소프트 액터-비평가:
확률론적 액터를 사용한 정책 외 최대 엔트로피 심층 강화 학습

투오마스 하르노야 [1] **아우릭 저우** [1] **피터 아벨** [1] **세르게이 레빈** [1]

초록

모델 없는 심층 강화 학습(RL) 알고리즘은 다양한 까다로운 의사 결정 및 제어 작업에서 그 성능이 입증되었습니다. 그러나 이러한 방법은 일반적으로 매우 높은 샘플 복잡성과 깨지기 쉬운 수렴 특성이라는 두 가지 주요 과제를 안고 있어 세심한 하이퍼파라미터 튜닝이 필요합니다. 이 두 가지 문제는 복잡한 실제 도메인에 이러한 방법을 적용하는 데 심각한 제약이 됩니다. 이 백서에서는 최대 엔트로피 강화 학습 프레임워크에 기반한 오프 정책 액터 크리틱 딥 RL 알고리즘인 소프트 액터 크리틱을 제안합니다. 이 프레임워크에서 행위자는 기대 보상을 극대화하는 동시에 엔트로피를 극대화하는 것을 목표로 합니다. 즉, 가능한 한 무작위적으로 행동하면서 과제를 성공적으로 수행하는 것을 목표로 합니다. 이 프레임워크에 기반한 기존의 딥러닝 방식은 Q-러닝 방식으로 공식화되었습니다. 이 방법은 정책 외 업데이트와 안정적인 확률적 액터 크리티컬 포뮬레이션을 결합하여 다양한 연속 제어 벤치마크 작업에서 최첨단 성능을 달성함으로써 기존의 온-정책 및 오프-정책 방법을 능가하는 성능을 발휘합니다. 또한, 다른 정책 외 알고리즘과 달리 우리의 접근 방식은 매우 안정적이며 다양한 무작위 시드에 걸쳐 매우 유사한 성능을 달성한다는 것을 입증했습니다.

1. 소개

모델 없는 심층 강화 학습(RL) 알고리즘은 게임(Mnih 외., 2013; Silver 외., 2016)에서 로봇 제어(Schulman 외., 2015)에 이르기까지 다양하고 까다로운 영역에 적용되어 왔습니다. RL과 신경망과 같은 고용량 함수 근사기의 조합은 광범위한 의사 결정 및 제어 작업을 자동화할 수 있는 가능성을 가지고 있지만, 아직 널리 채택되지는 않았습니다.

[1]버클리 인공 지능 연구, 캘리포니아 대학교 버클리, 미국. 대응 담당자: 투오마스 하르노야 <haarnoja@berkeley.edu>.

이러한 방법을 실제 도메인에 적용하는 데는 두 가지 주요 과제가 있습니다. 첫째, 모델 없는 딥러닝 방법은 샘플의 복잡성 측면에서 비용이 많이 드는 것으로 악명이 높습니다. 비교적 간단한 작업이라도 수백만 단계의 데이터 수집이 필요할 수 있으며, 고차원적인 관찰이 필요한 복잡한 행동의 경우 훨씬 더 많은 데이터가 필요할 수 있습니다. 둘째, 이러한 방법은 학습 속도, 탐색 상수 및 기타 설정을 문제 설정에 따라 신중하게 설정해야 좋은 결과를 얻을 수 있는 등 초매개변수와 관련하여 취약한 경우가 많습니다. 이 두 가지 문제는 모델 없는 딥러닝의 실제 작업 적용 가능성을 심각하게 제한합니다.

딥러닝 방법론의 샘플 효율성이 떨어지는 원인 중 하나는 온-정책 학습입니다. 가장 일반적으로 사용되는 딥러닝 알고리즘인 TRPO(Schulman et al., 2015), PPO(Schulman et al., 2017b), A3C(Mnih et al., 2016)는 각 그라데이션 단계마다 새로운 샘플을 수집해야 하기 때문입니다. 효과적인 정책을 학습하는 데 필요한 그라데이션 단계 수와 단계당 샘플 수가 작업 복잡성에 따라 증가하기 때문에 비용이 급격히 증가하게 됩니다. 오프 정책 알고리즘은 과거의 경험을 재사용하는 것을 목표로 합니다. 이는 기존의 정책 그라데이션 공식으로는 직접적으로 실현 가능하지 않지만, Q-러닝 기반 방법에서는 비교적 간단합니다(Mnih et al., 2015). 안타깝게도 정책 외 학습과 고차원 비선형 함수 근사치를 신경망과 결합하는 것은 안정성과 수렴에 대한 주요 과제를 제시합니다(Bhatnagar et al., 2009). 이러한 문제는 연속 상태 및 행동 공간에서 더욱 악화되는데, 이 경우 별도의 액터 네트워크를 사용하여 Q-러닝에서 최대화를 수행하는 경우가 많습니다. 이러한 환경에서 일반적으로 사용되는 알고리즘인 심층 결정적 정책 그라디언트(DDPG)(릴리크랩 등, 2015)는 샘플을 효율적으로 학습할 수 있지만, 극도의 취성 및 과매개변수 민감도로 인해 사용하기가 까다롭기로 악명이 높습니다(Duan 등, 2016; Henderson 등, 2017).

연속 상태 및 행동 공간에 대해 효율적이고 안정적인 모델 프리 딥러닝 알고리즘을 설계하는 방법을 살펴봅니다. 이를 위해 표준 최대 보상 강화 학습 목표를 엔트로피 최대화 조건으로 보강하는 최대 엔트로피 프레임워크를 활용합니다(Ziebart 외., 2008; Toussaint, 2009; Rawlik 외.),

2012; Fox et al., 2016; Haarnoja et al., 2017). 최대 엔트로피 강화 학습은 RL 목표를 변경하지만, 원래 목표는 온도 매개변수를 사용하여 복구할 수 있습니다(Haarnoja et al., 2017). 더 중요한 것은 최대 엔트로피 공식이 탐색과 견고성을 크게 향상시킨다는 점입니다. Ziebart(2010)가 논의한 것처럼 최대 엔트로피 정책은 모델 및 추정 오류에도 견고하며, (Haarnoja 외, 2017)가 입증한 것처럼 다양한 행동을 획득함으로써 탐색을 개선합니다. 선행 연구에서는 엔트로피 극대화를 통해 온정책 학습을 수행하는 모델 프리 딥 RL 알고리즘(O'Donoghue et al., 2016)과 소프트 Q러닝 및 그 변형에 기반한 오프정책 방법(Schulman et al., 2017a; Nachum et al., 2017a; Haarnoja et al., 2017)이 제안되었습니다. 그러나 온정책 변형은 위에서 설명한 이유로 샘플 복잡성이 낮은 반면, 오프정책 변형은 연속적인 행동 공간에서 복잡한 근사 추론 절차를 필요로 합니다.

이 백서에서는 샘플 효율적인 학습과 안정성을 모두 제공하는 소프트 액터 크리틱(SAC)이라고 하는 오프 정책 최대 엔트로피 액터 크리틱 알고리즘을 고안할 수 있음을 보여줍니다. 이 알고리즘은 21개의 행동 차원을 가진 휴머노이드 벤치마크(Duan et al., 2016)와 같이 매우 복잡하고 고차원적인 작업으로 쉽게 확장할 수 있으며, DDPG와 같은 정책 외 방법은 일반적으로 좋은 결과를 얻기 어렵습니다(Gu et al., 2016). 또한 SAC는 소프트 Q-러닝을 기반으로 하는 기존의 정책 외 최대 엔트로피 알고리즘에서 근사 추론과 연관된 복잡성과 잠재적 불안정성을 피할 수 있습니다(Haarnoja et al., 2017). 본 논문에서는 최대 엔트로피 프레임워크에서 정책 반복을 위한 수렴 증명을 제시하고, 심층 신경망으로 실질적으로 구현할 수 있는 이 절차에 대한 근사치에 기반한 새로운 알고리즘을 소프트 액터-크리틱이라고 부르는 새로운 알고리즘을 소개합니다. 소프트 액터 크리틱이 오프 정책 및 온 정책 사전 방법보다 성능과 샘플 효율성 모두에서 상당한 개선을 달성한다는 것을 보여주는 실증적 결과를 제시합니다. 또한 DDPG를 크게 개선한 결정론적 알고리즘을 제안하는 동시 연구인 트윈 지연 심층 결정론적(TD3) 정책 그라데이션 알고리즘(후지모토 외., 2018)과도 비교합니다.

2. 관련 작업

트위터의 소프트 액터 크리틱 알고리즘은 정책과 가치 함수 네트워크가 분리된 액터 크리틱 아키텍처, 효율성을 위해 이전에 수집한 데이터를 재사용할 수 있는 오프 정책 공식화, 안정성과 탐색을 위한 엔트로피 최대화라는 세 가지 핵심 요소를 통합합니다. 이 섹션에서는 이러한 아이디어 중 일부를 활용한 선행 연구를 검토합니다. 액터 크리티컬 알고리즘은 일반적으로 정책의 가치 함수를 계산하는 *정책* 반복과 *정책의* 가치 함수를 *계산하는 정책 평가를* 번갈아 수행하는 정책 반복에서 시작하여 도출됩니다.

그리고 더 나은 정책을 얻기 위해 가치 함수를 사용하는 *정책 개선*(Barto et al., 1983; Sutton & Barto, 1998). 대규모 강화 학습 문제에서는 일반적으로 수렴을 위해 이러한 단계 중 하나를 실행하는 것이 비현실적이며, 대신 가치 함수와 정책을 함께 최적화합니다. 이 경우 정책을 액터라고 하고 가치 함수를 비평가라고 합니다. 많은 액터-비판 알고리즘은 표준 온-정책 정책 기울기 공식에 기반하여 액터를 업데이트하며(Peters & Schaal, 2008), 이들 중 다수는 정책의 엔트로피도 고려하지만 엔트로피를 최대화하는 대신 정규화 도구로 사용합니다(Schul- man et al., 2017b; 2015; Mnih et al., 2016; Gruslys et al., 2017). 정책 내 훈련은 안정성을 향상시키는 경향이 있지만 샘플 복잡성이 떨어지는 결과를 초래합니다.

