Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Multi-agent evaluation for energy management by practically scaling α-rank*

Yiyun SUN†[1,2], Senlin ZHANG†[1,2], Meiqin LIU†‡[3,2,1], Ronghao ZHENG[1,2], Shanling DONG[1,2], Xuguang LAN[3]

[1]*National Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China*

[2]*College of Electrical Engineering, Zhejiang University, Hangzhou 310027, China*

[3]*National Key Laboratory of Human–Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University, Xi'an 710049, China*

†E-mail: 12110066@zju.edu.cn; slzhang@zju.edu.cn; liumeiqin@zju.edu.cn

Received June 24, 2023; Revision accepted Dec. 21, 2023; Crosschecked May 23, 2024

Abstract: Currently, decarbonization has become an emerging trend in the power system arena. However, the increasing number of photovoltaic units distributed into a distribution network may result in voltage issues, providing challenges for voltage regulation across a large-scale power grid network. Reinforcement learning based intelligent control of smart inverters and other smart building energy management (EM) systems can be leveraged to alleviate these issues. To achieve the best EM strategy for building microgrids in a power system, this paper presents two large-scale multi-agent strategy evaluation methods to preserve building occupants' comfort while pursuing system-level objectives. The EM problem is formulated as a general-sum game to optimize the benefits at both the system and building levels. The α-rank algorithm can solve the general-sum game and guarantee the ranking theoretically, but it is limited by the interaction complexity and hardly applies to the practical power system. A new evaluation algorithm (TcEval) is proposed by practically scaling the α-rank algorithm through a tensor complement to reduce the interaction complexity. Then, considering the noise prevalent in practice, a noise processing model with domain knowledge is built to calculate the strategy payoffs, and thus the TcEval-AS algorithm is proposed when noise exists. Both evaluation algorithms developed in this paper greatly reduce the interaction complexity compared with existing approaches, including ResponseGraphUCB (RG-UCB) and αInformationGain (α-IG). Finally, the effectiveness of the proposed algorithms is verified in the EM case with realistic data.

Key words: Energy management; Multi-agent deep reinforcement learning; Strategy evaluation; Power grid system

https://doi.org/10.1631/FITEE.2300438 **CLC number:** TP181

1 Introduction

Currently, decarbonization has become an emerging trend in the power system arena. However, integrating intermittent renewable energy, which has limited controllability and partial predictability, into the power system poses a significant challenge (Pigott et al., 2022). To address this challenge, the urgent development and deployment of flexible technologies, such as storage and demand-side response, are necessary to effectively balance the variability of intermittent renewable energy resources. Smart building microgrids are constructed based on this concept as the promising option for optimal energy management

‡ Corresponding author

* Project supported by the National Key R&D Program of China (No. 2021ZD0112700), the Zhejiang Provincial Natural Science Foundation of China (No. LZ22F030006), and the Fundamental Research Funds for the Central Universities, China (No. xtr072022001)

ORCID: Yiyun SUN, https://orcid.org/0009-0007-5593-8767; Meiqin LIU, https://orcid.org/0000-0003-0693-6574

(EM) from sustainable energy sources (Zhang et al., 2023). Building electricity consumption accounts for 70% of total electricity consumption (Su and Wang, 2012). With the notable growth in energy storage capacity at the distribution level, enabled by batteries and electric vehicles, implementing suitable control and coordination strategies can synchronize peak demand with peak renewable energy generation (Vincent et al., 2020; Cai et al., 2023). Building microgrid energy management is a comprehensive arrangement of diverse technologies involving conversion, distribution, and storage. Its objective is to optimize the synergistic effect among different energy carriers within a multi-energy system. By doing so, it aims to fulfill the energy requirements of building users (building level) while maintaining voltage and power stability in the overall power distribution network (system level) (Zhao et al., 2022).

Multi-agent reinforcement learning (MARL) methods have shown great potential in power system control and EM tasks (Claessens et al., 2018; Xu et al., 2020; Tong et al., 2023). MARL sets each building microgrid as a control agent, which learns control strategies by interacting with the environment. However, energy regulation in the large-scale power grid system is considered a general-sum game. As the number of agents increases, most MARL algorithms have difficulty in guaranteeing the efficiency of each agent's strategy (Dong et al., 2022). As a result, the evaluation of joint strategies becomes complicated because each smart building microgrid controlled by an RL agent is also influenced by the strategies of other RL agents. The α-rank method (Omidshafiei et al., 2019) enables investigation of the strengths and weaknesses of these joint strategies using game-theoretic evaluation techniques (wherein joint strategies are the combinations of strategies chosen by each agent). In this paper, we investigate the problem of applying α-rank to the evaluation of the EM strategy. An excellent joint strategy is obtained from ranking, which improves the stability and reliability of the overall power system.

However, there is a practical problem with using the α-rank for the EM problem in the multi-agent system. Each agent requires numerous interactions to estimate a credible payoff. For large-scale multi-agent systems, the cost of agents' interactions is high (Silver et al., 2016; Tuyls et al., 2018). It is difficult to meet the practical needs of large-scale strategy evaluation for an EM task. To reduce the number of interactions and the computational cost, the ResponseGraphUCB (RG-UCB) algorithm sets confidence intervals for the strategy payoffs (Rowland et al., 2019). αInformationGain (α-IG) determines the payoff model by maximizing the information gain from the α-rank belief (Rashid et al., 2021). Both methods require a comparison of all joint strategies, which is still expensive. This inspires us to improve the sampling efficiency during strategy evaluation.

In this paper, TcEval is proposed to alleviate the problem of high sampling complexity during evaluation, reducing the interaction complexity from $O(n^k)$ to $O(nr^{k-1}\text{poly}\ln n)$ (where k is the number of agents, n is the number of strategies owned by each agent, and r is the rank of the payoff tensor). Considering the unavoidable noise in practical power systems, TcEval-AS is proposed to achieve better sampling efficiency and evaluation performance compared with existing approaches, including RG-UCB and α-IG. The feasibility and superiority of the two proposed algorithms are verified in a simulated power grid system.

The contributions of this paper are as follows:

1. We innovatively apply the α-rank-based evaluation method to obtain the ranking of all reconstituted joint strategies at both the building and system levels, and thereafter, we obtain the optimal energy regulation strategy for each agent.

2. The evaluation method based on practically scaling the α-rank through low-rank tensor completion is proposed. Unlike existing methods (Rowland et al., 2019; Rashid et al., 2021) that require all joint strategy payoffs, we propose an approach to accurately estimate complete payoffs from a subset of key observed payoffs.

3. Considering the treatment of observed payoffs with noise in real scenarios, a payoff prediction model is presented in TcEval-AS to accurately predict the real payoff in an EM task with a few interactions.

2 System modeling and problem description

In this study, EnergyGrid is developed as an adaptation of the GridLearn environment (Pigott et al., 2022). As shown in Fig. 1, compared with the GridLearn environment, the EnergyGrid environment achieves the objectives of both demand

response and grid stability instead of focusing on only grid voltage stability. The EnergyGrid environment provides an energy model of multiple building microgrids in a mixed-use zone, connected by a distribution grid modeled with alternating current (AC) power flows. The details are described as follows.

2.1 System modeling

In this study, we use the IEEE-13-bus network model and select k building microgrids to connect to it. We develop building microgrid models based on the data from the prototype buildings provided by the U.S. Department of Energy (https://www.energycodes.gov/prototype-building-models). Each agent (building) controls multiple components; the details of these components are as follows:

1. The heating, ventilation, and air conditioning (HVAC) model is a simplified model of the electrical system, which consists of an air conditioning and ventilation device.

2. Photovoltaic (PV) and energy storage (ES) models are two components controlled by the agent, which assist in modifying the power consumption of a building.

In the EnergyGrid environment, agents are encouraged to achieve their objectives while maintaining voltage stability of the system. For ease of reference, the common symbols are summarized in Table 1. The specific observation space, action space, and reward function of each internal device of the agent are shown in Table 2.

2.2 Problem description

Stochastic games have long been used in MARL to model interactions among agents. This study focuses on analyzing interactions at the meta level. A meta-game in MARL concentrates on meta-strategies with different game styles, other than atomic actions (Omidshafiei et al., 2019; Muller et al., 2020). Specifically, a meta-strategy is a rule of choosing an appropriate action any time based on the agent's information, which is initiated by different strategy networks.

