Learning in Stackelberg Markov Games

Jun He[1], Andrew L. Liu[1], Yihsu Chen[2]

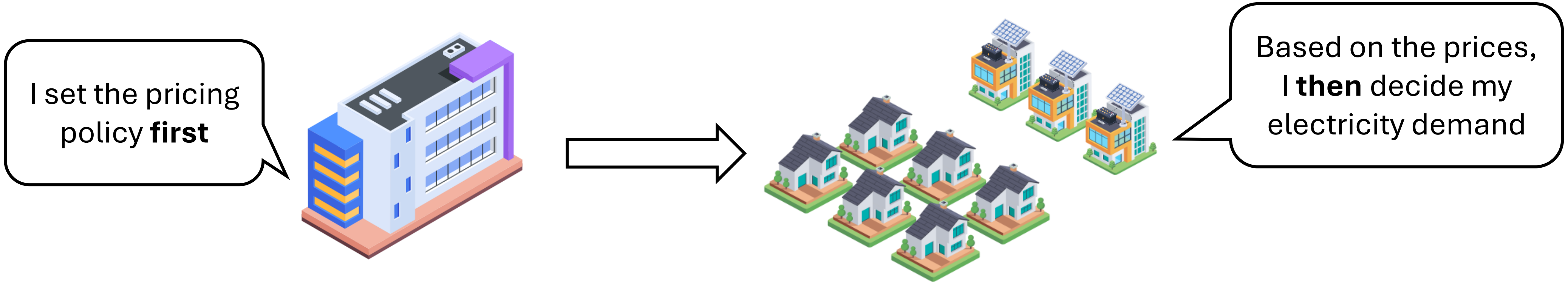
Emails: he184@purdue.edu, andrewliu@purdue.edu, yihsuchen@ucsc.edu

[1]Edwardson School of Industrial Engineering, Purdue University

[2]Electrical and Computer Engineering, University of California, Santa Cruz

Motivation

- Many real-world policy and market problems follow a leader-follower structure (Stackelberg Games).



- Dynamic, uncertain environments:
  - **Learning Algorithm:** for leader-follower best-response in Stackelberg Markov Games – dynamic leader-follower interactions with discounted rewards over infinite horizon.