정책 외 표본을 통합하고 고차 분산 감소 기법을 사용하여 견고성을 유지하면서 표본 효율성을 높이려는 노력이 있었습니다(O'Donoghue et al., 2016; Gu et al., 2016). 하지만 완전히 정책에서 벗어난 알고리즘이 여전히 더 나은 효율성을 달성합니다. 특히 널리 사용되는 비정책 행위자-비판적 방법인 결정론적 정책 그라데이션(Silver et al., 2014) 알고리즘의 심층 변형인 DDPG(Lillicrap et al., 2015)는 Q-함수 추정기를 사용하여 비정책 학습을 가능하게 하고, 이 Q-함수를 최대화하는 결정론적 행위자를 사용합니다. 따라서 이 방법은 결정적 행위자-비판적 알고리즘이자 근사 Q-학습 알고리즘으로 볼 수 있습니다. 안타깝게도 결정적 행위자 네트워크와 Q-함수 간의 상호작용으로 인해 DDPG는 일반적으로 안정화가 매우 어렵고 파라미터 설정이 과도할 경우 깨지기 쉽습니다(Duan et al., 2016; Henderson et al., 2017). 결과적으로 DDPG를 복잡하고 고차원적인 작업으로 확장하는 것은 어렵고, 온-정책 정책 그라데이션 방식은 여전히 이러한 설정에서 최상의 결과를 생성하는 경향이 있습니다(Gu et al., 2016). 대신 우리의 방법은 정책 외 행위자-비평가 훈련을 *확률적* 행위자와 결합하고, 나아가 엔트로피 최대화 목표를 통해 이 행위자의 엔트로피를 최대화하는 것을 목표로 합니다. 실제로 이 방법을 사용하면 훨씬 더 안정적이고 확장 가능한 알고리즘을 얻을 수 있으며, 실제로 DDPG의 효율성과 최종 성능을 모두 능가하는 결과를 얻을 수 있습니다. 유사한 방법을 확률적 값 그라데이션(SVG(0))의 0단계 특수한 경우로 도출할 수 있습니다(Heess et al., 2015). 그러나 SVG(0)은 표준 최대 기대 수익률만을 최적화한다는 점에서 우리의 방법과 다르며, 별도의 가치 네트워크를 사용하지 않는다는 점에서 학습이 더 안정적이라는 것을 알 수 있었습니다.

최대 엔트로피 강화 학습은 정책의 기대 수익률과 예상 엔트로피를 모두 극대화하기 위해 정책을 최적화합니다. 이 프레임워크는 역강화학습(Ziebart et al., 2008)에서 최적 제어(Todorov, 2008; Toussaint, 2009; Rawlik et al., 2012)에 이르기까지 다양한 맥락에서 사용되어 왔습니다. 유도 정책 탐색(Levine & Koltun, 2013; Levine et al., 2016)에서는 최대 엔트로피 분포가 정책 학습을 안내하는 데 사용됩니다.

높은 보상 영역으로 향하도록 유도합니다. 최근에는 여러 논문에서 최대 엔트로피 학습의 프레임워크에서 Q-러닝과 정책 그라데이션 방법 간의 연관성에 대해 언급했습니다 (O'Donoghue 외., 2016; Haarnoja 외., 2017; Nachum 외., 2017a; Schulman 외., 2017a). 대부분의 선행 모델 프리 연구는 이산 행동 공간을 가정하지만, Nachum 등(2017b)은 가우시안으로 최대 엔트로피 분포를 근사화하고 Haarnoja 등(2017)은 최적의 정책에서 샘플을 추출하도록 훈련된 샘플링 네트워크를 사용합니다. Haarnoja 등(2017)이 제안한 소프트 Q-러닝 알고리즘은 가치 함수와 행위자 네트워크를 가지고 있지만, Q-함수는 최적의 Q-함수를 추정하고 행위자는 데이터 분포를 통하지 않고는 Q-함수에 직접적인 영향을 주지 않기 때문에 진정한 행위자 비판적 알고리즘이 아닙니다. 따라서 Haarnoja 등(2017)은 행위자 비판 알고리즘에서 행위자가 아닌 근사 샘플러로서 행위자 네트워크를 모티브로 삼았습니다. 결정적으로, 이 방법의 수렴은 이 샘플러가 실제 후방에 얼마나 잘 근사화하느냐에 달려 있습니다. 반면, 저희는 정책 매개변수화에 관계없이 주어진 정책 클래스에서 최적의 정책으로 수렴하는 방법을 증명합니다. 이러한 사전 최대 엔트로피 방법은 일반적으로 처음부터 학습할 때 DDPG와 같은 최신 오프 정책 알고리즘의 성능을 능가하지 못하지만, 전처리 개선 및 미세 조정의 용이성과 같은 다른 이점이 있을 수 있습니다. 실험 결과, 유니티의 소프트 액터 크리티컬 알고리즘은 실제로 기존의 최신 오프-폴리시 딥러닝 방법의 성능을 큰 폭으로 능가하는 것으로 나타났습니다.

3. 예선전

먼저 표기법을 소개하고 표준 및 최대 엔트로피 강화 학습 프레임워크를 요약합니다.

3.1. 표기법

우리는 연속 행동 공간에서 정책 학습을 다룹니다. 상태 공간과 행동 공간이 연속적인 튜플 $(\mathcal{S}, \mathcal{A}, p, r)$과 미지의 상태 전이 확률 p로 정의되는 무한 지평선 마르코프 결정 과정 (MDP)을 고려합니다:

$[0, \infty)$은 현재 상태 $\mathbf{s}_t$ 와 액션 $\mathbf{a}_t$ 가 주어졌을 때 다음 상태 $\mathbf{s}_{t+1}$ 의 확률 밀도를 나타냅니다. 환경은 한정된 보상 r 을 방출합니다:

$[r_{\min}, r_{\max}]$를 사용합니다. 정책 $\pi(\mathbf{a}_t|\mathbf{s}_t)$에 의해 유도된 궤적 분포의 상태 및 상태-행동 한계값을 나타내기 위해 $\rho_\pi(\mathbf{s}_t)$와 $\rho_\pi(\mathbf{s}_t, \mathbf{a}_t)$를 사용합니다.

3.2. 최대 엔트로피 강화 학습

표준 RL은 예상 보상의 합을 최대화합니다 $\sum_t \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t)\sim\rho_\pi}[r(\mathbf{s}_t, \mathbf{a}_t)]$. 목표를 보강하여 확률적 정책을 선호하는 보다 일반적인 최대 엔트로피 목표(예: Ziebart (2010) 참조)를 고려할 것입니다.

정책의 예상 엔트로피를 $\rho_\pi(\mathbf{s}_t)$로 계산합니다:

$$J(\pi) = \sum_{t=0}^{T} \mathrm{E}_{(\mathbf{s}\,t,at)\sim\rho\pi}[r(\mathbf{s}_t, \mathbf{a}_t) + \alpha \mathrm{H}(\pi(\cdot\,|\mathbf{s}_t))] \,. \quad (1)$$

온도 매개변수 α는 보상 대비 엔트로피 항의 상대적 중요도를 결정하며, 따라서 최적 정책의 확률성을 제어합니다. 최대 엔트로피 목표는 기존 강화 학습에서 사용되는 표준 최대 기대 보상 목표와 다르지만, 기존 목표는 α 0으로 한계에서 복구할 수 있습니다. 이 백서의 나머지 부분에서는 온도를 항상 α로 스케일링하여 보상에 포함시킬 수 있으므로 명시적으로 온도를 작성하는 것을 생략하겠습니다[-1] .

이 목표에는 여러 가지 개념적, 실용적 이점이 있습니다. 첫째, 이 정책은 가능성이 없는 길을 포기하면서 더 폭넓게 탐색하도록 인센티브를 제공합니다. 둘째, 이 정책은 최적에 가까운 행동의 여러 모드를 포착할 수 있습니다. 여러 행동이 똑같이 매력적으로 보이는 문제 설정에서 정책은 해당 행동에 동일한 확률의 질량을 투입합니다. 마지막으로, 이전 연구에서는 이 목표를 통해 탐색을 개선하는 데 기여했으며(Haarnoja 외., 2017; Schulman 외., 2017a), 실험 결과, 일반적인 RL 목적 함수를 최적화하는 최신 방법보다 학습 속도가 상당히 향상되는 것을 관찰할 수 있었습니다. 기대 보상과 엔트로피의 합이 유한하다는 것을 보장하기 위해 할인 계수 γ를 도입하여 목표를 무한 지평선 문제로 확장할 수 있습니다. 무한 지평선 할인 사례에 대한 최대 엔트로피 목표를 적는 것은 더 복잡하므로 (Thomas, 2014) 부록 A로 미룹니다.

기존 방법들은 최적 정책을 복구할 수 있는 최적 Q-함수를 직접 푸는 방법을 제안했습니다(Ziebart et al., 2008; Fox et al., 2016; Haarnoja et al., 2017). 이 글에서는 정책 반복 공식화를 통해 소프트 액터-크리틱 알고리즘을 고안하는 방법을 논의할 것이며, 대신 현재 정책의 Q-함수를 평가하고 *정책 외 기울기* 업데이트를 통해 정책을 업데이트할 것입니다. 이러한 알고리즘은 기존 강화 학습에서도 제안된 바 있지만, 저희가 아는 바로는 최대 엔트로피 강화 학습 프레임워크에서 오프 정책 행위자 비판 방법은 처음입니다.

Σ

4. 소프트 정책 반복에서 소프트 액터-크리틱까지

오프 정책 소프트 액터-크리틱 알고리즘은 정책 도출 방법의 최대 엔트로피 변형에서 시작하여 도출할 수 있습니다. 먼저 이 유도 방법을 제시하고, 해당 알고리즘이 밀도 클래스에서 최적의 정책에 수렴하는지 검증한 다음, 이 이론에 기반한 실용적인 심층 강화 학습 알고리즘을 제시합니다.

