Pusan National University
Computer Science and Engineering
Technical Report 2024-10

강화학습 기반 V2G 전력거래 이익 최대화

202155517 김도균
201924488 배레온

황원주

목 차

1. 서론

1.1 연구 배경

온실가스 배출량의 지속적인 증가와 전례 없는 기상이변 들로 기후변화에 대한 위기의식이 고조되었다. 2018년에는 기후변화에 관한 정부 간 협의체 (Inter-governmental Panel on climate Change, IPCC)에서 2050년 까지 세계 탄소 순 배출량이 0이 되는 것을 목표로 삼고 있는 탄소 중립을 제안하였다. 2020년 1월 세계 경제 포럼에서는 세계가 직면한 가장 큰 위협으로 기후변화를 지목했으며 각국의 온실가스 감축 노력이 강화되어야 한다고 강조하였다. 게다가 2020년에는 선진국과 개도국이 함께 온실가스 배출 감축을 위해 맞은 파리협정 하에 신기후체제로 전환되었으며, 한국은 탄소 중립에 참가하였다. 한국은 초기에 2050 탄소중립 시나리오로 3개의 안을 공개하였다. 1안은 기존 체계와 구조를 최대한 활용하는 것이고 2안은 기술 발전으로 생활양식 변화를 3안은 화석연료 소비를 과감하게 줄이고 수소를 공급하는 방안이다. 이 중 심의를 거쳐 확정된 2안과 3안에서는 재생에너지 비중이 매우 확대된다.

제주의 경우에는 2030 탄소 없는 섬으로 비전을 발표하며 탄소제로를 목표로 신재생 에너지 사용 비중을 전체 중 16% 이상이라는 높은 비율로 책정하여 에너지 정책을 시행하였고, 이에 따라 신재생 에너지 발전시설이 급증하게 되었다. 이에 따라 제주지역에서 수요에 비해 전력 공급이 많아지는 현상이 발생했으며, 풍력발전기 가동을 중단하는 등 출력제한 문제가 발생하였다.

호남 · 영남권 지역의 경우에는 땅값이 상대적으로 싸 태양광 발전설비가 급격하게 증가하였고, 전력이 과잉 공급되게 되었다. 하지만, 상대적으로 전력 수요가 높은 수도권으로 전력을 전송할 전력망 인프라가 부족한 상황으로 2023년 산업부는 공공기관이 보유한 설비를 우선으로 출력제한을 실시하였다. 이에 따라, 태양광 발전 사업자들은 출력제한에 따른 금전적인 손실이 갈수록 늘어나고 있다. 이러한 흐름은 신규 사업자들의 사업 의지를 꺾을 수 있으며, 신재생 에너지의 보급 확대에 문제가 될 수 있다.

1.2 연구 목표

[그림 1]. 제주의 출력제한 해결 방안 (출처 : [1])

제주에서는 출력제한 해결을 위한 다양한 방안을 [그림 1]과 같이 내놓았다. 이중 가까운 미래에 적용하기 가장 적합한 것으로 보이는 대안은 미활용 전력 전기에너지를 저장하는 방안이다.

전력 과잉 공급은 한국만의 문제가 아니다. 미국의 재생에너지 생산량은 2020년 21%에서 2050년에 42%로 증가할 것으로 예상하였고, 필요 이상의 재생에너지가 생산되는 상황이 발생하였다. 이에 과다 생산된 전력을 저장하는 ESS 도입의 중요성이 대두되었다. 재생에너지의 생산 비중이 높은 캘리포니아는 2020년 기준 506MW의 ESS가 운영되고 있으며 추가로 1,027MW 규모의 도입을 준비하고 있다. 뉴욕은 2020년 기준 93MW 규모의 ESS를 도입하였으며, 1,076MW 규모의 ESS를 준비하고 있다.

중국은 2025년까지 ESS를 구축한 후, 본격적인 상용화에 접어들 것이라고 밝혔으며, 핵심 기술 및 장비의 자주화를 위해 중국 배터리 제조사인 CATL와 BYD에 유리한 조건을 내걸었다.

그 외에도 일본의 경우에는 대지진에 대비하여 비상 전원 확보 차원에서 ESS에 투자하고 있으며, 설치 보조금을 지원하고 세제 혜택을 주는 등의 정책을 실시하고 있다. 호주의 경우에는 16년의 대정전 이후 안정성 있는 국가 에너지 정책 추진을 위해 세계 최대 규모의 ESS를 구축하고 있으며, 프랑스는 태양광 발전 자가소비용 ESS 설치 자금을 지원하는 등의 정책을 실시하고 있다. 하지만, 이러한 ESS는 정작 한국에서는 많은 화재 사고와 재산 피해로 외면받고 있다.

이에 반하여 V2G는 전기차의 보급 확대에 따라 주목받는 시장으로 떠오르고 있다. 실제로 [그림 2]의 도표에 따르면, 국내 전기차 보급은 꾸준히 늘어나고 있으며, 전기차 충전 인프라 또한 매년 늘어나고 있다.

[그림 2]. 연도별 국내 전기차 · 충전기 보급 현황 (출처 : [2])

V2G는 전기차와 전력망을 양방향으로 연결하여, 전기차의 배터리를 전력망의 에너지원으로 활용하는 기술이며 이는 전기차가 단순한 이동 수단을 넘어, ESS의 역할을 할 수 있게 한다. V2G 시스템을 통해 전기차는 필요시 전력을 전력망에 공급하거나, 과잉 전력을 저장할 수 있는데 이러한 시스템은 사회 전반적 전력 비용 절감, 전력망의 안정성 유지, 에너지 관리의 최적화 등 지속 가능한 에너지 시스템을 운영하는 데 있어서 필수적이다.

하지만, 아직은 V2G 시스템을 통해 얻을 수 있는 경제적 이익이 충분히 크지 않으며 전기차 소유자가 V2G를 통해 얻는 금전적 혜택이 배터리 수명 단축으로 인한 비용을 상쇄할 만큼 충분하지 않은 경우가 많으므로 본팀은 V2G의 이익을 최대화할 수 있는 알고리즘을 개발하여 V2G가 사회 인프라에 적극적으로 도입될 수 있는 근거를 제공하고자 한다.

1.3 문제 분석

1.3.1 국내 ESS 시장의 위축

국내의 ESS는 연이은 화재 사고와 재산 피해로 많이 위축된 상황이다. ESS 운전 요금 지원의 종료로 시장 규모는 더욱 위축되었으며, 2021년까지의 손해액은 약 466억 원에 달한다. ESS 도입이 활발한 글로벌 시장과 달리 국내의 ESS는 외면받고 있다.

