Off-Policy Deep Reinforcement Learning with Analogous Disentangled Exploration

Anji Liu, Yitao Liang, Guy Van den Broeck

Computer Science Department, UCLA
Motivation: Why Reinforcement Learning

Is capable of solving large-scale and complex problems.

Learning through a trial-and-error process with little supervision.

Video games

Autonomous driving

Go
Background: Markov Decision Process

Environment

At time step $t$:
- state $s_t$
- action $a_t$
- reward $r_t$
- next state $s_{t+1}$

Agent

Policy

Determine which action to take given a state.

Value function

Measures discounted long-term reward

$$Q(s, a) = \mathbb{E}_{a_t \sim \pi, s_{t+1} \sim P, r_t \sim R} \left[ \sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, a_0 = a \right]$$

Goal of RL

Find a policy $\pi$ that maximizes the expected long-term reward.

$$\mathbb{E}_{s \sim \rho_0, a \sim \pi} [Q(s, a)]$$

where $\rho_0$ is the initial state distribution.
The Exploration-Exploitation Tradeoff

Exploration

Experiencing new states/rewards to potentially find better policies.

\[ \downarrow \]

expressiveness, stochastic

Exploitation

Utilize existing knowledge to behave optimally.

\[ \downarrow \]

optimality, deterministic

The necessity of two policies\(^1\): target policy \( \pi \) and behavior policy \( \mu \)

\(^1\)For simplicity, sometimes only one policy is explicitly constructed.
Goals

What we want to achieve?

Leverage the flexibility to **explicitly design two policies** to better balance the exploration-exploitation tradeoff.

How we achieve it?

Analogous Disentangled Actor Critic

- Restricting the disentangled behavior policy (**policy co-training**).
- Restricting value updates (**critic bounding**).
Combating the Exploration-Exploitation Dilemma

Exploration

Expressive behavior policies

e.g. maximum entropy RL objective
\[ \mu(a \mid s) \propto \exp(-E(s, a)) \]

Potential failure on exploitation 😞

Exploitation

Exploitative target policies

Always optimal w.r.t. current knowledge
\[ \pi(a \mid s) = \arg\max_a Q(s, a) \]

Usually requires uninformative noise-based exploration. 😞

Achieving better tradeoff through the flexibility to design separated policies?
Analogous Disentangled Actor-Critic (ADAC)

Environment → Collect → Replay Buffer

Sample collection
Model update

Interaction (during training)

Policy co-training with shared network

Interaction (during testing)

Replay Buffer → Sample → Batch of samples

$\mu$ → $\pi$

Analogous Disentangled Actor

$Q_{R'}^{\pi}$ → $Q_R^{\pi}$

Policy gradient

Joint update with critic bounding

Interaction
Stabilizing Policy Updates by Policy Co-training

Motivation

Expressiveness of the behavior policy

- Improve efficiency of exploration 😊
- Increased distance between target policy and behavior policy

Stability problems during learning 🙁

Obtain expressive while stable and optimal behavior policy.
Stabilizing Policy Updates by Policy Co-training

Key observation

- Classic policy optimization
  \[ \pi(a \mid s) = \arg \max_a Q(s, a) \]
  Deterministic policy gradient

- Maximum-entropy policy optimization
  \[ \mu(a \mid s) \propto \exp(Q(s, a)) \]
  Deterministic policy gradient + entropy regularization

**Joint learning** of behavior/target policy in a **shared neural network**.
Analogous Disentangled Behavior Policy

\[
\begin{align*}
\text{state} & \quad s \\
\text{target policy} & \quad \xi = [0, 0, \ldots, 0]^T \\
\text{behavior policy} & \quad \xi \sim \mathcal{N}(0, I)
\end{align*}
\]

Benefits

Increase the effectiveness of the behavior policy
- Reduce the distance between two policies and stabilize the learning process.
- Allowing expressive behavior policy for efficient exploration.
**Incorporating Intrinsic Reward via CriticBounding**

**Motivation**

Intrinsic reward

Guided and efficient exploration

Alter the environment-defined objective

Letting intrinsic reward to **only affect the behavior policy.**
Incorporating Intrinsic Reward via Critic Bounding

\[ Q^\pi_R \] value function w.r.t. the target policy and the environment-defined reward

\[ Q^\pi_{R'} \] value function w.r.t. the target policy and the enhanced reward, where \( R' = R + R^{in} \)

\[ \rightarrow \] Adopted to improve the target policy.

\[ \rightarrow \] Adopted to improve the behavior policy.

Theoretical justification

- Bounded training stability.
- Bounded training effectiveness.
Summary of Advantages

Policy co-training

- Binds the target policy and the behavior policy to stabilize the training process.
- Allowing expressive behavior policy.

Critic bounding

- Incorporate intrinsic reward for effective exploration.
- Has no effect on the optimality of the target policy.
Analysis of Analogous Disentangled Behavior Policy

Key results

Behavior policy
- Act curiously at the beginning.
- Focus on potentially rewarding actions after obtaining preliminary understanding of the environment.

