38 강화학습 기반 V2G 전력거래 이익 최대화

소속 정보컴퓨터공학부 분과 C 팀명 노란전력 참여학생 김도균, 배레온 지도교수 황원주

과제 개요

과제 목표

V2G 전력 거래 이익 최대화 문제를 강화학습을 사용하여 해결한다

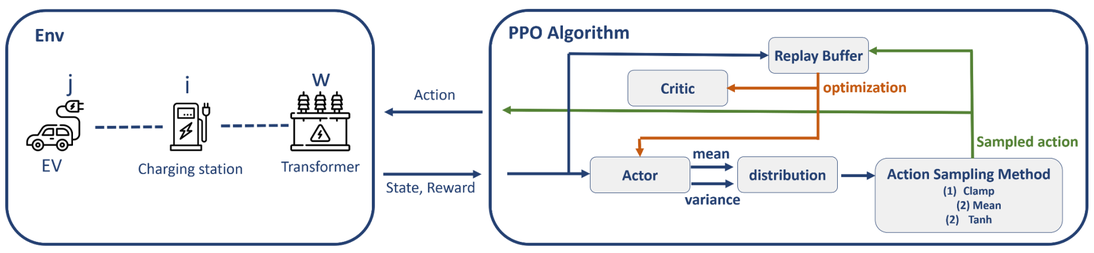
1) V2G 환경에서 강화학습을 통한 문제 해결의 가능성을 제공한다
2) V2G 환경에 적합한 강화학습 알고리즘을 제안한다

결과 요약

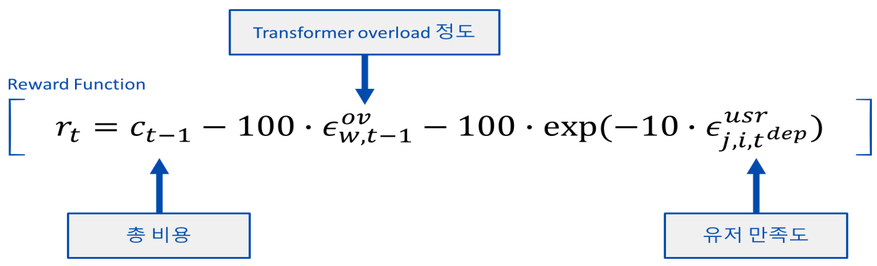
1) 강화학습 알고리즘은 V2G에서의 충방전을 충분히 고려하였으며, 학습 이후의 예측연산이 상당히 빨랐다
2) 기존의 SOTA 알고리즘 중에서는 TRPO가 좋은 성능을 보였다
3) Action Sampling 개선은 Tanh와 Clamp가 좋은 성능을 보였다
4) 대체로 TRPO와 PPO_TANH가 좋은 성능을 보이나 문제가 복잡할 수록 계산이 단순한 PPO_CLAMP가 좋은 성능을 보인다.

상세 내용

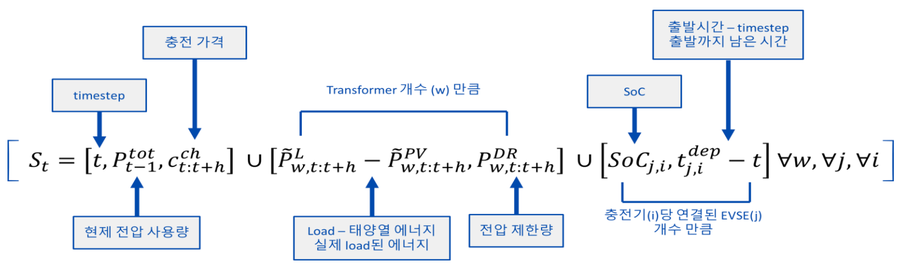
V2G 환경에서의 PPO



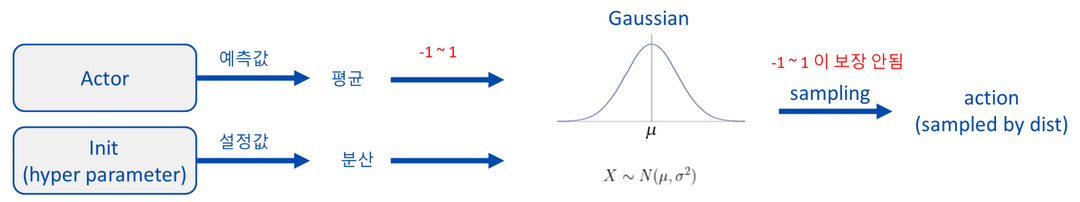
Reward



State



Action Sampling 기법

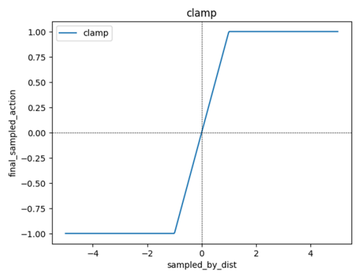


탐색을 위해 Action Sampling을 하는 과정에서 Sampling된 Action이 -1 ~ 1의 범위를 벗어나면서 학습에 부정적인 영향을 끼친다.