이에 반하여, 전기차 시장은 점차 확대되고 있으며, 충전기 또한 늘어나고 있다. 이에 전기차를 일종의 ESS로 사용하는 V2G를 도입하여 ESS와 연계하면, ESS의 수요 조절과 함께 ESS의 인식을 개선하고 편익도 늘리는 등의 다양한 이점을 기대한다.

1.3.2 기존 스케줄 방식의 한계

전기차 시장의 경제 효율성을 위해 탄생한 양방향 충전 시스템은 현재 대부분 소규모 충전소의 EMS에서 고전적인 수학적 알고리즘 혹은 모델 예측 제어(MPC)를 통해 충전 스케줄을 생성한다. 이때 EMS는 전력비용을 최소화하며, ESS의 충전 및 방전 시점을 최적화해 전력망의 부하를 줄이는데 이러한 알고리즘은 충전소 규모가 작은 덜 복잡한 환경에서 최적화 문제를 해결하는 데 효과적이다. 하지만 도착하는 전기차, 출발하는 전기차 수, 전기차 배터리 용량, 충전 수요, 에너지 가격, 충전소의 ESS 개수, ESS 용량 등의 다양한 제약조건을 고려해야 하는 위 방식은 제약조건에 따라 EMS 프로그래밍은 어려움을 겪게 된다. 마찬가지로 충전소 규모가 커지거나 다른 요인으로 복잡한 환경의 충전소에서는 기존의 알고리즘을 채택하기 무척 어려워진다. 따라서 이러한 제약조건에 따른 복잡한 문제도 효율적으로 해결하며 다양한 조건 속에서 이익을 최대화할 방안을 모색하고자 한다.

1.3.3 시뮬레이션의 필요성

개발한 알고리즘의 평가를 위하여 실제 전기차 충전소와 전기차를 활용하는 것은 현실적으로 어렵다. 이에, 시뮬레이션 환경을 활용하여 이익을 최대화할 방안을 찾고자 한다.

TABLE I
OVERVIEW OF EXISTING EV SIMULATORS FOCUSING ON SMART CHARGING STRATEGIES.

Simulator Name	V2G	Power Network Impact	EV Models	EV Behavior	Charging Stations	Available Baseline Algorithms	RL Ready	Programming Language	Comments
V2G-Sim [12]	Yes	Partial	Diverse	Real Charging Transaction Probability Distributions	Uniform	- Heuristics, Mathematical Programming	No	Not Open-source	- Customizable V2G simulations.
EVLibSim [13]	Yes	No	Diverse	Randomized	Uniform	- Heuristics, Mathematical Programming	No	Java	- Easily customizable simulations with a visual interface.
EV-EcoSim [14]	Yes	Complete	Uniform	Randomized	Uniform	- Mathematical Programming	No	Python	- Grid-Impact analysis
evsim [15]	No	Partial	Diverse	Real Charging Transaction Probability Distributions	Uniform	- Heuristics, Mathematical Programming	No	R	- Simulate and analyze the charging behavior of EV users.
OPEN [16]	Yes	Complete	Uniform	Randomized	Uniform	- Heuristics, Mathematical Programming	No	Python	- Modelling, control & simulations for smart local energy systems.
ACN-Sim [17]	No	Complete	Uniform	Real Charging Transactions	Uniform	- Heuristics, MPC, RL, Mathematical Programming	Yes	Python	- Designing a complete simulator framework.
SustainGym [18]	No	Complete	Uniform	Real Charging Transactions	Uniform	- RL	Yes	Python	- Providing a benchmark for sustainable RL applications.
Chargym [19]	Yes	Very Limited	Uniform	Randomized	Uniform	- Heuristics, RL	Yes	Python	- Comparing RL algorithms for smart charging.
EV2Gym (Ours)	Yes	Partial	Diverse	Real Charging Transaction Probability Distributions	Diverse	- Heuristics, MPC, RL, Mathematical Programming	Yes	Python	- Comprehensive simulator for any control algorithm.

[그림 3]. EV 시뮬레이터의 비교

[그림 3]은 현재 존재하는 다양한 시뮬레이션을 비교하는 자료다.

- V2G-Sim
  EV 모델과 동작과 같은 풍부한 기능을 갖췄지만, 오픈소스가 아니며 강화학습 개발을 위한 환경을 제공하지 않는다.
- EVLibSim
  다양한 EV 모델을 제공하고 있지만, Java로 작성되었으며 그리드에 미치는 영향을 시뮬레이션하지 않는다.
- EV-EcoSim
  자세한 배터리 사용률 및 성능 저하 모델이 포함되어 있으나, 현실적인 EV 사양 및 동작이 아닌 배터리에 중점을 두고 있다.
- EVsim
  EV 사용자의 행동을 평가하고 분석하는 데에 중점을 두었으며 사실적인 EV 행동 데이터로 구축되어 있으나, V2G의 영향을 조사할 수 있는 옵션이 포함되어 있지 않다
- OPEN
  EV를 포함하여 스마트 지역 에너지 시스템을 위한 통합 모델링, 제어 및 시뮬레이션을 위한 프레임워크로 배전 시스템 수준에서의 전력 흐름을 해결할 수 있다. 하지만, 통일된 EV 사양과 동작 모델만 있으며, 표준화된 Gym 환경이 없고 지나치게 단순화하여 강화학습 개발에 필요한 깊이가 부족하다.

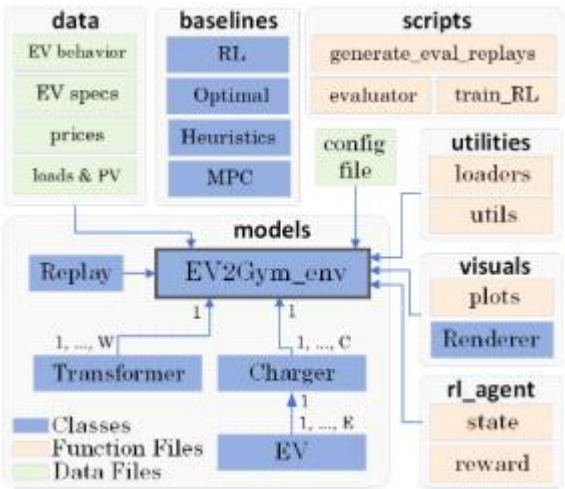
- ACN-Sim
  표준화된 Gym 환경이 포함되어 있으며, EV 주차장 특성에 따라 개발되었다. 가장 확립된 EV 시뮬레이터 플랫폼 중 하나이며 다른 오픈소스 그리드 시뮬레이터를 사용하여 전력 네트워크 계산을 지원한다, 하지만, V2G 지원을 염두에 두지 않아 V2G 연구에는 부적합하다.
- SustainGym
  ACN-Sim에서 state와 action space를 정의하여 EV 최적 충전 문제를 강화학습 벤치마크로 표준화하였지만, 그 외의 실질적인 기능 추가가 없어 ACN-Sim과 같은 문제를 공유한다.
- Chargym
  비용 및 페널티 설계에 중점을 둔 강화학습 알고리즘을 개발하는 데 사용한다. 모든 EV와 충전기가 동일한 사양을 가지고 EV 동작이 실제 데이터를 기반으로 하지 않는 등 기본 모델이 매우 단순하여 강화학습 연구에 부적합하다.