Target policy
- Remains optimal w.r.t. the current knowledge.

Target policy & behavior policy
- Bias between them significantly reduced.
Comparison with the State of the Art

Out-performs state-of-the-art methods in 10 out of 14 benchmark environments.
- ADAC consistently out-performs its base model, and retains the benefits of improvements developed by the base models.
- Comparison with SAC reveals the benefit brought by the disentangled structure.

<table>
<thead>
<tr>
<th>Environment</th>
<th>ADAC (DDPG)</th>
<th>ADAC (TD3)</th>
<th>TD3</th>
<th>DDPG</th>
<th>SAC</th>
<th>PPO</th>
</tr>
</thead>
<tbody>
<tr>
<td>RoboschoolAnt</td>
<td>2219 ± 373</td>
<td>2299† ± 333</td>
<td>2903†</td>
<td>450.0 ± 27.9</td>
<td>2726 ± 652</td>
<td>1280 ± 71</td>
</tr>
<tr>
<td>RoboschoolHopper</td>
<td>1578† ± 166</td>
<td>1711* ± 95</td>
<td>2302†</td>
<td>543.8 ± 307</td>
<td>2089 ± 657</td>
<td>1229 ± 345</td>
</tr>
<tr>
<td>RoboschoolHalfCheetah</td>
<td>3353 ± 847</td>
<td>1769† ± 452</td>
<td>955.1 ± 146.3</td>
<td>208.7 ± 137.1</td>
<td>1021 ± 263</td>
<td>578.9 ± 231.3</td>
</tr>
<tr>
<td>RoboschoolAtlasForwardWalk</td>
<td>234.6 ± 55.7</td>
<td>186.7* ± 37.9</td>
<td>607.2 ± 246.2</td>
<td>441.6 ± 120.4</td>
<td>807.0 ± 252.6</td>
<td>1225 ± 184.2</td>
</tr>
<tr>
<td>RoboschoolWalker2d</td>
<td>1679† ± 452</td>
<td>1564* ± 651</td>
<td>10526† ± 2367</td>
<td>208.7 ± 137.1</td>
<td>1021 ± 263</td>
<td>578.9 ± 231.3</td>
</tr>
<tr>
<td>Ant</td>
<td>3598† ± 374</td>
<td>374.5* ± 36.5</td>
<td>2845 ± 609</td>
<td>1009 ± 49</td>
<td>11541 ± 2989</td>
<td>881.7 ± 10.1</td>
</tr>
<tr>
<td>Hopper</td>
<td>9392 ± 199</td>
<td>2238* ± 40</td>
<td>10526† ± 2367</td>
<td>1009 ± 49</td>
<td>11541 ± 2989</td>
<td>881.7 ± 10.1</td>
</tr>
<tr>
<td>HalfCheetah</td>
<td>5122† ± 1314</td>
<td>1291* ± 42</td>
<td>4630† ± 778</td>
<td>186.2 ± 33.3</td>
<td>4067 ± 1211</td>
<td>1146 ± 368</td>
</tr>
<tr>
<td>Walker2d</td>
<td>1000† ± 0</td>
<td>1000* ± 0</td>
<td>1000† ± 0</td>
<td>1000* ± 0</td>
<td>1000* ± 0</td>
<td>98.90 ± 2.08</td>
</tr>
<tr>
<td>InvertedPendulum</td>
<td>9359† ± 0.17</td>
<td>9334* ± 1.39</td>
<td>7665 ± 566</td>
<td>27.20 ± 2.61</td>
<td>9353 ± 2896</td>
<td>98.90 ± 5.88</td>
</tr>
<tr>
<td>InvertedDoublePendulum</td>
<td>309.8† ± 15.6</td>
<td>-52.77* ± 1.94</td>
<td>288.4* ± 51.25</td>
<td>-123.90 ± 11.17</td>
<td>307.2* ± 57.92</td>
<td>266.9 ± 28.52</td>
</tr>
<tr>
<td>BipedalWalker</td>
<td>-10.76† ± 27.70</td>
<td>-98.52* ± 3.21</td>
<td>-57.97 ± 21.08</td>
<td>-50.05* ± 10.27</td>
<td>-127.4 ± 45.2</td>
<td>-105.3 ± 22.2</td>
</tr>
<tr>
<td>BipedalWalkerHardcore</td>
<td>290.0† ± 50.9</td>
<td>85.67* ± 23.42</td>
<td>289.7† ± 54.1</td>
<td>-65.89 ± 96.48</td>
<td>283.3 ± 69.29</td>
<td>59.32 ± 68.44</td>
</tr>
</tbody>
</table>

Off-Policy Deep Reinforcement Learning with Analogous Disentangled Exploration May 2020 15
Why intrinsic reward sometimes harms RL

Environment-defined reward stays the same while intrinsic reward keep growing.
ADAC out-performs baseline methods on challenging sparse-reward tasks when using intrinsic reward.
Thank You