→

4.1. 소프트 정책 반복 도출

먼저 최대 엔트로피 프레임워크에서 정책 평가와 정책 개선을 번갈아 가며 최적의 최대 엔트로피 정책을 학습하기 위한 일반적인 알고리즘인 소프트 정책 반복을 도출합니다. 이론적 분석( )과 수렴 보장을 위해 표 형식의 설정을 기반으로 도출했으며, 다음 섹션에서는 이 방법을 일반적인 연속 설정으로 확장합니다. 예를 들어 매개변수화된 밀도 집합에 해당하는 정책 집합 내에서 소프트 정책 반복이 최적의 정책으로 수렴한다는 것을 보여 드리겠습니다.

소프트 정책 반복의 정책 평가 단계에서는 방정식 1의 최대 엔트로피 목표에 따라 정책 π의 값을 계산하고자 합니다. 고정된 정책의 경우, 소프트 Q값은 다음과 같이 주어진 함수 $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$에서 시작하여 수정된 벨만 백업 연산자 $\mathcal{T}^{\pi}$를 반복적으로 적용하여 반복적으로 계산할 수 있습니다.

$$\mathcal{T}^{\pi}Q(\mathbf{s}_t, \mathbf{a}_t) \triangleq r(\mathbf{s}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim p}[V(\mathbf{s}_{t+1})], \quad (2)$$

어디

$$V(\mathbf{s}_t) = \mathbb{E}_{\mathbf{a}_t \sim \pi}[Q(\mathbf{s}_t, \mathbf{a}_t) - \log \pi(\mathbf{a}_t|\mathbf{s}_t)] \quad (3)$$

는 소프트 상태 값 함수입니다. $\mathcal{T}^{\pi}$를 로 반복적으로 적용하여 아래에 공식화하면 모든 정책 π에 대한 소프트 상태 값 함수를 구할 수 있습니다.

정리 1 (소프트 정책 평가). *방정식 2의 소프트 벨만 백업 연산자 $\mathcal{T}^{\pi}$와 매핑을 고려합니다.*
$Q^0 : |\mathcal{A}| < \infty$ *인* $\mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ *을 정의하고* $Q^{k+1} = \mathcal{T}^{\pi} Q^k$ *로 정의합니다.*
그러면 시퀀스 Q^k *는 다음과 같은 소프트 Q 값으로 수렴합니다.*
π*는* $k \rightarrow \infty$입니다.

증거. 부록 B.1을 참조하세요. □

정책 개선 단계에서는 정책을 업데이트하여 새로운 Q 함수의 지수를 보호합니다. 이 특정 업데이트 선택은 소프트 값 $\mathbf{s}_t$에 $\exp(Q^{\pi_{\text{old}}}(\mathbf{s}_t, \cdot))$ 제한된 정책을 π_{old}로 정할 수 있습니다. 실제로는 추적 가능한 정책을 선호하므로, 예를 들어 가우스 분포와 같이 매개변수화된 분포에 대응할 수 있는 일부 정책 집합 Π로 정책을 추가로 제한할 것입니다. $\pi \in \Pi$라는 제약 조건을 고려하기 위해 개선된 정책을 원하는 정책 집합에 투영합니다. 원칙적으로 어떤 투영법이라도 선택할 수 있지만, 쿨백-라이블러 발산으로 정의된 정보 투영법을 사용하는 것이 편리할 것입니다. 즉, 정책 개선 단계에서는 각 상태에 대해 다음에 따라 정책을 업데이트합니다.

$$\pi_{\text{new}} = \arg\min_{\pi' \in \Pi} D_{\text{KL}}\left(\pi'(\cdot|\mathbf{s}_t) \,\middle\|\, \frac{\exp(Q^{\pi_{\text{old}}}(\mathbf{s}_t, \cdot))}{Z^{\pi_{\text{old}}}(\mathbf{s}_t)}\right), \quad (4)$$

분할 함수 $Z^{\pi_{\text{old}}}(\mathbf{s}_t)$는 분포를 정규화하며, 일반적으로는 다루기 어렵지만 새 정책에 대한 기울기에 기여하지 않으므로 다음 섹션에서 설명하는 것처럼 무시할 수 있습니다. 이 예측의 경우, 식 1의 목표에 대해 새로운 예측 정책이 기존 정책보다 더 높은 값을 갖는다는 것을 보여줄 수 있습니다. 이 결과는 정리 2에서 공식화합니다.

정리 2 (소프트 정책 개선). $\pi_{\text{old}} \in \Pi$ *를* π *방정식* π_{new}*4에 정의된 최소화 문제의 최적화라고 하자. 그런 다음* $|\mathcal{A}| < \infty$ *인 모든* $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$에 *대해* $Q^{\pi_{\text{new}}}(\mathbf{s}_t, \mathbf{a}_t) \geq Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t)$.

증거. 부록 B.2를 참조하세요. □

전체 소프트 정책 반복 알고리즘은 소프트 정책 평가와 소프트 정책 개선 단계를 번갈아 가며 수행하며, Π의 정책 중 최적의 최대-최소 엔트로피 정책으로 수렴하는 것을 증명합니다(정리 1).

이 알고리즘은 최적의 해를 찾을 수 있지만 표 형식의 경우에만 정확한 형태로 수행 할 수 있습니다. 따라서 다음에서는 연속 도메인에 대한 알고리즘을 근사화하여 함수를 사용해야 합니다.

근사값을 나타내는데, 수렴할 때까지 두 단계를 실행하면 계산 비용이 너무 많이 듭니다. 이 근사치는 소프트 액터 크리틱이라는 새로운 실용적인 알고리즘을 탄생시켰습니다.

정리 1 (소프트 정책 반복). *임의의* $\pi \in \Pi$에서 *소프트 정책 평가와 소프트 정책 개선을 반복적으로 적용하면 정책* π^**를 수렴하여* $Q^{\pi^*}(\mathbf{s}_t, \mathbf{a}_t) \geq Q^{\pi}(\mathbf{s}_t, \mathbf{a}_t)$ $\pi \in \Pi$ *및* $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$에 대해 *가 됩니다.*

모든 $\pi \in \Pi$ *및* $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$에 *대한* $Q(\mathbf{s}_t, \mathbf{a}_t)$*는 다음과 같이 가정합니다.*
$|\mathcal{A}| < \infty$.

증거. 부록 B.3을 참조하세요. □

4.2. 소프트 액터-비평가

위에서 설명한 것처럼 대규모 연속 도메인은 소프트 정책 반복에 대한 실용적인 근사치를 도출해야 합니다. 이를 위해 Q-함수와 정책 모두에 대해 함수 근사치를 사용하고, 평가와 개선을 수렴으로 실행하는 대신 확률적 경사 하강으로 두 네트워크를 번갈아 가며 최적화합니다. 매개변수화된 상태 값 함수 $V_\psi(\mathbf{s}_t)$, 소프트 Q-함수 $Q_\theta(\mathbf{s}_t, \mathbf{a}_t)$ 및 추적 가능한 정책 $\pi_\varphi(\mathbf{a}_t \mathbf{s}_t)$를 고려하겠습니다. 이러한 네트워크의 파라미터는 ψ, ϑ, φ*입니다.* 예를 들어, 값 함수는 표현 신경망으로 모델링할 수 있고, 정책은 신경망에 의해 주어진 평균과 공분산을 가진 가우시안으로 모델링할 수 있습니다. 다음으로 이러한 파라미터 벡터에 대한 업데이트 규칙을 도출하겠습니다.

상태 값 함수는 소프트 값을 근사화합니다. 원칙적으로 별도의 함수를 포함할 필요는 없습니다.
는 방정식 3에 따라 Q 함수 및 정책과 관련이 있으므로 상태 값에 대한 이미터입니다. 이 수량은 다음과 같습니다.

의 단일 액션 샘플에서 추정된 값( )을 편향 없이 사용할 수 있지만, 실제로는 소프트 값에 대한 별도의 함수 근사치를 포함하면 학습을 안정화할 수 있고 다른 네트워크와 동시에 학습하는 것이 편리합니다. 소프트 값 함수는 제곱 잔차 오차를 최소화하도록 훈련됩니다.

$$J_V(\psi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}} \left[\frac{1}{2} \left(V_\psi(\mathbf{s}_t) - \mathbb{E}_{\mathbf{a}_t \sim \pi_\varphi} [Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \log \pi_\varphi(\mathbf{a}_t | \mathbf{s}_t)] \right)^2 \right] \tag{5}$$

는 이전에 샘플링된 상태 및 액션의 분포 또는 리플레이 버퍼입니다. 방정식 5의 기울기는 편향되지 않은 추정기를 사용하여 추정할 수 있습니다.

$$\hat{\nabla}_\psi J_V(\psi) = \nabla_\psi V_\psi(\mathbf{s}_t) \left(V_\psi(\mathbf{s}_t) - Q_\theta(\mathbf{s}_t, \mathbf{a}_t) + \log \pi_\varphi(\mathbf{a}_t | \mathbf{s}_t) \right), \tag{6}$$

리플레이 버퍼 대신 현재 폴링에 따라 액션이 샘플링됩니다. 소프트 벨맨 잔차를 최소화하도록 소프트 Q 함수 파라미터를 훈련할 수 있습니다.

$$J_Q(\theta) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \mathcal{D}} \left[\frac{1}{2} \left(Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - \hat{Q}(\mathbf{s}_t, \mathbf{a}_t) \right)^2 \right], \tag{7}$$

와 함

께

$$\hat{Q}(\mathbf{s}_t, \mathbf{a}_t) = r(\mathbf{s}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{\mathbf{s}_{t+1} \sim p} [V_{\bar{\psi}}(\mathbf{s}_{t+1})], \tag{8}$$

이 역시 확률적 그라데이션으로 최적화할 수 있습니다.

$$\hat{\nabla}_\theta J_Q(\theta) = \nabla_\theta Q_\theta(\mathbf{a}_t, \mathbf{s}_t) \left(Q_\theta(\mathbf{s}_t, \mathbf{a}_t) - r(\mathbf{s}_t, \mathbf{a}_t) - \gamma V_{\bar{\psi}}(\mathbf{s}_{t+1}) \right). \tag{9}$$

이 업데이트는 목표 값 네트워크 $V_{\bar{\psi}}$를 사용합니다.
$\bar{\psi}$는 기하급수적으로 이동하는 값의 평균일 수 있습니다.