그 외에도 다양한 EV 충전 관리 시뮬레이션이 있지만, 분야가 다르거나 구식이라는 문제점이 있다. 기존 강화학습을 도입한 시뮬레이션보다 현실적이고 표준화된 시뮬레이션 환경이 필요하였고, 이에 EV2Gym 시뮬레이션을 채택하였다.

2. 연구 방향

2.1 EV2Gym [3]

V2Gym은 V2G를 위한 오픈소스 시뮬레이터 환경이다. 강화학습 환경을 위한 파이썬 패키지 Gymnasium을 기반으로 구성되어 간소화된 강화학습 알고리즘 평가가 가능하다.



[그림 4]. EV2Gym 패키지의 폴더 및 파일 구조

[그림 4]는 EV2Gym 패키지의 폴더 및 파일 구조를 보여주며, 기본적인 클래스와 함수, 데이터 파일 등이 어떻게 상호 연결되어 있는지 보여준다. EV2Gym의 환경은 config file을 통해 정해진 변수로 이루어지며, Transformer, Charger, EV의 상호작용으로 동작한다. EV2Gym의 데이터는 실제 네덜란드의 오픈 데이터를 바탕으로 구성되어 있으며, EV의 행동과 스펙, 전기의 가격 등을 포함한다. 그 외에도 참고를 위한 baseline 알고리즘과 강화학습 state, reward 예시, 시각화를 위한 함수 등을 제공한다. 특히, EV2Gym은 오픈소스 겸 모듈식으로 필요에 따라 다양한 추가 기능을 추가 및 수정이 원활하게 가능하다는 장점이 있다.

Algorithm 1 Python code for running a simulation.

```
from EV2Gym.EV2Gym_env import EV2Gym
env← EV2Gym(config_file)                    ▷ Initialization
state, _ ← env.reset()
agent ← Algorithm(...)                      ▷ User-defined algorithm
for t = 1 to T do                           ▷ Simulation Phase
    actions ← agent.get_action(env, state)
    state, reward, done, _, stats ← env.step(actions)
end for
```

[그림 5]. EV2Gym의 실행 pseudocode

[그림 5]는 EV2Gym을 실행시키는 간단한 알고리즘을 보여준다. 시뮬레이션의 초기화 단계에서는 config file을 바탕으로 Charger, Transformer, EV의 개수 등 개별 시뮬레이션 환경의 다양한 매개 변수를 설정한다. [표 1]은 config file에 포함된 세부적인 매개 변수를 보여준다.

[표 1] config file의 매개변수

구성 요소	매개 변수	기호
시뮬레이션	- 시간 간격 (분 단위)	Δt
	- 시뮬레이션 길이 (step 단위)	T
	- 시작 날짜와 시간	
	- 충전 토폴로지 및 속성	
	- EV 속성	
	- EV 시나리오 (주거용, 직장, 공용)	
	- EV 충전소의 집합	C

	- Power Transformer의 집합	W
EV	- AC 충전 전력의 최대, 최소 (kW)	$\underline{P}_{AC}^{ch}, \overline{P}_{AC}^{ch}$
	- DC 충전 전력의 최대, 최소 (kW)	$\underline{P}_{DC}^{ch}, \overline{P}_{DC}^{ch}$
	- 방전 전력의 최대, 최소 (kW)	$\underline{P}^{dis}, \overline{P}^{dis}$
	- 배터리 용량의 최대, 최솟값 (kWh)	$\underline{E}, \overline{E}$
	- 충전 및 방전 효율	η^{ch}, η^{dis}
	- 도착 시 배터리 용량 (kWh)	E^{arr}
	- 출발 시 원하는 배터리 용량 (kWh)	E^{*}
	- CV/CC 전환 SoC (%)	τ
	- 도착 시간 & 출발 시간 (t)	t^{arr}, t^{dep}
Charging Station	- 충전 스테이션 전류의 최대, 최소 (A)	$\underline{I}^{cs}, \overline{I}^{cs}$
	- EVSE 충전 전류의 최대, 최소 (A)	$\underline{I}^{ch}, \overline{I}^{ch}$
	- EVSE 방전 전류의 최대, 최소 (A)	$\underline{I}^{dis}, \overline{I}^{dis}$
	- 전압 (V) & 위상	V, ϕ
	- EVSE의 집합	J
	- 충전기의 종류 (AC 또는 DC)	
	- 충·방전 가격 (€ / kWh)	c^{ch}, c^{dis}
Transformer	- 전력의 최대, 최소 (kW)	$\underline{P}^{tr}, \overline{P}^{tr}$
	- Inflexible Loads (kW)	P^{L}
	- 태양광 발전 (kW)	P^{PV}
	- Demand Response Event (kW)	P^{DR}
	- 연결된 충전소 집합	C_w

2.2 강화학습 알고리즘

2.2.1 TRPO [4]

TRPO는 정책 경사도 방법을 효율적으로 사용하기 위해 제안되었다. 정책 경사도에서 step-size는 중요한 역할을 하는데, 정책 경사도가 너무 커지면, 과장된 향상 방향을 따라 발산할 위험이 있고, 너무 작아지면, 학습 속도가 느려져 학습 효율이 떨어지는 문제가 있다.

TRPO는 정책 향상이 보장되는 이론적인 알고리즘을 바탕으로 근사하여 최

적 정책으로 수렴시킨다. 이때, 본래의 목적에 따라 step-size를 KL divergence로 제한한다. 하지만, 모든 state에 대한 KL divergence 계산을 하는 것은 현실적으로 어려움이 있기에 몬테카를로 기법을 이용하여 계산하게 된다. 즉, 2차 미분을 통해 Optimize가 이루어진다.