We set each building microgrid model as an agent. Each agent $i \in K$ ($K = \{1, 2, ..., k\}$) has a finite set of pure control strategies denoted by

Table 1 Common symbols and their interpretations

Symbol	Interpretation
t	Time period, $t \in \{1, 2, ..., T\}$
i	Smart building microgrid unit, $i \in \{1, 2, ..., k\}$
n	Number of strategies owned by each agent
S^i	Pure control strategy set of unit i
$\boldsymbol{s}$	Joint strategy, where $\boldsymbol{s} \in S^1 \times S^2 \times ... \times S^k$
$\boldsymbol{M}$	Complete observed payoff tensor
r^{comfort}	Indoor comfort measurement value
P_{comsume}	Energy consumption value
V_{d}	Voltage deviation of the connected bus
t_{set}	The temperature setpoint for air conditioning
δ	A factor for PV real power injection
S_{PV}	Real power injection for the PV system
P_{E}	A factor of fully charging and discharging for the ES system
S_{E}	State of charge for the ES system
r^{HVAC}	Reward value of HVAC (building level)
r^{sys}	Reward value of grid systems (system level)

PV: photovoltaic; ES: energy storage; HVAC: heating, ventilation, and air conditioning

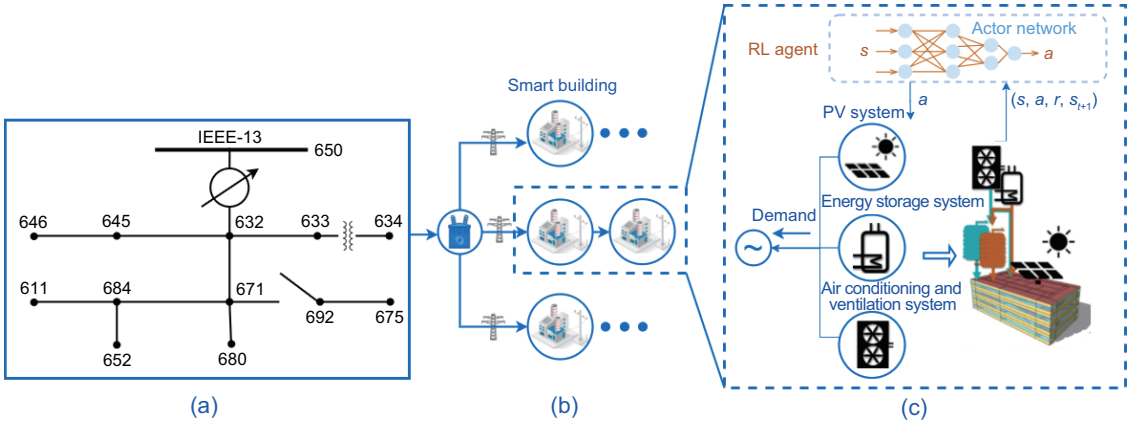

Fig. 1 Structure of EnergyGrid: (a) distributed generation network; (b) smart building microgrids; (c) details of the smart building microgrid with RL agents (RL: reinforcement learning; PV: photovoltaic)

Table 2 Parameters of each agent in EnergyGrid

Component device	Parameter		
	Observation space	Action space	Reward function
HVAC systems	r^{comfort}, P_{consume}	t_{set}	$r^{\text{HVAC}} = \alpha \cdot P_{\text{consume}} + \beta \cdot r^{\text{comfort}}$
Solar energy systems	S_{PV}	δ	0
Energy storage systems	S_{E}	P_{E}	0
Others	V_{d}		$r^{\text{sys}} = \gamma \cdot V_{\text{d}}$

HVAC: heating, ventilation, and air conditioning; PV: photovoltaic

$S^i = \{S_1^i, S_2^i, ..., S_n^i\}$, wherein each strategy S_z^i ($z = 1, 2, \ldots, n$) is represented by the z^{th} actor network $\mu_\phi^{i,z}$, which has been trained using the multi-agent deep deterministic policy gradient (MADDPG) algorithm (Lowe et al., 2017).

The purpose of the EM task is to achieve the rational strategy, further preserving building occupants' comfort while pursuing system-level objectives. This EM problem can be formulated as a general-sum game, which can be solved by using the idea of α-rank (Omidshafiei et al., 2019).

Denote the space of control strategy profiles as $S = \prod_{i=1}^{k} S^i$. Each agent selects strategies from S to form a joint strategy $\boldsymbol{s}_\sigma = \left(s_\sigma^1, s_\sigma^2, ..., s_\sigma^k\right) \in S$. Interaction of each joint strategy $\boldsymbol{s}_\sigma$ in S with the EnergyGrid yields the observation payoff $\boldsymbol{M}(\boldsymbol{s}_\sigma) = \left(M^1(\boldsymbol{s}_\sigma), M^2(\boldsymbol{s}_\sigma), ..., M^k(\boldsymbol{s}_\sigma)\right)$. Besides, the payoff tensor $\boldsymbol{M} = \left(\boldsymbol{M}(\boldsymbol{s}_1), \boldsymbol{M}(\boldsymbol{s}_2), ..., \boldsymbol{M}(\boldsymbol{s}_{|S|})\right)$ can be constructed based on the payoffs of all joint strategies.

Then, we calculate the state transfer matrix $\boldsymbol{C}$. We define $\boldsymbol{s}_\sigma, \boldsymbol{s}_\tau \in S$. Let $\boldsymbol{s}_\sigma$ and $\boldsymbol{s}_\tau$ differ only in one strategy of the single agent i, which uses strategy $\tau \in S^i$ instead of σ. Therefore, the transfer probability from $\boldsymbol{s}_\sigma$ to $\boldsymbol{s}_\tau$ of agent i can be expressed as follows:

$$C_{\boldsymbol{s}_\sigma, \boldsymbol{s}_\tau} = \begin{cases} \eta \dfrac{1 - \exp\left(\alpha\left(M^i(\boldsymbol{s}_\sigma) - M^i(\boldsymbol{s}_\tau)\right)\right)}{1 - \exp\left(-\xi\alpha\left(M^i(\boldsymbol{s}_\sigma) - M^i(\boldsymbol{s}_\tau)\right)\right)}, \\ \quad \text{if } M^i(\boldsymbol{s}_\sigma) \neq M^i(\boldsymbol{s}_\tau), \\ \dfrac{\eta}{\xi}, \quad \text{otherwise}, \end{cases}$$

where $M^i(\boldsymbol{s}_\sigma)$ is the expected payoff of agent i when using the joint strategy $\boldsymbol{s}_\sigma$, η is the reciprocal of the total number of valid profile transitions from a given strategy profile, α is the ranking strength, and ξ is a hyperparameter indicating the population size.

Then, the steady-state distribution π is calculated from the state transfer matrix $\boldsymbol{C}$, and the values of the transferred probability in π can be derived to obtain the ranks R of all joint strategies. Furthermore, top-rank joint strategy $\boldsymbol{s}^*$ is achieved from R.

Accordingly, the state transfer matrix $\boldsymbol{C}$ can be calculated from the complete observed payoff tensor $\boldsymbol{M}$. This means that $O(n^k)$ payoffs can be obtained to compute the top-rank joint strategy profile. First, each joint strategy requires numerous interactions to obtain the expected payoff in noisy cases. Second, exhaustive evaluation of all joint strategy profiles further increases computation burden. Therefore, this study attempts to achieve credible $\boldsymbol{C}$ from a small number of interactions.

3 Strategy evaluation of power grid by practically scaling α-rank with tensor completion

In this section, a large-scale strategy evaluation algorithm (TcEval) is proposed, and its specific process is shown in Fig. 2.

3.1 Estimating α-rank with tensor completion

In the grid system, the correlation among joint strategies in S leads to similar payoffs obtained. Therefore, strategy evaluation can be achieved by using low-rank attributes and incomplete payoff tensor complements (Czarnecki et al., 2020).

Suppose that we can obtain the precise value of payoff. TcEval (Algorithm 1) is proposed to precisely evaluate all joint strategies through a small portion of joint strategies. To be specific, TcEval uses an active query method to construct the sampling operator Ω, wherein the joint strategies with key payoffs are selected in an active manner. For each chosen joint strategy $\boldsymbol{s} = (s^1, s^2, ..., s^k) \in \Omega$, we can query its real value of the payoff and construct the incomplete payoff tensor $\boldsymbol{M}^\Omega$. Then, $\boldsymbol{M}^\Omega$ can be recovered through tensor completion, and α-rank

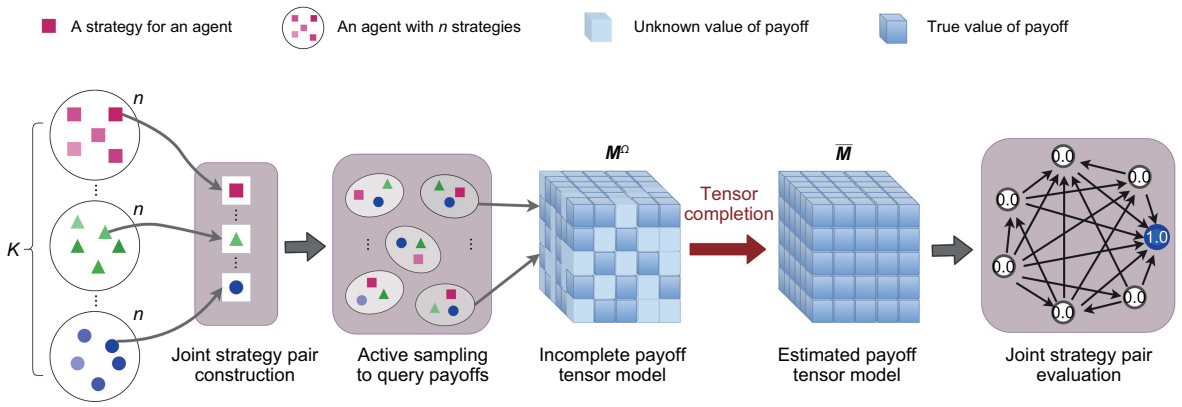

Fig. 2 Overview of TcEval. TcEval constructs n^k joint strategies for k agents (each with n strategies), estimates the complete payoff tensor $\overline{M}$ based on PS-HOSVD, and derives the steady-state strategy distribution $\bar{\pi}$ from $\overline{M}$

evaluation is performed on the recovered payoff tensor $\overline{M}$. The steady-state distribution of all joint strategies $\overline{\pi}$ can be obtained by calculating $\overline{M}$. In this study, the PS-HOSVD algorithm is chosen as the tensor complement module in TcEval (Xia et al., 2021) to recover the payoff tensor.

