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?

1. Offline Learning

- off-policy data
- expert demos
- prior runs of RL

\[ D = \{(s, a, s', r)\}_j \]

Update

\[ \pi_\theta \quad Q_\phi \]

2. Online Fine-tuning

\[ p(s'|s, a) \]

Update

\[ \pi_\theta(a|s) \quad (s, a, s', r) \]

\[ \pi_\theta \quad Q_\phi \]
1. On-policy fine-tuning methods exhibit slow online improvement.

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

Challenges Fine-tuning with Existing Methods

Policy Improvement Step

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

Maximize estimated returns

Trained by supervised learning

while staying close to the data distribution

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.

Advantage Weighted Actor Critic (AWAC)

Policy Improvement Step

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

s.t. \( D(\pi(\cdot|s) \| \pi_\beta(\cdot|s)) \leq \epsilon \)

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

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