강화학습 논문 리뷰 스터디 10기

SUNRISE
A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning

Lee, Kimin, Michael Laskin, Aravind Srinivas, Pieter Abbeel

UC Berkeley (ICML 2021)

발표자: 이동진
2023년 03월 06일

About authors
주저자

Kimin Lee

FOLLOW

Research Scientist, Google
Verified email at google.com - Homepage

Machine learning Reinforcement learning Deep learning

TITLE	CITED BY	YEAR
A simple unified framework for detecting out-of-distribution samples and adversarial attacks K Lee, K Lee, H Lee, J Shin Advances in neural information processing systems 31	1160	2018
Training confidence-calibrated classifiers for detecting out-of-distribution samples K Lee, H Lee, K Lee, J Shin International Conference on Learning Representations	669	2017
Using pre-training can improve model robustness and uncertainty D Hendrycks, K Lee, M Mazeika International Conference on Machine Learning, 2712-2721	506	2019
Reinforcement learning with augmented data M Laskin, K Lee, A Stooke, L Pinto, P Abbeel, A Srinivas Advances in neural information processing systems	382	2020
Decision transformer: Reinforcement learning via sequence modeling L Chen, K Lu, A Rajeswaran, K Lee, A Grover, M Laskin, P Abbeel, ... Advances in neural information processing systems	370	2021

About authors
주저자

Kimin Lee

Research Scientist, Google
Verified email at google.com - Homepage

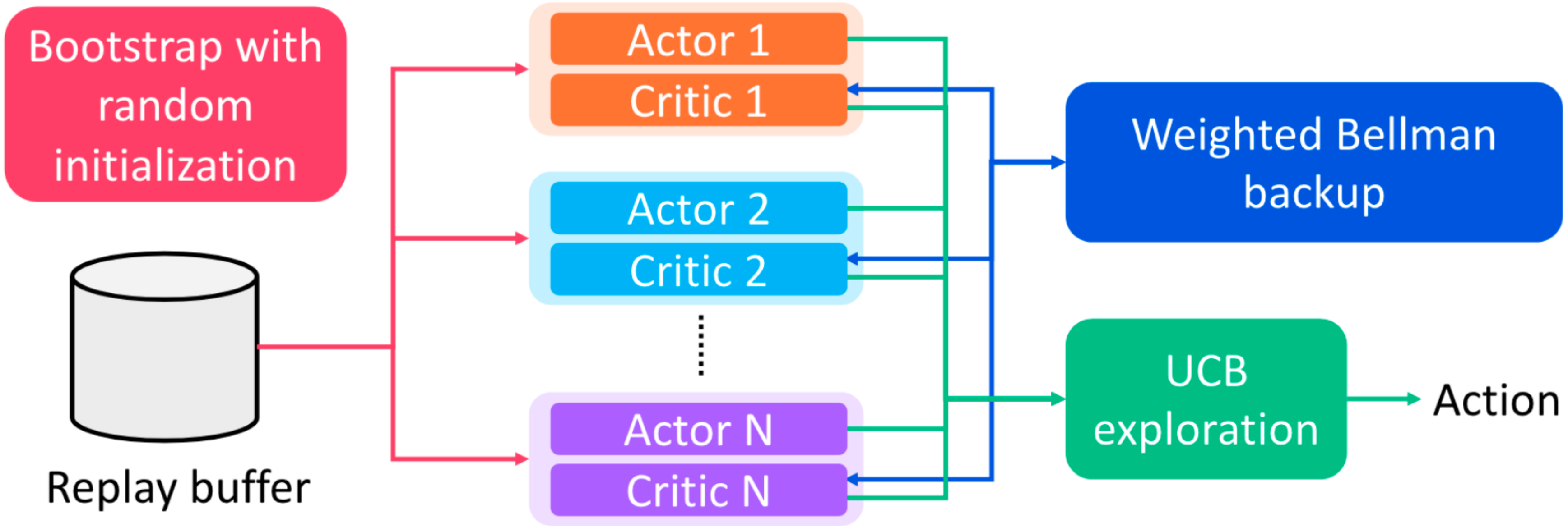
Machine learning Reinforcement learning Deep learning