3.2 Active sampling to query key payoffs

To achieve the effectiveness of TcEval, we expect to obtain a distribution that is closest to the real strategy distribution.

In the process of joint strategy transfer, the low-value "useless" joint strategy will be transferred to the high-value "useful" joint strategy, which means that $\overline{C}$ is determined mainly by these high-value joint strategies. Therefore, to obtain more "useful" joint strategies for the tensor complement, we propose an active query algorithm (Algorithm 2) based on Gaussian process regression (GPR) (Williams and Rasmussen, 1995). Its purpose is to obtain the joint strategies that have a greater impact on the final distribution.

Algorithm 1 TcEval: multi-agent evaluation under estimating α-rank with tensor completion

Input: active sampling operator, Ω; chosen rank, r^*; joint strategy, $\boldsymbol{s} \in S$.

Output: steady-state strategy distribution, $\bar{\pi}$; the ranks of all joint strategies, $\overline{R}$.

Construct n^k joint strategies for k agents, each with n strategies
Use the active sampling method to select s_{num} pairs with the maximum information gain by using Ω
Estimate the complete payoff tensor $\overline{M}$ based on the PS-HOSVD tensor completion algorithm with rank r^*
Calculate the Markov chain $\overline{C}$ from $\overline{M}$
Obtain steady-state strategy distribution $\bar{\pi}$ by using $\overline{C}$
Calculate $\overline{R}$ of all joint strategies
return $\bar{\pi}$, $\overline{R}$

Algorithm 2 Active sampling

Input: initial weight coefficient, γ; iterative weight coefficient, β; size of the active sampling operator, Size(Ω).

Output: active sampling operator, Ω.

Initialize the size of the active sampling operator Size(Ω)
Initialize the forecasting dataset $X_{\text{test}} = \{s_1, s_2, ..., s_{n^k}\}$
Calculate the number of initial training set samples $Z = \gamma \cdot \text{Size}(\Omega)$
Perform uniform sampling to obtain the training dataset X_{train}
Query payoffs to obtain the training label set Y_{train}
Move X_{train} out of X_{test}
while len(X_{train}) < Size(Ω) **do**
    Perform Gaussian process regression (GPR) based on Y_{train} and X_{train}
    Rank the prediction results and select the strategy pairs X_{iter} corresponding to the top $\beta \cdot (\text{Size}(\Omega) - \text{len}(X_{\text{train}}))$ predicted payoffs
    Update X_{train} by adding X_{iter}
    Query payoffs from X_{iter} and update Y_{train}
    Move X_{iter} out of X_{test}
end while
Copy joint strategies from X_{train} to Ω
return Ω

3.3 Complexity analysis of TcEval

TcEval practically scales α-rank evaluation through tensor complement algorithms. The payoff tensor completion infers an unknown value from the sampled data, predicated on the low-rank assumption. Based on the proof of Xia et al. (2021), accurate α-rank distribution can be obtained with $O(nr^{k-1}\text{poly}\ln n)$ sampled payoffs, where $\text{poly}\ln n$ is the certain polynomial of the logarithmic function.

Although most agents' interaction payoff tensors are of low rank, there are still payoff tensors of high or even full rank. Du et al. (2021) have proven that as long as the high-rank tensor is not too distant from a low-rank tensor, payoff tensors based on estimated low-rank recovery maintain the trend of the real payoff tensor, and TcEval can still work.

4 Payoff predictive model for a noisy case

In this section, we construct the payoff predictive model for the noisy case and propose TcEval-AS, which is shown in Fig. 3.

TcEval-AS uses focal sampling for estimating $\hat{M}_{\omega_i}$ and traverses the joint strategies in Ω to obtain $\hat{M}_{\omega_i}$. Then the PS-HOSVD tensor complement algorithm is applied to obtain the recovered matrix $\overline{\hat{\boldsymbol{M}}}$. Finally, the ranking of all joint strategies is obtained by performing α-rank analysis on $\overline{\hat{\boldsymbol{M}}}$.

4.1 Payoff predictive model based on focal sampling

In most EM environments, agents using joint strategy ω_i ($\forall\omega_i \in \Omega$) need to interact a sufficient number of times to obtain the expected payoff $\hat{M}_{\omega_i}$.

To solve the problem that we can access only noisy payoff, the payoff probabilistic predictor can be represented based on the mathematical prior model as follows:

$$p(\boldsymbol{x}|L,\boldsymbol{\psi}^*)=\frac{p(\boldsymbol{x}|\boldsymbol{\psi}^*)P(L|\boldsymbol{x})}{P(L|\boldsymbol{\psi}^*)}, \tag{1}$$

where $p(y|\boldsymbol{x})$ is expressed as the probabilistic predictor, the payoff observation is expressed as $\boldsymbol{x}$, and the observation probability is expressed as y. In this predictor model, the maximum probability payoff can be calculated as the expected payoff. For searching the maximum probability, we define $L=\{y|y\geq\lambda,\ \lambda<y_{\max},\ \lambda\to y_{\max}\}$. Prior generative model $p(\boldsymbol{x}|\boldsymbol{\psi}^*)$ is established to represent the noisy (empirical) payoff distribution $v(s)$, and $\boldsymbol{\psi}^*$ is expressed as the fitted parameters through training.

As $P(L|\boldsymbol{\psi}^*)$ is difficult to find, a search model $q(\boldsymbol{x}|\boldsymbol{\vartheta})$ is established to find the optimal parameter $\boldsymbol{\vartheta}^*$:

$$\begin{aligned}\boldsymbol{\vartheta}^*&=\arg\min_{\boldsymbol{\vartheta}} D_{\mathrm{KL}}\big(p(\boldsymbol{x}|L,\boldsymbol{\psi}^*)||q(\boldsymbol{x}|\boldsymbol{\vartheta})\big)\\&=\arg\max_{\boldsymbol{\vartheta}}\mathbb{E}_{p(\boldsymbol{x}|\boldsymbol{\psi}^*)}\Big[\ln\big(q(\boldsymbol{x}|\boldsymbol{\vartheta})\big)P(L|\boldsymbol{x})\Big],\end{aligned} \tag{2}$$

where $D_{\mathrm{KL}}(\cdot)$ is defined as the Kullback–Leibler (KL) divergence between the target conditional. The

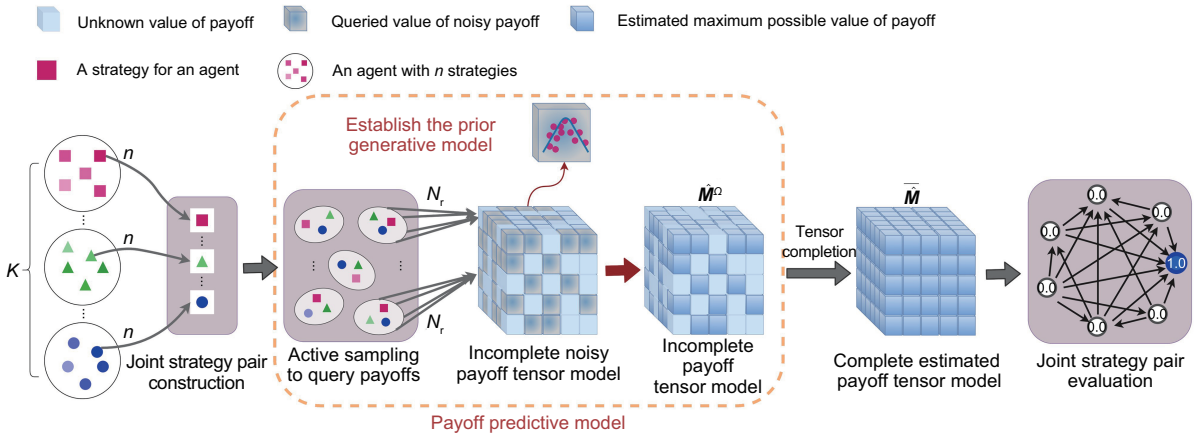

Fig. 3 Overview of TcEval-AS. TcEval-AS constructs n^k joint strategies for k agents (each with n strategies), builds the incomplete payoff tensor $\hat{M}$ with repeated interactions (N_r times for each joint strategy), estimates the complete payoff tensor $\overline{\hat{M}}$ based on tensor complementation, and calculates the steady-state strategy distribution π from $\overline{\hat{M}}$

data in set L are difficult to obtain. Therefore, $P(L|\boldsymbol{x})$ will be vanishingly small for most $\boldsymbol{x}$ sampled from $p(\boldsymbol{x}|\boldsymbol{\psi}^*)$, which will cause the high variance of approximation. An adaptive sampling algorithm based on the focal region is proposed to achieve accurate estimation of the expected payoff by using a small amount of the sampling data (Brookes and Listgarten, 2018; Brookes et al., 2019).

