불완전한 시연 데이터를 이용한 오프라인-온라인 강화학습을 위한 심층 반복 규제 기반 정책 최적화

장현성, 김정현

세종대학교

23012724@sju.ac.kr, j.kim@sejong.ac.kr

Deep Iterative Regularized Policy Optimization for Offline-to-Online Reinforcement Learning with Imperfect Demonstrations

Hyeonseong Chang, Junghyun Kim

Sejong Univ.

요 약

본 연구에서는 불완전한 시연 데이터를 활용하여 효율적으로 최적 정책을 학습하는 심층 반복 제약 오프라인-온라인 (offline-to-online) 강화학습 모델을 제안한다. 제안 모델은 강화학습에 특화된 심층 신경망 구조인 SimBa를 도입하여 모델의 표현력을 강화하고, 학습 진행도에 따라 정책 차이 규제 강도를 점진적으로 완화하는 어닐링 (annealing) 기법을 적용하였다. 4가지의 다양한 로봇 제어 시뮬레이션 환경에서 실험을 수행한 결과, 제안 모델은 기존 모델 대비 모든 환경에서 큰 폭의 성능 향상을 달성하였다.

Ⅰ. 서론

최근 로봇 제어 분야에서는 사전에 수집된 데이터셋을 활용하여 로봇을 학습시키는 데이터 기반 (data-driven) 접근법인 오프라인 강화학습이 차세대 핵심 기술로 주목받고 있다. 오프라인 강화학습의 핵심은 시연 (demonstration) 데이터셋 내의 행동 분포를 효과적으로 모델링하는 데 있다. 에이전트의 최종 성능은 학습 데이터의 품질과 커버리지 (coverage)에 크게 의존한다. 그러나 로봇 제어와 같은 고차원 환경에서 고품질의 전문가 시연 데이터를 확보하는 것은 비용이 크다. 이러한 문제를 해결하기 위해, 최근에는 오프라인 강화학습으로 얻은 정책을 초기 모델로 삼아, 실제 환경에서의 추가적인 상호작용을 통해 성능을 향상시키는 오프라인-온라인 (offline-to-online) 강화학습이 활발히 연구되고 있다 [1].

특히, 상대적으로 품질이 낮은 시연 데이터로 학습을 시작하더라도 점진적으로 최적 성능에 도달할 수 있는 방법론인 반복 규제 정책 최적화 (Iterative Regularized Policy Optimization, IRPO) [2]가 제안된 바 있다. IRPO는 오프라인 단계에서 행동 모방을 통해 기준 정책 (baseline policy)을 형성하고, 온라인 단계에서는 이 기준 정책과의 쿨백-라이블러 발산 (Kullback-Leibler divergence, KLD)을 제약 조건으로 두어 급격한 정책 변화를 방지하면서 최적 정책을 탐색한다. IRPO는 근위 정책 최적화 (Proximal Policy Optimization, PPO) [3]를 기반으로 동작한다.

IRPO는 3계층 내외의 얕은 신경망 (shallow network) 구조에 의존하고 있어, 복잡한 로봇의 동역학이나 방대한 상태 공간을 정교하게 표현하는 데 한계가 있다. 이에 본 연구에서는 시연 데이터의 행동 분포를 보다 정밀하게 근사하고 최적 정책의 수렴성을 높이기 위해, 강화학습에 특화된 심층 신경망 아키텍처인 SimBa [4]를 도입하여 모델의 표현력을 개선하고자 한다. 또한, 학습 초기에는 강력한 규제를 통해 안정성을 확보하고 후반부에는 규제를 완화하여 탐색을 촉진할 수 있도록, KLD 제약 계수를 점진적으로 감소시키는 어닐링 (annealing) 기법을 적용하였다. 4가지의 강화학습 시뮬레이션 환경에서 실험을 수행한 결과, 제안 모델은 기존 모델인 IRPO 대비 큰 폭의 성능 향상을 달성하였다.

Ⅱ. 본론

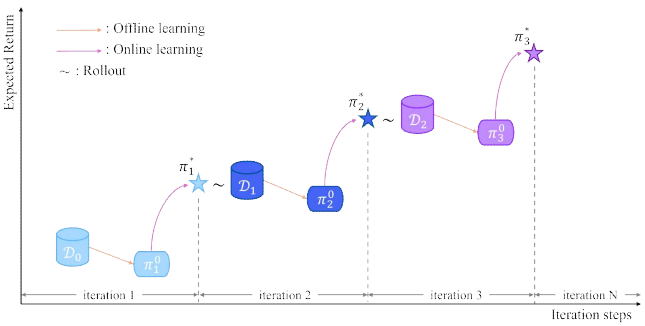

그림 1. 오프라인-온라인 강화학습 반복 순서.

