Contrastive UCB: Provably Efficient Contrastive Self-Supervised Learning in Online Reinforcement Learning

Shuang Qiu[1] Lingxiao Wang[2] Chenjia Bai[3] Zhuoran Yang[4] Zhaoran Wang[2]

[1] University of Chicago [2] Northwestern University [3] Shanghai AI Lab [4] Yale University

ICML 2022

Contrastive Learning in RL

- Deep reinforcement learning (RL)
  - Representation power of the neural networks
  - Challenge: millions of interactions with the environment
- Low-dimensional representation learning
  - Improve sample efficiency
  - Learn representation via solving auxiliary problems
  - Contrastive self-supervised learning

Motivation

- Empirical studies on contrastive learning in RL:
  - Temporal information
  - Local spatial structure
  - Image augmentation
  - Return feedback
  - ...
- Our work: theoretical understanding of contrastive learning in RL
  - Temporal information

Contribution

- The first provable UCB-based RL algorithm that incorporates a contrastive loss
- Prove that our algorithm recovers the true representations via contrastive learning and simultaneously achieves sample efficiency
- Provide empirical studies to show the efficacy of the UCB-based RL method with contrastive learning inspired by our theory
- Extend our findings to zero-sum Markov games (MGs) which reveals a new direction

Problem Setting

- Episodic MDP
  - ε-suboptimal policy π

$$\max_{\pi'} V_1^{\pi'}(s_1, r) - V_1^{\pi}(s_1, r) \leq \varepsilon$$

- Episodic zero-sum MG
  - ε-approximate Nash equilibrium (NE) (π, ν)

$$\max_{\pi'} V_1^{\pi',\nu}(s_1, r) - \min_{\nu'} V_1^{\pi,\nu'}(s_1, r) \leq \varepsilon$$

- Low-rank transition dynamics

$$\mathbb{P}_h(s'|z) = \psi_h^*(s')^\top \phi_h^*(z)$$

  - Both ψ_h^* and ϕ_h^* are unknown, different from the linear MDP setting
  - $z = (s, a)$ for MDPs and $z = (s, a, b)$ for MGs

Algorithms

- Contrastive UCB
  - UCB-based value iteration + contrastive loss minimization
  - Contrastive loss

$$\mathcal{L}_h(\psi, \phi; \mathcal{D}_h^k) := \mathbb{E}_{(s,a,s',y)\sim\mathcal{D}_h^k}\left[y \log(1 + 1/\psi(s')^\top \phi(s,a)) + (1-y)\log(1 + \psi(s')^\top \phi(s,a))\right]$$

    - ★ Negative sample distribution $\mathcal{P}_{\mathcal{S}}^{-}(\cdot)$
    - ★ Temporal contrastive data $y \sim \mathrm{Bernoulli}(1/2)$, $s' \sim \mathbb{P}_h(\cdot|s,a)$ if $y = 1$ and $s' \sim \mathcal{P}_{\mathcal{S}}^{-}(\cdot)$ otherwise
    - ★ Function spaces: $\phi \in \Phi$ and $\psi \in \Psi$
  - Exploration via UCB bonus: constructed based on the learned $\phi(s,a)$
- Contrastive ULCB
  - ULCB-based value iteration + contrastive loss minimization

Theoretical Results

Theorem 1

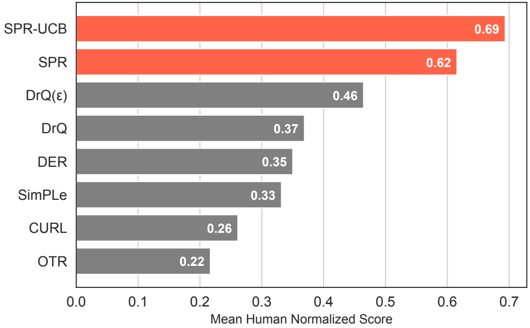
Setting proper parameters, with high probabilities, our algorithms ensure

- *the learned representations recover the true transitions,*
- *after K rounds, the generated policy is*
  - $\widetilde{\mathcal{O}}(\sqrt{\log(|\Psi||\Phi|)/K})$*-suboptimal policy for single-agent MDPs,*
  - $\widetilde{\mathcal{O}}(\sqrt{\log(|\Psi||\Phi|)/K})$*-approximate NE for Markov games.*

- Function space complexity: $\log(|\Psi||\Phi|)$
- Sample complexity: $\widetilde{\mathcal{O}}(1/\varepsilon^2)$ to achieve ε-suboptimal policy or ε-approximate NE

Proof-of-Concept Experiments

- Setup
  - SPR architecture
  - UCB bonus: the last layer $\phi(s, a)$
  - SPR-UCB: SPR + UCB bonus
  - Atari-100K benchmark
  - https://github.com/Baichenjia/Contrastive-UCB
- Results
  - SPR-UCB outperforms SPR and other baseline algorithms.
  - More results in our paper



Thank you!