강화학습 논문 리뷰 스터디 9기

Hindsight Credit Assignment

Anna Harutyunyan, Will Dabney, Thomas Mesnard, Mohammad Azar, Bilal Piot, Nicolas Heess, Hado van Hasselt, Greg Wayne, S atinder Singh, Doina Precup, Remi Munos

DeepMind (NeurIPS 2019)

발표자: 이동진
2022년 11월 21일

About authors

교신 저자

Rémi Munos

FOLLOW

DeepMind
Verified email at inria.fr - Homepage

Reinforcement learning   deep learning   bandit theory   statistical learning

TITLE	CITED BY	YEAR
Bootstrap your own latent-a new approach to self-supervised learning JB Grill, F Strub, F Altché, C Tallec, P Richemond, E Buchatskaya, ... Advances in neural information processing systems 33, 21271-21284	2403	2020
Unifying count-based exploration and intrinsic motivation M Bellemare, S Srinivasan, G Ostrovski, T Schaul, D Saxton, R Munos Advances in neural information processing systems 29	1175	2016
A distributional perspective on reinforcement learning MG Bellemare, W Dabney, R Munos International Conference on Machine Learning, 449-458	1105	2017
Impala: Scalable distributed deep-rl with importance weighted actor-learner architectures L Espeholt, H Soyer, R Munos, K Simonyan, V Mnih, T Ward, Y Doron, ... International conference on machine learning, 1407-1416	1077	2018
Sample efficient actor-critic with experience replay Z Wang, V Bapst, N Heess, V Mnih, R Munos, K Kavukcuoglu, ... arXiv preprint arXiv:1611.01224	772	2016

About authors

주저자1

Anna Harutyunyan

FOLLOW

DeepMind
Verified email at google.com - Homepage
Reinforcement learning Machine learning Graph theory

TITLE	CITED BY	YEAR
Safe and efficient off-policy reinforcement learning R Munos, T Stepleton, A Harutyunyan, M Bellemare Advances in neural information processing systems 29	557	2016
Reinforcement learning from demonstration through shaping T Brys, A Harutyunyan, HB Suay, S Chernova, ME Taylor, A Nowé Twenty-fourth international joint conference on artificial intelligence	216	2015
Expressing Arbitrary Reward Functions as Potential-Based Advice A Harutyunyan, S Devlin, P Vrancx, A Nowé Twenty-Ninth Conference on Artificial Intelligence (AAAI)	79	2015
Multi-objectivization of reinforcement learning problems by reward shaping T Brys, A Harutyunyan, P Vrancx, ME Taylor, D Kudenko, A Nowé 2014 international joint conference on neural networks (IJCNN), 2315-2322	78	2014
Policy Transfer using Reward Shaping T Brys, A Harutyunyan, ME Taylor, A Nowé Fourteenth International Conference on Autonomous Agents and Multi-Agent …	77	2015
Q(λ) with Off-Policy Corrections A Harutyunyan, MG Bellemare, T Stepleton, R Munos International Conference on Algorithmic Learning Theory, 305-320	72	2016

About authors

주저자2

Will Dabney

FOLLOW

DeepMind
Verified email at google.com - Homepage

Reinforcement Learning   Machine Learning   Artificial Intelligence

TITLE	CITED BY	YEAR
Rainbow: Combining improvements in deep reinforcement learning M Hessel, J Modayil, H Van Hasselt, T Schaul, G Ostrovski, W Dabney, ... Thirty-second AAAI conference on artificial intelligence	1741	2018
A distributional perspective on reinforcement learning MG Bellemare*, W Dabney*, R Munos arXiv preprint arXiv:1707.06887	1105	2017
Distributed distributional deterministic policy gradients G Barth-Maron, MW Hoffman, D Budden, W Dabney, D Horgan, D Tb, ... arXiv preprint arXiv:1804.08617	422	2018
Distributional reinforcement learning with quantile regression W Dabney, M Rowland, M Bellemare, R Munos Proceedings of the AAAI Conference on Artificial Intelligence 32 (1)	414	2018
Successor features for transfer in reinforcement learning A Barreto, W Dabney, R Munos, JJ Hunt, T Schaul, HP van Hasselt, ... Advances in neural information processing systems 30	408	2017
Implicit quantile networks for distributional reinforcement learning W Dabney, G Ostrovski, D Silver, R Munos International conference on machine learning, 1096-1105	324	2018