FOLLOW

TITLE	CITED BY	YEAR
Overcoming catastrophic forgetting with unlabeled data in the wild K Lee, K Lee, J Shin, H Lee Proceedings of the IEEE/CVF International Conference on Computer Vision, 312-321	129	2019
Network randomization: A simple technique for generalization in deep reinforcement learning K Lee, K Lee, J Shin, H Lee International Conference on Learning Representations	128	2019
Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning K Lee, M Laskin, A Srinivas, P Abbeel International Conference on Machine Learning, 6131-6141	111	2021
Robust inference via generative classifiers for handling noisy labels K Lee, S Yun, K Lee, H Lee, B Li, J Shin International conference on machine learning, 3763-3772	86	2019
Context-aware dynamics model for generalization in model-based reinforcement learning K Lee, Y Seo, S Lee, H Lee, J Shin International Conference on Machine Learning, 5757-5766	63	2020

논문을 선택한 동기

- 지도학습에서 앙상블은 성능 향상의 치트키
- 강화학습에서 앙상블이란 무엇일까?
  - Replay buffer를 공유하는 N개의 actor와 N개의 critic



(a) SUNRISE: actor-critic version

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

SUNRISE Overview

- 알고리즘 분류
  - Model-free, off-policy learning, continuous/discrete action space

SUNRISE Overview

- 알고리즘 분류
  - Model-free, off-policy learning, continuous/discrete action space
- 기반 알고리즘
  - Soft Actor-Critic (연속 행동 공간), Rainbow (이산 행동 공간)
  - N개의 actor와 N개의 critic

SUNRISE Overview

- 알고리즘 분류
  - Model-free, off-policy learning, continuous/discrete action space
- 기반 알고리즘
  - Soft Actor-Critic (연속 행동 공간), Rainbow (이산 행동 공간)
  - N개의 actor와 N개의 critic
- 사용한 환경
  - Mujoco, DeepMind Control Suite, Atari

N개의 actor와 critic이 있으면 좋은 점

- 현재 주어진 상태 s_t에서 N개의 행동 후보 $\{a_{t,i}\}_{t=1}^{N}$ 생성
- 각 행동마다 N개의 행동 가치 함수가 있어서 분포를 생각할 수 있음
  - Weighted Bellman backups
  - Bootstrap with random initialization
  - UCB exploration

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

Single model-free/off-policy 알고리즘의 문제점

- Sample inefficient
  - 경험 데이터를 충분히 많이 활용하지 못함
- Error propagation in Q-learning
  - Function approximation error 및 overestimation을 내재한 타겟을 사용하여 행동 가치 함수 학습
- Exploration Exploitation trade-off
  - 주로 랜덤에 의한 exploration을 함

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

앙상블을 사용한 문제 완화

- Sample inefficient
  - 경험 데이터를 충분히 많이 활용하지 못함 ➠ 경험 데이터를 많은 네트워크 학습에 사용
- Error propagation in Q-learning
  - Function approximation error 및 overestimation을 내재한 타겟을 사용하여 행동 가치 함수 학습 ➠ Double Q-learning으로 완화 (예) TD3, Double DQN
- Exploration Exploitation trade-off
  - 주로 랜덤에 의한 exploration을 함 ➠ 여러 네트워크로부터 예측 불확실성을 고려

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

Weighted Bellman backups (1)

- 일반적인 soft Q-learning

$$\mathcal{L}_Q(\theta) = (Q_\theta(s_t, a_t) - r_t - \gamma \bar{V}(s_{t+1}))^2$$