네트워크 가중치, 이는 기차를 안정화시키는 것으로 나타났습니다.
(Mnih et al., 2015). 또는

목표 가중치를 주기적으로 현재 값 함수 가중치와 일치시키도록 설정할 수 있습니다(부록 E 참조). 마지막으로, 정책 파라미터는 방정식 4에서 예상되는 KL-분산을 직접 최소화하여 학습할 수 있습니다:

$$\cdot \exp\left(Q_\theta(\mathbf{s}_t, \cdot) - \right)$$

알고리즘 1 소프트 액터-크리틱

파라미터 벡터 $\psi, \bar{\psi}, \vartheta, \varphi$를 초기화합니다.
각 **반복에 대해**
  각 환경 단계에 **대해 다음을 수행합니다.**
    $\mathbf{a}_t \sim \pi_\varphi(\mathbf{a}_t | \mathbf{s})_t$
    $\mathbf{S}_{t+1} \sim P(\mathbf{S}_{t+1} | \mathbf{S}_t, \mathbf{A})_t$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$
  에 대한 종료
  각 그라데이션 **단계에 대해**
    $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$
    $\vartheta_i \leftarrow \vartheta_i - \lambda_Q \hat{\nabla}_{\vartheta_i} J_Q(\vartheta_i)$ for $i \in \{1, 2\}$
    $\varphi \leftarrow \varphi - \lambda_\pi \hat{\nabla}_\varphi J_\pi(\varphi)$
    $\bar{\psi} \leftarrow \tau\psi + (1 - \tau)\bar{\psi}$
  에 대한 종료
에 대한 종료

여기서 ϵ_t 는 구형 가우시안과 같은 고정 분포에서 샘플링된 입력 노이즈 벡터입니다. 이제 방정식 10의 목표를 다음과 같이 다시 작성할 수 있습니다.

$$J_\pi(\varphi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}, \epsilon_t \sim \mathcal{N}} [\log \pi_\varphi(f_\varphi(\epsilon_t; \mathbf{s}_t) | \mathbf{s}_t) - Q_\theta(\mathbf{s}_t, f_\varphi(\epsilon_t; \mathbf{s}_t))], \tag{12}$$

여기서 π_φ 는 f_φ 의 관점에서 암시적으로 정의되며, 파티션 함수는 φ 와 독립적이며 다음과 같이 정의할 수 있습니다.

는 생략합니다. 방정식 12의 기울기는 다음과 같이 대략적으로 구할 수 있습니다.

$$\hat{\nabla}_\varphi J_\pi(\varphi) = \nabla_\varphi \log \pi_\varphi(\mathbf{a}_t | \mathbf{s})_t + (\nabla_{\mathbf{a}_t} \log \pi_\varphi(\mathbf{a}_t | \mathbf{s}_t) - \nabla_{\mathbf{a}_t} Q(\mathbf{s}_t, \mathbf{a}_t)) \nabla_\varphi f_\varphi(\epsilon_t; \mathbf{s}_t), \tag{13}$$

여기서 $\mathbf{a}_t$에서평가됩니다. 이션 $(\epsilon_t; \mathbf{s}_t)$. 이 편향되지 않은 그라데이션 추정기는 DDPG 스타일의 정책 그라데이션(Lillicrap et al., 2015)을 추적 가능한 확률론적 정책으로 확장합니다.

또한 트위터의 알고리즘은 가치 기반 방법의 성능을 저하시키는 것으로 알려진 정책 개선 단계의 긍정적 편향성을 완화하기 위해 두 개의 Q 함수를 사용합니다(Hasselt,

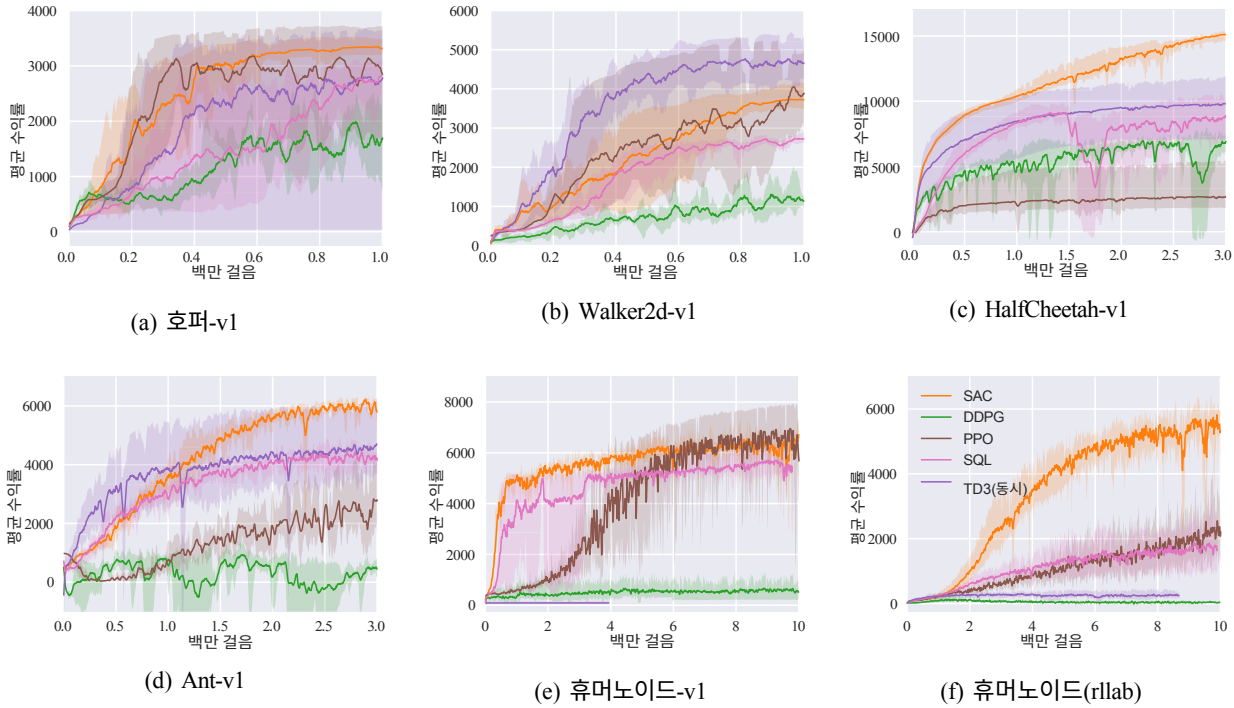
$$J_\pi(\varphi) = \mathbb{E}_{\mathbf{s}_t \sim \mathcal{D}}\Big[D_{\mathrm{KL}}\Big(\pi_\varphi(\cdot|\mathbf{s}_t) \,\Big\|\, \frac{\exp(Q_\theta(\mathbf{s}_t,\cdot))}{Z_\theta(\mathbf{s}_t)} \Big) \Big], \tag{10}$$

J_π를 최소화하기 위한 몇 가지 옵션이 있습니다. 정책 기울기 방법의 일반적인 해결책은 정책과 목표 밀도 네트워크를 통해 기울기를 역전파할 필요가 없는 가능성 비율 기울기 추정기(Williams, 1992)를 사용하는 것입니다. 그러나 우리의 경우 목표 밀도는 신경망으로 표현되는 Q-함수이며, 이는 미분할 수 있으므로 대신 재모수화 기법을 적용하여 분산 추정기를 더 낮추는 것이 편리합니다. 이를 위해 신경망 변환을 사용하여 정책을 재매개변수화합니다.

$$\mathbf{a}_t = f_\varphi(\epsilon_t; \mathbf{s}_t), \tag{11}$$

2010; 후지모토 외., 2018). 특히 다음을 매개변수화합니다. 매개변수 ϑ_i를 사용하여 두 개의 Q 함수를 생성한 다음, $J_Q(\vartheta_i)$를 최적화하도록 간접적으로 훈련합니다. 그런 다음 최소값을 사용합니다.

방정식 6의 값 경사도와 방정식 13의 극빙 경사도에 대한 Q-함수는 후지모토 외(2018)가 제안한 바와 같습니다. 저희 알고리즘은 단 하나의 Q함수를 사용하여 21차원 휴머노이드를 포함한 까다로운 작업을 학습할 수 있지만, 특히 어려운 작업에서 두 개의 Q함수를 사용하면 학습 속도가 크게 빨라지는 것을 확인했습니다. 전체 알고리즘은 알고리즘 1에 설명되어 있습니다. 이 방법은 현재 정책으로 환경에서 경험을 수집하는 것과 리플레이 버퍼에서 샘플링한 배치의 확률적 기울기를 사용하여 함수 근사치를 업데이트하는 것을 번갈아 가며 수행합니다. 실제로는 단일 환경 단계에 이어 하나 또는 여러 개의 그라데이션 단계를 수행합니다(부록 D 참조).



(a) 호퍼-v1 (b) Walker2d-v1 (c) HalfCheetah-v1

(d) Ant-v1 (e) 휴머노이드-v1 (f) 휴머노이드(rllab)

그림 1. 연속 제어 벤치마크에 대한 학습 곡선. 소프트 액터 크리티컬(노란색)은 모든 작업에서 일관된 성능을 보이며 가장 까다로운 작업에서 온-정책 및 오프-정책 방법 모두보다 우수한 성능을 보입니다.

로 설정할 수 있습니다.) 리플레이 버퍼에서 정책 외 데이터를 사용하는 것은 값 추정기와 정책 외 데이터로만 학습할 수 있기 때문에 가능합니다. 알고리즘은 임의의 상태-행동 튜플에 대해 평가할 수 있는 한 정책의 매개변수화에 구애받지 않습니다.