Content

- Introduction
- Background (notation)
- Conditioning on the Future
- Algorithm
- Experiment

The instrumental learning object in RL

- 강화학습 알고리즘은 크게 exploration과 credit assignment을 반복
  - **Exploration** : 환경과 상호작용하며 데이터를 수집
  - **Credit assignment**
    - 각각의 행동들이 미래의 누적 보상에 기여한 정도 측정 [Johan Ferret *et al.*, 2019]
    - 주어진 미래의 누적 보상과 과거 행동 사이의 관련성 파악 [Anna Harutyunyan *et al.*, 2019]
    - 상태, 행동, 미래의 누적 보상 사이의 관계 파악 [Thomas Mesnard *et al.*, 2021]

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

The instrumental learning object in RL

- 강화학습 알고리즘은 크게 exploration과 credit assignment을 반복
  - **Exploration** : 환경과 상호작용하며 데이터를 수집
  - **Credit assignment** : 수집한 데이터로부터 상태, 행동, 누적 보상 사이의 관계 파악

- 행동가치함수
  - *How does choosing an action a in a state s affect future return?*
  - 상태 s에서 행동 a를 취했을 때 얻게 되는 기대 누적 보상은 얼마나 될까?
  - $\mathbb{E}_\pi[G_t|S_t = s, A_t = a]$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Issues in learning value function

- Issue 1: Variance
  - Monte Carlo 추정과 같은 방법은 trajectory에 들어있는 많은 randomness에 의해 추정치의 분산이 크다.

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Issues in learning value function

- Issue 1: Variance
- Issue 2: Partial observability (Bias)
  - 분산을 줄이기 위해 TD learning이 고안되었다. TD learning은 Markov property를 가정하기 때문에 POMDP에서 편향이 생긴다. 또한, function approximator와 같이 사용하는 이유로도 편향이 생긴다.

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Issues in learning value function

- Issue 1: Variance
- Issue 2: Partial observability (Bias)
- Issue 3: Time as Proxy
  - $\mathrm{TD}(\lambda)$가 bias와 variance의 trade-off를 조절할 수 있지만, 가까운 행동에 더 큰 credit을 부여한다.

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Issues in learning value function

- Issue 1: Variance
- Issue 2: Partial observability (Bias)
- Issue 3: Time as Proxy
- Issue 4: No counterfactuals
  - Trajectory에 있는 행동의 가치함수만 업데이트한다. 하지만, 하나의 trajectory로부터 모든 행동에 대한 credit assignment를 업데이트하고 싶다.
  - (ex) "행동을 취했더니 보상을 이만큼 받았어" 대신 "보상을 이만큼 받았을 때, 이 행동은 얼마나 관련 있을까? 저 행동을 얼마나 관련 있을까?"

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Hindsight conditioning

- Hindsight credit assignment
  - *Given the future outcome (reward or state), how relevant was the choice of a in x to achieve it?*
  - 기대 누적 보상이 z일 때, 상태 s에서 행동 a을 취한 것과 무슨 관련이 있을까?
  - $\mathbb{P}(a|S = s, Z = z; \pi)$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Hindsight conditioning

- Hindsight credit assignment
  - *Given the future outcome (reward or state), how relevant was the choice of a in x to achieve it?*
  - 기대 누적 보상이 z일 때, 상태 s에서 행동 a을 취한 것과 무슨 관련이 있을까?
  - $\mathbb{P}(a|S = s, Z = z; \pi)$

- 논문의 전개
  - **hindsight distribution 정의** : 미래 누적 보상가 조건부로 주어지는 행동의 확률 분포
  - **가치함수 재기술** : hindsight distribution 및 importance sampling 사용하여 가치함수 기술
  - **알고리즘 제안** : 다시 쓴 가치함수를 가지고 강화학습

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Notations

- MDP$(\mathcal{X}, \mathcal{A}, p, r, \gamma)$
- 정책 $\pi(a|x)$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Notations

