# Learning in Stackelberg Markov Games

**Jun He**
Edwardson School of Industrial Engineering
Purdue University
West Lafayette, IN 47906
he184@purdue.edu

**Andrew L. Liu**
Edwardson School of Industrial Engineering
Purdue University
West Lafayette, IN 47906
andrewliu@purdue.edu

**Yihsu Chen**
Electrical and Computer Engineering
University of California, Santa Cruz
Santa Cruz, CA 95064
yihsuchen@ucsc.edu

## Abstract

This paper studies a general framework for learning Stackelberg equilibria in dynamic and uncertain environments, where a single leader interacts with a population of adaptive followers. Motivated by equitable electricity rate design for customers with distributed energy resources, we formalize a class of Stackelberg Markov games and establish the learning framework for stationary equilibrium. We extend the framework to incorporate a continuum of agents via mean-field (MF) approximation. We validate the framework on an energy market, where a utility company sets electricity rates for a large population of households. Our results show that learned policies can achieve economic efficiency, equity across income groups, and stability in energy systems, while also encouraging renewable adoption and reducing reliance on fossil-fuel generation to mitigate climate change.

## 1 Introduction

Many real-world scenarios can be modeled as Stackelberg games, where a leader first commits to a strategy and followers respond rationally based on the leader's choice. Classical approaches to solving Stackelberg games often require explicit models of the follower's objective and best-response behavior, often through bilevel optimization techniques [1]. As a result, such methods are limited to stylized, static environments. In contrast, many policy design problems involve dynamic and stochastic environments, where agents adapt to the evolving system and the leader must learn a policy that effectively shapes long-run outcomes.

Recent advances in multi-agent reinforcement learning (RL) have opened up new possibilities for mechanism design. The AI Economist framework [2] exemplifies this by introducing a two-level RL approach, where one planner leader and economic followers co-adapt in a complex economic simulation. Theoretical work has also been done for learning Stackelberg equilibria, such as sample complexity under bandit feedback [3], local convergence of gradient-based dynamics [4], and Stackelberg-Nash equilibria with myopic followers [5].

We propose a general learning framework for Stackelberg Markov games with infinite-horizon discounted rewards. We first study the two-agent setting, and then extend the framework to incorporate a continuum of followers via MF approximation. To compute equilibria, we introduce a RL algorithm that alternates between follower and leader best-response learning, without requiring explicit knowledge of the follower's reward function. By focusing on equitable electricity rate design, our

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2025.

framework promotes solar PV and storage adoption without disproportionately burdening households who cannot afford such technologies. In turn, widespread adoption of renewables reduces dependence on fossil-fuel generation, contributing directly to climate change mitigation.

## 2 The Learning Framework for Stackelberg Markov Games

A Stackelberg game is a sequential-move game in which one agent (leader) commits to a strategy first, anticipating the other's response, and the second agent (follower) selects the best response after observing this commitment. We study Stackelberg interactions embedded in an infinite-horizon discounted Markov games. Notation-wise, let $\mathcal{I} = \{L, F\}$ denote the index for leader and follower, respectively. We write $-i$ as the opponent of $i$; that is, if $i = L$ then $-i = F$, and vice versa. For any sets $\mathcal{X}, \mathcal{Y}$, we use $\mathcal{X} \times \mathcal{Y}$ denotes the Cartesian product, $|\mathcal{X}|$ the cardinality if discrete, and $\mathcal{P}(\mathcal{X})$ the probability measure over measurable $\mathcal{X}$. We let $x \sim \mathcal{Q}$ indicate that $x$ follows distribution $\mathcal{Q}$.

### 2.1 The Single-Leader-Single-Follower Game

The definition of a Stackelberg Markov game is given below.

**Definition 2.1.** A *Stackelberg Markov game* with a single leader and a single follower is a tuple $\mathcal{G}_S := (\{\mathcal{S}_i, \mathcal{A}_i, P_i, r_i, \gamma_i\}_{i\in\mathcal{I}})$, where $\mathcal{S}_i$ is a (measurable) state spaces, and $\mathcal{A}_i$ is the action space of agent $i$. Agent $i$'s stochastic transition kernel $P_i(s_i, a_i, a_{-i})$ defines the probability distribution over next states, given current state $s_i$ and joint actions $(a_i, a_{-i})$. The reward functions $r_i : \mathcal{S}_i \times \mathcal{S}_{-i} \times \mathcal{A}_i \times \mathcal{A}_{-i} \to \mathbb{R}$ specify agent $i$'s one-step payoff, and $\gamma_i \in [0, 1)$ denotes its discount factor.

