Daogui Tang
Yi-Ping Fang
Enrico Zio
Laboratoire Génie Industriel, CentraleSupélec
Gif-sur-Yvette, France
Contact : daogui.tang@centralesupelec.fr

A zero-sum Markov defender-attacker game for modeling false pricing in smart grids and its solution by multi-agent reinforcement learning

1. Background

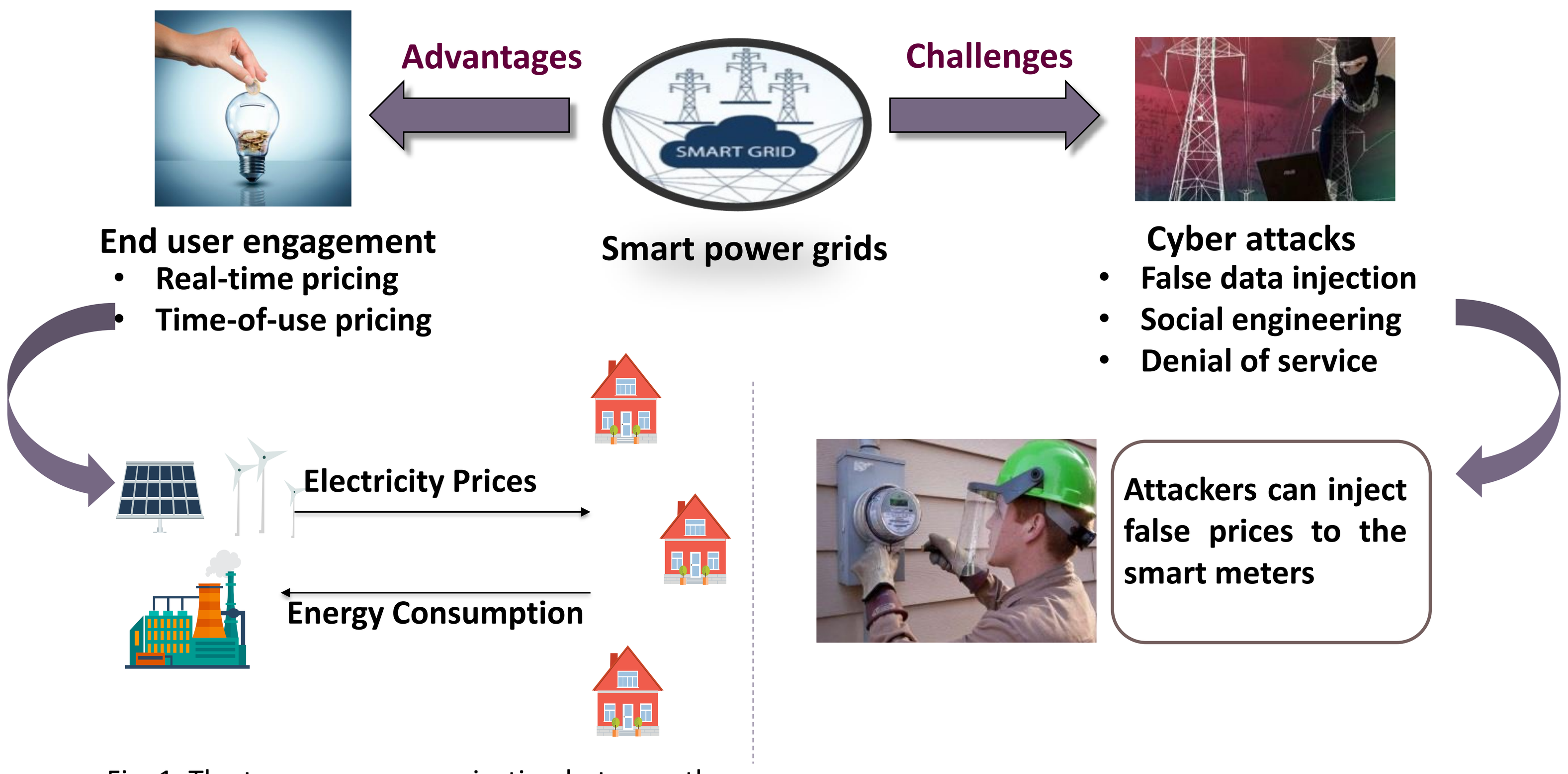
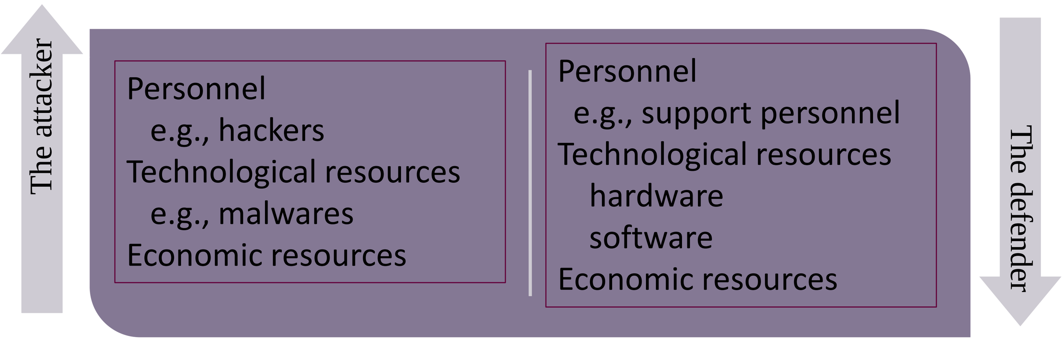