- MDP$(\mathcal{X}, \mathcal{A}, p, r, \gamma)$
- 정책 $\pi(a|x)$
- Trajectory $\tau = (X_k, A_k, R_k)_{k \in \mathbb{N}^+}$
  - $\tau \sim \mathcal{T}(x, \pi)$: 상태 $X_0 = x$에서 시작해서 정책을 따르며 만들어지는 trajectory
  - $\tau \sim \mathcal{T}(x, a, \pi)$: 상태 $X_0 = x$에서 행동 $A_0 = a$를 취한 후 정책을 따르며 만들어지는 trajectory

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Notations

- MDP$(\mathcal{X}, \mathcal{A}, p, r, \gamma)$
- 정책 $\pi(a|x)$
- Trajectory $\tau = (X_k, A_k, R_k)_{k \in \mathbb{N}^+}$
  - $\tau \sim \mathcal{T}(x, \pi)$: 상태 $X_0 = x$에서 시작해서 정책을 따르며 만들어지는 trajectory
  - $\tau \sim \mathcal{T}(x, a, \pi)$: 상태 $X_0 = x$에서 행동 $A_0 = a$를 취한 후 정책을 따르며 만들어지는 trajectory
- Return $Z(\tau) = \sum_{k \geq 0} \gamma^k R_k$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Notations

- MDP$(\mathcal{X}, \mathcal{A}, p, r, \gamma)$
- 정책 $\pi(a|x)$
- Trajectory $\tau = (X_k, A_k, R_k)_{k \in \mathbb{N}^+}$
  - $\tau \sim \mathcal{T}(x, \pi)$: 상태 $X_0 = x$에서 시작해서 정책을 따르며 만들어지는 trajectory
  - $\tau \sim \mathcal{T}(x, a, \pi)$: 상태 $X_0 = x$에서 행동 $A_0 = a$를 취한 후 정책을 따르며 만들어지는 trajectory
- Return $Z(\tau) = \sum_{k \geq 0} \gamma^k R_k$
- 가치함수
  - 상태가치함수 $V^\pi(x) = \mathbb{E}_{\tau \sim \mathcal{T}(x,\pi)}[Z(\tau)]$
  - 행동가치함수 $Q^\pi(x, a) = \mathbb{E}_{\tau \sim \mathcal{T}(x,a,\pi)}[Z(\tau)]$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Overview

- Hindsight distribution
  - Let $\tau \sim \mathcal{T}(x, \pi)$ be a trajectory starting from x and f be some function of it
  - Hindsight distribution: $h(a|x, f(\tau); \pi)$

- Importance sampling
  - 주어진 미래의 값 $f(\tau)$과 행동 a의 관계

$$\frac{h(a|x, f(\tau); \pi)}{\pi(a|x)}$$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Conditioning on future states

- State-conditional hindsight distribution

$$h_k(a|x,y;\pi) := \mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(A_0 = a|X_k = y)$$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Conditioning on future states

$$P(A|X) = \frac{P(X|A)P(A)}{P(X)}$$

- State-conditional hindsight distribution

$$h_k(a|x, y; \pi) \coloneqq \mathbb{P}_{\tau \sim \mathcal{T}(x,\pi)}(A_0 = a | X_k = y)$$

- 베이즈 정리를 사용한 의미 해석

$$\begin{aligned} h_k(a|x, y; \pi) &= \mathbb{P}(A_0 = a | X_0 = x, X_k = y; \pi) \\ &= \frac{\mathbb{P}(X_k = y | X_0 = x, A_0 = a; \pi)\mathbb{P}(A_0 = a | X_0 = x; \pi)}{\mathbb{P}(X_k = y | X_0 = x; \pi)} \end{aligned}$$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Conditioning on future states

- State-conditional hindsight distribution

$$h_k(a|x,y;\pi) \coloneqq \mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(A_0 = a|X_k = y)$$