2.2.2 PPO [5]

PPO는 TRPO와 같은 목적을 가지고 있다. 주어진 데이터를 가지고 현재 policy를 최대한 큰 step만큼 향상하면서도 발산할 정도로 크게는 업데이트하지 않는 것이 목적이다. PPO와 TRPO의 차이점은 Penalty Term 즉, 너무 크게 변경되지 않도록 제한하는 부분에서 차이를 보인다. TRPO처럼 KL divergence를 사용하는 대신에 Clipping 방식을 사용하였다. 이를 통해 TRPO와 같은 방식의 계산을 2차 미분이 아닌 1차 미분으로 계산할 수 있도록 하면서, 비교적 간단한 구현으로 실용적으로 사용할 수 있게 되었다.

Algorithm 1 PPO, Actor-Critic Style

for iteration=$1, 2, \ldots$ **do**
  for actor=$1, 2, \ldots, N$ **do**
    Run policy $\pi_{\theta_{\text{old}}}$ in environment for T timesteps
    Compute advantage estimates $\hat{A}_1, \ldots, \hat{A}_T$
  end for
  Optimize surrogate L wrt θ, with K epochs and minibatch size $M \leq NT$
  $\theta_{\text{old}} \leftarrow \theta$
end for

[그림 6]. PPO pseudocode

2.2.3 SAC [6]

SAC는 실제 환경에서 Model-free 알고리즘 적용이 어렵다는 문제에서 시작되었다. On-policy의 경우에는 갱신마다 샘플링 과정이 필요하기에 샘플 효율이 저하된다는 문제가 있고, Off-Policy의 경우에는 연속적인 state와 action에서 샘플의 복잡도가 높아진다는 문제가 있다. DDPG의 경우 이를 Off-policy에 정책 경사도 기반으로 해결했으나, 하이퍼파라미터 변화에 대한 민감성이 높고 수렴성이 약하다는 문제가 있었다. 이에 SAC는 Off-policy와 정책 경사도 기반을 따르면서, 하이퍼파라미터의 민감성이 낮고 수렴성이 높은 알고리즘을 제안하였다.

SAC는 기존 강화학습의 목표인 기대 보상의 최대화에 엔트로피 최대화를 사용하였다. 이를 통해 확률적인 최적 정책을 구성할 수 있도록 하였다. 또한, 엔트로피 최대화가 적용된 정책 반복알고리즘인 SPI (Soft Policy Iteration) 를 기반으로 동작하는데, SPI의 요구 연산량이 많고 Table 방식에서만 적용 가능하다는 단점을 근사화를 통해서 해결하였다.

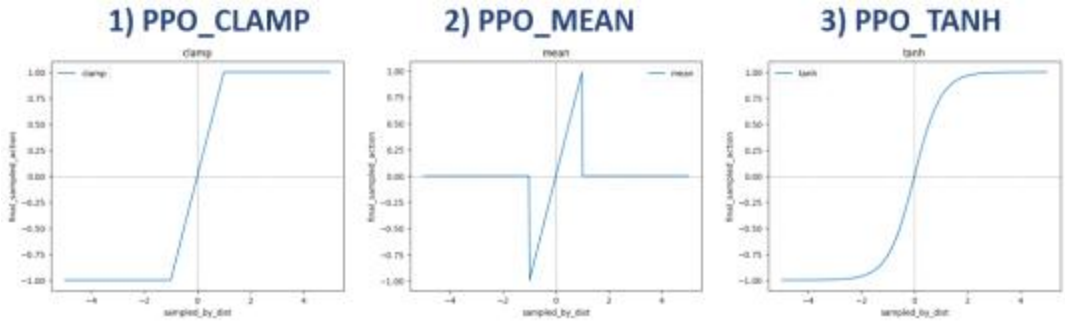
Algorithm 1 Soft Actor-Critic

Initialize parameter vectors ψ, $\bar{\psi}$, θ, ϕ.
for each iteration **do**
  for each environment step **do**
    $\mathbf{a}_t \sim \pi_\phi(\mathbf{a}_t|\mathbf{s}_t)$
    $\mathbf{s}_{t+1} \sim p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
    $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$
  end for
  for each gradient step **do**
    $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi J_V(\psi)$
    $\theta_i \leftarrow \theta_i - \lambda_Q \hat{\nabla}_{\theta_i} J_Q(\theta_i)$ for $i \in \{1, 2\}$
    $\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi J_\pi(\phi)$
    $\bar{\psi} \leftarrow \tau\psi + (1-\tau)\bar{\psi}$
  end for
end for

[그림 7]. SAC pseudocode

2.3 Action Sampling

PPO 알고리즘의 Actor가 예측을 할때, 강화학습의 탐험을 보장하기 위해서, 정확한 값을 사용하는 것이 아닌, Gaussian 분포하에 sampling된 값을 사용한다. 이때, Actor가 예측한 값은 원하는 action의 분포인 [-1, 1]을 만족하지만, Gaussian 분포를 거치면, 보장할 수 없게 된다. 이로 인하여, 실제로 11% 정도의 범위를 벗어난 action이 나오게 되었고, 이에 다른 Action Sampling 방법을 사용해 보았다.



[그림 8]. Action Sampling 방법별 그래프

2.3.1 Clamping

해당 방법은 범위를 벗어난 값을 범위안의 값으로 제한시키는 clamping 함수를 사용한 것이다. 가장 구현이 단순하면서, 기존의 값을 보존할 수 있는 방법이다.

2.3.2 Mean 값 사용

해당 방법은 Gaussian 분포상 가장 sampling 될 확률이 높은 mean 값을 사용하는 방식이다. 확률 분포를 가장 잘 보존할 수 있는 방법이다.

2.3.3 Tanh 함수 사용

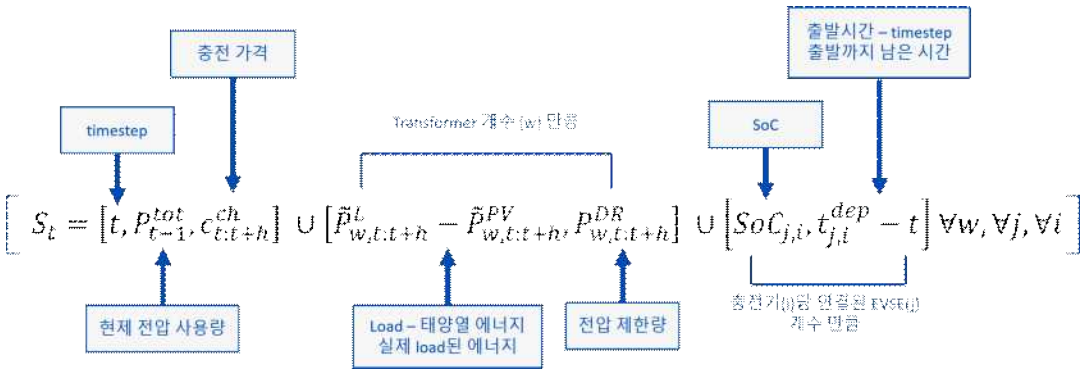
해당 방법은 Tanh 함수의 결과가 [-1, 1] 사이라는 점을 사용하여, sampling 된 값이 Tanh 함수를 한번 더 거치도록 한 방식이다. 급격한 변화를 주는 Clamping, Mean과는 다르게 부드럽게 치환할 수 있는 방법이다.