Stackelberg Markov Games

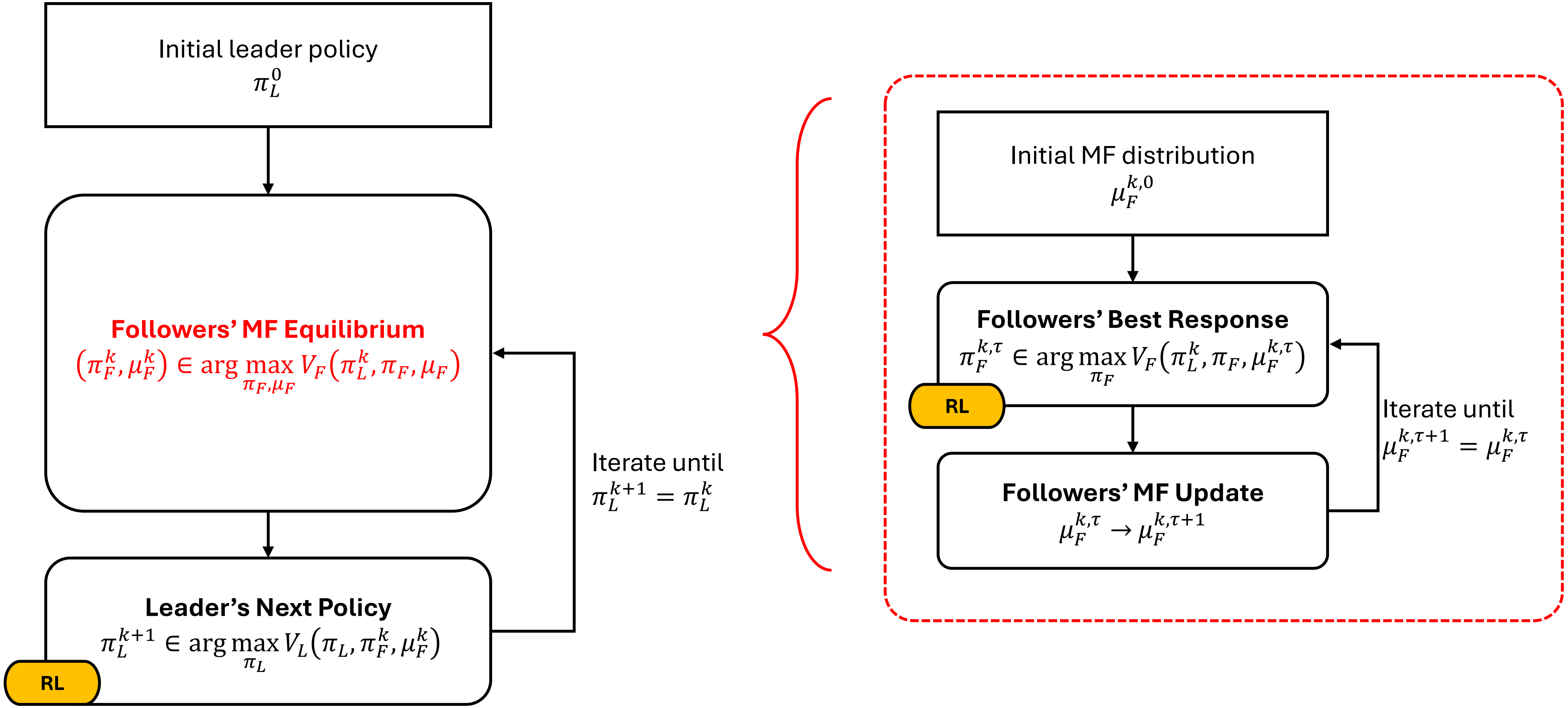
- Single Leader Single Follower:
  - Leader commits to a policy first: π_L
  - Follower then chooses a policy: π_F
- **Follower:** best response to any leader's policy:
  - Best response: $\mathrm{BR}_F(\pi_L) \coloneqq \arg\max_{\pi_F} V_F(\pi_L, \pi_F)$
- **Leader:** anticipates follower will take its best response to π_L
  - Leader's optimal policy $\pi_L^{\mathrm{SSE}} \in \arg\max_{\pi_L} V_L(\pi_L, \mathrm{BR}_F(\pi_L))$
  - As a result: $\pi_F^{\mathrm{SSE}} \in \mathrm{BR}_F\left(\pi_L^{\mathrm{SSE}}\right)$

Stationary Stackelberg Equilibrium (SSE)
$\left(\pi_L^{\mathrm{SSE}}, \pi_F^{\mathrm{SSE}}\right)$

Extension to Infinite Followers

- Single Leader Infinite Follower
  - **Mean field (MF):** followers' states and actions have a distribution
  - **Followers:** reduce to a single representative agent $\leftrightarrow$ MF distribution μ_F
  - **Followers' MF equilibrium:** for any π_L, they stabilize at $(\pi_F, \mu_F)(\pi_L)$
  - **Stationary Stackelberg MF equilibrium (SS-MFE):** $(\pi_L^*, \pi_F^*, \mu_F^*)$

Learning Framework

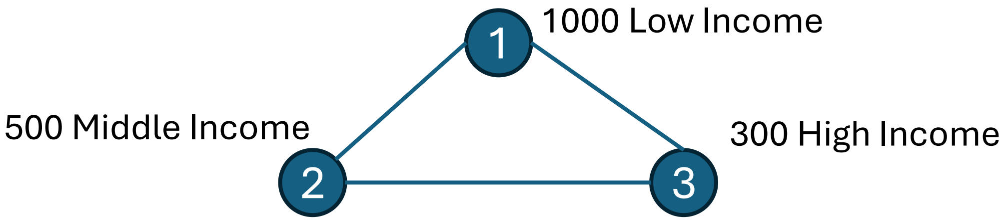


Numerical Experiment (3-Node Power Network)

- **Utility company (leader):** charge electricity rates
  - To minimize difference of ***energy expenditure incidence (EEI)*** between different income groups.