Denoting $r(\boldsymbol{x})$ as the importance sampling distribution, Eq. (2) can be expressed as

$$\boldsymbol{\vartheta}^* = \arg\max_{\boldsymbol{\vartheta}} \mathbb{E}_{r(\boldsymbol{x})} \left[\frac{p(\boldsymbol{x}|\boldsymbol{\psi}^*)}{r(\boldsymbol{x})} \ln\left(q(\boldsymbol{x}|\boldsymbol{\vartheta})\right) P(L|\boldsymbol{x}) \right]. \tag{3}$$

We construct a series of conditions $T_L^{(t)}$ and their corresponding distributions $r^t(\boldsymbol{x})$, as follows:

$$T_L^{(t)} : L \subset L^{t+1} \subset L^t, \tag{4}$$

where Eq. (4) ensures that L^t gradually approximates the set L as time t increases and it guarantees the existence of $\mathbb{E}_{r^t(\boldsymbol{x})}[P(L^t|\boldsymbol{x})]$.

In the subsequent time series, we use the search model to represent the focal sampling distribution $r^t(\boldsymbol{x}) = q(\boldsymbol{x}|\boldsymbol{\vartheta}^t)$. Therefore, Eq. (3) can be approximated as follows:

$$\boldsymbol{\vartheta}^{t+1} = \arg\max_{\boldsymbol{\vartheta}^t} \mathbb{E}_{q(\boldsymbol{x}|\boldsymbol{\vartheta}^t)} \left[\frac{p(\boldsymbol{x}|\boldsymbol{\psi}^*)}{q(\boldsymbol{x}|\boldsymbol{\vartheta}^t)} \ln\left(q(\boldsymbol{x}|\boldsymbol{\vartheta}^t)\right) P(L^t|\boldsymbol{x}) \right]. \tag{5}$$

The proposed algorithm is shown in Algorithm 3. In this study, we choose the Gaussian model as the search model (SearchProb), and the count interval (Inte = var_0/num; num = 8) is defined to manipulate the probability of noisy payoff for that interval.

To demonstrate the effectiveness of the adaptive sampling method in practice, we have conducted a simple simulation case study, as shown in Fig. 4. We ensure that our approach satisfactorily approximates the desired expected payoff under noisy conditions.

4.2 Complexity analysis of TcEval-AS

Based on the theory of tensor complement with noisy terms (Xia et al., 2021), we give our results in Theorems 1 and 2. The proof can be seen in the appendix.

Algorithm 3 Adaptive sampling for estimating probability maximization

Input: noisy payoff, $X_{\text{train}} = \{x_1, x_2, ..., x_{N_r}\}$.
Output: estimated incomplete payoff tensor, $\hat{\boldsymbol{M}}$.

for $a = 1, 2, ..., |S|$ **do**
  Construct the generative model GenModel$(\boldsymbol{x}, \boldsymbol{\psi})$
  $m_0 \leftarrow \text{means}(X_{\text{train}})$
  $\text{var}_0 \leftarrow \text{var}(X_{\text{train}})$
  Calculate the possible probabilities corresponding to each noisy payoff, $Y_{\text{train}} = \{y_1, y_2, ..., y_{N_r}\}$
  $\boldsymbol{\psi}^* \leftarrow \text{GenModel}(X_{\text{train}}, Y_{\text{train}})$
  $\boldsymbol{\vartheta}^1 \leftarrow \boldsymbol{\psi}^*$
  for $t = 1, 2, ..., T$ **do**
    Construct $\text{set}^t = X^t_{\text{score}} = \{x_1, x_2, ..., x_N\}$, where $x \in [m_0 - 5\text{var}_0, m_0 + 5\text{var}_0]$
    $Y^t_{\text{score}} \leftarrow h_{\text{pre}}(X^t_{\text{score}})$
    $\text{Rank}^t \leftarrow \text{rank}(X^t_{\text{score}})$
    $\lambda^t \leftarrow$ mean of the first q^{th} percentile of values in Rank^t
    $\text{FocalProb} \leftarrow \frac{\text{SearchProb}(\text{set}^t, \boldsymbol{\psi}^*)}{\text{SearchProb}(\text{set}^t, \boldsymbol{\vartheta}^t)} P(L^t|\text{set}^t)$
    $\boldsymbol{\vartheta}^t \leftarrow \text{GenModel}(X^t_{\text{score}}, Y^t_{\text{score}}, \text{FocalProb})$
  end for
  $\mathbb{E}(\boldsymbol{x}) \leftarrow \boldsymbol{x} \sim N(\boldsymbol{\vartheta}^{\text{T}})$
  $\hat{M}_a = \mathbb{E}(\boldsymbol{x})$
end for
return $\hat{\boldsymbol{M}}$

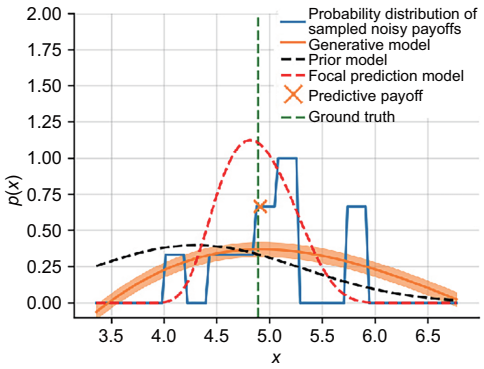

Fig. 4 An illustrative example showing that our final approximate payoff estimated using a small amount of sampling data is close to the real one

Theorem 1 (Approximate low-rank payoff tensor) Let $\hat{\boldsymbol{M}} = \boldsymbol{M} + \boldsymbol{Z}_\tau$ be the observed payoff tensor bounded in $[-M_{\max}, M_{\max}]$ and assume the observation noise $||\boldsymbol{Z}_\tau||_\infty < \tau$. Define Ω as the sampling operator with m groups of joint strategies actively sampled for evaluation. The observed payoff tensor selected by the sampling operator Ω is $\hat{\boldsymbol{M}}^\Omega = \boldsymbol{M}^\Omega + \boldsymbol{Z}^\Omega_\tau$. By performing the PS-HOSVD tensor completion algorithm on $\hat{\boldsymbol{M}}^\Omega$ to obtain $\overline{\hat{\boldsymbol{M}}}$, there exist constants $C_0 > 0$, $C' > 0$, and $\alpha > C_0$

such that if the number of actively selected samples satisfies

$$m \geq C_0\gamma^2\left(nr^{k-1}+r^{(k-1)/2}n^{k/2}\right)\text{poly}\ln n, \quad (6)$$

then $P\left(\max_{\boldsymbol{s}\in\prod_k S^k}\left|\overline{\hat{\pi}}(\boldsymbol{s})-\hat{\pi}(\boldsymbol{s})\right|\leq\xi\right)\geq 1-\frac{1}{n^2}$ can be obtained, where

$$\xi\in\left(0,18\times 2^{-n}\sum_{i=1}^{|S|-1}\binom{|S|}{i}i^{|S|}\right),$$

$$\tau=\frac{\xi g(\alpha,\eta,m,M_{\max})}{18L(\alpha,M_{\max})\sum_{i=1}^{|S|-1}\binom{|S|}{i}i^{|S|}n^{k/2}\Re},$$

$$\Re=C'\gamma^2\wedge_{\max}(\boldsymbol{M})/\wedge_{\min}(\boldsymbol{M})r^{1/2}n^{k/4}\ln^{k+2}(n)+1,$$

$$L(\alpha,M_{\max})=2\alpha e^{2\alpha_{\max}},$$

$$g(\alpha,\eta,m,M_{\max})=\eta\frac{e^{2\alpha_{\max}}-1}{e^{2m\alpha_{\max}}-1}.$$