3. 연구 내용

3.1 시나리오 정의

설계된 V2G 시나리오의 주요 참가자는 크게 EV 사용자와 V2G 그리드 관리자로 나누어진다. V2G 알고리즘은 이러한 환경에서 충전 요구사항, 그리드의 상태, 전기 가격 변동 등을 고려하여 최적의 충방전 일정을 만들어, 그리드의 안정성과 경제적 이익을 최대화하는 것을 목표로 한다. V2G 시스템은 전기 가격의 시세차이를 반영하여 전기 충전 가격을 절감하고 이는 특정 비율의 수수료를 제외한 후 EV 사용자에게 모두 반영된다. 그리드 관리자는 수수료를 통해 그리드의 운영비과 추가 수익을 가져가는 윈-윈 솔루션이 될 수 있다.

3.1.1 State와 action



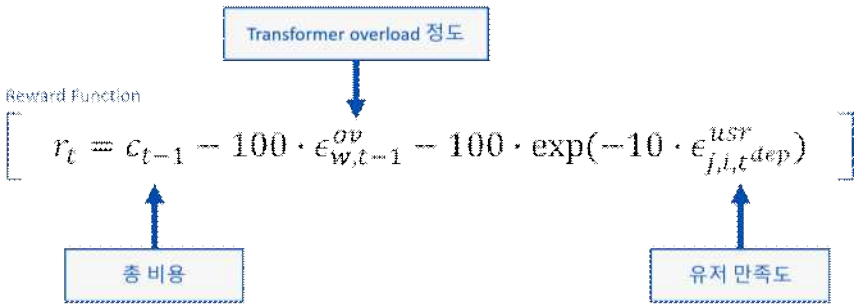
[그림 9]. State Function의 구조

State는 시뮬레이션의 상태를 나타내는 값으로 필요한 정보를 담은 vector이다. vector size는 로 는 horizon, 는 Transformer의 개수, 은 EVSE의 개수를 의미한다. 여기에 추가로 필요에 따라 calencer 기반의 배터리 열화와 cycling 기반의 배터리 열화 예측 값이 사용된다.

action은 각 EVSE의 충전량을 바탕으로 계산되며, 강화학습 중 제한된 동작을 보장하기 위해 [-1, 1]의 범위로 제한하여 계산한다. 각각, 1은 최대로 충전, -1은 최대로 방전을 의미하며, 총 EVSE의 개수만큼의 vertor space를 가진다.

3.1.2 기존의 리워드

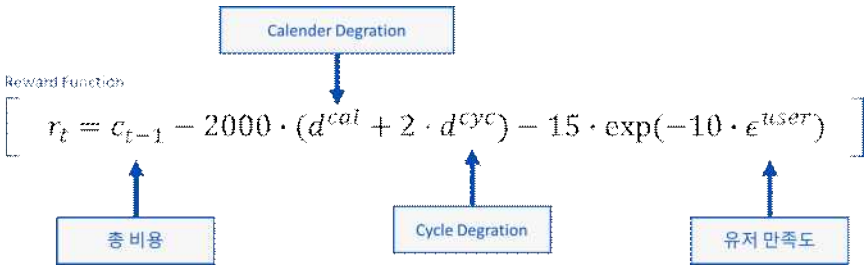
기존 Ev2Gym의 RL BaseLine에서는 다양한 리워드 함수가 제안된다. 그 중에서 V2G 환경에서의 이익 최대화에 관한 Reward Function은 ProfitMax TrPenalty UserIncentives 함수 이다. reward는 해당 action으로 얻을 수 있는 비용, 유저 만족도, Transformer의 overload 정도로 계산한다. 유저 만족도는 유저가 목표로 하는 SoC상태와 현제 SoC 상태의 비율로 계산하며, Transformer overload 정도는 Load된 Power량이 limit Power를 넘은 값 만큼을 사용한다. 해당 리워드 함수는 총 수익을 최적화 하면서 변압기 과부하 방지와 사용자 만족도를 함께 높이는 것을 목표로 한다. 또한 사용자 만족도에 지수 함수를 사용하여 더 낮은 만족도에서 더 큰 페널티를 부여한다.



[그림 10]. ProfitMax TrPenalty UserIncentives 함수의 구조

3.2 EV 사용자 시나리오

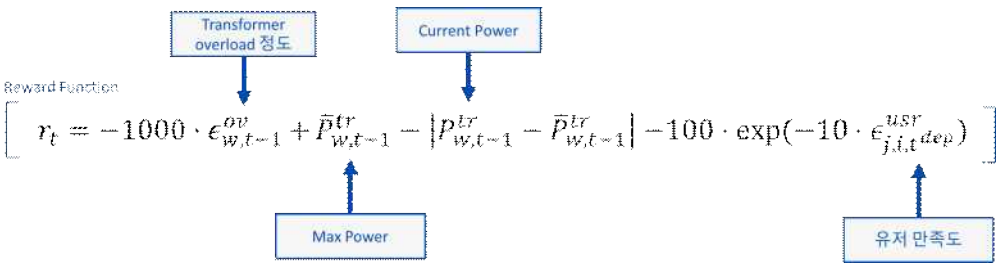
EV 사용자 입장에서 고려해야 할 문제는 경제적 이익과 충전 만족도 배터리 열화 방지가 있다. 경제적 이익의 경우에는 전기 판매가와 구매가의 차이를 통해 최종적으로 충전을 완료했을 때 기존에 비해 낮은 가격으로 충전을 하는 것이 목표이고, 충전 만족도의 경우 출발했을 때 목표 SOC에 도달하는 것이다. 배터리 열화 방지의 경우 과도한 충방전으로 인한 배터리 수명 단축을 방지하는 역할을 한다. 이 점들을 고려하여 다음과 같은 Reward Function을 설계하였다. 먼저, 총 비용을 advantage로 주고, 배터리 열화 정도에 따라 penalty를 준다. 이후, 유저 만족도에 따라 penalty를 더하여 리워드를 만든다. 기본적으로 매우 작은 소수값을 가지는 배터리 열화는 큰 가중치를 부여하여 충분한 영향력을 가지도록 하며, 유저 만족도는 지수함수를 사용하여 낮을수록 선형함수보다 더 큰 penalty를 가지도록 하였다. 사이클 기반 배터리 열화의 가중치를 시간 기반 배터리 열화보다 높게 설정하여, 사이클 기반 배터리 열화에 더욱 집중할 수 있도록 하였다. 자세한 가중치는 여러번의 실험을 통해 설정하였다.