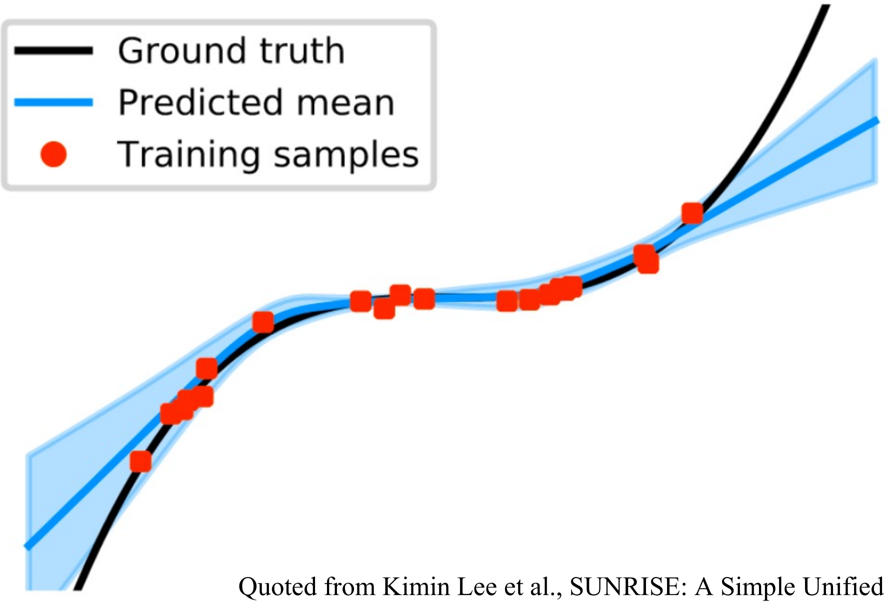
$$\bar{V}(s_t) = \mathbb{E}_{a_t \sim \pi_\phi}[Q_{\bar{\theta}}(s_t, a_t) - \alpha \log \pi_\phi(a_t|s_t)]$$

- Function approximation error가 내재된 target을 사용
  - ➠ inconsistency and unstable convergence

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

Weighted Bellman backups (2)

- N개의 actor와 critic $\{Q_{\theta_i}, \pi_{\phi_i}\}_{i=1}^{N}$ ※ 실험에서는 $N = 5$
- 행동 가치 함수 추정치 $\{Q_{\theta_i}(s,a)\}_{i=1}^{N}$의 분포를 고려
- Q-네트워크간 분산이 클수록 불확실하고 에러를 갖고 있는 추정치로 생각



Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

SUNRISE

Weighted Bellman backups (3)

- Weighted Bellman backups

$$\mathcal{L}_{WQ}(\theta_i) = w(s_{t+1}, a_{t+1})\left(Q_{\theta_i}(s_t, a_t) - r_t - \gamma \bar{V}(s_{t+1})\right)^2,$$

$$w(s,a) = \sigma(-\bar{Q}_{\text{std}}(s,a) * T) + 0.5,$$

  - $a_{t+1} \sim \pi_{\phi_i}(a|s_{t+1})$
  - σ: 시그모이드 함수
  - $\bar{Q}_{\text{std}}(s,a)$는 $Q_{\overline{\theta_i}}(s_t, a_t)$의 표본 표준편차
  - T: temperature

- $\bar{Q}_{\text{std}}(s,a)$가 클수록 Q-네트워크 학습에 적은 가중치를 부여

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

Bootstrap with random initialization

- 에이전트간 다양성 (diversity) 확보 방법 2가지
- 1. 파라미터 랜덤 초기화 $\{\theta_i, \phi_i\}_{i=1}^{N}$
- 2. Transition masking을 통해 서로 다른 데이터를 사용
  - Transition $\tau_t = (s_t, a_t, r_t, s_{t+1})$ 저장할 때, 에이전트마다 $m_{t,i} \sim \text{Bernourlli}(\beta)$도 함께 저장
  - Replay buffer는 공유하지만, 에이전트마다 각 transition을 사용할 수도 있고 아닐 수도 있음
  - 하지만, 실험에서는 $\beta = 1$일 때, 즉 데이터를 다 사용할 때 성능이 제일 좋았음.