In this paper, we focus on the case in which $\mathcal{S}_i$ and $\mathcal{A}_i$ are discrete and finite. Each agent's value function is defined as the discounted total expected return as $V_i(s_i, s_{-i}, \pi_i, \pi_{-i}) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma_i^t r_i(s_{i,t}, s_{-i,t}, a_{i,t}, a_{-i,t}) \mid s_{i,0} = s_i, s_{-i,0} = s_{-i}\right]$ subject to $s_{i,t+1} \sim P_i(s_i, a_i, a_{-i}), a_{i,t} \sim \pi_i(\cdot|s_{i,t})$. Provided that the other player chooses $\pi_{-i}$, the goal for agent $i$ is to find the best policy $\pi_i^*$ that maximizes its value function starting with $s_i \in \mathcal{S}_i$ such that $V_i(s_i, s_{-i}, \pi_i^*, \pi_{-i}) \geq V_i(s_i, s_{-i}, \pi_i, \pi_{-i}), \forall \pi_i$. For each agent $i$, at state $s_i$, given the opponent's policy $\pi_{-i}$ and state $s_{-i}$, the agent treats the opponent as part of the environment and solves a single-agent MDP to compute an optimal response $\pi_i^*$. This defines the best response mapping $\text{BR}_i : \mathcal{S}_i \times \mathcal{S}_{-i} \times \mathcal{P}(\mathcal{A}_{-i}) \to \mathcal{P}(\mathcal{A}_i)$ with which $\text{BR}_i(s_i, s_{-i}, \pi_{-i}) := \text{argmax}_{\pi_i} V_i(s_i, s_{-i}; \pi_i, \pi_{-i})$. For notation brevity, we omit the two state arguments and write $\text{BR}_i(\pi_{-i})$ to denote the best response mappings. To facilitate the analysis of optimal stationary (i.e., time-invariant, memoryless) policies [6, 7] policies, we introduce the following assumption, and define the stationary Stackelberg equilibrium (SSE) in the game $\mathcal{G}_S$.

**Assumption 2.1.** There exists a finite $R \geq 0$ such that $|r_i(s_i, s_{-i}, a_i, a_{-i})| \leq R, \forall s_i, a_i, a_{-i}, i \in \mathcal{I}$.

**Definition 2.2.** A policy pair $(\pi_L^{\text{SSE}}, \pi_F^{\text{SSE}})$ in $\mathcal{G}_S$ is an SSE if, for any states $s_L, s_F$, it satisfies that the leader finds the optimal policy $\pi_L^{\text{SSE}} \in \text{BR}_L(\text{BR}_F(\pi_L))$. As a result, $\pi_F^{\text{SSE}} \in \text{BR}_F(\pi_L^{\text{SSE}})$.

We now establish the existence and uniqueness of an SSE, which requires the following assumption:

**Assumption 2.2.** For each agent $i$, there exist constants $d_i \geq 0$ such that fo any policies $\pi_{-i}, \pi'_{-i} \in \mathcal{P}(\mathcal{A}_{-i})$, one has $D_H(\text{BR}_i(s_i, s_{-i}, \pi_{-i}) - \text{BR}_i(s_i, s_{-i}, \pi'_{-i})) \leq d_i \|\pi_{-i} - \pi'_{-i}\|_1$, where $D_H(A, B) := \max\{\sup_{a\in A} \inf_{b\in B} \|a - b\|_1, \sup_{b\in B} \inf_{a\in A} \|a - b\|_1\}$ is the Hausdorff distance to measure the distance between two nonempty sets $A, B \subseteq \Pi_i, \forall i \in \mathcal{I}$, endorsed by $\ell_1$-metric $\|\cdot\|_1$.