Fig. 1. The two-way communication between the utility company and consumers.

False pricing attacks: the attacker injects false prices to the smart meters so that a part of consumers change their energy consumption behavior, and, thus, potentially cause overload of some distribution lines.



The attacker and defender need to find the best policies of allocating their resources to maximize their benefit .

2. Problem formulation

System overview

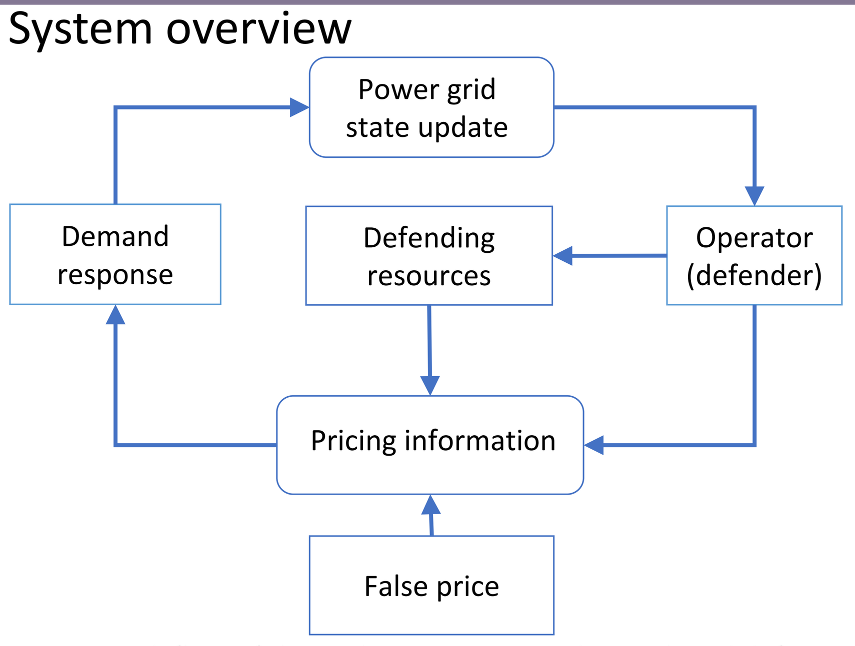


Fig. 2. Work flow of the real-time pricing with attack and defense.

- The operator sends electricity prices to the smart meters in consumers' houses
- The attacker can inject false prices to the smart meters
- The defender allocates protection resources to protect consumers
- Consumers respond to the electricity prices
- The operator receives consumption information and update real-time prices

The decision process of the attacker and defender can be modelled by a two-player zero-sum Markov Game, $MG = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R} \rangle$.