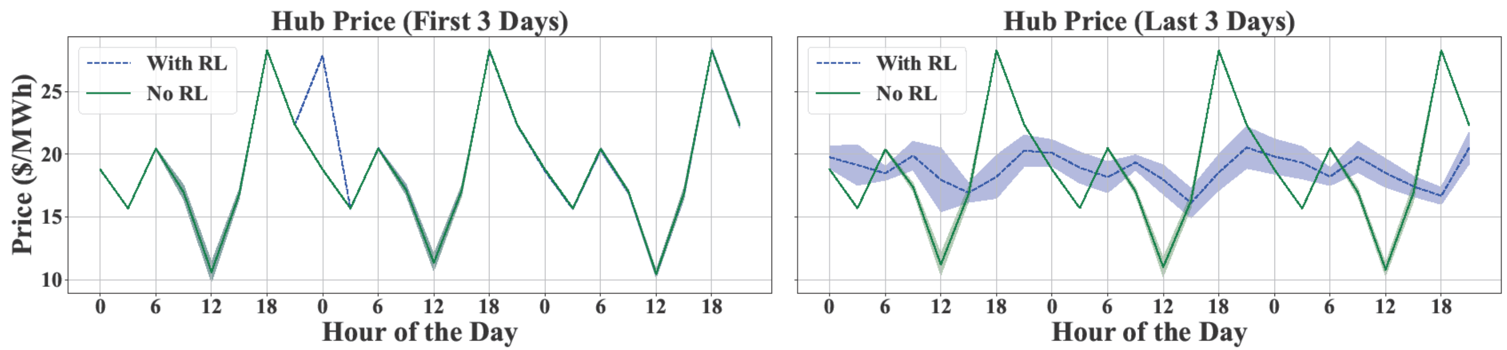
EEI = Energy Spending / Income

- **Prosumers (followers):** learns battery charge/discharge strategies based on price signals

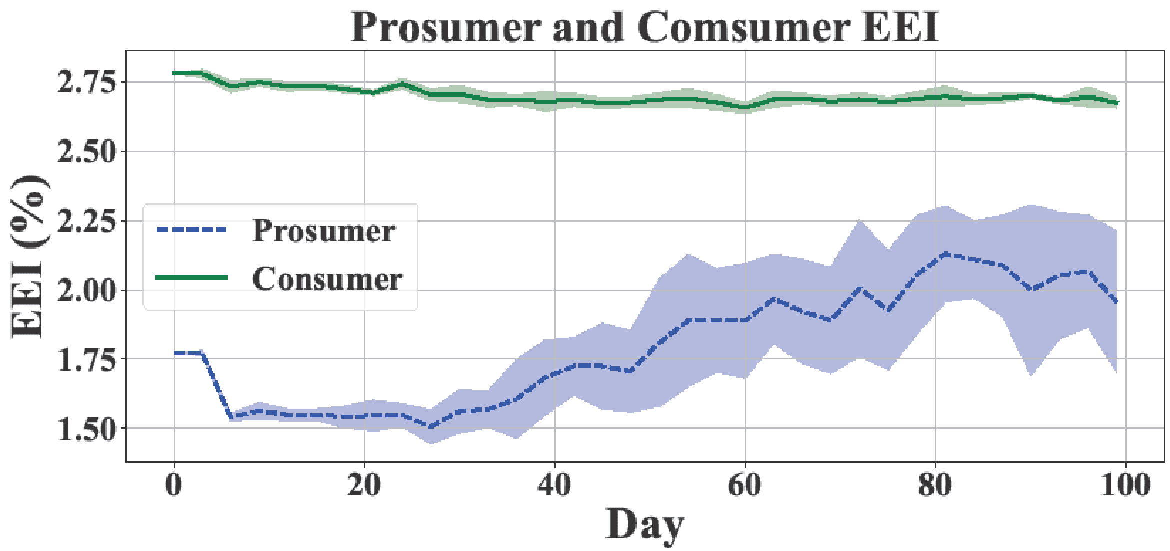


- **Consumers:** 3000 VERY Low Income per node
- RL Algorithm: PPO; Simulation: 100 Days

Results



Having renewables makes price changes less extreme



Ensures **equity** across income groups while **promoting renewable adoption**.

Thank you!