

# 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)

Coordinated Charging and Discharging Control of an EV-Based VPP Using Multi-Agent Reinforcement Learning  
Koshin Hayashi (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、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せ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 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、各エージェントが当該ステーション配下の複数 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。

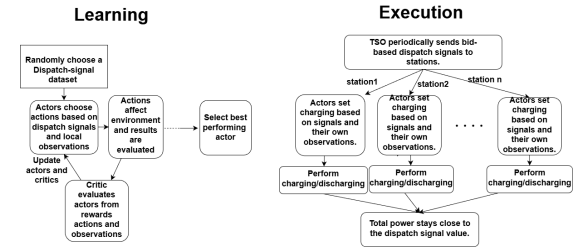


図 1 EV 充放電制御モデルの全体像

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、Zadaianchuk らの提唱する DE-MADDPG (4) を参考に、局所目的 (出発時 SoC) と全体協調 (指令追従) を分離して扱うため、ローカル critic と共有グローバル critic を併用する。ここでは通常の MADDPG と異なる数式を紹介する。なお以下では  $x \in \{L, G\}$  を用いてローカル/グローバルを統一的に表す。

$$J_i^x(\theta_i) = \mathbb{E}_{(\cdot) \sim \mathcal{D}} [Q^x(\zeta^x)], \quad (1)$$

$$y^x = r^x + \gamma Q^{x-}(\zeta^{x'}), \quad (2)$$

$$\theta_i \leftarrow \theta_i + \alpha \nabla_{\theta_i} \left( (1-w) J_i^L + w J_i^G \right), \quad 0 \leq w \leq 1. \quad (3)$$

式 (1) は actor の目的関数、式 (2) は critic 学習に用いる TD 目標、式 (3) はローカル目的とグローバル目的の更新を重み  $w$  で凸結合して actor を更新することを表す。ここで  $\zeta^L = (o_i, a_i)$ 、 $\zeta^G = (s, a)$  とし、 $o_i$  はステーション  $i$  の局所観測、 $s$  は全体状態、 $a = (a_1, \dots, a_N)$  は joint action、 $a_i = \mu_i(o_i; \theta_i)$  である。 $r^L$  は出発時 SoC に関する報酬、 $r^G$  は指令追従に関する報酬である。 $\mathcal{D}$  は replay buffer、 $Q^{x-}$  は target critic であり、 $\zeta^{x'}$  は次時刻の入力 (target actor により得た次時刻行動を含む) を表す。

## 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 の到着分布および充

電履歴等のデータには公開データセットを用いる (5–8)。

本ケーススタディでは、性能指標として (i) Target SoC satisfaction rate、(ii) charging dispatch tracking rate、(iii) discharging dispatch tracking rate を用いる。ここで、Target SoC satisfaction rate は、各 EV の出発時刻において SoC がユーザーの設定した目標値以上であった出発事象の割合として定義する。また tracking rate は、全 288 ステップの内、VPP 集約出力が指令値の許容帯 (本研究では日本の需給調整市場仕様に従い入札幅  $\times 0.1$  を許容幅とする) 内に収まったステップの割合として算出する。

提案手法は全 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。ターゲットネットワークは Polyak 平均で更新し、係数は actor/local critic で  $\tau = 0.02$ , global critic で  $\tau_G = 0.02$  とした。学習率は actor  $2 \times 10^{-5}$ , local critic  $1 \times 10^{-4}$ , global critic  $5 \times 10^{-5}$  とし、actor および local/global 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 ( $\theta = 0.15, \sigma = 0.5$ , gain = 3.0) と  $\epsilon$ -greedy を併用し、OU スケールはエピソード 1–500 で  $1.0 \rightarrow 0.2$ ,  $\epsilon$  はエピソード 1–100 で  $1.0 \rightarrow 0.05$  へ線形減衰させた。

#### 4 結果

学習の進行に伴う性能指標の推移を図 2 に示す。学習の初期段階から指令追従性能が改善し、収束近傍では安定した追従が観察される。収束付近の方策で評価した結果、Target SoC satisfaction rate は 63.5 %、charging dispatch tracking rate は 99.5 %、discharging dispatch tracking rate は 100.0 % であった。同条件の集中最適化では前述 3 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。

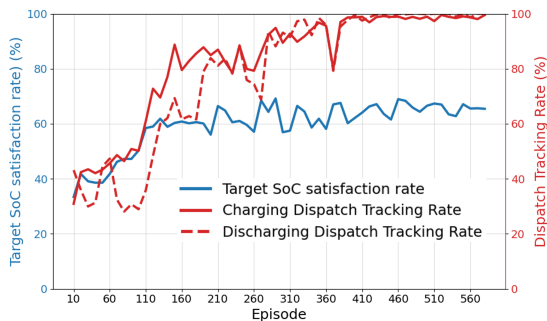


図 2 学習に伴う性能指標の推移

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指令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。

遅延観点では、本方式は実行時に各ステーションで actor の推論 (forward 計算) のみを行うため、通信遅延はかからない。また計算遅延も RTX-4080 を利用した実測値で数十 ms 程度に

抑えられ、低遅延制御が可能であると考えられる。

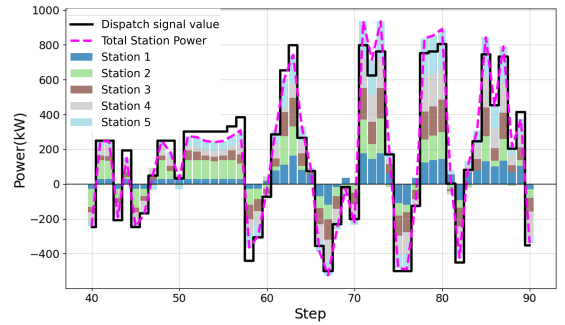


図 3 代表テストにおけるステーション電力

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、階層型 MADDPG 制御を示した。ローカル critic (ユーザー制約) と共有グローバル critic (指令追従) を併用し、両者の評価に基づく方策更新を重み付きに統合することで、分散実行のまま二目的を同時に扱う枠組みを与えた。結果として、集中最適化と同等性能の指令追従性能を達成したが、EV ユーザー満足度には改善の余地が残された。今後は報酬設計や学習手法の改良を通じて、両目的のさらなる両立を目指すと同時に、EV ユーザー満足度を 100 % に近づけるためにルールベース制御を組み合わせたハイブリッド制御の検討を進めるべきだ。また、更に実環境に近いシナリオでの評価や実証実験の展開も行う必要がある。

#### 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. 3 2020.
- (2) Keonwoo Park and Ilkyeong Moon. Multi-agent deep reinforcement learning approach for ev charging scheduling in a smart grid. *Applied Energy*, 328:120111, 12 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*, 14:8886, 7 2022.
- (4) Hassam Ullah Sheikh and Ladislau Bölöni. Multi-agent reinforcement learning for problems with combined individual and team reward. 3 2020.
- (5) California Institute of Technology. Acn-data: Ev charging dataset (jpl site). Technical report, California Institute of Technology, 2019.
- (6) EPFL Distributed Electrical Systems Laboratory (DESL). Level-3 ev charging dataset. Technical report, Distributed Electrical Systems Laboratory (DESL), EPFL, 6 2025.
- (7) L.L.C. PJM Interconnection. Rto regulation signal data. Technical report, PJM Interconnection, L.L.C., 10 2024.
- (8) ElaadNL. Dataset 2: Distribution of arrival times on weekdays (workplace). Technical report, ElaadNL.