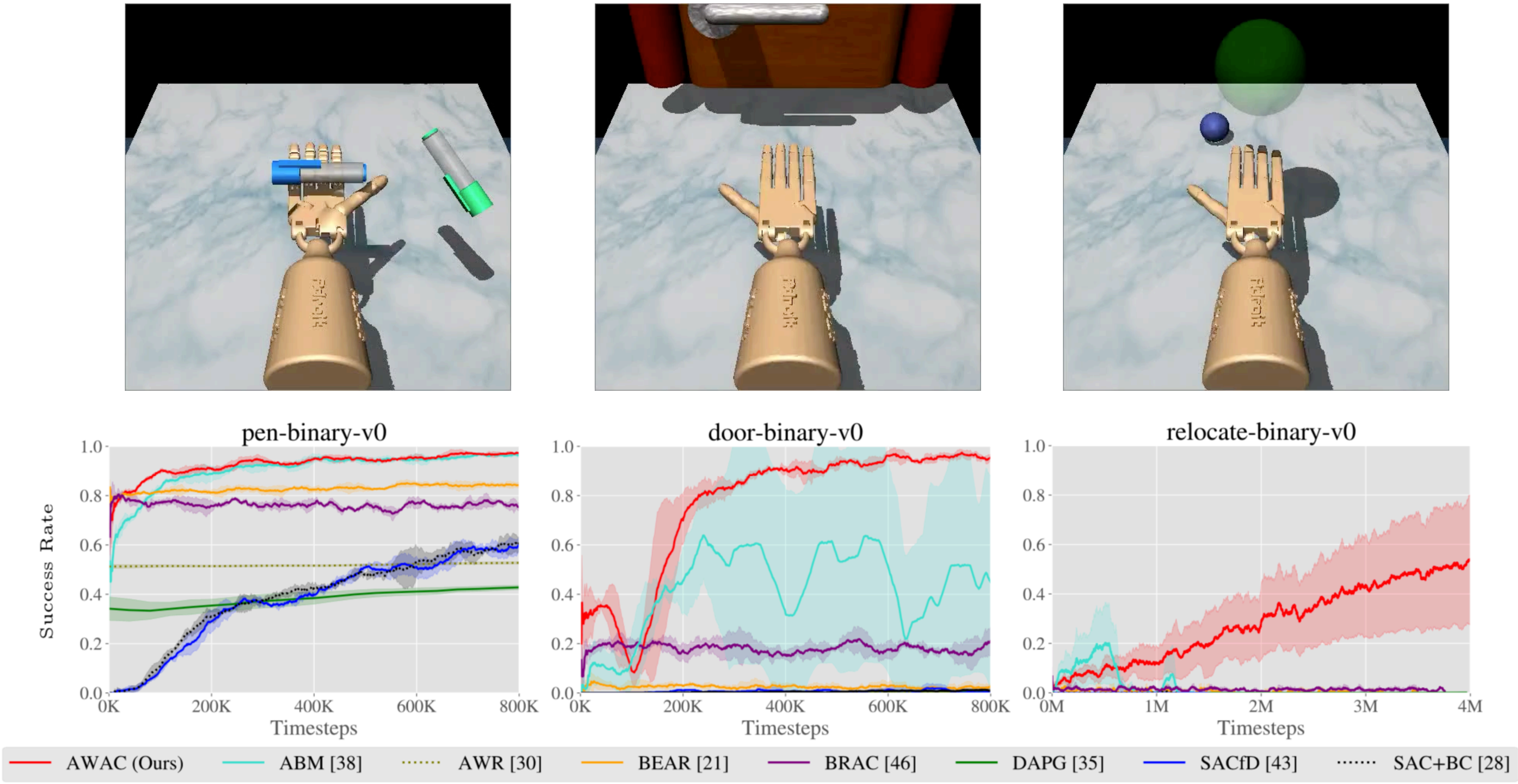
$$\pi_{k+1} = \arg\max_{\pi \in \Pi} \mathbb{E}_{\mathbf{a} \sim \pi(\cdot|\mathbf{s})}[A^{\pi_k}(\mathbf{s}, \mathbf{a})]$$
$$\text{s.t. } D(\pi(\cdot|\mathbf{s}) || \pi_\beta(\cdot|\mathbf{s})) \leq \epsilon$$

AWAC incorporates a KL constraint into the actor-critic framework implicitly

$$\theta_{k+1} = \arg\max_{\theta} \mathbb{E}_{\mathbf{s},\mathbf{a} \sim \beta}\left[\log \pi_\theta(\mathbf{a}|\mathbf{s}) \frac{1}{Z(\mathbf{s})} \exp\left(\frac{1}{\lambda} A^{\pi_k}(\mathbf{s}, \mathbf{a})\right)\right]$$

AWAC trains well offline, fine-tunes quickly online, and does not need to estimate a behavior model.

Dextrous Manipulation Tasks



Our algorithm can be used to solve difficult dextrous manipulation tasks - it solves door opening in **under 1 hour** of online interaction.

Accelerating Online Reinforcement Learning with Offline Datasets

Ashvin Nair, Murtaza Dalal, Abhishek Gupta, Sergey Levine
UC Berkeley