Theorem 2 (Noisy payoff) Suppose that the payoff tensor $\boldsymbol{M}\in\mathbb{R}^{n^1\times n^2\times\ldots\times n^k}$ is a low-rank tensor with rank $r=\max(r_1(\boldsymbol{A}),r_2(\boldsymbol{A}),...,r_n(\boldsymbol{A}))$. Define $L(\alpha,M_{\max})=2\alpha e^{2\alpha_{\max}}$ and $g(\alpha,\eta,m,M_{\max})=\eta\frac{e^{2\alpha_{\max}}-1}{e^{2m\alpha_{\max}}-1}$. Let $\xi\in(0,18\times 2^{-n}\sum_{i=1}^{|S|-1}\binom{|S|}{i}i^{|S|}\Big)$. Let Ω be the sampling operator with m groups of joint strategies actively sampled for evaluation. For each joint strategy $\omega_i\in\Omega$, let $\hat{a}_{\omega_i}$ be an empirical estimation payoff constructed by taking N_{r} interactions. $\overline{\hat{\pi}}$ is the steady-state strategy distribution obtained by computing the α-rank on $\hat{\boldsymbol{M}}$, achieved by using PS-HOSVD on $\hat{\boldsymbol{M}}^{\Omega}$. There exist constants $C_0>0$, $C'>0$, and γ such that if the number of selected samples m satisfies

$$m\geq C_0\gamma^2\left(nr^{k-1}+r^{(k-1)/2}n^{k/2}\right)\text{poly}\ln n \quad (7)$$

and N_{r} satisfies

$$N_{\mathrm{r}}\geq\frac{648M_{\max}^2m\ln(2mkn^2)L^2(\alpha,M_{\max})}{\xi^2g^2(\alpha,\eta,m,M_{\max})}\cdot\left(\sum_{i=1}^{|S|-1}\binom{|S|}{i}i^{|S|}\right)^2\Re^2, \quad (8)$$

then we can obtain

$$P\left(\max_{\boldsymbol{s}\in\prod_k S^k}\left|\overline{\hat{\pi}}(\boldsymbol{s})-\pi(\boldsymbol{s})\right|\leq\xi\right)\geq 1-\frac{2}{n^2}. \quad (9)$$

5 Simulations

In this section, we propose two scenarios based on EnergyGrid to validate the effectiveness of TcEval and TcEval-AS.

Scenario 1 constructs EnergyGrid with five buildings, where each agent controls five strategies of the EM task.

Scenario 2 constructs EnergyGrid with 10 buildings, where each agent controls three strategies of the EM task.

5.1 Baselines

We compare our proposed algorithms with the following algorithms:

1. RG-UCB (Rowland et al., 2019), which uses a sampling scheme based on confidence bounds. The joint strategy exceeding the confidence bound is moved out of the set at each sampling round, and a stopping condition $C(\delta)$ is used to control the number of observations per pair. δ is the confidence level of the estimation of the payoffs. Due to the high cost of each step of the interaction, we discuss its performance only in Scenario 1 with noisy payoff.

2. α-IG (Rashid et al., 2021), which uses an active sampling strategy to estimate the α-rank payoff tensor from as few samples as possible. The parameters of the payoff belief distribution are updated by choosing strategy payoff with the largest decrease in entropy. Due to the high cost of each step of the interaction (i.e., 400 interactions in the 4×4 game), we analyze its performance only in Scenario 1 with noisy payoff.

Moreover, four metrics are proposed for evaluating algorithm performance:

1. α-rank relevance J_{k} (Signorino and Ritter, 1999) is calculated by Kendall's tau-b correlation coefficient. A larger value of J_{k} means a better correlation between π and $\overline{\hat{\pi}}$, and it is denoted as follows:

$$J_{\mathrm{k}}=\frac{\sum_{(i,j)\in[|S|],i\neq j}\text{sign}(\pi_i-\pi_j)\cdot\text{sign}\left(\overline{\hat{\pi}}_i-\overline{\hat{\pi}}_j\right)}{\sqrt{\sum_{i\neq j}\text{sign}^2(\pi_i-\pi_j)\cdot\sum_{i\neq j}\text{sign}^2\left(\overline{\hat{\pi}}_i-\overline{\hat{\pi}}_j\right)}}.$$

2. Ranking correct J_{c} (Du et al., 2021) measures the correctness of each joint strategy ranking, which is denoted as follows:

$$J_{\mathrm{c}}=E_{|S|}\left[\sigma\left[\left|\pi_{s_i}-\overline{\hat{\pi}}_{s_i}\right|<1/\left(5n^k\right)\right]\right],$$

where the symbol σ[predicate] is calculated as one or zero depending on whether the predicate is true.

3. Distribution error π_{error} is expressed as the estimated strategy distribution $\overline{\hat{\pi}}$ error:

$$\pi_{\text{error}}=\max_{\boldsymbol{s}_i}\left|\overline{\hat{\pi}}-\pi\right|.$$

4. Payoff error M_{error} is represented as the recovered estimated payoff tensor $\overline{\hat{\boldsymbol{M}}}$ error:

$$M_{\text{error}} = \frac{1}{n^k} \left\| \overline{\hat{\boldsymbol{M}}} - \boldsymbol{M} \right\|_{\text{F}}.$$

5.2 Evaluation results

5.2.1 Evaluation performance of TcEval

To verify the effectiveness of TcEval, Fig. 5 shows the evaluation results in Scenario 1. It can be seen that TcEval obtains accurate ranking with as few as about 400 interactions, reducing the number of interactions by >60% compared to the original α-rank.

Furthermore, to explore the stability of TcEval with an increasing number of agents, we expand Scenario 1 to Scenario 2. Fig. 6 demonstrates that using the TcEval algorithm can reduce the number of sampled joint strategies to 3500, which represents a reduction of >90% compared to α-rank.

5.2.2 Evaluation performance of TcEval-AS

The evaluation performances of RG-UCB, α-IG, and TcEval-AS in Scenario 1 with noisy payoffs are shown in Fig. 7. Note that RG-UCB and α-IG require a large number of interactions, and the average number of interactions is 12 000 by five simulations, which is shown by two horizontal lines in Fig. 7. TcEval-AS achieves better performance in terms of all metrics, compared to RG-UCB and α-IG, with the same number of interactions. When RG-UCB and α-IG achieve stability during performance evaluation, TcEval-AS obtains similar J_{k} and J_{c} by 2000 interactions, reducing the number of interactions by >80%.

5.3 Control performance of the top-rank joint strategy

In this subsection, we measure the model–reality matching performance. For Scenario 2 with noisy payoffs, two metrics are proposed to describe the EM performance of the optimal joint strategies, which are the top-rank strategy calculated by TcEval-AS and TcEval (due to the excessive computational complexity involved in the sampling process, we cannot calculate the top-rank strategies provided by both RG-UCB and α-IG). These two metrics are also considered as follows:

1. Voltage deviation V_{d} measures the overall voltage stability of the power grid system, which is expressed as voltage deviation from 1 per unit each time. Let v_i denote the voltage of bus i; voltage deviation V_{d} can be denoted as $V_{\text{d}} = \sum (v_i - 1)^2$.

2. Building comfort R_{com} measures the temperature comfort indicator inside each building, which can be denoted as $R_{\text{com}} = r^{\text{comfort}}$.

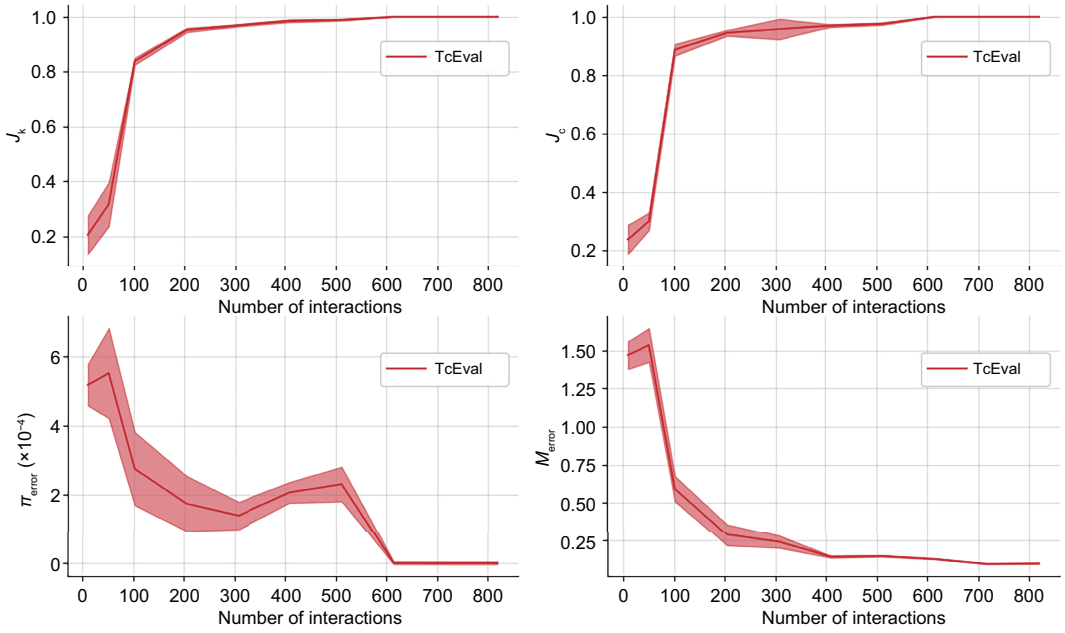

Fig. 5 Evaluation performance of TcEval with precise payoffs in Scenario 1 ($r = 2$)