5. 실험

실험적 평가의 목표는 우리 방법의 샘플 복잡성과 안정성이 이전의 오프 정책 및 온 정책 심층 강화 학습 알고리즘과 어떻게 비교되는지 이해하는 것입니다. 우리는 OpenAI 체육관 벤치마크 제품군의 다양한 까다로운 연속 제어 작업(Brock-man 외., 2016)과 휴머노이드 작업(Duan 외., 2016)의 rllab 구현에서 우리의 방법을 이전 기법과 비교합니다. 쉬운 작업은 다양한 알고리즘으로 해결할 수 있지만, 21차원 휴머노이드(rllab)와 같이 복잡한 벤치마크는 정책 외 알고리즘으로 해결하기가 매우 어렵습니다(Duan et al., 2016). 알고리즘의 안정성 또한 성능에 큰 영향을 미칩니다. 작업이 쉬울수록 좋은 결과를 얻기 위해 하이퍼파라미터를 조정하는 것이 더 실용적인 반면, 가장 어려운 벤치마크( )에서 더 민감한 알고리즘( )의 경우 이미 좁은 유효 하이퍼파라미터의 범위가 엄청나게 작아져 성능이 저하될 수 있습니다(Gu et al., 2016).

저희는 저희의 방법을 보다 효율적인 오프 정책 딥 RL 방법 중 하나로 간주되는 알고리즘인 심층 결정론적 정책 그래디언트(DDPG)(Lillicrap 외, 2015), 안정적이고 효과적인 온 정책 정책 그래디언트 알고리즘인 프록시멀 정책 최적화(PPO)(Schulman 외, 2017b), 최대 엔트로피 정책을 학습하는 최근 오프 정책 알고리즘인 소프트 Q-러닝(SQL)(Haarnoja 외, 2017)과 비교하고 있습니다. SQL 구현에는 대부분의 환경에서 성능을 향상시키는 것으로 확인된 두 가지 Q 함수도 포함되어 있습니다. 저자가 제공한 구현을 사용하여 트윈 지연 심층 결정론적 정책 그라데이션 알고리즘(TD3)(후지모토 외, 2018)과 추가적으로 비교했습니다. 이는 저희 방법과 동시에 제안된 DDPG의 확장으로, 다른 개선 사항과 함께 이중 Q-러닝 트릭을 연속 제어에 처음 적용했습니다. 부록 E에는 신뢰 경로 일관성 학습(Trust-PCL)(Nachum et al., 2017b)과 두 가지 다른 SAC 변형을 포함했습니다. DDPG와 PPO에 대한 평가를 위해 탐색 노이즈를 해제했습니다. 탐색 노이즈를 명시적으로 주입하지 않는 최대 엔트로피 알고리즘의 경우, 탐색 노이즈(SQL)로 평가하거나 평균 동작(SAC)을 사용했습니다. SAC 구현의 소스 코드[1] 및 동영상[2] 는 온라인에서 확인할 수 있습니다.

[1]github.com/haarnoja/sac
[2]sites.google.com/view/soft-actor-critic

5.1. 비교 평가

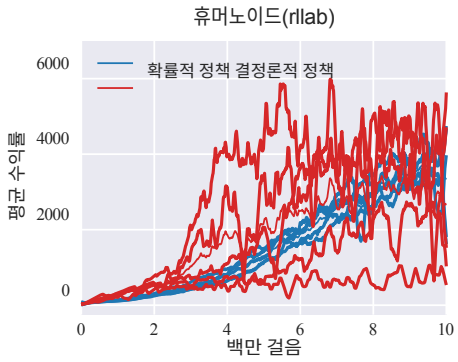
그림 1은 DDPG, PPO 및 TD3에 대한 훈련 중 평가 롤아웃의 총 평균 수익률을 보여줍니다. 각 알고리즘에 대해 서로 다른 무작위 시드를 사용하여 5개의 서로 다른 인스턴스를 학습하고, 각 인스턴스는 1000개의 환경 단계마다 한 번의 평가 롤아웃을 수행합니다. 실선 곡선은 평균에 해당하며 음영 영역은 다섯 번의 시도에서 최소 및 최대 수익률에 해당합니다.

그 결과, 전반적으로 SAC는 쉬운 작업에서는 기준 방법과 비슷한 성능을 보였으며, 어려운 작업에서는 큰 차이로 기준 방법을 능가하는 것으로 나타났습니다.

학습 속도와 최종 성능의 차이가 있습니다. 예를 들어, DDPG는 Ant-v1, 휴머노이드- v1, 휴머노이드(rllab)에서 아무런 진전을 보이지 못했으며, 이는 이전 연구(Gu et al., 2016; Duan et al., 2016)를 통해 입증된 결과입니다. 또한 SAC는 더 고차원적이고 복잡한 작업에서 안정적으로 학습하는 데 필요한 큰 배치 크기로 인해 PPO보다 훨씬 빠르게 학습합니다. 또 다른 최대 엔트로피 RL 알고리즘인 SQL도 모든 작업을 학습할 수 있지만, SAC보다 속도가 느리고 점근 성능이 떨어집니다. 본 실험에서 SAC가 달성한 정량적 결과는 이전 연구에서 다른 방법으로 보고된 결과와 매우 유리하게 비교되며(Duan et al., 2016; Gu et al., 2016; Henderson et al., 2017), 이는 이러한 벤치마크 과제에서 SAC의 샘플 효율성과 최종 성능이 모두 최신 기술을 뛰어넘는다는 것을 나타냅니다. 이 실험에 사용된 모든 하이퍼파라미터는 부록 D에 나열되어 있습니다.

5.2. 절제 연구

이전 섹션의 결과는 최대 엔트로피 원리에 기반한 알고리즘이 휴머노이드 작업과 같은 까다로운 작업에서 기존 RL 방법을 능가할 수 있음을 시사합니다. 이 섹션에서는 SAC의 어떤 특정 구성 요소가 우수한 성능에 중요한지 자세히 살펴봅니다. 또한 보상 스케일링과 목표 값 업데이트 평활화 상수 등 가장 중요한 하이퍼파라미터에 SAC가 얼마나 민감한지도 살펴봅니다.



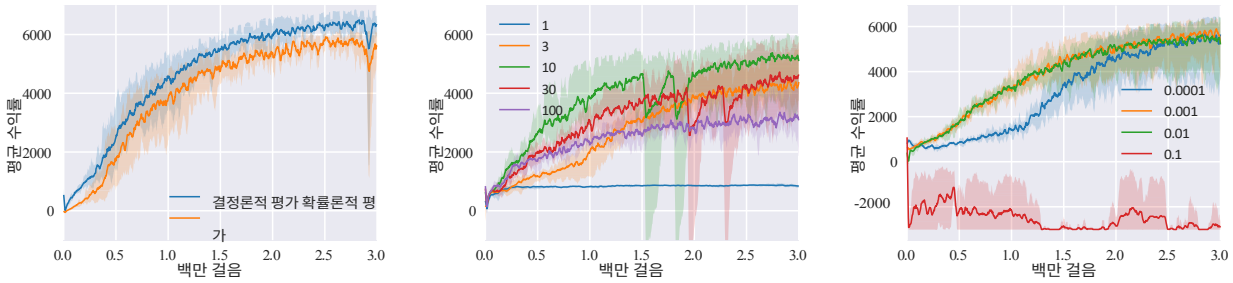
확률론적 정책과 결정론적 정책. 소프트 액터 크리틱은 최대 엔트로피 객체를 통해 확률적 정책을 학습합니다. 엔트로피는 정책과 가치 함수 모두에 나타납니다. 정책에서는 정책 분산이 조기에 수렴하는 것을 방지합니다(방정식 10). 가치 함수에서는 엔트로피가 높은 행동을 유발하는 상태 공간 영역의 값을 증가시켜 탐색을 장려합니다(방정식 5). 정책의 확률성과 엔트로피 극대화가 성능에 어떤 영향을 미치는지 비교하기 위해 엔트로피를 최대화하지 않고, 두 개의 Q 함수를 사용하고, 하드 타겟 업데이트를 사용하고, 별도의 타겟 액터를 사용하지 않고, 학습된 탐색 노이즈가 아닌 고정 노이즈를 사용한다는 점을 제외하면 DDPG와 매우 유사한 SAC의 결정론적 변형과 비교합니다. 그림 2는 서로 다른 랜덤으로 초기화된 두 가지 변형을 사용한 5개의 개별 실행을 비교한 것입니다.

그림 2. 휴머노이드(rllab) 벤치마크에서 개별 랜덤 시드의 안정성 측면에서 SAC(파란색)와 결정론적 변형 SAC(빨간색)의 비교. 이 비교는 결정론적 정책에서 시드 간의 변동성이 훨씬 높아짐에 따라 확률론이 학습을 안정화할 수 있음을 나타냅니다.

시드. 소프트 액터 크리틱은 훨씬 더 일관된 성능을 보이는 반면, 결정론적 변형은 시드에 따라 매우 높은 변동성을 보여 안정성이 상당히 떨어집니다. 그림에서 알 수 있듯이 엔트로피 극대화를 통해 확률적 정책을 학습하면 학습을 크게 안정화할 수 있습니다. 이는 하이퍼파라미터를 조정하기 어려운 어려운 작업에서 특히 중요합니다. 이 비교에서는 현재 값 네트워크와 일치하도록 목표 네트워크 파라미터를 주기적으로 덮어쓰는 하드 업데이트를 통해 목표 값 네트워크 가중치를 업데이트했습니다(모든 벤치마크 작업의 평균 성능 비교는 부록 E 참조).

정책 평가. SAC는 확률적 정책으로 수렴하기 때문에 최상의 성능을 위해 최종 정책을 최소로 억제하는 것이 유리한 경우가 많습니다. 평가를 위해 정책 분포의 평균을 선택하여 최대 사후 조치를 근사화합니다. 그림 3(a)는 훈련 수익률과 이 전략으로 얻은 평가 수익률을 비교한 것으로, 결정론적 평가가 더 나은 성과를 낼 수 있음을 보여줍니다. 모든 훈련 곡선은 보상의 합을 나타내며, 이는 정책의 엔트로피도 최대화하는 SQL 및 Trust-PCL을 비롯한 다른 최대 엔트로피 RL 알고리즘 및 SAC에 의해 최적화된 목표와 다르다는 점에 유의해야 합니다.