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

UCB exploration (1)

- Upper confidence bound 방법
  - Multi-armed bandit (MAB) 문제에서 밴딧이 줬던 보상의 평균 뿐만 아니라 밴딧을 선택한 횟수까지 고려하여 밴딧을 선택하는 알고리즘.
  - $A_t = \mathrm{argmax}_a \mathrm{Q(a)} + \mathrm{c}\sqrt{{\log t}/{N_t(a)}}$

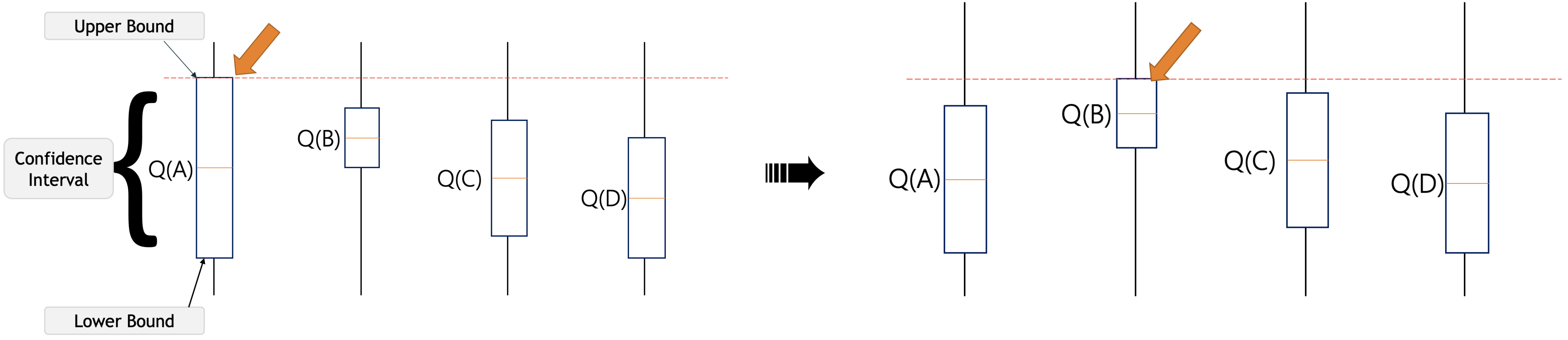


Image from https://www.geeksforgeeks.org/upper-confidence-bound-algorithm-in-reinforcement-learning/

UCB exploration (2)

- 현재 주어진 상태 s_t에서 N개의 행동 후보 $\{a_{t,i}\}_{t=1}^{N}$ 생성
- 각 행동마다 N개의 행동 가치 함수가 있어서 분포를 생각할 수 있음
- UCB가 가장 큰 행동을 선택

$$a_t = \max_a Q_{\text{mean}}(s_t, a) + \lambda Q_{\text{std}}(s_t, a)$$

  - 구현에서는 $a \in \{a_{t,i}\}_{t=1}^{N}$

Image from https://www.geeksforgeeks.org/upper-confidence-bound-algorithm-in-reinforcement-learning/

SUNRISE

Algorithm

$$\mathcal{L}^{\texttt{SAC}}_{\texttt{actor}}(\phi) = \mathbb{E}_{s_t \sim \mathcal{B}}\Big[\mathcal{L}_\pi(s_t, \phi)\Big], \tag{3}$$

$$\mathcal{L}_\pi(s_t, \phi) = \mathbb{E}_{a_t \sim \pi_\phi}\Big[\alpha \log \pi_\phi(a_t|s_t) - Q_\theta(s_t, a_t)\Big]. \tag{4}$$