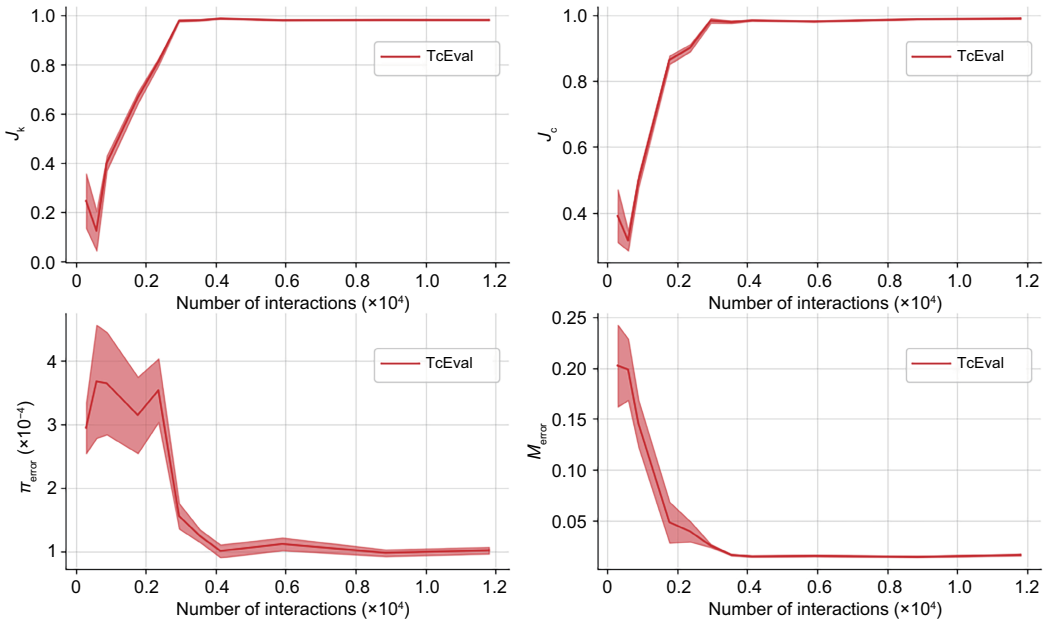

Fig. 6 Evaluation performance of TcEval with precise payoffs in Scenario 2 ($r = 2$)

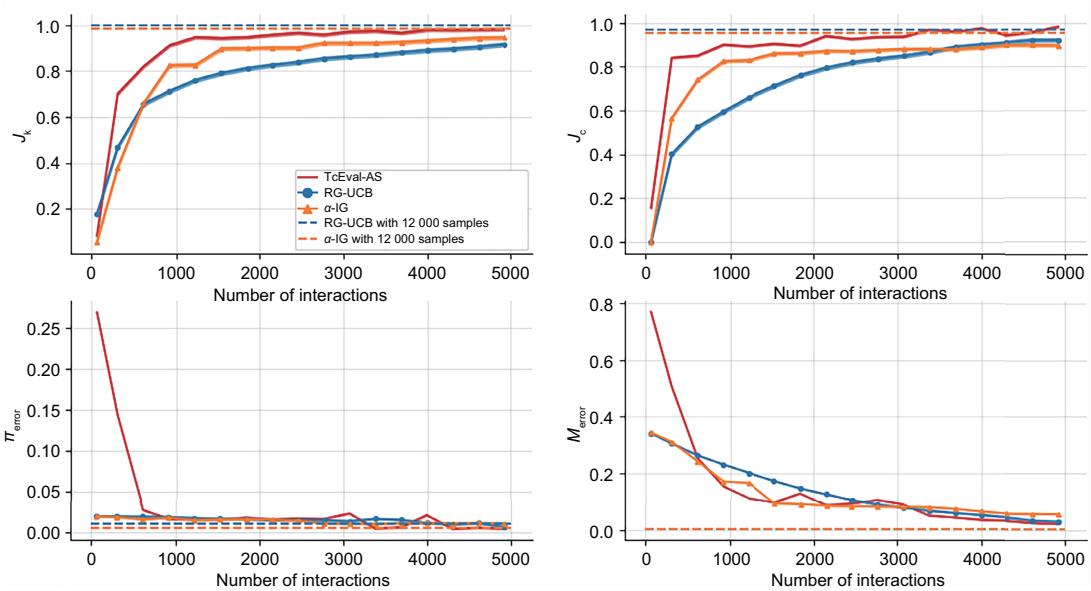

Fig. 7 Evaluations of Scenario 1 with noise by using TcEval-AS, RG-UCB, and α-IG ($r = 2$, $N_r = 6$; the confidence of RG-UCB is set to 0.01)

It can be observed from Fig. 8 that compared to the baseline joint strategy, the top-rank joint strategy evaluated by TcEval-AS usually effectively reduces the overall voltage deviation, especially during high deviations, which ensures the stability of the power system at the system level. TcEval shows improvements in the voltage deviation metric compared to the directly trained original algorithm (MADDPG), but it exhibits suboptimal performance during the peak time period when compared to TcEval-AS. By examining the average rewards of daily comfort performance for each agent in Fig. 9, it can be observed that the baseline joint strategy does not perform well on some agents, possibly because all agents converge on similar control strategies. The top-rank joint strategy evaluated by TcEval-AS usually shows

better user comfort performance.

The results show that the combination of MARL methods with the strategy evaluation algorithm is an effective roadmap toward trustworthy MARL.

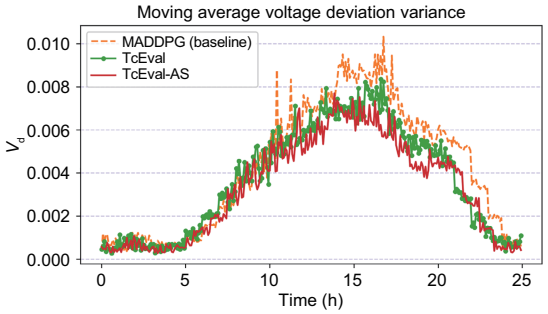

Fig. 8 Overall voltage deviation performance in Scenario 2 with strategies obtained from different algorithms, where the baseline is the joint strategy directly outputted by MADDPG

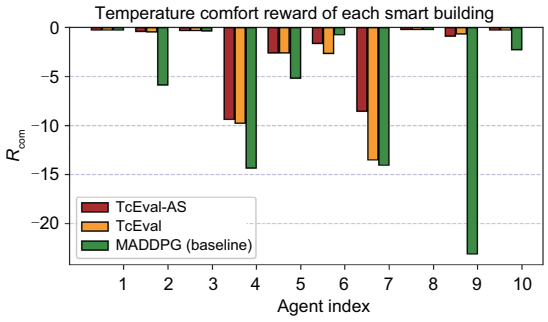

Fig. 9 Daily comfort performance in Scenario 2 with strategies obtained from different algorithms, where the baseline is the joint strategy directly outputted by MADDPG

Simultaneously, TcEval-AS outperforms TcEval in noisy scenarios, indicating that the proposed payoff predictive model can exhibit excellent predictive performance while reducing the sampling number.

5.4 Focusing on rank parameter discussion

To analyze the effect of the preset parameter r on the evaluation performance, Fig. 10 discusses the evaluation results in Scenario 1 with preset rank r =1, 2, 3, and 4. It can be seen that all metrics consistently exhibit robust performance across varying levels of r. Besides, the complementary payoff tensor with a choice of $r = 1$ achieves satisfactory evaluation results, indicating that the low-rank tensor can provide an accurate approximation compared to the high-rank payoff tensor.

5.5 Focusing on the payoff predictive model

TcEval-AS proposes a real payoff prediction method, aiming to mitigate the influence of noise in payoffs on evaluation results. To verify the validity of the proposed method, we compare it with the mean prediction method (Rowland et al., 2019). The simulation is set up for Scenario 2, and noisy observations are simulated by adding Gaussian noise to the observed payoffs for each interaction. As shown in Fig. 11, the proposed payoff prediction method achieves better accuracy and robustness than the traditional prediction method in the noisy case.