Zero-sum is a situation in game theory in which one person's gain is equivalent to another's loss, so the net change in wealth or benefit is zero.

- $\mathcal{S} = \{s_1, s_2, \ldots, s_t\}$: the finite set of environment **states**;

$$s = \begin{cases} s_1, & L = 0 \\ s_2, & L > 0 \end{cases}$$

- $\mathcal{A} = \{\mathcal{A}^a, \mathcal{A}^d\}$ represents the **joint action** of the attacker and defender
  - $\mathcal{A}^a = \{a_1^a, a_2^a, \ldots, a_{n_a}^a\}$: attacker's action space
  - $\mathcal{A}^d = \{a_1^d, a_2^d, \ldots, a_{n_d}^d\}$: defender's action space

- $\mathcal{T}$: the **state transition probability** function.

At a given state $s \in \mathcal{S}$, the probability of the environment change to $s' \in \mathcal{S}$ with the joint action $(a^a, a^d) \in \mathcal{A}$ can be defined as:

$$\mathcal{T}(s'|s, a^a, a^d) \doteq \Pr(s_{t+1} = s'|s_t = s, a_t^a = a, a_t^d = a^d)$$

- $\mathcal{R} = \{\mathcal{R}_a, \mathcal{R}_d\}$: the player's **immediate reward** function.
  - $\mathcal{R}_a = \{r_1^a, r_2^a, \ldots, r_t^a\}$: attacker's reward
  - $\mathcal{R}_d = \{r_1^d, r_2^d, \ldots, r_t^d\}$: defender's rewards

References

[1] A.-H. Mohsenian-Rad and A. Leon-Garcia, "Distributed internet-based load altering attacks against smart power grids," IEEE Transactions on Smart Grid, vol. 2, pp. 667-674, 2011.

[2] D. Tang, Y.-P. Fang, E. Zio, and J. E. Ramirez-Marquez, "Resilience of Smart Power Grids to False Pricing Attacks in the Social Network," IEEE Access, vol. 7, pp. 80491-80505, 2019.

[3] L. Wei, A. I. Sarwat, W. Saad, and S. Biswas, "Stochastic games for power grid protection against coordinated cyber-physical attacks," IEEE Transactions on Smart Grid, vol. 9, pp. 684-694, 2018.

[4] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," Cambridge, USA: The MIT Press, 2011.

3. Multi-agent reinforcement learning

The Markov game cannot be solved by traditional methods because the consumption behavior of consumers is unknown to the players, so the reward and transition probability are unavailable.

Temporal-Difference (TD) multi-agent reinforcement learning
--learn directly from raw experience without a model of the environment's dynamics;
--update estimates based in part on other learned estimates

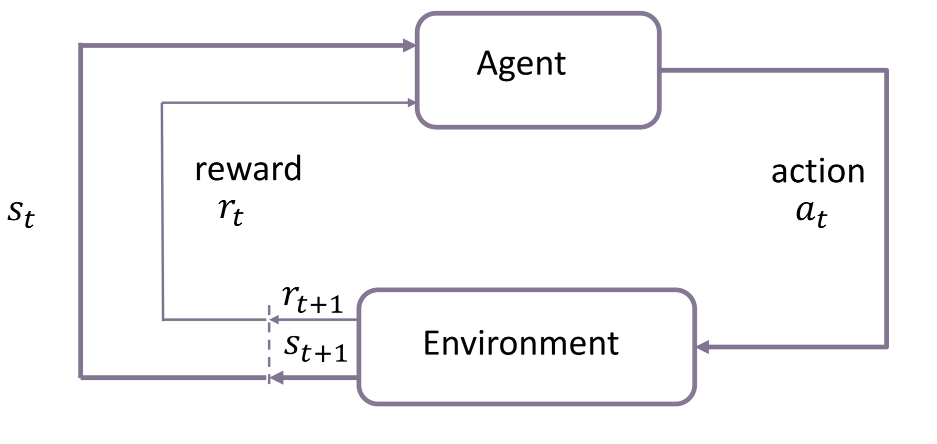


Fig. 3. The agent-environment interaction in a Markov decision process.

