マルチエージェント強化学習によるEV群VPPの充放電協調制御

林 弘辰 * (筑波大学)、 Xue Sihui (筑波大学)、 小平　大輔 (筑波大学)

Coordinated Charging and Discharging Control of an EV-Based VPP Using Multi-Agent Reinforcement Learning

Koshin Hayashi* (University of Tsukuba) , Sihui Xue (University of Tsukuba), Daisuke Kodaira (University of Tsukuba)

1　まえがき

バーチャルパワープラント（VPP）の需給調整市場参入では、リソースアグリゲーターが系統運用者から与えられる指令信号に対して、リアルタイムかつ低遅延でリソースを制御し追従する能力が求められる場合がある。一方、電気自動車（EV）を多数束ねた VPP を集中最適化で制御する場合、個別 EV 状態の収集・通信およびオンライン計算がボトルネックとなり、応答レイテンシ制約下では運用が困難になり得る。

一方、Multi-Agent Deep Deterministic Policy Gradient (MADDPG) に代表される MARL（Multi-Agent Reinforcement Learning）は CTDE（Centralized Training and Decentralized Execution）の性質を持ち、実行時に頻繁な全体通信や大規模最適化を要さずに協調方策を実装できる [1]。EV 群制御に MARL を適用した先行研究として、ユーザ側 QoS（充電費用や充電状態（State of Charge: SoC）等）に重点を置くもの [2] や、系統側の指標に重点を置くもの [3] があるが、双方を同一の MARL 枠組みで明示的に最適化・評価する設計は限定的である。

ここで、本稿の貢献は次の 2 点である。(1) 需給調整市場への参加を想定し、低遅延かつスケーラブルな階層型 MADDPG 制御を設計した。(2) 日本の需給調整市場を想定したシミュレーション環境にてケーススタディを行い、指令追従率および出発時 SoC 満足度の両方を評価指標として有効性を検討する。

2　提案手法

<2·1>　制御モデルの全体像　本研究では、VPP を構成する複数の充電ステーションをエージェントとし、各エージェントが配下の複数 EV の充放電出力を決定する階層型の構造を採る。図 1 のように、学習時は critic が actor を評価し actor が方策を更新するため中央集権的である。しかし実行時には、各ステーションは（i）自ステーションに接続中 EV の状態（SoC、残滞在時間、目標 SoC 等）と（ii）市場から与えられる指令信号のみを用いて行動し、中央集権的な critic はこの時利用されていないため、ステーション間の高頻度な個別通信を前提としない。

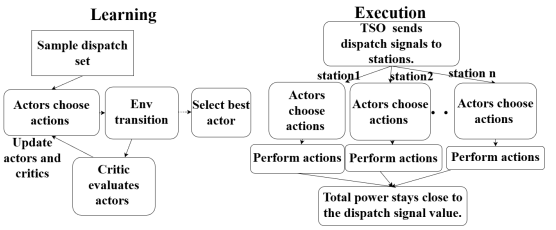

図 1 EV 充放電制御モデルの全体像
Fig. 1. Overview of the EV charging/discharging control model

<2·2>　階層型 MADDPG の実装　本モデルは MADDPG [1] を基礎とするが、Sheikh らの提唱する Decomposed Multi-Agent Deep Deterministic Policy Gradient (DE-MADDPG) [4] を参考に、局所目的達成度を評価するローカル critic と全体目的を評価するグローバル critic を併用する。通常の MADDPG と異なる点は、critic 入力を 2 系統に分解し、ローカル側は局所観測のみに基づく方策評価、グローバル側は全体情報に基づく方策評価を同時に学習する点である。ここでは $x \in \{\mathrm{L}, \mathrm{G}\}$ を用いてローカル／グローバルを統一的に表す。

ローカル critic は $\zeta^{\mathrm{L}} = (o_i, a_i)$ を入力とし、ステーション i の局所観測 o_i と行動 $a_i = \mu_i(o_i;\theta_i)$ のみから、局所目標に対する方策の価値 $Q^{\mathrm{L}}(\zeta^{\mathrm{L}})$ を推定する。一方、グローバル critic は $\zeta^{\mathrm{G}} = (s, a)$ を入力とし、全体状態 s と joint action $a = (a_1, \ldots, a_N)$ に基づいて、全体目標に対する価値 $Q^{\mathrm{G}}(\zeta^{\mathrm{G}})$ を推定する。報酬設計としては、r^{L} に出発時の目標 SoC 達成度（未達分が大きいほど低評価）を中心に与え、学習を安定化させるために充放電量に比例した整形項を併用する。r^{G} は指令追従を目的として、要求電力と全ステーション合計電力の偏差に基づき、許容帯域内では正の報酬、帯域外では偏差に応じたペナルティを与える。

actor 更新は DE–MADDPG の基本形に従い、ローカル目的とグローバル目的から得られる更新方向を重み w で凸結合する：

$$\theta_i \leftarrow \theta_i + \alpha \nabla_{\theta_i} \Big((1-w) J_i^{\mathrm{L}} + w J_i^{\mathrm{G}} \Big), \qquad 0 \le w \le 1. \quad (1)$$

ここで J_i^{L} と J_i^{G} は、それぞれローカル critic とグローバル critic に基づく（標準的な）方策勾配の目的関数であり、replay buffer $\mathscr{D}$ と target network を用いて学習する。

3　ケーススタディ

本研究は、日本の需給調整市場における二次調整力相当の参加を想定し、指令信号が 5 分ごとに到来する離散時間環境（1 ステップ=5 分）で評価する。1 日を 288 ステップとして 1 エピソードを構成し、VPP は 5 ステーション、各 10 台の普通充電器（合計 50 台）からなるものとする。EV 到着時刻は ElaadNL–職場分布 [5]、到着時 SoC は DESL-EPFL の lv3 充電器データセット [6]、目標 SoC・滞在時間は ACN（JPL）[7]（電池容量情報がないため EV 容量を 100kWh と仮定）、ディスパッチ信号は PJM RTO Regulation Signal（2024 年 10 月）[8] を用いた。

