Learning in Stackelberg Markov Games: Equitable Energy Price Design

Jun He, Edwardson School of Industrial Engineering, Purdue University
Andrew L. Liu, Edwardson School of Industrial Engineering, Purdue University
Yihsu Chen, Electrical and Computer Engineering, University of California, Santa Cruz

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation & Problem Formulation

Tackling Climate Change via Hierarchical Decision Making

- Climate and energy systems involve a leader (utility company) and followers (prosumers, consumers).
- Coordinating such systems efficiently requires anticipatory strategies – i.e., leaders account for how others respond.
- Example: utilities set energy prices → households respond by adjusting charging, usage, or storage.

Research Question

- *Can we learn equilibrium policies for hierarchical, multi-agent systems that model real-world climate and energy interactions?*

Stackelberg Markov Game Formulation

- **Two levels:** leader-follower agents.
- **State dynamics:** energy demand, renewable generation, price dynamics.
- **Objective:** design equitable electricity rates for different income groups to promote renewable energy adoptions.

Stackelberg Markov Games

What is a Stackelberg Game?

- **Sequential-move game:**
- **Leader:** commits a strategy first.
- **Followers:** commit strategies after knowing leader's strategy.

Classical vs. Markovian Formulations

- **Classical:** (i) static, one-shot interaction, (ii) complete information, and (iii) equilibrium over single move
- **This work:** (i) infinite-horizon Markov game, (ii) stochastic, partially observed dynamics, and (iii) repeated interactions via state transitions.

Single Leader Single Follower

- **Index:** $i \in \{L, F\}$, and $-i$ refers to the opponent of i
- **State & action spaces:** S_i, A_i
- **Probability transition kernels:** $P_i: S_i \times A_i \times A_{-i} \to S_i$
- **Rewards:** $r_i: S_i \times S_{-i} \times A_i \times A_{-i} \to \mathbb{R}$
- **Discount factors:** $\gamma_i \in (0,1]$
- **Policies:** $\pi_i: S_i \to \mathbb{P}(A_i)$
- **Value functions:** $V_i(s_i, s_{-i}, \pi_i, \pi_{-i}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma_i^t r_i(s_{i,t}, s_{-i,t}, a_{i,t}, a_{-i,t}) \middle| s_{i,0} = s_i, s_{-i,0} = s_{-i}\right]$

Follower's Best Response

- Given leader's policy π_L, follower finds $\mathrm{BR}_F(\pi_L) \coloneqq \arg\max_{\pi_F} V_F(\pi_L, \pi_F)$

Leader's Optimal Policy (Stationary Stackelberg Equilibrium / SSE)

- Leader anticipates follower will take its best response to π_L
- Leader finds optimal policy $\pi_L^{\mathrm{SSE}} \in \arg\max_{\pi_L} V_L(\pi_L, \mathrm{BR}_F(\pi_L))$
- As a result: $\pi_F^{\mathrm{SSE}} \in \mathrm{BR}_F(\pi_L^{\mathrm{SSE}})$

Equilibrium Learning Framework

Initial leader policy π_L^0

↓

Follower's Best Response (using RL)
$\pi_F^k \in \arg\max_{\pi_F} V_F(\pi_L^k, \pi_F)$

↓

Leader's Next Policy (using RL)
$\pi_L^{k+1} \in \arg\max_{\pi_L} V_L(\pi_L, \pi_F^k)$

Iterate until $\pi_L^{k+1} = \pi_L^k$

Solving with Reinforcement Learning (RL)

- **Boltzmann policy:** $\pi_i \coloneqq \mathrm{softmax}_{\alpha_i}(\cdot \mid s_i) = \frac{e^{\alpha_i Q^{*,\pi_{-i}}(s_i,\cdot)}}{\sum_{a_i} e^{\alpha_i Q^{*,\pi_{-i}}(s_i,a_i)}}$
- **ε-net:** discretize $\mathbb{P}(A_i)$ into $\mathcal{N}_i^{\varepsilon}$ & project $\mathrm{proj}_{\varepsilon}(\pi_i) \coloneqq \arg\min_{\pi_i' \in \mathcal{N}_i^{\varepsilon}} \|\pi_i - \pi_i'\|_1$
- **Follower's best response:** $\hat{\pi}_F^k = \mathrm{proj}_{\varepsilon}\left(\mathrm{softmax}_{\alpha_F}\left(\hat{Q}^{*,\hat{\pi}_L^k}\right)\right)$
- **Leader's next policy:** $\hat{\pi}_L^{k+1} = \mathrm{proj}_{\varepsilon}\left(\mathrm{softmax}_{\alpha_L}\left(\hat{Q}^{*,\hat{\pi}_F^k}\right)\right)$

Convergence Guarantee (Sketch)

- Under mild assumptions (reward & transition continuity, boundedness, Lipschitz condition best response with $d_L d_F \leq 1$), if $\alpha_i = \log(1/\varepsilon)/\phi(\varepsilon)$ and when $K \geq \log_{1/d_L d_F}(2/\varepsilon)$, the leader's policy satisfies $\left\|\hat{\pi}_L^K - \pi_L^{\mathrm{SSE}}\right\|_1 \leq O(\varepsilon)$.