The policy of an agent: Associate different probabilities to each action in a state to maximize the return.

Discounted reward:

$$Q := r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

State-value function
--The expected return of the state following the policy π is:

$$V_\pi = E_\pi\left\{\sum_{k=0}^{\infty} \gamma^k r_{t+k+1} | s_t = s\right\}$$

Action-value function
--The expected return of the action a^a given the action of the opponent in state s following the policy π:

$$Q^\pi(s, a^a, a^d) = E_\pi\left\{\sum_{0}^{k+1} \gamma^k r_{t+k+1} | s_t = s, a_t = a^a, a_t^d = a^d\right\}$$

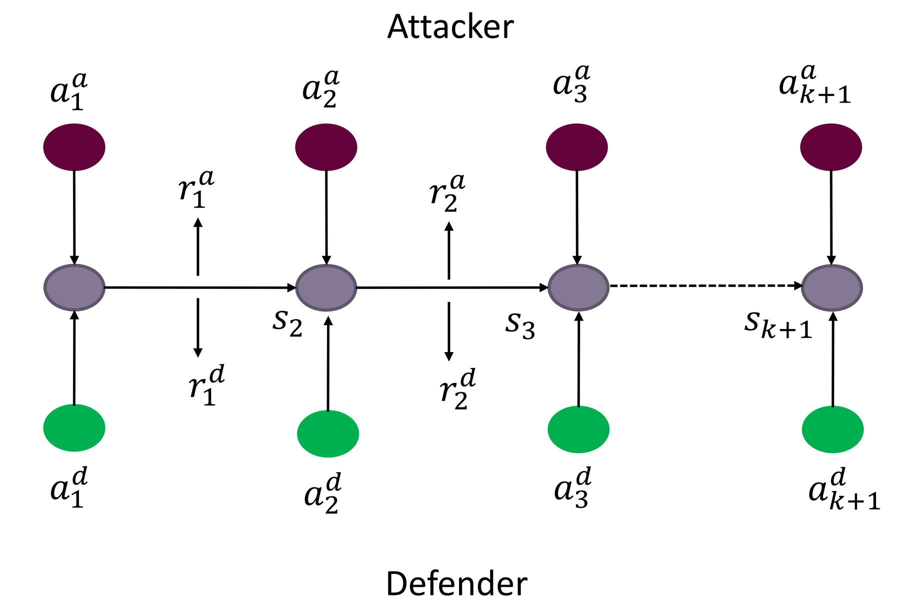


Fig. 4. The learning process of players in the proposed two-player zero-sum Markov game through interaction with environment .

Exploit & explore – ε-greedy
- Explore: the agent randomly choose actions with the probability ε
- Exploit : the agent exploit the learned policy with the probability 1-ε

***Minimax-Q* learning**
--maximize one's benefit under the worst-case assumption that the opponent will always endeavor to minimize it.

$$Q(s, a^a, a^d) = Q(s, a^a, a^d) + \alpha \cdot (R'(s, a^a, a^d) + \gamma \cdot Val(s') - Q(s, a^a, a^d))$$

$$Val(s') = \max_{\pi_a} \min_{a^d \in \mathcal{A}^d} \sum_{a^a \in \mathcal{A}^a} Q(s, a^a, a^d)\, \pi_a$$

4. Case study

A modified IEEE 13 node test feeder and the IEEE 34 node test feeder are adopted as case study:

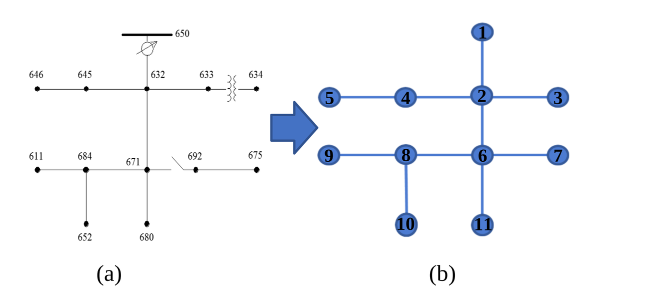


Fig. 5. 11 node test feeder here considered.