Algorithm 1 SUNRISE: SAC version

```
for each iteration do
  for each timestep t do
    // UCB EXPLORATION
    Collect N action samples: A_t = {a_{t,i} ~ π_{φ_i}(a|s_t) | i ∈ {1, ..., N}}
    Choose the action that maximizes UCB: a_t = arg max_{a_{t,i} ∈ A_t} Q_mean(s_t, a_{t,i}) + λ Q_std(s_t, a_{t,i})
    Collect state s_{t+1} and reward r_t from the environment by taking action a_t
    Sample bootstrap masks M_t = {m_{t,i} ~ Bernoulli(β) — i ∈ {1, ..., N}}
    Store transitions τ_t = (s_t, a_t, s_{t+1}, r_t) and masks in replay buffer B ← B ∪ {(τ_t, M_t)}
  end for
  // UPDATE AGENTS VIA BOOTSTRAP AND WEIGHTED BELLMAN BACKUP
  for each gradient step do
    Sample random minibatch {(τ_j, M_j)}_{j=1}^B ~ B
    for each agent i do
      Update the Q-function by minimizing (1/B) Σ_{j=1}^B m_{j,i} L_WQ(τ_j, θ_i) in (5)
      Update the policy by minimizing (1/B) Σ_{j=1}^B m_{j,i} L_π(s_j, φ_i) in (4)
    end for
  end for
end for
```

Experimental results

OpenAI Gym Mujoco

➢ 로봇의 각 관절의 각도 및 각속도를 나타내는 벡터 입력

	Cheetah	Walker	Hopper	Ant	SlimHumanoid-ET
PETS	2288.4 ± 1019.0	282.5 ± 501.6	114.9 ± 621.0	1165.5 ± 226.9	2055.1 ± 771.5
POPLIN-A	1562.8 ± 1136.7	-105.0 ± 249.8	202.5 ± 962.5	1148.4 ± 438.3	-
POPLIN-P	4235.0 ± 1133.0	597.0 ± 478.8	2055.2 ± 613.8	2330.1 ± 320.9	-
METRPO	2283.7 ± 900.4	-1609.3 ± 657.5	1272.5 ± 500.9	282.2 ± 18.0	76.1 ± 8.8
TD3	3015.7 ± 969.8	-516.4 ± 812.2	1816.6 ± 994.8	870.1 ± 283.8	1070.0 ± 168.3
SAC	4474.4 ± 700.9	299.5 ± 921.9	1781.3 ± 737.2	979.5 ± 253.2	1371.8 ± 473.4
SUNRISE	4501.8 ± 443.8	1236.5 ± 1123.9	2643.2 ± 472.3	1502.4 ± 483.5	1926.6 ± 375.0

Table 1. Performance on OpenAI Gym at 200K timesteps. The results show the mean and standard deviation averaged over ten runs. For baseline methods, we report the best number in prior works (Wang & Ba, 2020; Wang et al., 2019).

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

Experimental results

DeepMind Control Suite

➢ 이미지 입력

500K step	PlaNet	Dreamer	SLAC	CURL	DrQ	RAD	SUNRISE
Finger-spin	561 ± 284	796 ± 183	673 ± 92	926 ± 45	938 ± 103	975 ± 16	983 ±1
Cartpole-swing	475 ± 71	762 ± 27	-	845 ± 45	868 ± 10	873 ± 3	876 ± 4
Reacher-easy	210 ± 44	793 ± 164	-	929 ± 44	942 ± 71	916 ± 49	982 ± 3
Cheetah-run	305 ± 131	570 ± 253	640 ± 19	518 ± 28	660 ± 96	624 ± 10	678 ± 46
Walker-walk	351 ± 58	897 ± 49	842 ± 51	902 ± 43	921 ± 45	938 ± 9	953 ± 13
Cup-catch	460 ± 380	879 ± 87	852 ± 71	959 ± 27	963 ± 9	966 ± 9	969 ± 5
100K step							
Finger-spin	136 ± 216	341 ± 70	693 ± 141	767 ± 56	901 ± 104	811 ± 146	905 ± 57
Cartpole-swing	297 ± 39	326 ± 27	-	582 ± 146	759 ± 92	373 ± 90	591 ± 55
Reacher-easy	20 ± 50	314 ± 155	-	538 ± 233	601 ± 213	567 ± 54	722 ± 50
Cheetah-run	138 ± 88	235 ± 137	319 ± 56	299 ± 48	344 ± 67	381 ± 79	413 ± 35
Walker-walk	224 ± 48	277 ± 12	361 ± 73	403 ± 24	612 ± 164	641 ± 89	667 ± 147
Cup-catch	0 ± 0	246 ± 174	512 ± 110	769 ± 43	913± 53	666 ± 181	633 ± 241