Reward Regularization

- **Add a strongly concave function:** $r_i^{REG} = r_i + H(\pi_i(\cdot \mid s_i))$
- **Smooth & unique best response:** Lipschitz continuity guaranteed.
- Policy updates converge to **a unique fixed point.**

Extension to Infinite Followers

Single Leader Infinite Followers

- Real-world scenarios with a central authority interacting with a large agent population.
- Each follower's impact is negligible; collective behavior shapes dynamics.
- Followers are assumed **homogeneous & interchangeable.**
- In the limit as population size → ∞: each agent faces a **mean-field (MF) distribution** over states and actions; interactions reduce to a representative agent and the MF.

Initial leader policy π_L^0

↓

Followers' MF Equilibrium
$(\pi_F^k, \mu_F^k) \in \arg\max_{\pi_F, \mu_F} V_F(\pi_L^k, \pi_F, \mu_F)$

↓

Leader's Next Policy
$\pi_L^{k+1} \in \arg\max_{\pi_L} V_L(\pi_L, \pi_F^k, \mu_F^k)$

Initial MF $\mu_F^{k,0}$

↓

Followers' Best Response
$\pi_F^{k,\tau} \in \arg\max_{\pi_F} V_F(\pi_L^k, \pi_F, \mu_F^{k,\tau})$

↓

Followers' MF Update
$\mu_F^{k,\tau} \to \mu_F^{k,\tau+1}$

Until μ_F converges

Numerical Experiment

IEEE 3-Node Example

Motivation

- **Real-world challenge:** electricity pricing in presence of Distributed Energy Resources (DERs).
- Risk of a **utility death spiral**: wealthier prosumers reduce grid usage, leaving higher costs for low-income users.
- **Goal:** Learn tariffs to promote equity, stability & efficiency, and eventually renewable adoption.

Model Overview

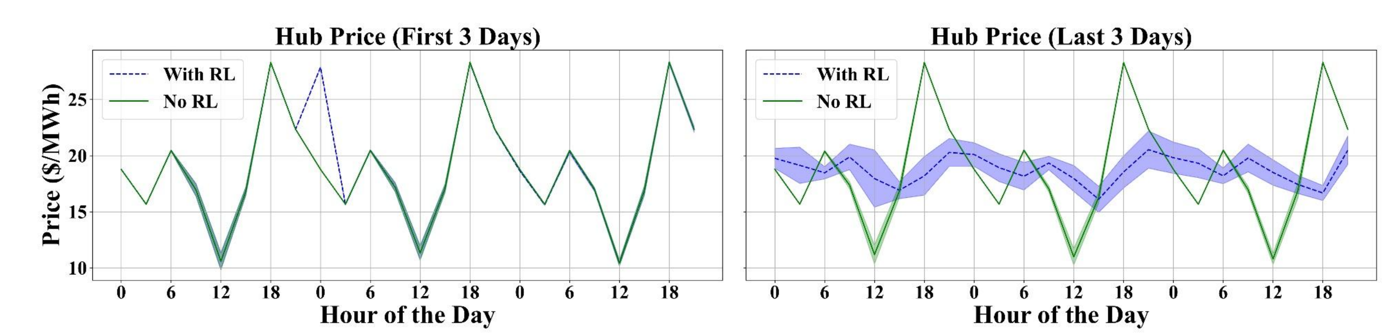
- **Leader:** utility sets rates per-MWh and fixed charges.
- **Followers:** 3 aggregators, manage prosumers + consumers.
- **Aggregators** learn battery policies under MF game setup.
- **Learning with PPO:** 8 timesteps/day; 100 days; 5 random seeds

Power Network Configuration & Population

- 3-node grid with 4 generators and 3 transmission lines.
- Each node has 3,000 pure **consumers** (income: $15k).
- **Prosumers:** 1000 low- ($25k), 500 middle- ($45k), and 300 high-income ($65k) at Node 1, 2 and 3, respectively.
- **Objective:** the leader minimizes inequality in **energy expenditure incidence (EEI)** = Electricity spending ÷ Household income.

Results: Price Stability

- RL-based learning **reduces volatility** in nodal prices.
- Daily LMP patterns become more stable over time



Results: Learned Tariffs

- Per-MWh rates stabilize.
- Fixed charges increase with income.
- Promotes **fairness**: align cost burden with ability to pay (EEI difference shrinks).