- 베이즈 정리를 사용한 의미 해석

$$\begin{aligned}
h_k(a|x,y;\pi) &= \mathbb{P}(A_0 = a|X_0 = x, X_k = y;\pi) \\
&= \frac{\mathbb{P}(X_k = y|X_0 = x, A_0 = a;\pi)\mathbb{P}(A_0 = a|X_0 = x;\pi)}{\mathbb{P}(X_k = y|X_0 = x;\pi)} \\
\frac{h_k(a|x,y;\pi)}{\pi(a|x)} &= \frac{\mathbb{P}(X_k = y|X_0 = x, A_0 = a;\pi)}{\mathbb{P}(X_k = y|X_0 = x;\pi)} \\
&= \frac{\mathbb{P}_{\tau\sim\mathcal{T}(x,a,\pi)}(X_k = y)}{\mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(X_k = y)}
\end{aligned}$$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Conditioning on future states

- State-conditional hindsight distribution

$$h_k(a|x, y; \pi) \coloneqq \mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(A_0 = a|X_k = y)$$

- 행동가치함수

$$Q^\pi(x, a) = \mathbb{E}_{\tau\sim\mathcal{T}(x,\pi)}[\sum_{k\geq 1} \gamma^k \frac{h_k(a|x, X_k; \pi)}{\pi(a|x)} R_k]$$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Conditioning on future states

- State-conditional hindsight distribution

$$h_k(a|x, y; \pi) \coloneqq \mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(A_0 = a|X_k = y)$$

- 행동가치함수

$$Q^\pi(x, a) = \mathbb{E}_{\tau\sim\mathcal{T}(x,\pi)}[\sum_{k\geq 1} \gamma^k \frac{h_k(a|x, X_k; \pi)}{\pi(a|x)} R_k]$$

- Advantage 함수

$$A^\pi(x, a) = r(x, a) - r^\pi(x) + \mathbb{E}_{\tau\sim\mathcal{T}(x,\pi)}\left[\left(\frac{h_k(a|x, X_k; \pi)}{\pi(a|x)} - 1\right)\gamma^k R_k\right],$$

where $r^\pi(x) = \sum_{a\in\mathcal{A}} \pi(a|x) r(x, a)$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Conditioning on future returns

- Return-conditional hindsight distribution

$$h_z(a|x, z; \pi) := \mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(A_0 = a|Z(\tau) = z)$$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Conditioning on future returns

- Return-conditional hindsight distribution

$$h_z(a|x,z;\pi) \coloneqq \mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(A_0 = a|Z(\tau) = z)$$

- 상태가치함수

$$V^\pi(x) = \mathbb{E}_{\tau\sim\mathcal{T}(x,a,\pi)}[Z(\tau)\frac{\pi(a|x)}{h_z(a|x,Z(\tau);\pi)}]$$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Conditioning on future returns

- Return-conditional hindsight distribution

$$h_z(a|x,z;\pi) \coloneqq \mathbb{P}_{\tau\sim\mathcal{T}(x,\pi)}(A_0 = a|Z(\tau) = z)$$

- 상태가치함수

$$V^\pi(x) = \mathbb{E}_{\tau\sim\mathcal{T}(x,a,\pi)}[Z(\tau)\frac{\pi(a|x)}{h_z(a|x,Z(\tau);\pi)}]$$

- Advantage 함수

$$A^\pi(x,a) = \mathbb{E}_{\tau\sim\mathcal{T}(x,a,\pi)}[(1-\frac{\pi(a|x)}{h_z(a|x,Z(\tau);\pi)})Z(\tau)]$$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Policy gradient

- State-conditional hindsight distribution

Theorem 3. *Let π_θ be the policy parameterized by θ, and $\beta = \gamma$. Then, the gradient of the value at some state x_0 is:*

$$\nabla_\theta V^{\pi_\theta}(x_0) = \mathbb{E}_{\tau \sim \mathcal{T}(x_0, \pi_\theta)} \Big[\sum_{k \geq 0} \gamma^k \sum_a \nabla \pi_\theta(a|X_k) Q^x(X_k, a) \Big] \tag{7}$$

$$= \mathbb{E}_{\tau \sim \mathcal{T}(x_0, \pi_\theta)} \Big[\sum_{k \geq 0} \gamma^k \nabla \log \pi_\theta(A_k|X_k) A^z(X_k, A_k) \Big], \tag{8}$$

$$Q^x(X_k, a) \stackrel{def}{=} r(X_k, a) + \sum_{t \geq k+1} \gamma^{t-k} \frac{h_\beta(a|X_k, X_t)}{\pi_\theta(a|X_k)} R_t,$$

$$A^z(x, a) \stackrel{def}{=} \left(1 - \frac{\pi_\theta(a|x)}{h_z(a|x, Z(\tau_{k:\infty}))}\right) Z(\tau_{k:\infty}).$$