보상 규모. 소프트 액터 크리틱은 에너지 기반 최적 정책의 온도 역할을 하여 확률성을 제어하기 때문에 보상 신호의 스케일링에 특히 민감합니다. 보상 규모가 클수록 입력값이 낮아집니다. 그림 3(b)는 보상 규모가 달라질 때 학습 성능이 어떻게 변하는지를 보여줍니다: 보상 규모가 작을 경우, 정책이 거의 균일해져 보상 신호를 활용하지 못하고 결과적으로 성능이 크게 저하됩니다. 보상 규모가 큰 경우 모델은 처음에 빠르게 학습합니다,



(a) 평가 (b) 보상 규모 (c) 타겟 평활화 계수(τ)

그림 3. Ant-v1 작업에서 선택된 하이퍼파라미터에 대한 소프트 액터-크리틱의 민감도. (a) 평균 행동을 사용하여 정책을 평가하면 일반적으로 더 높은 수익을 얻을 수 있습니다. 이 정책은 엔트로피도 최대화하도록 학습되며, 평균 행동은 일반적으로 최대 수익률 목표에 대한 최적의 행동과 일치하지 않는다는 점에 유의하십시오. (b) 소프트 액터-크리틱은 최적 정책의 온도와 관련이 있기 때문에 보상 스케일링에 민감합니다. 최적의 보상 규모는 환경에 따라 다르며 각 작업에 대해 개별적으로 조정해야 합니다.
(c) 목표값 평활화 계수 τ는 훈련을 안정화시키는 데 사용됩니다. 빠르게 움직이는 목표(큰 τ)는 불안정(빨간색)을 초래할 수 있으며, 느리게 움직이는 목표(작은 τ)는 훈련 속도를 느리게 만듭니다(파란색).

하지만 정책이 거의 결정론적이 되어 적절한 탐색 부족으로 인해 국소 최소값이 낮아질 수 있습니다. 보상 규모를 적절히 조정하면 모델은 탐사와 착취의 균형을 유지하여 학습 속도가 빨라지고 점근 성능이 향상됩니다. 실제로 보상 스케일은 튜닝이 필요한 유일한 하이퍼파라미터이며, 최대 엔트로피 프레임워크에서 온도의 역수라는 자연스러운 해석은 이 파라미터를 조정하는 방법에 대한 좋은 직관력을 제공합니다.

목표 네트워크 업데이트. 안정성을 향상시키기 위해 실제 값 함수를 천천히 추적하는 별도의 목표 값 네트워크를 사용하는 것이 일반적입니다. 저희는 평활 상수 τ *와* 함께 지수 이동 평균을 사용하여 이전 작업에서 흔히 사용되는 목표 값 네트워크 가중치를 업데이트합니다(Lill- icrap et al., 2015; Mnih et al., 2015). 값이 1이면 매 반복마다 가중치를 직접 복사하는 하드 업데이트에 응답하고 0이면 타겟을 전혀 업데이트하지 않습니다. 그림 3(c)는 τ*가* 변할 때 SAC의 성능을 비교한 것입니다. τ*가* 크면 불안정해질 수 있고 τ*가* 작으면 훈련 속도가 느려질 수 있습니다. 그러나 τ *의* 적절한 값 범위가 비교적 넓다는 것을 알았고 모든 과제에서 동일한 값(0.005)을 사용했습니다. 그림 4(부록 E)에서는 지수 이동 평균을 사용하는 대신 현재 네트워크 가중치를 1000개의 기울기 단계마다 타르 겟 네트워크에 직접 복사하는 또 다른 SAC 변형과도 비교합니다. 이 변형은 환경 단계 사이에 그라데이션 단계를 두 개 이상 수행하여 성능을 향상시킬 수 있지만 계산 비용이 증가하는 이점이 있습니다.

6. 결론

우리는 엔트로피 최대화와 안정성의 이점을 유지하면서 샘플 효율적인 학습을 제공하는 오프 정책 최대 엔트로피 심층 강화 학습 알고리즘인 소프트 액터-크리틱(SAC)을 소개합니다. 이론적 결과를 통해 소프트 정책 반복을 도출하고, 이를 통해 최적의 정책으로 수렴하는 것을 보여줍니다. 이 결과를 통해 소프트 액터 크리티컬 알고리즘을 공식화할 수 있으며, 이 알고리즘이 오프 폴리시 DDPG 알고리즘과 온 폴리시 PPO 알고리즘을 포함한 최첨단 모델 프리 딥 RL 방법보다 성능이 뛰어나다는 것을 실증적으로 보여줍니다. 실제로 이 접근법의 샘플 효율은 DDPG의 샘플 효율을 크게 상회합니다. 우리의 결과는 확률론적 엔트로피 최대화 강화 학습 알고리즘이 견고성과 안정성을 개선할 수 있는 유망한 방법을 제공할 수 있음을 시사하며, 2차 정보(예: 신뢰 영역(Schulman et al., 2015))를 통합하는 방법이나 보다 표현적인 정책 클래스를 포함한 최대 엔트로피 방법에 대한 추가 탐색은 향후 연구의 흥미로운 방향입니다.

감사

알고리즘을 구현하는 데 통찰력 있는 토론과 도움을 주고 DDPG 기준 코드를 제공한 Vitchyr Pong, Trust-PCL 실험을 실행하는 데 지원을 제공한 Ofir Nachum, 이 백서의 초기 버전에 대한 귀중한 피드백을 제공한 George Tucker에게 감사의 말씀을 전합니다. 이 작업은 지멘스와 버클리 딥드라이브의 지원을 받았습니다.

참조

Barto, A. G., Sutton, R. S. 및 Anderson, C. W. 어려운 학습 문제를 해결할 수 있는 뉴런과 같은 적응 요소 trol 문제를 해결할 수 있습니다. *시스템, 인간 및 사이버네틱스에 관한 IEEE 트랜잭션*, 834-846, 1983.

Bhatnagar, S., Precup, D., Silver, D., Sutton, R. S., Maei, H. R., 및 Szepesva´ri, C. 임의의 평활 함수 근사치를 사용한 수렴적 시간-차이 학습. *신경 정보 처리 시스템의 발전 (NIPS)*, 1204-1212, 2009.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., 및 Zaremba, W. OpenAI gym.

Duan, Y., Chen, X. Houthooft, R., Schulman, J., Abbeel, P. 지속적인 제어를 위한 심층 강화 학습 벤치마킹. *국제 머신 러닝 컨퍼런스(ICML)*, 2016.

Fox, R., Pakman, A. 및 Tishby, N. 소프트 업데이트를 통한 강화 학습의 노이즈 길들이기. *인공 지능의 불확실성에 관한 컨퍼런스(UAI)*, 2016.

후지모토, S., 반 후프, H., 및 메거, D. 액터 크리티컬 방법의 함수 근사치 오류 해결. *arXiv preprint arXiv:1802.09477*, 2018.

Gruslys, A., Azar, M. G., Bellemare, M. G., and Munos, R. The reactor: 샘플 효율적인 액터-비평 아키텍처. *arXiv 사전 인쇄물 arXiv:1704.04651*, 2017.

구, S., 릴리크랩, T., 가라마니, Z., 터너, R. E., 및 레빈, S. Q-prop: 정책 비평가를 통한 표본 효율적 정책 그라데이션. *arXiv 사전 인쇄물 arXiv:1611.02247*, 2016.

Haarnoja, T., Tang, H., Abbeel, P., Levine, S. 심층 에너지 기반 정책을 통한 강제 학습. In *국제 기계 학습 컨퍼런스(ICML)*,
pp. 1352-1361, 2017.

Hasselt, H. V. 이중 Q- 학습. *신경 정보 처리 시스템(NIPS)의 발전*, 2613-2621쪽, 2010.

히스, N., 웨인, G., 실버, D., 릴리크랩, T., 에레즈, T., 타사, Y. 스토캐스에 의한 연속 제어 정책 학습- 틱 값 그래디언트. *신경 정보 처리 시스템의 발전 (NIPS)*, 2944-2952, 2015.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., Meger, D. 중요한 심층 강화 학습. *arXiv preprint arXiv:1709.06560*, 2017.

Kingma, D. 및 Ba, J. Adam: 확률적 최적화를 위한 방법. *국제 학습 프레젠테이션 컨퍼런스(ICLR)*, 2015.

Levine, S. and Koltun, V. 가이드 정책 검색. *국제 기계 학습 컨퍼런스(ICML)*, 1-9쪽, 2013.

Levine, S., Finn, C., Darrell, T. 및 Abbeel, P. 심층 시각 운동 정책의 엔드투엔드 학습. *기계 학습 연구 저널*, 17(39):1-40, 2016.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D. 및 Wierstra, D. 심층 강화 학습을 통한 연속 제어. *arXiv preprint arXiv:1509.02971*, 2015.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. 및 Riedmiller, M. 심층 강화 학습으로 아타리 연주하기. *arXiv preprint arXiv:1312.5602*, 2013.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidje- land, A. K., Ostrovski, G. 외. 심층 강화 학습을 통한 인간 수준의 제어. *Nature*, 518(7540): 529-533, 2015.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T. P., Harley, T., Silver, D., Kavukcuoglu, K. 심층 강화 학습을 위한 비동기적 방법. *국제 기계 학습 컨퍼런스(ICML)*, 2016.

Nachum, O., Norouzi, M., Xu, K., Schuurmans, D. 가치와 정책 기반 재강화 학습 간의 격차 해소. *신경 정보 처리 시스템의 발전(NIPS)*, 2772-2782, 2017a.