本ケーススタディでは、性能指標として（i）Target SoC satisfaction rate、（ii）Dispatch tracking rate を用いる。ここで (i) は、各 EV の出発時刻において SoC が目標値以上であった出発事象の割合として定義する。(ii) は、全ステップの内 VPP 集約出力が指令値の許容帯（日本の市場仕様に従い入札幅 × 0.1 を許容幅とする）内であったステップの割合として算出する。

提案手法は 600 エピソードで学習し，バッチサイズ 1024，割引率 $\gamma = 0.975$ とした。actor/critic の隠れ層サイズはいずれも 256 とした。actor 更新では local/global critic の評価を重み $w = 0.3$ で凸結合して用いた。探索は OU ノイズと ε-greedy を併用し，強度は学習の進行に合わせて減衰させた。

4　結果・考察

学習の進行に伴う性能指標の推移を図 2 に示す。学習の初期段階では両目標が拮抗し、収束近傍では安定して 80 %以上の指令追従が観察された。収束付近 (660) の方策で評価した結果、Target SoC satisfaction rate は 67.1 %、Dispatch tracking rate は 88.5 % であった。これは密で低分散な報酬を与える global の勾配が支配しやすく、疎で遅延のある local は信用割当が難しく学習信号が弱くなるため、結果として global 側に偏ると考えられる。同条件の集中最適化では前述 2 指標とも 100 %を達成することを考えると、本手法は一定の成果を示したが、さらなる改善の余地があると考えられる。

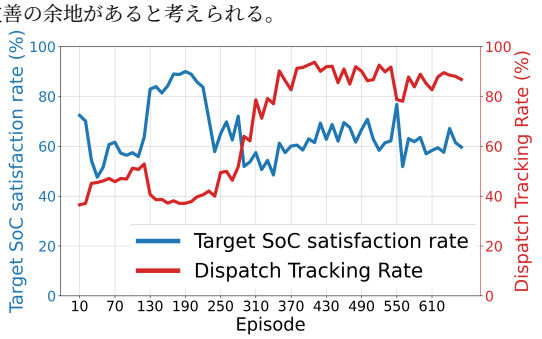

図 2 学習に伴う性能指標の推移
Fig. 2. Performance metrics over training episodes

代表的なテストエピソードにおけるステーション出力を図 3 に示す。ステーション単位の分散実行でありながら、紫の破線により表現される集約出力が、黒の実線で示される系統運用者から与えられる指令信号値に追従する協調挙動が得られていることが確認できる。

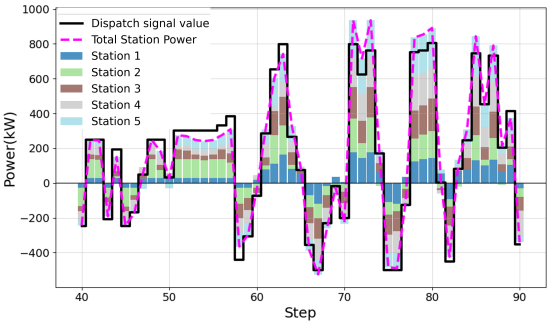

図 3 代表テストにおけるステーション電力
Fig. 3. Station power in a representative test episode

5　まとめ

本稿では、需給調整市場における EV–VPP の低遅延制御を目的として、階層型 MADDPG に基づく制御枠組みを提案した。ユーザ需要（出発時 SoC 確保）と指令追従の二つの目標を、それぞれの評価に基づく方策更新により扱うことで、分散実行下で両目的を同時に扱う構成を示した。シミュレーションでは指令追従が集中最適化に匹敵する一方、ユーザ需要満足率の向上には課題が残った。今後は、報酬設計および学習率・割引率・目的重み係数等のハイパーパラメータ調整により、両目的の達成水準の向上を図るとともに、ユーザ満足率を 100 %に近づけるため、ルールベースの安全層を併用したハイブリッド制御を検討する。具体的には、残時間と現在 SoC から目標 SoC 達成が物理的に困難と判定される場合に限り、actor 出力を上書きして最大充電を指令し、出発直前の未充足を回避する。さらに、EV・充電器の異質性、予測誤差、通信遅延等を含む実環境に近い条件での追加評価と実証が必要である。

文献

[1] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pages 6379–6390, 2017.

[2] Keonwoo Park and Ilkyeong Moon. Multi-agent deep reinforcement learning approach for EV charging scheduling in a smart grid. *Applied Energy*, 328:120111, December 2022.

[3] Peixiao Fan, Jia Hu, Song Ke, Yuxin Wen, Shaobo Yang, and Jun Yang. A frequency–pressure cooperative control strategy of multi-microgrid with an electric–gas system based on MADDPG. *Sustainability*, 14(14):8886, July 2022.

[4] Hassam Ullah Sheikh and Ladislau Bölöni. Multi-agent reinforcement learning for problems with combined individual and team reward, 2020. Preprint (DE-MADDPG).

[5] ElaadNL. Dataset 2: Distribution of arrival times on weekdays (workplace). Technical report, ElaadNL, The Netherlands, 2018. Available via ElaadNL open data platform.

[6] EPFL Distributed Electrical Systems Laboratory (DESL). Level-3 EV charging dataset. Dataset repository, June 2025.

[7] California Institute of Technology. ACN-Data: EV charging dataset. Dataset portal, 2019.

[8] PJM Interconnection, L.L.C. RTO regulation signal data (sample). PJM ancillary services data (sample file), October 2024.