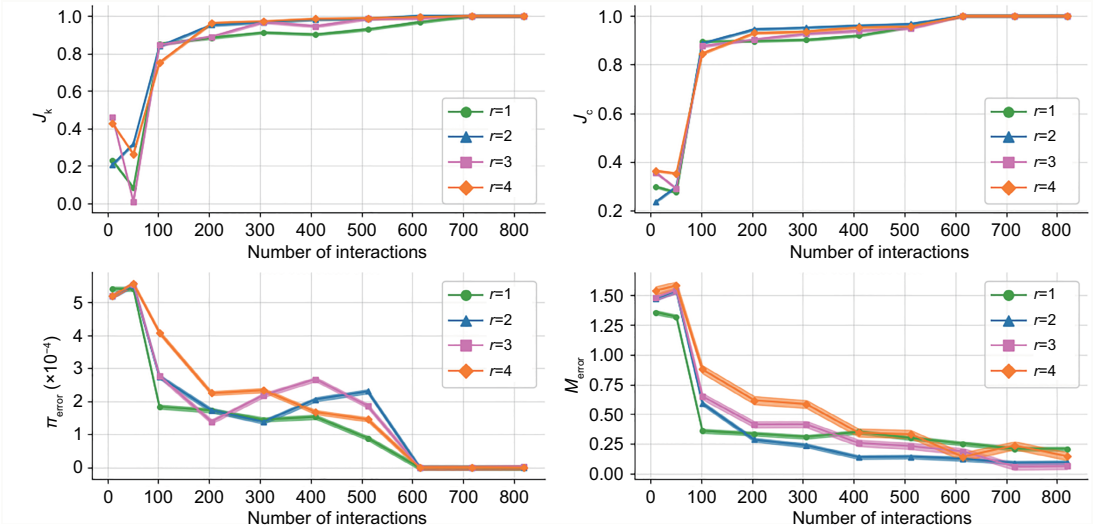

Fig. 10 Evaluation results for different ranks in Scenario 1 with $\alpha = 0.01$

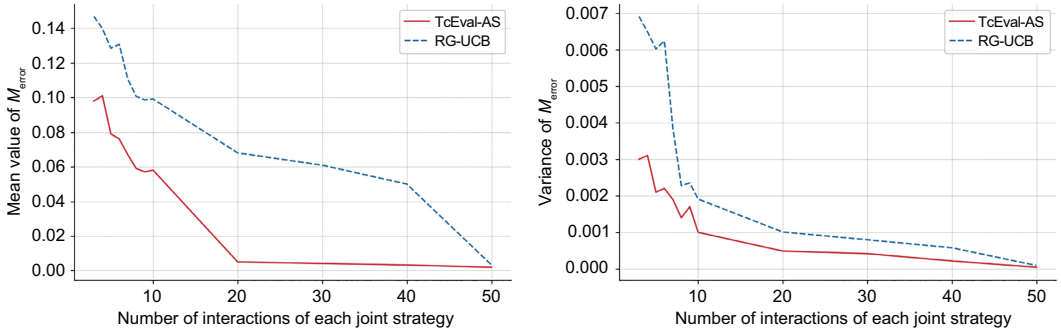

Fig. 11 Mean value and variance of payoff error M_{error} in Scenario 2 for different methods

6 Conclusions

In this paper, two algorithms, namely, TcEval and TcEval-AS, are proposed for the EM task in a simulated power grid system. TcEval is developed to achieve large-scale strategy evaluation by estimating the α-rank with tensor complement. Moreover, considering the noise prevalent in practice, the proposed TcEval-AS adds a noise processing module based on TcEval, using a mathematical model to fit the noisy payoff distribution and predict the real payoff. Joint strategy evaluation is performed in two cases based on EneryGrid, and it is shown that the proposed algorithms can achieve better performance than RG-UCB and α-IG methods while reducing the number of interactions.

Contributors

Yiyun SUN designed the research. Yiyun SUN and Meiqin LIU processed the data. Yiyun SUN drafted the paper. Yiyun SUN, Meiqin LIU, Senlin ZHANG, Ronghao ZHENG, Shanling DONG, and Xuguang LAN revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Brookes DH, Listgarten J, 2018. Design by adaptive sampling. https://arxiv.org/pdf/1810.03714v4

Brookes DH, Park H, Listgarten J, 2019. Conditioning by adaptive sampling for robust design. Proc 36th Int Conf on Machine Learning, p.773-782.

Cai WQ, Kordabad AB, Gros S, 2023. Energy management in residential microgrid using model predictive control-based reinforcement learning and Shapley value. *Eng Appl Artif Intell*, 119:105793. https://doi.org/10.1016/j.engappai.2022.105793

Claessens BJ, Vrancx P, Ruelens F, 2018. Convolutional neural networks for automatic state-time feature extraction in reinforcement learning applied to residential load control. *IEEE Trans Smart Grid*, 9(4):3259-3269. https://doi.org/10.1109/TSG.2016.2629450

Czarnecki WM, Gidel G, Tracey B, et al., 2020. Real world games look like spinning tops. Proc 34th Int Conf on Neural Information Processing Systems, Article 1463.

Dong Q, Wu ZY, Lu J, et al., 2022. Existence and practice of gaming: thoughts on the development of multi-agent system gaming. *Front Inform Technol Electron Eng*, 23(7):995-1001. https://doi.org/10.1631/FITEE.2100593

Du YL, Yan X, Chen X, et al., 2021. Estimating α-rank from a few entries with low rank matrix completion. Proc 38th Int Conf on Machine Learning, p.2870-2879.

Lowe R, Wu Y, Tamar A, et al., 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. Proc 31st Int Conf on Neural Information Processing Systems, p.6382-6393.

Muller P, Omidshafiei S, Rowland M, et al., 2020. A generalized training approach for multiagent learning. Proc 8th Int Conf on Learning Representations.

Omidshafiei S, Papadimitriou C, Piliouras G, et al., 2019. α-rank: multi-agent evaluation by evolution. *Sci Rep*, 9(1):9937. https://doi.org/10.1038/s41598-019-45619-9

Pigott A, Crozier C, Baker K, et al., 2022. GridLearn: multiagent reinforcement learning for grid-aware building energy management. *Electr Power Syst Res*, 213:108521. https://doi.org/10.1016/j.epsr.2022.108521

Rashid T, Zhang C, Ciosek K, 2021. Estimating α-rank by maximizing information gain. Proc AAAI Conf on Artificial Intelligence, p.5673-5681. https://doi.org/10.1609/aaai.v35i6.16712

Rowland M, Omidshafiei S, Tuyls K, et al., 2019. Multiagent evaluation under incomplete information. Proc 33rd Int Conf on Neural Information Processing Systems, Article 1101.

Shalev-Shwartz S, Ben-David S, 2014. Understanding Machine Learning: from Theory to Algorithms. Cambridge University Press, Cambridge, UK.

Signorino CS, Ritter JM, 1999. Tau-b or not tau-b: measuring the similarity of foreign policy positions. *Int Stud Q*, 43(1):115-144. https://doi.org/10.1111/0020-8833.00113

Silver D, Huang A, Maddison CJ, et al., 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484-489. https://doi.org/10.1038/nature16961

Su WC, Wang JH, 2012. Energy management systems in microgrid operations. *Electr J*, 25(8):45-60. https://doi.org/10.1016/j.tej.2012.09.010

Tong Z, Li N, Zhang HM, et al., 2023. Dynamic user-centric multi-dimensional resource allocation for a wide-area coverage signaling cell based on DQN. *Front Inform Technol Electron Eng*, 24(1):154-163. https://doi.org/10.1631/FITEE.2200220

Tuyls K, Perolat J, Lanctot M, et al., 2018. A generalised method for empirical game theoretic analysis. Proc 17th Int Conf on Autonomous Agents and Multiagent Systems, p.77-85.

Vincent R, Ait-Ahmed M, Houari A, et al., 2020. Residential microgrid energy management considering flexibility services opportunities and forecast uncertainties. *Int J Electr Power Energy Syst*, 120:105981. https://doi.org/10.1016/j.ijepes.2020.105981

Williams CKI, Rasmussen CE, 1995. Gaussian processes for regression. Proc 8th Int Conf on Neural Information Processing Systems, p.514-520.

Xia D, Yuan M, Zhang CH, 2021. Statistically optimal and computationally efficient low rank tensor completion from noisy entries. *Ann Stat*, 49(1):76-99. https://doi.org/10.1214/20-AOS1942

Xu HC, Domínguez-García AD, Sauer PW, 2020. Optimal tap setting of voltage regulation Transformers using batch reinforcement learning. *IEEE Trans Power Syst*, 35(3):1990-2001. https://doi.org/10.1109/TPWRS.2019.2948132

Zhang YY, Rao XP, Liu CY, et al., 2023. A cooperative EV charging scheduling strategy based on double deep Q-network and prioritized experience replay. *Eng Appl Artif Intell*, 118:105642. https://doi.org/10.1016/j.engappai.2022.105642

Zhao LY, Yang T, Li W, et al., 2022. Deep reinforcement learning-based joint load scheduling for household multi-energy system. *Appl Energy*, 324:119346. https://doi.org/10.1016/j.apenergy.2022.119346

Appendix: Proofs of theories

We begin by giving the necessary lemmas for the proofs of Theorems 1 and 2.