Table 2. Performance on DeepMind Control Suite at 100K and 500K environment steps. The results show the mean and standard deviation averaged five runs. For baseline methods, we report the best numbers reported in prior works (Kostrikov et al., 2021).

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

Ablation study

Weighted Bellman backups 효과 검증 (1)

- Mujoco 환경의 보상에 노이즈가 추가하여 에이전트 학습. 평가는 원래 보상으로.
  - $r'(s,a) = r(s,a) + \mathcal{N}(0,1)$

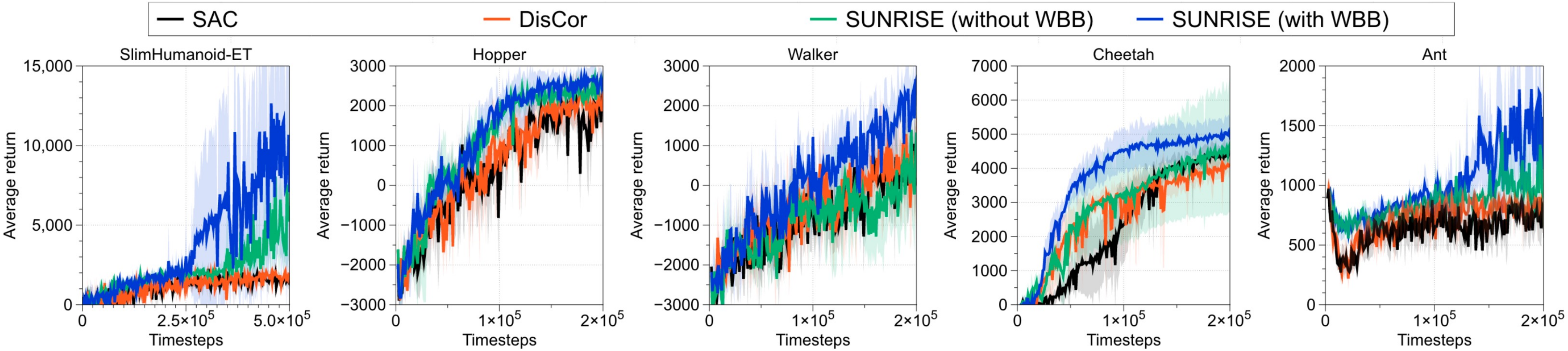


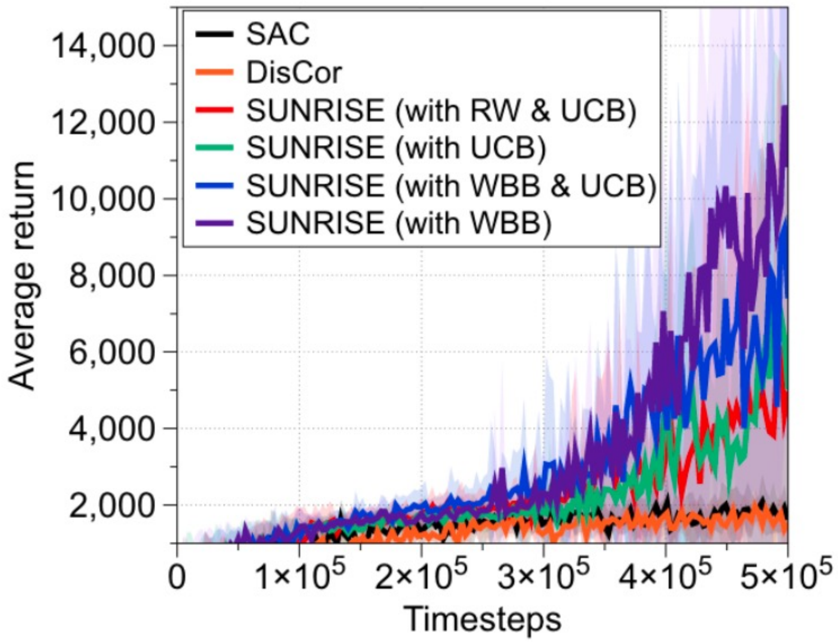
Figure 2. Learning curves on OpenAI Gym with noisy rewards. To verify the effects of the weighted Bellman backups (WBB), we