**Theorem 2.1.** *Given Assumptions 2.1 and 2.2, when $d_L d_F < 1$, there exists an SSE to $\mathcal{G}_S$.*

### 2.2 General RL Framework

We now introduce a general RL framework for computing an SSE. At each round $k$, we first fix the leader's policy $\pi_L^k$, and compute the following until $\pi_L^{k+1} = \pi_L^k$ for some $k > 0$: (i) $\pi_F^{k*} \in \text{BR}_F(\pi_L^k)$, and (ii) $\pi_L^{k+1} \in \text{BR}_L(\pi_F^{k*})$. To implement this procedure in an RL setting, we let $\mathbf{s}_i = (s_i, s_{-i}, a_{-i})$ and $\boldsymbol{P}_i = (P_i, P_{-i}, \pi_{-i})$, and define the Q-function as $Q_i^{\pi_i, \pi_{-i}}(\mathbf{s}_i, a_i) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma_i^t r_i(\mathbf{s}_{i,t}, a_{i,t}) \,\middle|\, \mathbf{s}_{i,0} = \mathbf{s}_i, a_{i,0} = a_i\right]$, and optimal Q-function satisfies the Bellman equation for agent $i$ given $\pi_{-i}$ as $Q_i^{*,\pi_{-i}}(\mathbf{s}_i, a_i) = r_i(\mathbf{s}_i, a_i) + \gamma_i \max_{a'_i} \mathbb{E}_{\mathbf{s}'\sim\boldsymbol{P}_i}\left[Q_i^{*,\pi_{-i}}(\mathbf{s}'_i, a'_i)\right]$. However, this general approach does not guarantee convergence unless the best-response mapping satisfies

strong regularity conditions such as Lipschitz continuity. The first approach is to use Boltzmann policy, which uses the softmax operator: $\pi_i := \text{softmax}_{\alpha_i}(\cdot|s_i) = \frac{\exp(\alpha_i Q^{*,\pi_{-i}}(\mathbf{s}_i,\cdot))}{\sum_{a_i}\exp(\alpha_i Q^{*,\pi_{-i}}(\mathbf{s}_i,a_i))}, \forall i \in \mathcal{I}$, with hyperparameter $\alpha_i > 0$. It has been proved in [8] that softmax is Lipschitz continuous. Following [9], we use a finite $\varepsilon$-nets to bound the approximation error to the argmax operator. That is, for a given policy $\pi_i$, we define a finite cover $\mathcal{N}_i^\varepsilon = \{\hat{\pi_i}^{(1)}, \hat{\pi_i}^{(2)}, \cdots, \hat{\pi_i}^{(N_i^\varepsilon)}\} \subset \mathcal{P}(\mathcal{A}_i)$ such that for any $\pi_i$, there exists $\hat{\pi}_i \in \mathcal{N}_i^\varepsilon$ with $\|\pi_i - \hat{\pi}_i\|_1 \leq \varepsilon$. The projection of $\pi_i$ onto the net is defined as $\text{proj}_\varepsilon(\pi_i) := \text{argmin}_{\pi_i' \in \mathcal{N}_i^\varepsilon}\|\pi_i - \pi_i'\|_1$. We also define the *action gap* at $\mathbf{s}_i$ as $\delta_{\mathbf{s}_i}(Q^{*,\hat{\pi}_{-i}^{(j)}}) := \min_{a_i \in \mathcal{A}_i \setminus \text{argmax} Q^{*,\hat{\pi}_{-i}^{(j)}}(\mathbf{s}_i,\cdot)} \left(\max_{a_i' \in \mathcal{A}_i} Q^{*,\hat{\pi}_{-i}^{(j)}}(\mathbf{s}_i, a_i') - Q^{*,\hat{\pi}_{-i}^{(j)}}(\mathbf{s}_i, a_i)\right)$, for all $j = 1, \cdots, N_i^\varepsilon$. Then, for any $\varepsilon > 0$, there exists a positive function $\phi(\varepsilon)$ and an $\varepsilon$-net $\mathcal{N}_i^\varepsilon$ such that for all $Q^{*,\hat{\pi}_{-i}^{(j)}}$ and at any state $\mathbf{s}_i$, $\delta_{\mathbf{s}_i}(Q^{*,\hat{\pi}_{-i}^{(j)}}) \geq \phi(\varepsilon)$. Specifically, for $k = 0, 1, \ldots$, the policies are now updated as: $\hat{\pi}_F^k = \text{proj}_\varepsilon(\text{softmax}_{\alpha_F}(\hat{Q}^{*,\hat{\pi}_L^k}))$ and $\hat{\pi}_L^{k+1} = \text{proj}_\varepsilon(\text{softmax}_{\alpha_L}(\hat{Q}^{*,\hat{\pi}_F^k}))$.