Nachum, O., Norouzi, M., Xu, K., Schuurmans, D. Trust-PCL: 지속적인 제어를 위한 정책 외 신뢰 영역 방법. *arXiv preprint arXiv:1707.01891*, 2017b.

O'Donoghue, B., Munos, R., Kavukcuoglu, K., and Mnih, V. PGQ: 정책 그라데이션과 Q-러닝 결합. *arXiv 사전 인쇄물 arXiv:1611.01626*, 2016.

Peters, J. and Schaal, S. 정책 그라데이션을 통한 운동 기술의 강화 학습. *신경망*, 21(4):682- 697, 2008.

Rawlik, K., Toussaint, M. 및 Vijayakumar, S. 확률론적 최적 제어 및 근사 추론에 의한 강화 학습에 대해. *Robotics: 과학과 시스템 (RSS)*, 2012.

슐만, J., 레빈, S., 아빌, P., 조던, M. I., 모리츠, P. 신뢰 지역 정책 최적화. In *Inter- national Conference on Machine Learning (ICML)*, pp. 1889-1897, 2015.

Schulman, J., Abbeel, P., Chen, X. 정책 그라데이션과 소프트 Q-러닝 간의 동등성. *arXiv preprint arXiv:1704.06440*, 2017a.

슐만, J., 울스키, F., 다리왈, P., 래드포드, A., 클리모프, O. 근거리 정책 최적화 알고리즘. *arXiv preprint arXiv:1707.06347*, 2017b.

Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D. 및 Riedmiller, M. 결정론적 정책 그라디언트 알고리듬. *국제 기계 학습 컨퍼런스(ICML)*, 2014.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. 심층 신경망과 트리 검색으로 바둑 게임 마스터하기. *Nature*, 529(7587):484-489, Jan 2016. ISSN 0028-0836. Article.

서튼, R. S. 과 바토, A. G. *강화 학습: 소개*, 볼륨 1. MIT press Cambridge, 1998.

Thomas, P. 자연 행위자 비평 알고리즘의 편향성. *기계 학습에 관한 국제 컨퍼런스(ICML)*, 441-448쪽, 2014.

토도로프, E. 최적 제어와 추정 사이의 일반적인 이중성. *IEEE 의사 결정 및 제어 컨퍼런스(CDC)*, 4286-4292쪽. IEEE, 2008.

투생, M. 근사 짝 추론을 사용한 로봇 궤적 최적화. *국제 기계 학습 컨퍼런스 (ICML)*, 1049-1056쪽. ACM, 2009.

Williams, R. J. 연결주의 강화 학습을 위한 간단한 통계적 기울기 추종 알고리즘. *기계 학습*, 8(3-4):229-256, 1992.

최대 인과 엔트로피 원리로 의도적 적응 행동 모델링. 카네기 멜론 대학교, 2010.

Ziebart, B. D., Maas, A. L., Bagnell, J. A. 및 Dey, A. K. 최대 엔트로피 역 강화 학습. *인공 지능에 관한 AAAI 컨퍼런스 (AAAI)*, 1433-1438, 2008.

A. 최대 엔트로피 목표

할인된 최대 엔트로피 목표의 정확한 정의는 정책 그라데이션 방식에 할인 계수를 사용할 때 일반적으로 상태 분포는 할인하지 않고 보상만 할인하기 때문에 복잡합니다. 그런 의미에서 할인된 정책 그라디언트는 일반적으로 실제 할인된 목표를 최적화하지 않습니다. 대신, Thomas(2014)가 논의한 것처럼 할인이 분산을 줄이는 역할을 하면서 평균 보상을 최적화합니다. 그러나 할인 계수 하에서 최적화되는 목표는 다음과 같이 정의할 수 있습니다.

$$J(\pi) = \sum_{t=0}^{\infty} \mathbb{E}_{(\mathbf{s},\mathbf{a})\sim\rho_\pi} \left[\sum_{l=t}^{\infty} \gamma^{l-t} \mathbb{E}_{\mathbf{s}\sim p, \mathbf{a}\sim\pi} \left[r(\mathbf{s}_t, \mathbf{a}_t) + \alpha \mathcal{H}(\pi(\cdot|\mathbf{s}_t)) | \mathbf{s}_t, \mathbf{a}_t \right] \right]. \tag{14}$$

이 목표는 현재 정책에 따라 확률 ρ_π 로 가중치를 부여한 모든 상태-행동 튜플$(\mathbf{s}_t, \mathbf{a}_t)$에서 발생하는 미래 상태에 대한 할인된 기대 보상과 엔트로피를 최대화하는 것에 해당합니다.

B. 증명

B.1. 정리 1

정리 1 (소프트 정책 평가). *방정식 2의 소프트 벨맨 백업 연산자* $\mathcal{T}^\pi$*와* $|\mathcal{A}| < \infty$*인 매핑* $Q^0 : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$*을 고려하고* $Q^{k+1} = \mathcal{T}^\pi Q^k$ *를 정의합니다. 그러면 수열* Q^k *은* $k \rightarrow \infty$로서 π*의 소프트 Q-값으로 수렴합니다.*

증명. 엔트로피 증강 보상을 $r_\pi(\mathbf{s}_t, \mathbf{a}_t)$, $r(\mathbf{s}_t, \mathbf{a}_t) + \mathbb{E}_{\mathbf{s}_{t+1}\sim p}[\mathcal{H}(\pi(\mathbf{s}_{t+1}))]$ 로 정의하고 업데이트 규칙을 다음과 같이 재작성합니다.

$$Q(\mathbf{s}_t, \mathbf{a}_t) \rightarrow r_\pi(\mathbf{s}_t, \mathbf{a}_t) + \gamma \mathbb{E}_{\mathbf{s}_{t+1}\sim p, \mathbf{a}_{t+1}\sim\pi}[Q(\mathbf{s}_{t+1}, \mathbf{a}_{t+1})]] \tag{15}$$

를 계산하고 정책 평가를 위한 표준 수렴 결과를 적용합니다(Sutton & Barto, 1998). *가정* $|\mathcal{A}| < \infty$ 은 엔트로피 증강 보상이 한계가 있다는 것을 보장하기 위해 필요합니다. □

B.2. 정리 2

정리 2 (소프트 정책 개선). $\pi_{\text{old}} \in \Pi$라고 *하고* π *를 방정식* $_{\text{new}}$*4에 정의된 최소화 문제의 최적화라고 합니다. 그런 다음* $|\mathcal{A}| < \infty$*인 모든* $(\mathbf{s}_t, \mathbf{a}_t) \in \mathcal{S} \times \mathcal{A}$*에 대해* $Q^{\pi_{\text{new}}}(\mathbf{s}_t, \mathbf{a}_t) \geq Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t)$*.*

증명. $\pi_{\text{old}} \in \Pi$를 π라고 하고 $Q^{\pi_{\text{old}}}$와 $V^{\pi_{\text{old}}}$를 해당 소프트 상태 동작 값과 소프트 상태 값이라고 하고 π_{new}로 정의할 수 있습니다.

$$\begin{aligned} \pi_{\text{new}}(\cdot|\mathbf{s}_t) &= \arg\min_{\pi'\in\Pi} D_{\text{KL}}\left(\pi'(\cdot|\mathbf{s}_t) \,\|\, \frac{\exp(Q^{\pi_{\text{old}}}(\mathbf{s}_t, \cdot))}{Z^{\pi_{\text{old}}}(\mathbf{s}_t)}\right) \\ &= \arg\min_{\pi'\in\Pi} J_{\pi_{\text{old}}}(\pi'(\cdot|\mathbf{s}_t)). \end{aligned} \tag{16}$$

항상 $\pi_{\text{new}} = \pi_{\text{old}} \in \Pi$를 선택할 수 있기 때문에 $J_{\pi_{\text{old}}}(\pi_{\text{new}}(\cdot|\mathbf{s}_t)) \leq J_{\pi_{\text{old}}}(\pi_{\text{old}}(\cdot|\mathbf{s}_t))$가 되어야 합니다. 따라서

$$\mathbb{E}_{\mathbf{a}_t\sim\pi_{\text{new}}}[\log \pi_{\text{new}}(\mathbf{a}_t|\mathbf{s}_t) - Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t) + \log Z^{\pi_{\text{old}}}(\mathbf{s}_t)] \leq \mathbb{E}_{\mathbf{a}_t\sim\pi_{\text{old}}}[\log \pi_{\text{old}}(\mathbf{a}_t|\mathbf{s}_t) - Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t) + \log Z^{\pi_{\text{old}}}(\mathbf{s}_t)], \tag{17}$$

파티션 함수 Z^π 이전은 상태에만 의존하므로 부등식은 다음과 같이 감소합니다.

$$\mathbb{E}_{\mathbf{a}_t\sim\pi_{\text{new}}}[Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t) - \log \pi_{\text{new}}(\mathbf{a}_t|\mathbf{s}_t)] \geq V^{\pi_{\text{old}}}(\mathbf{s}_t). \tag{18}$$

다음으로 소프트 벨맨 방정식을 생각해 보겠습니다: $\mathbf{s}_{t+1} \sim p\,[V^{\pi_{\text{old}}}(\mathbf{s}_{t+1})]$

$$
\begin{aligned}
Q^{\pi_{\text{old}}}(\mathbf{s}_t, \mathbf{a}_t) &= r(\mathbf{s}_t, \mathbf{a}_t) + \gamma\, \mathbb{E}_{\mathbf{s}} \\
&\leq r(\mathbf{s}_t, \mathbf{a}_t) + \gamma\, \mathbb{E}_{\mathbf{s}_{t+1} \sim p,\, \mathbf{a}_{t+1} \sim \pi_{\text{new}}} \left[Q^{\pi_{\text{old}}}(\mathbf{s}_{t+1}, \mathbf{a}_{t+1}) - \log \pi_{\text{new}}(\mathbf{a}_{t+1} | \mathbf{s}_{t+1})\right] \\
&\leq Q^{\pi_{\text{new}}}(\mathbf{s}_t, \mathbf{a}_t), \qquad (19)
\end{aligned}
$$