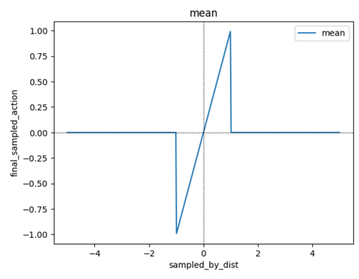
이를 해결하기 위해 Gaussian에서 Sampling된 Action에 하나의 함수를 추가하여 Action이 -1 ~ 1 범위 안으로 들어오도록 한다.

아래의 그래프와 같은 **3가지 방식 (Clamp, Mean, TanH)** 으로 PPO의 Action Sampling을 개선하였다.

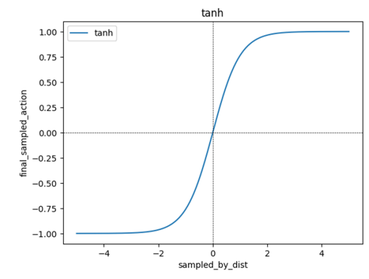
1) PPO_CLAMP



2) PPO_MEAN



3) PPO_TANH



실험 결과

EV 충전기 10개

Table 1. V2G Profit Maximization Problem with 10 EV Chargers (1000 episode)

Env	Algorithm	Profits/costs ¢	ϵ^{usr} (%)	Energy Ch. (kWh)	Energy Disch. (kWh)	Tr. Ov. (kWh)	Execution Time (s)	Reward ($\times 10^{-3}$)
Env1	PPO	7.60	65.86	74.56	91.32	42.69	0.05	-4.30
	PPO_CLAMP	17.66	53.53	48.94	110.11	31.10	0.05	-3.15
	PPO_MEAN	9.89	62.61	63.82	92.35	29.63	0.04	-3.00
	PPO_TANH	22.83	49.16	56.42	134.75	9.77	0.05	**-1.01**
	SAC	16.08	56.36	73.23	123.65	31.26	0.02	-3.16
	TRPO	24.75	48.25	65.34	145.06	8.08	0.02	**-0.83**
Env2	PPO	8.84	64.12	69.84	92.67	40.15	0.05	-4.04
	PPO_CLAMP	17.76	53.41	49.59	111.23	34.29	0.05	-3.47
	PPO_MEAN	10.03	62.39	63.22	92.72	31.42	0.05	-3.18
	PPO_TANH	24.70	46.40	48.69	136.63	19.44	0.05	**-1.98**
	SAC	15.50	56.98	75.10	123.11	29.03	0.02	-2.94
	TRPO	25.63	46.66	59.28	144.35	8.13	0.02	**-0.84**

Blue is Best Reward, **Green** is Second Reward

EV 충전기 20개

Table 2. V2G Profit Maximization Problem with 20 EV Chargers (1000 episode)

Env	Algorithm	Profits/costs ¢	ϵ^{usr} (%)	Energy Ch. (kWh)	Energy Disch. (kWh)	Tr. Ov. (kWh)	Execution Time (s)	Reward ($\times 10^{-3}$)
Env1	PPO	51.49	43.93	67.69	258.57	231.99	0.06	-23.28
	PPO_CLAMP	37.17	52.41	95.98	225.90	43.19	0.06	**-4.39**
	PPO_MEAN	51.78	43.86	66.28	260.57	83.50	0.06	-8.43
	PPO_TANH	33.17	54.16	95.81	210.41	62.66	0.06	**-6.34**
	SAC	6.14	72.08	222.08	203.01	135.64	0.03	-13.62
	TRPO	9.92	67.84	151.00	162.67	73.74	0.02	-7.43
Env2	PPO	51.50	43.93	67.60	258.50	160.68	0.06	-16.15
	PPO_CLAMP	36.99	52.45	93.70	223.39	31.85	0.06	**-3.26**
	PPO_MEAN	51.71	43.67	65.80	261.44	39.73	0.06	**-4.06**
	PPO_TANH	32.96	54.24	96.10	210.01	48.31	0.06	-4.90
	SAC	6.59	73.24	252.86	224.68	177.95	0.03	-17.84
	TRPO	9.03	68.39	165.05	172.67	73.18	0.02	-7.37

Blue is Best Reward, **Green** is Second Reward