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Algorithm

Return-conditional HCA

- Trajectory $\tau = (X_i, A_i, R_i)_{i \in \mathbb{N}^+}$로부터,
  - hindsight distribution을 업데이트하고,
  - advantage 함수를 계산하여 policy gradient 계산에 사용

Algorithm 2 Return-conditional HCA

Given: Initial π, h_z, V

1: **for** $k = 1, \ldots$ **do**
2: Sample $\tau = X_0, A_0, R_0, \ldots$ from π
3: **for** $i = 0, 1, \ldots$ **do**
4: Compose the return $Z(\tau_{i:\infty})$ starting from X_i
5: Train $h_z(A_i | X_i, Z_i)$ via cross-entropy
6: $Z_h \leftarrow \left(1 - \frac{\pi(A_i|X_i)}{h_z(A_i|X_i, Z(\tau_{i:\infty}))}\right) Z(\tau_{i:\infty})$
7: Follow the gradient $\nabla \log \pi(A_i|X_i) Z_h$
8: **end for**
9: **end for**

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Algorithm

State-conditional HCA

Algorithm 1 State-conditional HCA

Given: Initial π, h_β, V, $\hat{r}$; horizon T

for $k = 1, \ldots$ **do**
    Sample $\tau = X_0, A_0, R_0, \ldots, R_T$ from π
    for $i = 0, \ldots, T-1$ **do** ▷ Train hindsight distribution
        for $j = i, \ldots, T$ **do**
            Train $h_\beta(A_i | X_i, X_j)$ via cross-entropy
        end for
    end for
    for $i = 0, \ldots, T-1$ **do** ▷ Train baseline and reward predictor
        $Z = 0$
        for $j = i, \ldots, T-1$ **do**
            $Z \leftarrow Z + \gamma^{j-i} R_j$
        end for
        $Z \leftarrow Z + \gamma^{T-i} V(X_T)$
        Update $V(X_i)$ towards Z
        Update $\hat{r}$ towards R_i
    end for
    for $i = 0, \ldots, T-1$ **do** ▷ Train policy of all actions with the hindsight-conditioned return
        for all actions a **do**
            $Z_h = \pi(a|X_i, a)\hat{r}(X_i, a)$
            for $j = i+1, \ldots, T-1$ **do**
                $Z_h \leftarrow Z_h + \gamma^{j-i} \frac{h_\beta(a|X_i, X_j)}{\pi(a|X_i)} R_j$
            end for
            $Z_{h,a} \leftarrow Z_h + \gamma^{T-i} \frac{h_\beta(a|X_i, X_T)}{\pi(a|X_i)} V(X_T)$
        end for
        Follow the gradient $\sum_a \nabla \pi(a|X_i) Z_{h,a}$
    end for
end for

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Learning hindsight distribution

- Trajectory를 사용해서 supervised learning으로 학습
  - $h_\beta(a|x,y;\pi)$: 두 상태 x, y를 입력 받아서 행동 a를 출력
  - $h_z(a|x,z;\pi)$: 상태 x와 실수 z를 입력 받아서 행동 a를 출력