여기서 소프트 벨만 방정식과 방정식 18의 바운드를 적용하여 RHS에서 Q^{π} 이전 를 반복적으로 확장했습니다. $Q^{\pi_{\text{new}}}$ 로의 수렴은 정리 1에서 따릅니다. □

B.3. 정리 1

정리 1 (소프트 정책 반복). *모든* $\pi \in \Pi$에 *대한 소프트 정책 평가 및 소프트 정책 개선의 반복 적용*
는 $|A| < \infty$라고 *가정할 때 모든* $\pi \in \Pi$ *및* $(\mathbf{s}_t, \mathbf{a}) \in S \times A$에 *대해* $Q^{\pi^*}(\mathbf{s}, \mathbf{a}) \geq_t Q^{\pi_t}(\mathbf{s}, \mathbf{a})$*가 되는 정책* π^* *로 수렴합니다.*

증명. π_i 를 반복 i*에서의* 정책이라고 가정합니다. 정리 2에 따르면, 수열 Q^{π} i 는 단조롭게 증가합니다. $\pi \in \Pi$에 대해 Q^{π} 는 위에서 바운드되므로(보상과 엔트로피가 모두 바운드됩니다), 수열은 어떤 π^* 로 수렴합니다. 우리는 여전히 π^* 가 실제로 최적임을 보여줄 필요가 있습니다. 수렴 시 모든 $\pi \in \Pi, \pi \neq \pi^*$ 에 대해 $J_{\pi}(\pi^*(\cdot|\mathbf{s}_t)) < J_{\pi}(\pi(\cdot|\mathbf{s}_t))$의 경우 여야 합니다. 정리 2의 증명에서와 동일한 반복 논증을 사용하면, 모든 $(\mathbf{s}, \mathbf{a}) \in S \times A$에 대해 $Q^{\pi}(\mathbf{s}, a) > Q^{\pi}(\mathbf{s}, \mathbf{a})$ 를 구할 수 있습니다,
즉, Π에 있는 다른 정책의 소프트 값이 수렴된 정책의 소프트 값보다 낮습니다. 따라서 π^* 는 Π에서 최적입니다. □

C. 액션 바운드 적용

여기서는 액션 분포로 바운드되지 않은 가우시안 함수를 사용합니다. 그러나 실제로는 액션을 유한한 간격으로 제한해야 합니다. 이를 위해 가우스 샘플에 역 스쿼싱 함수(tanh)를 적용하고 변수 변경 공식을 사용하여 제한된 액션의 가능성을 계산합니다. 즉, $\mathbf{u} \in \mathbb{R}^D$ 을 확률 변수이고 $\mu(\mathbf{u}|\mathbf{s})$를 무한 지지를 갖는 해당 밀도라고 가정합니다. 그러면 $\mathbf{a} = \tanh(\mathbf{u})$는 원소 단위로 적용되는 tanh가 (-1, 1)에서 지지되는 확률 변수이며 밀도는 다음과 같이 주어집니다.

$$\pi(\mathbf{A}|\mathbf{S}) = \mu(\mathbf{U}|\mathbf{S}) \det \left| \frac{d\mathbf{a}^{-1}}{d\mathbf{u}} \right|. \tag{20}$$

자코비안 $\mathbf{da}/\mathbf{du} = \text{diag}(1 - \tanh^2(\mathbf{u}))$는 대각선이기 때문에 로그 가능성은 간단한 형태를 갖습니다.

$$\text{로그 } \pi(\mathbf{a}|\mathbf{s}) = \text{로그 } \mu(\mathbf{u}|\mathbf{s}) - \sum_{i=1} \log\left(1 - \tanh^2(u_i)\right), \tag{21}$$

여기서 u_i 는 $\mathbf{u}$의 i^{th} 요소입니다.

D. 하이퍼파라미터

표 1에는 그림 1과 그림 4의 비교 평가에 사용된 일반적인 SAC 파라미터가 나와 있습니다. 표 2에는 각 환경에 맞게 조정된 보상 규모 매개변수가 나열되어 있습니다.

표 1. SAC 하이퍼파라미터

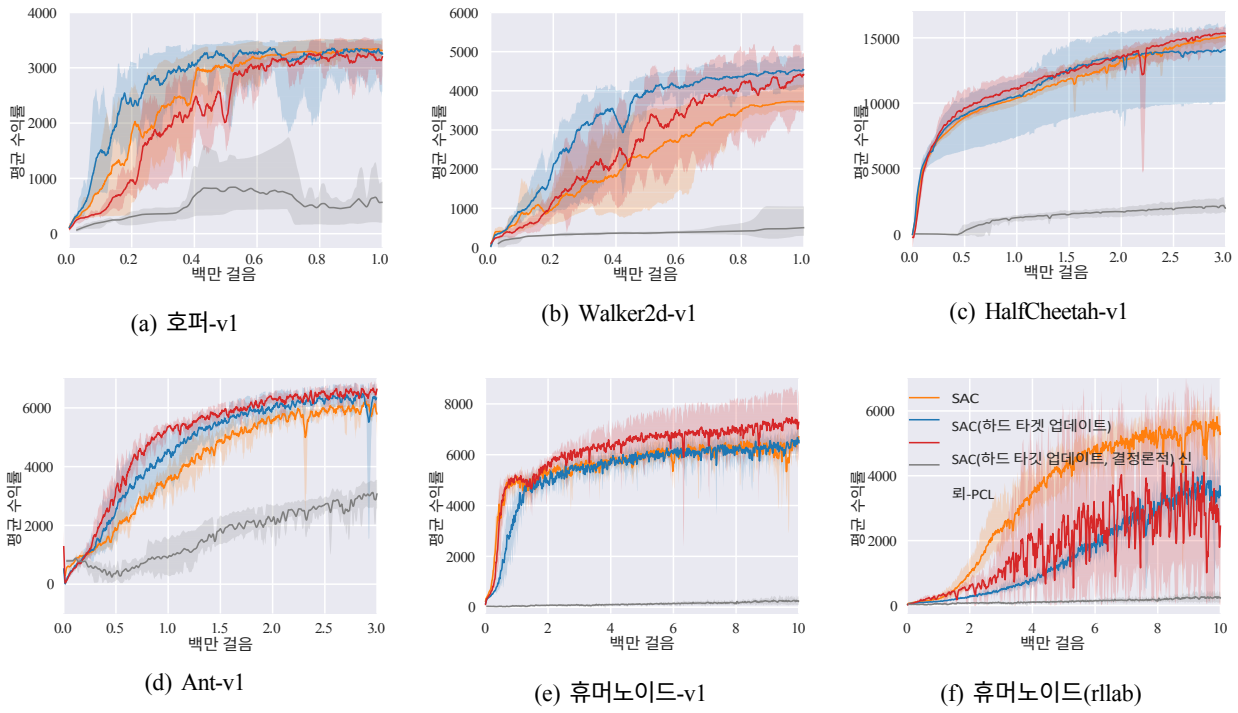
매개변수	값
공유	
옵티마이저	아담 (킹마 & 바, 2015)
학습률	$3 \cdot 10^{-4}$
할인(γ) 리플	0.99
레이 버퍼 크기	10^6
숨겨진 레이어 수(모든 네트워크)	2
레이어당 숨겨진 유닛 수	256
미니배치당 샘플 수	256
비선형성	ReLU
SAC	
목표 평활화 계수 (τ)	0.005
목표 업데이트 간격	1
그라데이션 단계	1
SAC(하드 타겟 업데이트)	1000
목표 평활화 계수 (τ)	4
목표 업데이트 간격	1
그라데이션 단계(휴머노이드 제외)	
그라데이션 단계(휴머노이드)	

표 2. SAC 환경별 매개변수

환경	작업 차원	보상 규모
호퍼-v1	3	5
Walker2d-v1	6	5
HalfCheetah-v1	6	5
Ant-v1	8	5
휴머노이드-v1	17	20
휴머노이드(rllab)	21	10

E. 추가 기준선 결과

그림 4는 SAC와 Trust-PCL을 비교한 것입니다(그림 4. Trust-PC는 주어진 환경 단계 수 내에서 대부분의 작업을 해결하지 못하지만, 더 오래 실행하면 결국 더 쉬운 작업(Nachum et al., 2017b)을 해결할 수 있습니다. 또한 지수 이동 평균을 사용하는 대신 목표 값 네트워크 가중치를 주기적으로 직접 복사하는 변형과 값 업데이트(방정식 6)와 정책 업데이트(방정식 13)에서 결정론적 정책을 가정하는 결정론적 제거는 두 개의 Q 함수가 있고, 하드 목표 업데이트를 사용하며, 별도의 목표 액터가 없고, 학습이 아닌 고정 탐색 노이즈를 사용한다는 점을 제외하면 DDPG와 매우 유사합니다. 이 두 가지 방법 모두 모든 작업을 학습할 수 있으며, SAC가 가장 빠른 휴머노이드(rllab) 작업을 제외한 모든 작업에서 SAC와 비슷한 성능을 보입니다.



(a) 호퍼-v1 (b) Walker2d-v1 (c) HalfCheetah-v1

(d) Ant-v1 (e) 휴머노이드-v1 (f) 휴머노이드(rllab)

그림 4. 추가 기준선(Trust-PCL) 및 두 가지 SAC 변형에 대한 학습 곡선. 하드 타겟 업데이트가 포함된 소프트 액터 크리틱(파란색)은 가중치의 지수 평활 평균을 사용하는 대신 1000회 반복마다 가치 함수 네트워크 가중치를 직접 복사한다는 점에서 표준 SAC와 다릅니다. 결정론적 제거(빨간색)는 가우시안 탐색 노이즈가 고정된 결정론적 정책을 사용하고, 값 함수를 사용하지 않으며, 액터 및 비평가 함수 업데이트에서 엔트로피 항을 삭제하고, 타깃 Q-함수에 하드 타깃 업데이트를 사용합니다. 이는 두 개의 Q-함수와 하드 타깃 업데이트를 사용하고 타깃 액터를 제거하는 DDPG와 동일합니다.