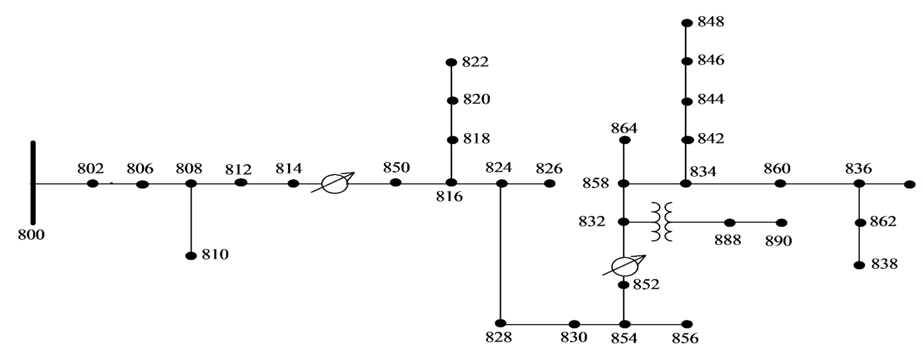
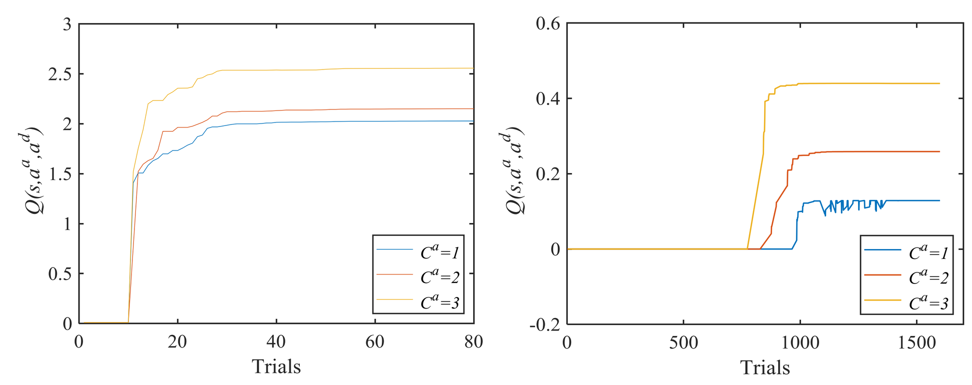
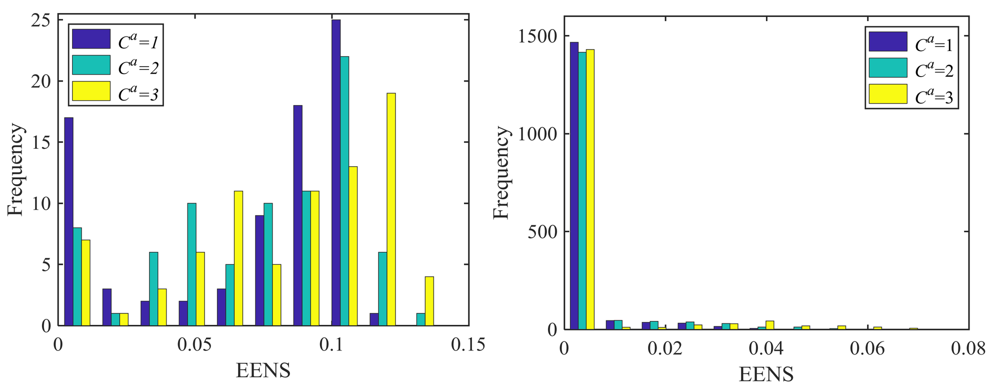


Fig. 6. IEEE 34 node test feeder.