[그림 11]. EV 사용자 시나리오 Reward Function의 구조

3.3 그리드 관리자 시나리오

그리드 관리자 입장에서는 수수료 방식의 비즈니스 모델을 설계 했기에 전기 기 시세차이에 따른 경제적 이익은 고려되지 않아도 된다. 하지만, 일정 수준의 사용자 만족도를 고려해야 참가자를 모을 수 있기에 사용자 만족도는 고려되어야 한다. 또한, 그리드의 안정적인 운영을 위하여 전력 설정값 추적 오차와 변압기의 과부하 방지를 고려해야한다. 여기에 변압기의 용량을 최대한으로 활용할 수 있도록 변압기 용량 활용에 따른 보상을 추가한다. 즉, 변압기를 최대한 활용하면서도 과부하는 방지하여 효율적인 충방전을 목표로 하는 Reward Function을 설계하였다. 변압기의 과부하는 안정적인 운영에서 제일 기피해야할 사항이기에 비교적 높은 가중치를 주었다. 과부하만 방지하면 변압기 사용을 기피할 수 있기에 가능한 변압기 활용할 수 있도록 사용한 변압기 양을 보상으로 추가하였다. 하지만, 너무 과한 변압기 사용은 기피하여야 하기에 Max Power에 가까울수록 큰 reward가 주어지도록 하였다. 또한, 유저 만족도를 추가하여 방전에 의한 변압기 사용을 기피하도록 하였다. 세부적인 가중치는 여러번의 실험을 통해 정하였다.



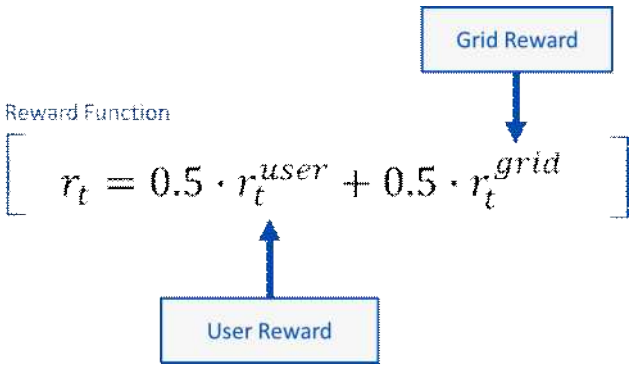
[그림 12]. 그리드 관리자 시나이로 Reward Function의 구조

3.4 통합 시나리오

최종적으로는 EV 사용자와 그리드 참여자 모두를 만족시킬 수 있는 시나리오를 목표로 한다. 이미 두 역할에 따른 Reward Function을 설계하였기에 통합 시나리오에서는 두 Reward Function을 조합하여 설계한다. Reward를 합치는 방식에는 여러가지가 있다.

3.4.1 동일 가중치 합

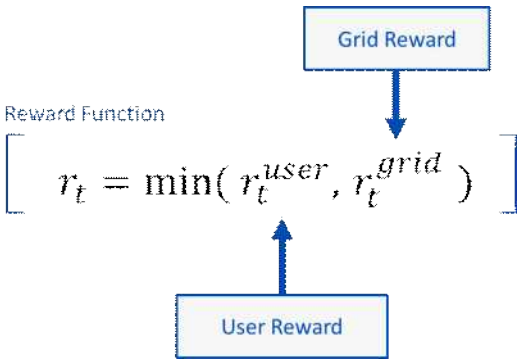
동일 가중치 합은 모든 Reward에 동일한 가중치를 두고 합산하여 설계하는 방식이다. 단순한 계산으로 두 Reward를 균형 있게 반영할 수 있다는 장점이 있다.



[그림 13]. ProfitMax TrPenalty UserIncentives 함수의 구조

3.4.2 Minimum

두 Reward의 값을 계산한 후 최솟값을 사용하는 방식이다. 둘중 더 낮은 성과를 보이는 항목을 우선적으로 개선할 수 있도록하며, 약한 부분에 집중하여 전체 성능을 고르게 향상시킨다.



[그림 14]. ProfitMax TrPenalty UserIncentives 함수의 구조

4. 연구 결과 분석 및 평가

4.1 기존의 Reward Function

[표 2]. V2G Profit Maximization 충전기 10개 (1000 에피소드)

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

[표 3]. V2G Profit Maximization 충전기 20개 (1000 에피소드)

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

가장 좋았던 리워드는 파란색, 그 다음으로 좋았던 리워드는 초록색으로 표시하였다. 대부분의 경우에서는 TRPO가 가장 좋은 성능을 보였으나, 충전기의 개수가 늘어나면 늘어날수록 즉, 문제가 복잡해질수록 TRPO 보단 Action Sampling 방식을 변형한 PPO들의 성능이 좋았다. Action Sampling 방식 중에서는 Tanh 함수를 사용하는 것이 대부분 좋은 성능을 보였으나, TRPO와 같이 문제가 복잡해질수록 연산이 단순한 Clamping이 오히려 좋은 결과를 보여주었다. 리워드 함수에서 Transformer OverLoad에 과도한 과중치가 부과되어 리워드가 높을수록, 충전에 소극적인 모습을 보여주었다.