Lemma A1 (Rowland et al., 2019) Suppose payoffs are bounded in $[-M_{\max}, M_{\max}]$. The distribution $\hat{\pi}$ obtained from the empirical payoff matrix $\hat{\boldsymbol{M}}$ satisfies $\max_{\boldsymbol{s}\in\Pi_k S^k}\left|\pi(\boldsymbol{s})-\overline{\hat{\pi}}(\boldsymbol{s})\right|\le\xi$ with a probability of at least $1-\delta$, if

$$\sup_{\boldsymbol{s}\in S}\left|\overline{\hat{\boldsymbol{M}}_{\boldsymbol{s}}}-\boldsymbol{M}_{\boldsymbol{s}}\right|\le\frac{\xi g(\alpha,\eta,m,M_{\max})}{36\alpha \mathrm{e}^{2\alpha_{\max}}\sum_{i=1}^{|S|-1}\binom{|S|}{i}i^{|S|}},$$

where $g(\alpha,\eta,m,M_{\max})=\eta\dfrac{\mathrm{e}^{2\alpha_{\max}}-1}{\mathrm{e}^{2m\alpha_{\max}}-1}$, $\delta\in(0,1)$, and $\xi\in\left(0,18\times 2^{-n}\sum_{i=1}^{|S|-1}\binom{|S|}{i}i^{|S|}\right)$.

Lemma A2 (Xia et al., 2021) Suppose that the payoff tensor $\boldsymbol{M}\in\mathbb{R}^{n^1\times n^2\times\ldots\times n^k}$ is a low-rank tensor with rank $r=\max(r_1(\boldsymbol{A}),r_2(\boldsymbol{A}),...,r_n(\boldsymbol{A}))$. Define $\hat{\boldsymbol{M}}=\boldsymbol{M}+\boldsymbol{Z}_\tau$ to be the observed payoff tensor bounded in the interval $[-M_{\max},M_{\max}]$ and assume the observation noise $||\boldsymbol{Z}_\tau||_\infty<\tau$. Let Ω be the sampling operator with m groups of joint strategies, and $\hat{\boldsymbol{M}}^\Omega=\boldsymbol{M}^\Omega+\boldsymbol{Z}_\tau^\Omega$. There exists constant $C_0>0$ such that if the sampling number m satisfies

$$m\ge C_0\gamma^2\left(nr^{k-1}+r^{(k-1)/2}n^{k/2}\right)\mathrm{poly}\ln n,$$

then we can obtain

$$\left\|\overline{\hat{\boldsymbol{M}}}-\boldsymbol{M}\right\|_2\le C'\gamma^2\Gamma r^{1/2}\frac{n^{3k/4}}{m^{1/2}}\ln^{k+2}(n)||\boldsymbol{Z}_\tau^\Omega||_2$$

with a probability of at least $1-\dfrac{1}{n^2}$, for constant $C'>0$, and $\Gamma=\wedge_{\max}(\boldsymbol{M})/\wedge_{\min}(\boldsymbol{M})$.

Lemma A3 (Xia et al., 2021) For any tensor $\boldsymbol{M}$ and set $\Omega\subseteq[n^1]\times[n^2]\times\cdots\times[n^k]$, there exists $||\boldsymbol{M}^\Omega||_2\le k\dfrac{m^{1/2}}{n^{k/2}}\max_{\boldsymbol{s}\in S}|\boldsymbol{M}_{\boldsymbol{s}}|$.

Now we provide the proof of Theorem 1.

Proof of Theorem 1 According to Lemmas A2 and A3, we can obtain

$$\begin{aligned}\left\|\overline{\hat{\boldsymbol{M}}}-\hat{\boldsymbol{M}}\right\|_2&\le\left\|\overline{\hat{\boldsymbol{M}}}-\boldsymbol{M}\right\|_2+||\boldsymbol{M}-\hat{\boldsymbol{M}}||_2\\&\le C'\gamma^2 r^{1/2}\frac{n^{3k/4}}{m^{1/2}}\ln^{k+2}(n)||\boldsymbol{Z}_\tau^\Omega||_2+||\boldsymbol{Z}_\tau||_2\\&\le\left(C'\gamma^2 r^{1/2}\frac{\ln^{k+2}(n)}{n^{k/4}}+1\right)n^{k/2}|\boldsymbol{Z}_\tau|_{\max}.\end{aligned}$$

Thus, we can obtain

$$\sup_{\boldsymbol{s}\in S}\left|\overline{\hat{\boldsymbol{M}}}_{\boldsymbol{s}}-\hat{\boldsymbol{M}}_{\boldsymbol{s}}\right|\le\frac{\xi g(\alpha,\eta,m,M_{\max})}{18L(\alpha,M_{\max})\sum_{i=1}^{|S|-1}\binom{|S|}{i}i^{|S|}}.$$

Next, we provide the proof of Theorem 2.

Proof of Theorem 2 Define $\boldsymbol{Z}_\tau=\hat{\boldsymbol{M}}-\boldsymbol{M}$, $\hat{\boldsymbol{M}}_{\boldsymbol{s}}=f\left(\hat{M}_{\boldsymbol{s}}^1,\hat{M}_{\boldsymbol{s}}^2,...,\hat{M}_{\boldsymbol{s}}^{N_r}\right)$ and $\tau=\dfrac{\xi g(\alpha,\eta,m,M_{\max})}{18L(\alpha,M_{\max})\sum_{i=1}^{|S|-1}\binom{|S|}{i}i^{|S|}n^{k/2}\Re}$, where $f(\boldsymbol{M})$ denotes the payoff prediction process. Then, we can

obtain

$$
\begin{aligned}
&P(||\boldsymbol{Z}_{\tau}^{\Omega}||_2>\tau)\\
\leq&P\left(k\frac{m^{1/2}}{n^{k/2}}\max_{\boldsymbol{s}\in S}|M_{\boldsymbol{s}}|>\tau\right)\quad\text{(by Lemma A3)}\\
=&P\left(\exists\boldsymbol{s}\in S:|\hat{\boldsymbol{M}}_{\boldsymbol{s}}-\boldsymbol{M}_{\boldsymbol{s}}|>\frac{\tau n^{k/2}}{km^{1/2}}\right)\qquad\text{(A1)}\\
\leq&\sum_{s\in S}P\left(|\hat{\boldsymbol{M}}_{s}-\boldsymbol{M}_{s}|>\frac{\tau n^{k/2}}{km^{1/2}}\right)\qquad\text{(A2)}\\
&\left(\text{since }N_{\mathrm{r}}>\frac{2M_{\max}^2m\ln(2mkn^2)}{\tau^2n^k}\right)\\
\leq&\sum_{\boldsymbol{s}\in S}\frac{1}{mkn^2}=\frac{1}{kn^2}<\frac{1}{n^2}.
\end{aligned}
$$

Here (A1) holds because of the union bound theorem (Shalev-Shwartz and Ben-David, 2014), and (A2) holds because of Hoeffding's inequality. Let $x_1,x_2,...,x_n$ be the variables bounded in $[t_{\min},t_{\max}]$; then, for any $\epsilon>0$, there exists

$$
\begin{aligned}
&P(|f(x_1,x_2,...,x_{N_{\mathrm{r}}})-\mathbb{E}(x_i)|>\epsilon)\\
\leq&P\left(\left|\frac{1}{N_{\mathrm{r}}}\sum_{i=1}^{N_{\mathrm{r}}}\left(x_i-\mathbb{E}(x_i)\right)\right|>\epsilon\right)\\
\leq&2\mathrm{e}^{-2N_{\mathrm{r}}\epsilon^2/(t_{\max}-t_{\min})^2}.
\end{aligned}
$$

So, with a probability of at least $1-\frac{1}{n^2}$, we have

$$
\tau=\frac{\xi g(\alpha,\eta,m,M_{\max})}{18L(\alpha,M_{\max})\sum_{i=1}^{|S|-1}\binom{|S|}{i}i^{|S|}n^{k/2}\Re}.
$$

Combined with Lemma A2 and the union bound, with a probability of at least $1-\frac{2}{n^2}$, we can obtain

$$
\sup_{\boldsymbol{s}\in S}\left|\overline{\hat{\boldsymbol{M}}}_{\boldsymbol{s}}-\boldsymbol{M}_{\boldsymbol{s}}\right|\leq\frac{\xi g(\alpha,\eta,m,M_{\max})}{18L(\alpha,M_{\max})\sum_{i=1}^{|S|-1}\binom{|S|}{i}i^{|S|}}.
$$

Using Lemma A1, $\max_{\boldsymbol{s}\in\prod_k S^k}\left|\overline{\hat{\pi}}(\boldsymbol{s})-\pi(\boldsymbol{s})\right|\leq\xi$ can be obtained. Thus, the proof is completed.