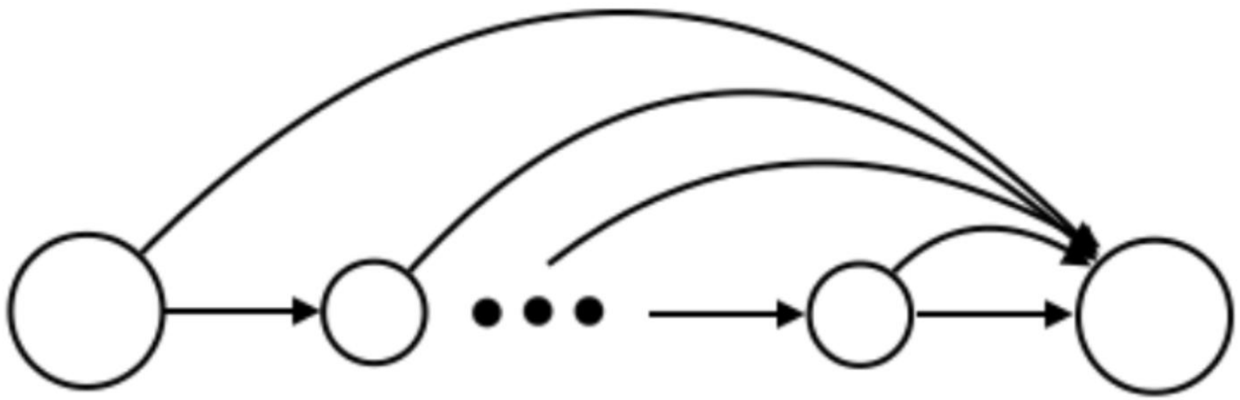
- 가치함수를 학습하는 것보다 더 쉬운 문제

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Experiment

Shortcut environment

- n=5개의 상태
- 2개의 행동
  - 하나의 행동은 바로 final로 전이, 다른 행동은 다음 상태로 전이
  - 10%을 확률로 바로 final로 전이
- final 상태에서 보상 +1, 이외 상태에서는 -1



Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Shortcut environment – results

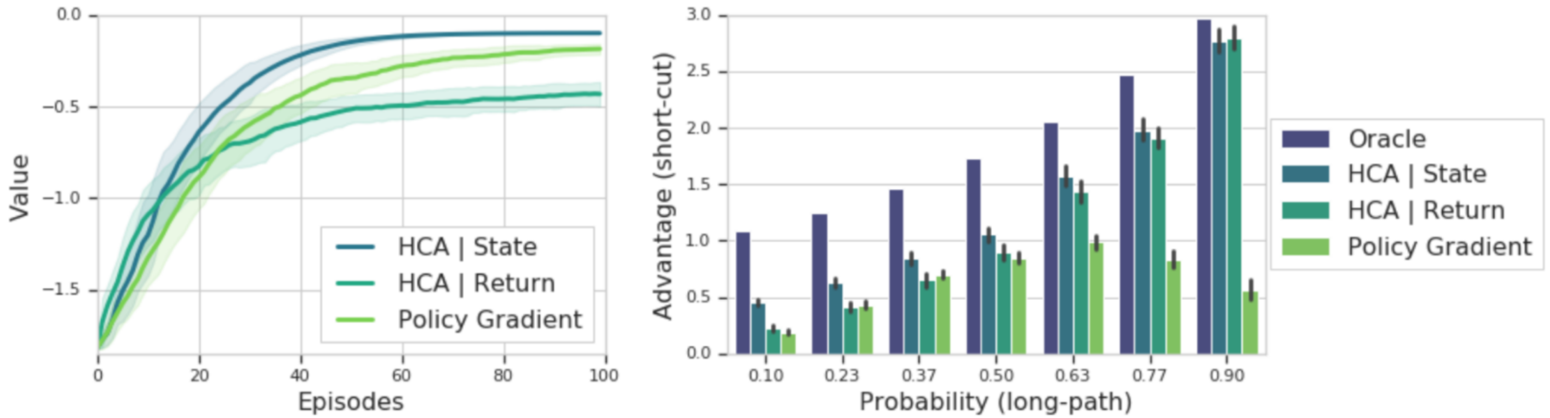
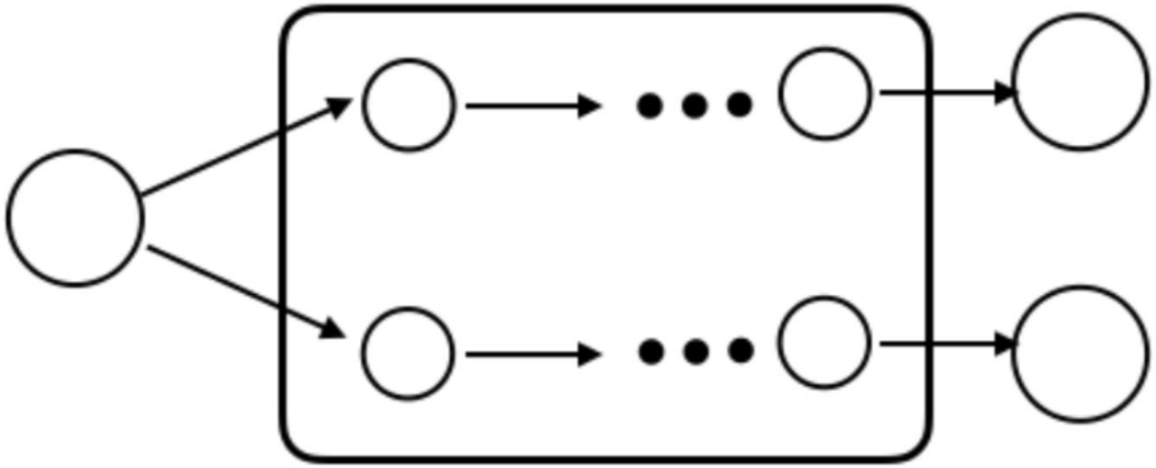