consider SUNRISE with WBB and without WBB. The solid line and shaded regions represent the mean and standard deviation, respectively, across four runs.

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

Ablation study

Weighted Bellman backups 효과 검증 (2)

- 제일 복잡한 Humanoid 환경에 더 큰 노이즈를 추가한 보상으로 실험
  - $r'(s,a) = r(s,a) + \mathcal{N}(0,5)$



(a) Large noise

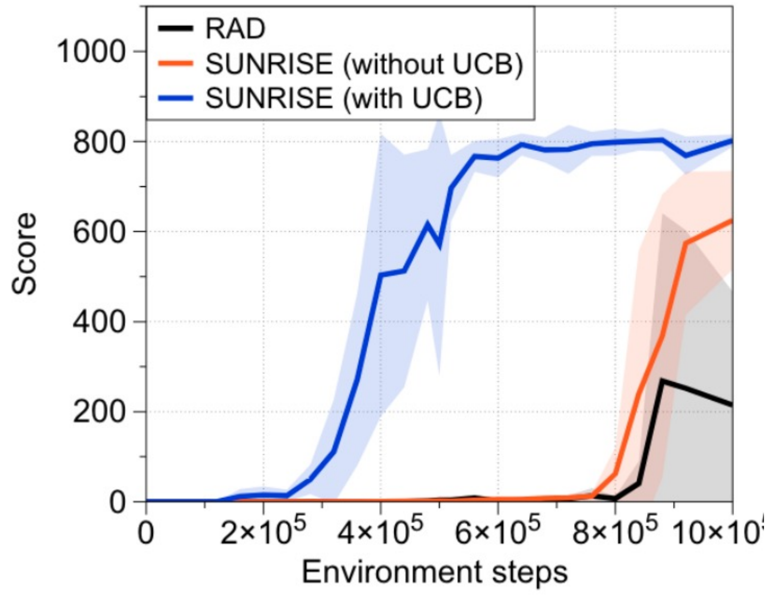
Mujoco의 Humanoid 환경
RW: Random weight
WBB: Weighted Bellman backups

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

Ablation study

UCB exploration의 효과

- CartPole with sparse reward에 실험
- Single agent가 더 많이 학습되면 SUNRISE를 뛰어 넘을 수 있는지 확인