그림 1은 오프라인-온라인 강화학습의 반복 순서를 나타낸 그림이다. 에이전트는 $i \geq 0$번째 불완전 시연 데이터셋 D_i의 행동 분포를 모방하도록 학습한다. 에이전트는 오프라인 강화학습으로 수식 (1)과 같은 손실함수를 사용하는 행동 모방 (behavior cloning) 알고리즘을 사용한다.

$$L_{BC}(\theta) = -E_{(s,a)\sim D_i}[\log \pi_\theta(a|s)], \quad (1)$$

여기서 $(s,a) \sim D_i$는 i번째 데이터셋의 상태, 행동 쌍을 의미하고, π_θ는 신경망으로 근사하여 찾고자 하는 행동 정책 (policy)이다. 신경망으로 근사하여 얻은 행동 모방 정책 (π^0_{i+1})를 기준점으로 잡는다. 이후, KLD 제약을 추가하여 수식 (2)와 같은 손실함수를 사용하는 온라인 강화학습을 통해 최적 정책을 근사한다.

$$L_{KLD}(\theta) = E[\sum_{t=0}^{H-1} \gamma^t (r(s_t,a_t) - \lambda_k \log(\frac{\pi_\theta(a_t|s_t)}{\pi_k(a_t|s_t)})], \quad (2)$$

여기서 t는 타임 스텝, H는 에피소드 길이 그리고 k는 반복 순서를 의미한다. γ는 할인율로 미래 보상의 영향을 조절하는 역할을 한다. KLD는 급진적인 정책 변화를 방지하기 위해 현재 정책과 행동 모방 정책 간의 분포 차이를 규제한다. KLD 규제 정도는 하이퍼파라미터 λ_k로 이루어진다. 수식 (2)에서 π_θ는 최적 정책을 찾기 위해 보상 $r(s_t,a_t)$를 최대화하는 방향으로 업데이트된다. 다음 반복을 위해 k번째 반복의 최적 정책 π^*_k

표 1. 제안 모델과 IRPO의 환경 별 성능 비교.

모델	KL 계수 조절	Halfcheetah [5]	Hopper [5]	Reach [6]	Attitude Control [7]
IRPO [3]	Iterative	7678.4	3044.7	-5.32	0.54
제안 모델	Annealed	10879.4	3307.1	-3.77	0.76
향상률(%)	-	**41.7**	**8.6**	**29.1**	**40.7**

로부터 불완전 시연 데이터셋 D_{i+1}을 수집 (rollout)한다. 위 과정을 정책이 수렴할 때까지 반복한다.

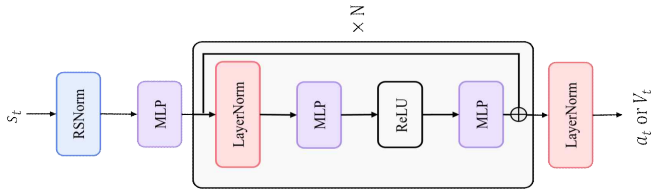

그림 2. 강화학습 특화 심층화 구조: SimBa

PPO 신경망을 안정적으로 심층화할 수 있는 구조인 SimBa는 그림 2와 같다. SimBa는 입력값의 분산이 높은 강화학습의 특징을 고려하여 이동 통계 정규화 (running statistics normalization, RSNorm)와 레이어 정규화 (layer normalization, LayerNorm), 잔차 연결을 다층 퍼셉트론 (multilayer perceptron, MLP) 사이에 적절히 배치하여 설계되었다. 본 논문에서는 블록 개수로 N=2를 사용하였다.

실험 환경은 d4rl [5]의 halfcheetah 환경과 hopper 환경, panda-gym [6]의 reach 환경 그리고 fly-craft [7]의 attitude control 환경을 사용하였다. Halfcheetah와 hopper 환경에서 사용한 불완전 시연 데이터셋은 d4rl에서 제공하는 halfcheetah-medium-v2와 hopper-medium-v2 데이터셋을 사용하였다. Panda-gym의 reach 환경은 비례-적분-미분 (proportional-integral-derivative, PID) 제어기를 사용하여 생성한 데이터를 사용하였다. 마지막으로 fly-craft의 attitude control 환경에서도 PID 제어기를 통해 생성한 불완전 시연 데이터를 사용하였다 [2].

실험 환경별로 λ_k는 0.2에서 0.01로 점차 줄이는 어닐링 방식을 사용하였다. 단, attitude control은 저품질 시연 데이터의 영향을 줄이기 위해 첫 번째 반복에서만 $\lambda_1 = 0.001$로 고정하여 실험하였다. 환경별 성능 평가 지표로 halfcheetah, hopper 그리고 reach는 에피소드 평균 보상 (mean reward)을 사용하였고, attitude control은 성공률을 사용하였다. 모든 실험은 오프라인-온라인 학습 반복을 3회 수행하였다.