Figure 3: Shortcut. **Left:** learning curves for $n = 5$ with the policy between long and short paths initialized uniformly. Explicitly considering the likelihood of reaching the final state allows state-conditioned HCA to more quickly adjust its policy. **Right:** the advantage of the shortcut action estimated by performing 1000 rollouts from a fixed policy. The x-axis depicts the policy probabilities of the actions on the long path. The oracle is computed analytically without sampling. When the shortcut action is unlikely and rarely encountered, it is difficult to obtain an accurate estimate of the advantage. HCA is consistently able to maintain larger (and more accurate) advantages.

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Delayed effect environment

- 초기 상태에서 2가지 행동을 취할 수 있고,
  - 그에 따라 두 가지 경로로 나뉘지만, 두 가지 경로의 representation은 똑같음 (pomdp)
  - 각 경로의 final state에서 보상을 각각 1과 -1
- 사실 초기 상태에서의 선택에 따라 최종 보상이 결정



Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Delayed effect environment - results

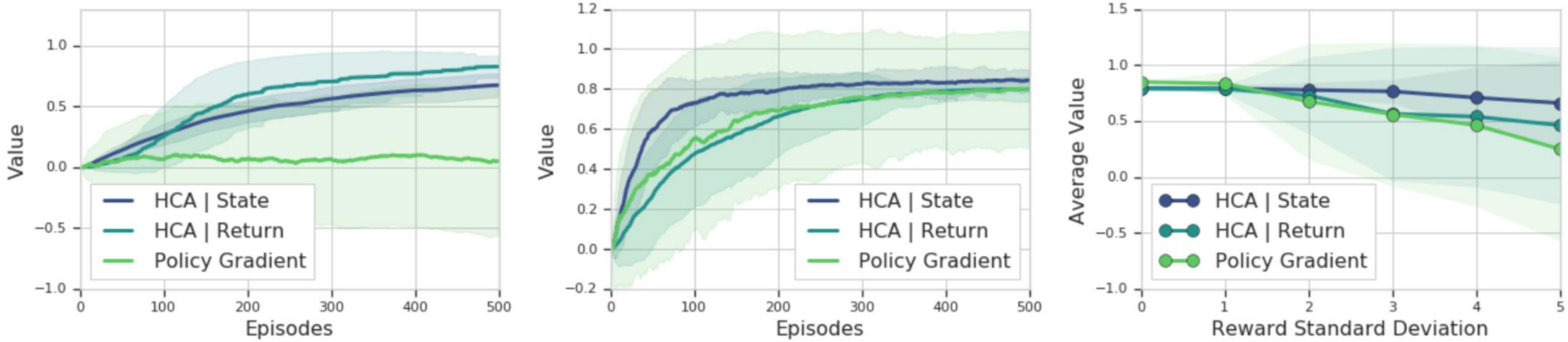
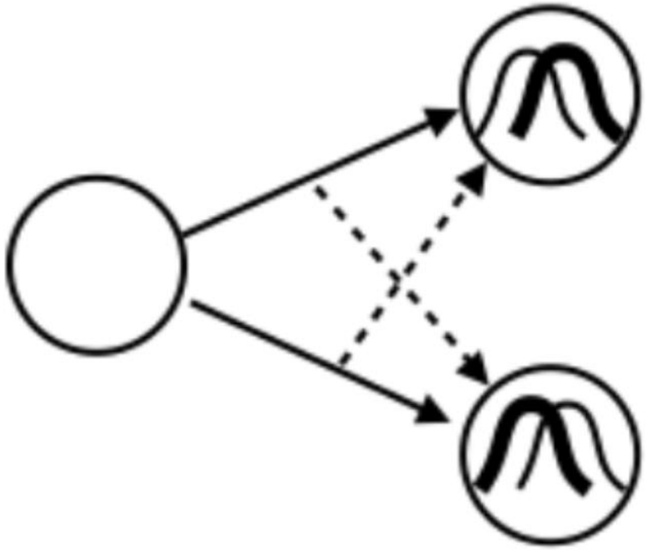