4.2 EV 사용자 시나리오

[표 4]. EV 사용자 시나리오 충전기 10개 (1000 에피소드)

Env	Algorithm	Profits/costs ₵	ϵ^{usr} (%)	Energy Ch.(kWh)	Energy Disch.(kWh)	Tr. Ov.(kWh)	Degration Calender (e^{-3})	Degration Cycle (e^{-3})
Env1	PPO	2.75	70.64	72.35	71.08	42.19	0.0141	0.0215
	PPO_CLAMP	8.77	64.37	63.38	86.23	14.22	0.0133	0.0220
	PPO_MEAN	-9.17	83.42	89.00	38.72	57.72	0.0158	0.0179
	PPO_TANH	-7.61	82.01	90.28	45.66	40.46	0.0156	0.0189
	TRPO	-15.28	92.74	124.26	38.31	125.55	0.0175	0.0202
Env2	PPO	2.74	70.64	72.38	71.10	42.61	0.0141	0.0215
	PPO_CLAMP	8.37	64.76	63.19	84.55	14.42	0.0133	0.0217
	PPO_MEAN	-9.14	83.39	88.77	38.61	57.30	0.0158	0.0179
	PPO_TANH	-7.78	82.18	90.26	45.01	41.50	0.0156	0.0188
	TRPO	-15.73	92.51	122.69	37.35	125.36	0.0175	0.0201

EV 사용자 입장의 리워드를 사용했을 때, TRPO가 사용자 만족도를 높이는 데에는 좋았으나, 방전량이 많이 줄어든 것을 보아 V2G 활용면에서는 부적합했던 것으로 보인다. 오히려 80% 이상의 만족도를 유지하면서도 충방전을 적당히 조절한 PPO_TANH가 적합해보인다. 자연적인 배터리 수명 저하를 의미하는 Calender Degration의 값은 약간 늘었으나 충방전 횟수에 의한 수명 저하인 Cycle Degration은 줄어든 것을 보면 기존 리워드 수식과 비교했을 때, 사용자의 만족도를 높이면서도 사용자 배터리의 수명 저하를 방지할 수 있는 충방전을 한다는 목표는 이룬것으로 보인다.

4.3 그리드 관리자 시나리오

[표 5]. 그리드 관리자 시나리오 충전기 10개 (1000 에피소드)

Env	Algorithm	Profits/costs ₵	ϵ^{usr} (%)	Energy Ch.(kWh)	Energy Disch. (kWh)	Tr. Ov. (kWh)
Env1	PPO	10.49	61.62	59.43	90.47	23.30
	PPO_CLAMP	25.15	45.38	37.22	129.40	33.78
	PPO_MEAN	16.81	54.46	50.20	107.90	10.39
	PPO_TANH	15.84	56.46	56.53	108.04	10.62
	TRPO	22.91	50.84	86.01	158.58	6.96
Env2	PPO	10.45	61.60	59.55	90.67	36.72
	PPO_CLAMP	25.12	45.38	37.23	129.41	42.42
	PPO_MEAN	15.00	56.84	55.69	105.06	21.48
	PPO_TANH	12.26	60.43	61.87	99.08	12.61
	TRPO	16.30	57.43	93.89	142.58	13.59

그리드 관리자 입장의 리워드를 사용했을 때, EV 사용자 입장과는 반대로 충전보다 방전에 집중하는 모습을 모였다. 그럼에도 TRPO와 PPO_TANH의 경우에는 유저 만족도가 오히려 올랐다. 하지만, 유저 만족도가 오를수록 변압기의 과부하 정도도 함께 오르는 모습을 보였다. 이에, 기존보다 변압기의 과부하를 방지하려는 목표는 이루지 못하였으나, 늘어난 과부하에 비하여 유저 만족도가 많이 늘어난 모습을 보였다.

4.4 통합 시나리오

4.4.1 동일 가중치 합

[표 6]. 통합 시나리오 (동일 가중치 합) 충전기 10개 (1000 에피소드)

Env	Algorithm	Profits/costs ₵	ϵ^{usr} (%)	Energy Ch.(kWh)	Energy Disch. (kWh)	Tr. Ov. (kWh)	Degration Calender (e^{-3})	Degration Cycle (e^{-3})
Env1	PPO	10.29	62.20	62.66	91.62	14.51	0.0128	0.0233
	PPO_CLAMP	10.55	61.62	59.39	90.46	18.62	0.0128	0.0226
	PPO_MEAN	20.04	50.68	43.35	115.75	16.58	0.0114	0.0239
	PPO_TANH	20.20	51.12	47.01	117.68	9.67	0.0115	0.0247
	TRPO	27.68	45.14	58.70	152.36	6.36	0.0116	0.0325
Env2	PPO	10.52	61.71	60.65	91.39	37.46	0.0128	0.0231
	PPO_CLAMP	10.46	61.60	59.39	90.51	44.60	0.0129	0.0226
	PPO_MEAN	19.85	50.80	43.37	115.21	25.26	0.0114	0.0238
	PPO_TANH	20.98	50.11	46.60	121.03	19.95	0.0115	0.0254
	TRPO	25.19	47.32	58.97	144.50	19.02	0.0118	0.0310

동일 가중치 합을 사용한 통합은 오히려 기존의 reward 함수보다 좋지 않은

결과를 가져왔다. 유저 만족도는 비슷하나, 변압기의 과부하 정도가 오히려 늘었다. Cycle 기반의 배터리 부하는 줄어든 것으로 보아 불필요한 충방전을 피하려는 의도와 변압기에서 최대한 많은 power를 사용하려는 의도와 합쳐져 한번에 많은 power를 사용하게 된것으로 보인다. 이는 오히려 좋지 못한 결과를 가져왔다.

4.4.2 Minimum

[표 7]. 통합 시나리오 (Minimum) 충전기 10개 (1000 에피소드)

Env	Algorithm	Profits/costs ₵	ϵ^{usr} (%)	Energy Ch.(kWh)	Energy Disch. (kWh)	Tr. Ov. (kWh)	Degration Calender (e^{-3})	Degration Cycle (e^{-3})
Env1	PPO	9.52	63.07	65.89	91.78	22.08	0.0131	0.0238
	PPO_CLAMP	23.47	47.38	42.97	127.61	35.82	0.0108	0.0264
	PPO_MEAN	9.95	62.53	60.63	89.40	18.60	0.0130	0.0227
	PPO_TANH	17.42	53.48	48.14	109.35	8.74	0.0119	0.0238
	TRPO	18.41	53.58	55.89	118.13	8.29	0.0120	0.0266
Env2	PPO	9.59	63.14	67.48	92.95	29.22	0.0132	0.0242
	PPO_CLAMP	24.13	46.48	39.84	128.10	47.72	0.0108	0.0260
	PPO_MEAN	9.96	62.53	60.62	89.40	26.63	0.0130	0.0226
	PPO_TANH	17.42	53.48	48.14	109.35	16.04	0.0119	0.0238
	TRPO	19.77	53.31	75.37	137.48	15.64	0.0121	0.0312

Minimum을 사용한 통합의 경우에는 전체적으로 개선된 결과를 보였다. 변압기의 과부하 정도도 올랐으나, 사용자 만족도와 profit이 함께 올라 충분한 보상이 되었으며, Cycle 기반의 배터리 부하가 줄어들어 적은 횟수로 더 효율적인 충방전을 진행한 것으로 보인다. 변압기 과부하가 늘어난 점은 동일 가중치 합과 같으나 추가적인 개선이 없었던 동일 가중치 합과는 달리 다른 부분에서 많은 개선이 있었기에 충분히 효율적인 방식중 하나로 보인다.