정량적 실험 결과는 표 1에 제시하였다. 제안 모델은 기존 모델인 IRPO에 비해 각각의 환경에서 성능이 41.7%, 8.6%, 29.1%, 40.7% 향상되었다. 이러한 결과는 SimBa 구조로 인해 높은 표현력을 얻었기 때문으로 해석할 수 있다. 특히, hopper와 같은 저차원 상태 공간보다 나머지 환경과 같은 고차원 상태 공간에서의 성능 향상 폭이 더 큰 것을 볼 수 있다.

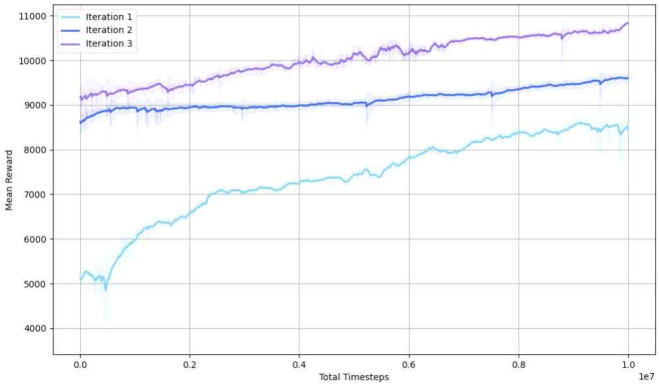

그림 3. Halfcheetah 환경 기준 제안 모델의 반복 횟수별 학습 곡선.

제안 모델의 학습 곡선 예시를 시각적으로 나타낸 결과는 그림 3에 제시하였다. 그림 2는 halfcheetah 환경에 대해 어닐링을 사용하여 학습한 결과를 그래프로 나타내었다. 어닐링을 통해 에이전트의 성능이 점진적으로 향상되도록 설계하였다. 그림 2를 보면 각 반복(iteration)별로 성능이 안정적으로 향상되는 것을 확인할 수 있다.

Ⅲ. 결론

본 연구에서는 저품질의 시연 데이터를 활용하여 최적 정책을 탐색하는 IRPO 알고리즘의 성능 한계를 극복하기 위해, SimBa 구조 도입과 KLD 제약 계수 어닐링 기법을 제안하였다. 제안하는 모델은 신경망의 심층화를 통해 복잡한 로봇 제어 환경에서의 상태 표현력을 강화하였으며, KLD 계수 어닐링을 통해 점진적이고 안정적인 학습을 유도하였다. 4가지 시뮬레이션 실험 결과, 제안 모델은 기존 모델 대비 모든 환경에서 큰 폭의 성능 향상을 기록하며 그 유효성을 입증하였다. 향후 연구로는 고정된 KLD 계수 스케줄링 대신, 학습 상태에 따라 KLD 제약 강도를 적응적으로 조절하는 자동 최적화 기법에 대해 탐구할 계획이다.

ACKNOWLEDGMENT

이 논문은 2025년도 교육부 및 서울특별시의 재원으로 서울RISE 센터의 지원을 받아 수행된 지역혁신중심 대학지원체계(RISE)의 결과입니다. (2025-RISE-01-019-04)

참 고 문 헌

[1] Li, J., Hu, X., Xu, H., Liu, J., Zhan, X. and Zhang, Y. Q., "Proto: Iterative policy regularized offline-to-online reinforcement learning," *arXiv preprint* arXiv:2305.15669, 2023.

[2] Xudong, G., Dawei, F., Xu, K., Zhai, Y., Yao, C., Wang, W., Ding, B. and Wang, H., "Iterative regularized policy optimization with imperfect demonstrations," in *Proc. Int. Conf. Machine Learning* (ICML), 2024.

[3] Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O., "Proximal policy optimization algorithms," *arXiv preprint* arXiv:1707.06347, 2017.

[4] Lee, H., Lee, Y., Seno, T., Kim, D., Stone, P. and Choo, J., "Hyperspherical normalization for scalable deep reinforcement learning," in *Proc. Int. Conf. Machine Learning* (ICML), 2025.

[5] Fu, J., Kumar, A., Nachum, O., Tucker, G. and Levine, S., "D4RL: Datasets for deep data-driven reinforcement learning," *arXiv preprint* arXiv:2004.07219, 2020.

[6] Gallouédec, Q., Cazin, N., Dellandréa, E. and Chen, L., "Panda-Gym: Open-source goal-conditioned environments for robotic learning," *arXiv preprint* arXiv:2106.13687, 2021.

[7] Gong, X., Dawei, F., Xu, K., Sun, Z., Zhou, X., Zheng, S. et al., "VVC-Gym: A fixed-wing UAV reinforcement learning environment for multi-goal long-horizon problems," in *Proc. Int. Conf. Learning Representations* (ICLR), 2025.