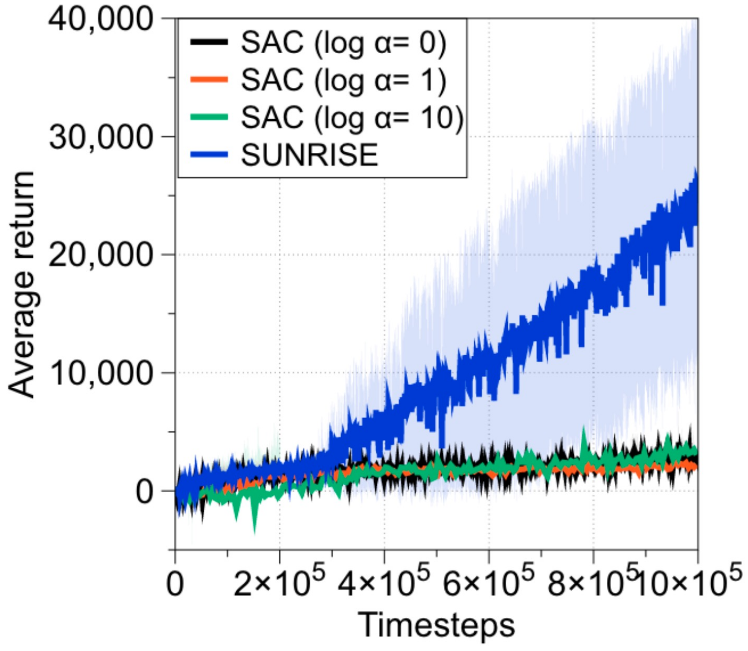
(b) Sparse reward

DMC의 CartPole with sparse reward

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

더 많이 학습한 single agent와 비교

- 네트워크가 N배 더 많은 만큼 더 많은 파라미터 업데이트를 한다.
- Single agent가 더 많이 학습되면 SUNRISE를 뛰어 넘을 수 있는지 확인



(c) Gradient update

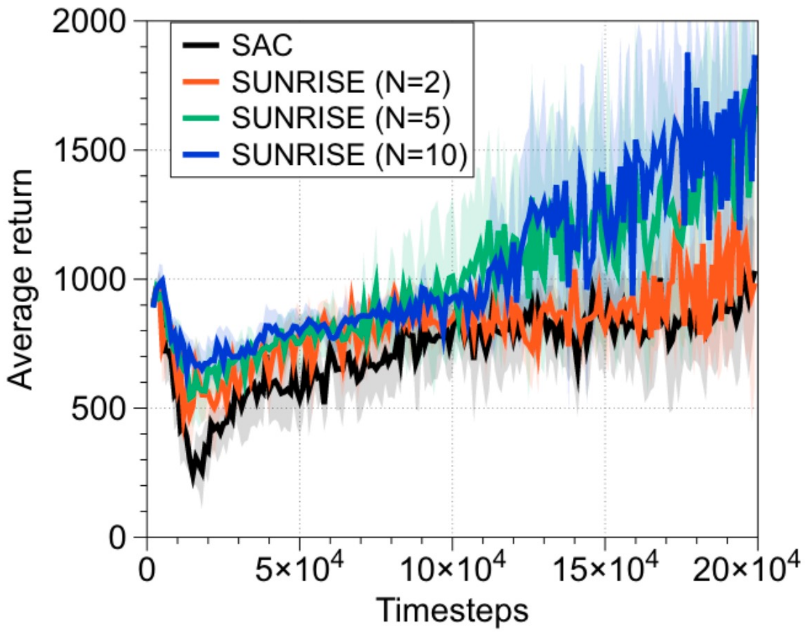
Mujoco의 Humanoid 환경

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

Ablation study

네트워크 개수의 영향

- 네트워크 개수가 증가할 수록 성능이 증가하지만,
- $N = 5$에서 saturation 된다.



(d) Ensemble size

Mujoco의 Ant 환경

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)

Discussion
Computation overhead

- 각 네트워크 forward pass가 병렬 처리할 수 있기 때문에 충분히 efficient 하다고 주장
- But, 시간에 따른 실험 결과는 없음

Quoted from Kimin Lee et al., SUNRISE: A Simple Unified Framework for Ensemble Learning in Deep Reinforcement Learning. (2021)