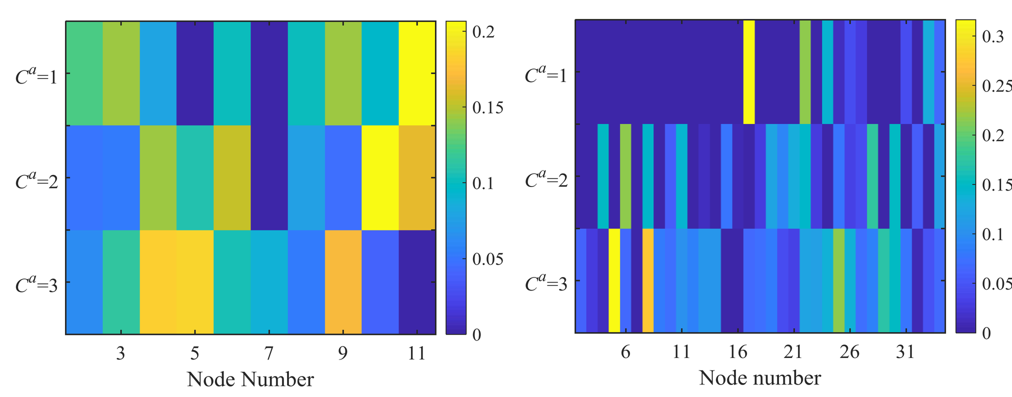
a. The 11 node test feeder
b. The IEEE 34 node test feeder

Fig. 7. Result of learned Q value.



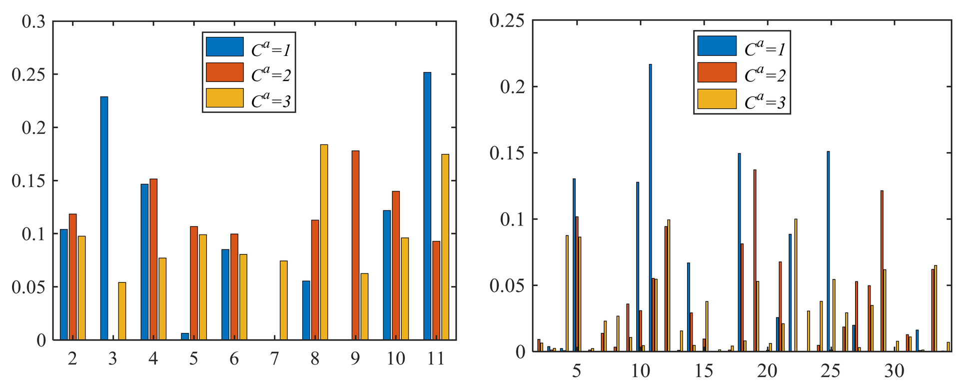
a. The 11 node test feeder
b. The IEEE 34 node test feeder

Fig. 8. The distribution of the expected energy not supplied (EENS).



a. The 11 node test feeder
b. The IEEE 34 node test feeder

Fig. 9. Policy of the attacker.



a. The 11 node test feeder
b. The IEEE 34 node test feeder

Fig. 10. Policy of the defender.

- Discussion
  - In the first beginning, the Q value remains the initial value, the agent nearly searches the policy randomly and the leaning efficiency is quite low.
  - The learning process will converge more quickly if the attacker has more resources.
  - For the IEEE 34 node test feeder, attacking one node only is hard to cause load shedding, so it's difficult to converge.
  - In general, more attack resources lead to more serious impact.
  - When the attacker has few resources, he/she will focus on several specific nodes.
- Summary
  - This work introduces a real-time pricing model that considers the uncertainty of consumers' demand response behavior based on welfare maximization . The operator is assumed to have no knowledge of consumers' responsive behavior to electricity prices.
  - A framework is established to analyze the dynamic decision process of the attacker and defender , both of whom have little knowledge of the consumers' behavior mechanism.
  - The model-free multi-agent reinforcement learning is proposed to identify the vulnerabilities and find the best defending policies under different attack resources.

ESREL 2019
09.2019, Hannover