Figure 4: Delayed effect. **Left:** Bootstrapping. The learning curves for $n = 5$, $\sigma = 0$, and a 3-step return, which causes the agent to bootstrap in the partially observed region. As expected, naive bootstrapping is unable to learn a good estimate. **Middle:** Using full Monte Carlo returns (for $n = 3$) overcomes partial observability, but is prone to noise. The plot depicts learning curves for the setting with added white noise of $\sigma = 2$. **Right.** The average performance w.r.t. different noise levels – predictably, state HCA is the most robust.

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Ambiguous bandit environment

- 초기 상태에서 2가지 행동을 할 수 있음
- 2개 행동에 대한 결과의 보상은 평균은 각 1과 2이고 표준편차가 1.5인 분포에서 샘플링



Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Experiment

Ambiguous bandit environment - results

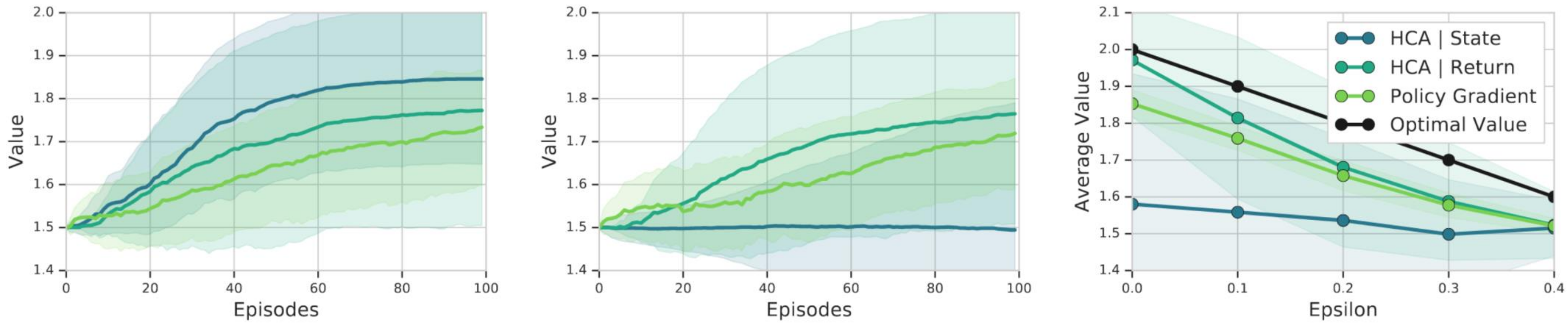


Figure 5: Ambiguous bandit with Gaussian rewards of means 1, 2, and standard deviation 1.5. **Left:** The state identity is observed. Both HCA methods improve on PG. **Middle:** The state identity is hidden, handicapping state HCA, but return HCA continues to improve on PG. **Right:** Average performance w.r.t. different ϵ-s with Gaussian rewards of means 1, 2, and standard deviation 0.5. Note that the optimal value itself decays in this case.

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)

Conclusion
Closing

- **Main idea** : Hindsight distribution 정의

- **Main contribution** : 가치함수를 Hindsight distribution을 사용하여 재기술

- **Cons**
  - Weak implementation
  - 생각보다 credit assignment 문제에 대한 명확한 기술 및 구체적인 예시 부족

Quoted from Anna Harutyunyan et al., Hindsight Credit Assignment. (2019)