5. 결론 및 향후 연구 방향

본 연구는 V2G 전력거래 이익 최대화를 위한 강화학습 기반 알고리즘 개발과 평가를 주제로 하였다. 이를 위해 EV2Gym이라는 시뮬레이션 환경을 활용하였다. EV2Gym은 V2G를 위한 오픈소스 시뮬레이터로, Gymnasium을 기반으로 구성되어 있어 강화학습 알고리즘의 평가가 용이하고, 이 실제 네덜란드의 오픈 데이터를 바탕으로 구성되어 있어, EV의 행동과 스펙, 전기 가격 등을 현실적으로 반영할 수 있었다.

연구에서는 여러 가지 SOTA(State-of-the-Art) 알고리즘을 EV2Gym 환경에서 실험하였다. TRPO, PPO, SAC 알고리즘을 사용하였으며, 각 알고리즘의 성능을 다양한 시나리오에서 비교 분석하였다. 특히 PPO 알고리즘의 경우, Action Sampling 방식에 따른 성능 차이를 조사하였다. 기존의 Gaussian 분포를 이용한 sampling 외에도 Clamping, Mean 값 사용, Tanh 함수 사용 등 다양한 방식을 시도하였다. 이를 통해 복잡한 환경에서는 Clamping 방식이, 그 외의 경우에는 Tanh 함수를 사용한 방식이 대체로 좋은 성능을 보인다는 것을 확인하였다.

또한, 더 정확한 목표를 이루기 위하여 Reward 함수를 개선하고자 하였다.

Reward 함수의 경우, EV 사용자와 그리드 관리자의 관점을 모두 고려한 새로운 함수들을 설계하였다. EV 사용자 시나리오에서는 경제적 이익, 충전 만족도, 배터리 열화 방지를 고려한 reward 함수를 사용하였으며, 그리드 관리자 시나리오에서는 변압기의 과부하 방지, 변압기 용량 활용, 전력 설정값 추적 오차 등을 고려하였다. 이 두 시나리오를 통합하는 방식으로 동일 가중치 합과 Minimum 방식을 시도하였으며, Minimum 방식이 더 효과적임을 확인하였다. Minimum 방식을 사용함으로서 어느정도의 과부하가 진행되더라도, 유저 만족도와 이익을 늘리고, 배터리 열화 방지를 이뤄냈다.

하지만, 본 연구에는 몇 가지 주요한 한계점이 존재한다. 우선, 화재 안정성에 대한 정밀한 모델링과 분석이 부족하다. V2G 시스템에서 배터리의 잦은 충방전은 화재 위험을 증가시킬 수 있는데, 이에 대한 정량적인 분석이 충분히 이루어지지 않았다. 또한, 사용자 이득에 대한 세밀한 계산이 미흡하다. 장기적인 관점에서 V2G 참여로 인한 배터리 수명 감소와 전기요금 절감 효과를 종합적으로 고려한 실질적인 경제적 이득 분석이 부족하다. 더불어, 현재 연구에서 사용된 데이터가 네덜란드로 한정되어 있어, 다양한 지역적 특성과 전력 시스템 구조를 반영하지 못했다는 한계가 있다. 이는 연구 결과의 일반화 가능성을 제한하는 요인이 될 수 있다.

이러한 한계점들을 극복하기 위해선 향후 추가적인 연구가 필요하다. 먼저, 국내 지역 특성과 유사한 그리드 구조에서의 시뮬레이션이 필요하다. 연구 결과를 국내에서도 적용하기 위해서는 국내의 전력 수요 패턴, 신재생 에너지 비율, 전기차 보급 현황, 전력 가격 변동 등을 반영한 데이터를 활용하여 시뮬레이션을 수행해야 한다. 이를 통해 국내 상황에 더욱 적합한 V2G 운영 전략을 도출할 수 있을 것이며, 연구 결과의 실제 적용 가능성을 높일 수 있을 것이다.

또한, 화재 안정성에 대한 보다 정밀한 모델링과 사용자 이득에 대한 세밀한 계산이 필요하다. 배터리의 열화와 충방전 패턴이 화재 위험에 미치는 영향을 정량적으로 분석하고, 이를 강화학습 알고리즘의 제약 조건으로 반영해야 한다. 사용자의 실질적인 경제적 이득을 장기적 관점에서 계산하는 방법을 개발하고, 이를 리워드 함수에 적절히 반영해야 한다. 이를 통해 V2G의 안전성을 높이고 사용자들의 적극적인 참여를 유도할 수 있는 전략을 수립할 수 있을 것이다.

향후 연구를 통해 V2G의 실용성과 효율성을 크게 향상시킨다면, 국내 실정에 맞는 최적화된 V2G 운영 전략을 개발할 수 있을것으로 기대된다. 이를 통해, 국내 전력 시스템의 안정성 향상과 신재생 에너지 활용도 증대에 기여할 수 있을 것이다. 또한, 더욱 정교해진 안전성 분석과 경제성 평가는 V2G 시스템의 대중적 수용성을 높이는 데 중요한 역할을 할 것이다.

6. 참고 문헌

[1] E. C. Kang. (2021, Sep. 12). "CFI 제주 2030, 탄소중립·에너지전환 선도" [Online]. Available: https://www.kharn.kr/news/article.html?no=17250 (downloaded 2024, Oct. 21) (in Korean)

[2] "무공해차 통합누리" [Online]. Available: https://ev.or.kr/nportal/main.do# (downloaded 2024, Oct. 21) (in Korean)

[3] Stavros Orfanoudakis, Cesar Diaz-Londono, Yunus E. Yılmaz, Peter Palensky, and Pedro P. Vergara, "EV2Gym: A Flexible V2G Simulator for EV Smart Charging Research and Benchmarking," arXiv preprint arXiv:2404.01849v1, 2024.

[4] John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel, "Trust Region Policy Optimization," arXiv preprint arXiv:1502.05477v5, 2015.

[5] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov, "Proximal Policy Optimization Algorithms," arXiv preprint arXiv:1707.06347v2, Aug. 2017.

[6] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine, "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor," Proceedings of the 35th International Conference on Machine Learning, PMLR 80:1861-1870, 2018.