**Theorem 2.2.** *Let assumption 2.2 hold, and suppose that $d_L d_F < 1$. Fix $\varepsilon > 0$ and set $\alpha_L = \alpha_F = \log(1/\varepsilon)/\phi(\varepsilon)$. Let $(\hat{\pi}_L^k, \hat{\pi}_F^k)$ denote the policy iterates using projected Boltzmann policies with $\varepsilon$-net. Then, for any $K \geq \log_{1/(d_L d_F)}(2/\varepsilon)$, the leader's policy satisfies $\|\hat{\pi}_L^K - \pi_L^{SSE}\|_1 \leq \left(\frac{1+d_L+2|\mathcal{A}_L|+2d_L|\mathcal{A}_F|}{1-d_L d_F} + 1\right)\varepsilon = O(\varepsilon)$, where $\pi_L^{SSE}$ denotes the leader's SSE policy in $\mathcal{G}_S$.*

This bound shows that it can closely approximate the true best response, while preserving Lipschitz continuity. The second approach is to add a regularization term to the reward function, which is widely used in RL. We then analyze the game using the regularized value function for for each $s_i$ such that $V_i^{\text{reg}}(s_i, s_{-i}, \pi_i, \pi_{-i}) := \mathbb{E}\left[\sum_{t=0}^\infty \gamma_i^t r_i^{\text{reg}}(s_{i,t}, s_{-i,t}, a_{i,t}, a_{-i,t}) \middle| s_{i,0} = s_i, , s_{-i,0} = s_{-i}\right]$, where $r_i^{\text{reg}}(s_i, s_{-i}, a_i, a_{-i}) = r_i(s_i, s_{-i}, a_i, a_{-i}) + H(\pi_i(\cdot \mid s_i))$, and $H(\cdot)$ is a $\rho$-strongly concave function. In Theorem 4.3 (Appendix 4.3), we show that under standard continuity and boundedness conditions, the policy iterates converge to a fixed point under the regularized learning dynamics.

### 2.3 Extension to Stackelberg Games with MF Followers under Regularization

We now consider the extension where there is one leader but an infinite number of followers. To formalize this setting, we adopt a MF approach in which followers are modeled as homogeneous and interchangeable. In the limit as the number of followers approaches infinity, each individual has negligible influence on the aggregate behavior, which is captured by an MF distribution over states and actions. We can then focus on the interaction of a single representative follower responding to the MF. For notational consistency, we retain the index set $\mathcal{I} = \{L, F\}$. Let $\mu_{F,t} \in \mathcal{P}(\mathcal{S}_F \times \mathcal{A}_F)$ denote an MF distribution at time $t$, representing the joint distribution of the population's states and actions in the infinite-agent limit, defined as $\mu_{F,t}(s,a) := \lim_{N\to\infty} \frac{\sum_{j=1, j\neq i}^N \mathbb{1}_{(s_{F,t}^j, a_{F,t}^j) = (s,a)}}{N}, \forall s \in \mathcal{S}_F, a \in \mathcal{A}_F$, where $N$ is the number of followers, and $(s_{F,t}^j, a_{F,t}^j)$ denotes the $j$-th follower's state and action pair. The indicator function $\mathbb{1}_{(\ldots)} = 1$ if the condition is true and 0 if false. Each agent's value function is redefined with the addition MF argument: $V_i(s_i, s_{-i}, \pi_i, \pi_{-i}, \mu_F) := \mathbb{E}\left[\sum_{t=0}^\infty \gamma_i^t r_i^{\text{reg}}(s_{i,t}, s_{-i,t}, a_{i,t}, a_{-i,t}, \mu_F) \middle| s_{i,0} = s_i, s_{-i,0} = s_{-i}\right]$, subject to $s_{i,t+1} \sim P_i(s_i, a_i, a_{-i}, \mu_F), a_{i,t} \sim \pi_i(\cdot|s_{i,t}, \mu_F), \forall i \in \mathcal{I}$. Finally, the evolution of MF is a mapping $\Gamma : \mathcal{P}(\mathcal{S}_F \times \mathcal{A}_F) \times \mathcal{S}_L \times \mathcal{S}_F \to \mathcal{P}(\mathcal{S}_F \times \mathcal{A}_F)$, defined as $\mu_F' := \Gamma(\mu_F, \pi_L, \pi_F), \forall \mu_F, \pi_L, \pi_F$, as a new component to the game. The stationary Stackelberg MF equilibrium (SS-MFE) is defined as follows:

**Definition 2.3.** In a Stackelberg Markov game with MF followers, the tuple $(\pi_L^{\text{SE}}, \pi_F^{\text{SE}}, \mu_F^{\text{SE}})$ forms an SS-MFE, if for any $s_L, s_F$: (i) $V_F(s_F, s_L, \pi_F^{\text{SE}}, \pi_L^{\text{SE}}, \mu_F^{\text{SE}}) \geq V_F(s_F, s_L, \pi_F, \pi_L^{\text{SE}}, \mu_F^{\text{SE}}), \forall \pi_F$, (ii) $\mu_F^{\text{SE}} = \Gamma(\mu_F^{\text{SE}}, \pi_L^{\text{SE}}, \pi_F^{\text{SE}})$, and (iii) $V_L(s_L, s_F, \pi_L^{\text{SE}}, \pi_F^{\text{SE}}, \mu_F^{\text{SE}}) \geq V_L(s_L, s_F, \pi_L, \pi_F^{\text{SE}}, \mu_F^{\text{SE}}), \forall \pi_L$.

### 2.4 Learning Framework for SS-MFE

We re-define the best response mappings with the introduction of the MF: $\text{BR}_i : \mathcal{S}_i \times \mathcal{S}_{-i} \times \mathcal{P}(\mathcal{A}_{-i}) \times \mathcal{P}(\mathcal{S}_F \times \mathcal{A}_F) \mapsto \mathcal{P}(\mathcal{A}_i)$ for both $i \in \mathcal{I}$. Then, at each iteration $k$, given the leader's policy $\pi_L^k$, the follower and MF dynamics proceed through an inner loop with iterator $\tau = 0, 1, \cdots$, and

$\pi_F^{k,\tau+1} = \text{BR}_F(s_F, s_L, \pi_L^k, \mu_F^{k,\tau}), \mu_F^{k,\tau+1} = \Gamma(\mu_F^{k,\tau}, \pi_L^k, \pi_F^{k,\tau+1})$ until convergence to $(\pi_F^{k*}, \mu_F^{k*})$. The leader then updates its policy as $\pi_L^{k+1} = \text{BR}_L(s_L, s_F, \pi_F^{k*}, \mu_F^{k*})$.

**Theorem 2.3.** *Under the same assumptions for Theorem 2.1, there exists a unique stationary SS-MFE under regularization to $\mathcal{G}_{MF}$ if $d_\mu^\mu + d_\mu^F d_F^\mu < 1$ and $\frac{d_F^L + d_\mu^L}{1-(d_F^\mu + d_\mu^\mu + d_\mu^F)} \max\{d_L^F, d_L^\mu\} < 1$,*

where the $d$'s are Lipschitz constants defined in assumptions in appendix. The pseudocode for a general RL-based algorithm to solving the game is provided in Algorithm 1 in appendix.

## 3 Numerical Experiment

We apply our framework to a real-world electricity tariff design, motivated by the growing adoption of distributed energy resources (DERs), such as rooftop solar and battery storage. As higher-income households invest in DERs, they reduce grid dependence or export energy for profit, thereby lowering their net payments to the utility. Lower-income households, who are less likely to afford DERs, continue to rely on the grid and bear a disproportionate share of the infrastructure costs. This dynamic exacerbates energy inequity and raises serious concerns [10], described as the risk of a *utility death spiral*. We adopt the same test case and settings as in [10, 11]. The power network we consider consists of a 3-node grid with 4 generators and 3 transmission lines. The utility (leader) learns a pricing policy for per-kWh charges and fixed charges to recover maintenance costs, aiming to minimize inequality in *energy expenditure incidence* (EEI), defined as the percentage of household income spent on electricity. On the follower side, each node hosts 3,000 consumers with income \$15,000. The *prosumers* (who can produce and consume electricity) population varies by node: 1000 low-income (\$25,000), 500 middle-income (\$45,000), and 300 high-income (\$65,000) at Nodes 1, 2 and 3, respectively. We model 3 aggregators, each representing a node in the grid and managing a population of both prosumers and consumers. Each learns charging/discharging policies for its prosumers' solar and storage systems, and responds to both the utility's policy and real-time locational marginal prices (LMPs) determined by a system operator via economic dispatch. We use PPO [12] and set each simulated day to 8 time steps (3-hour intervals). The utility updates every 3 days, while aggregators update at every time step. The simulation runs for 100 days over 5 random seeds. Figure 1 compares wholesale prices at the start and end of training, with and without RL. Under RL, price volatility reduces significantly, and daily patterns stabilize. Figure 2 shows the learned per-kWh add-on rates and fixed charges. Over time, the utility's policy converges to a pricing structure in which higher-income groups have higher fixed charges, helping align payment responsibility with ability to pay and maintain energy equity. More results are shown in Appendix 4.5.

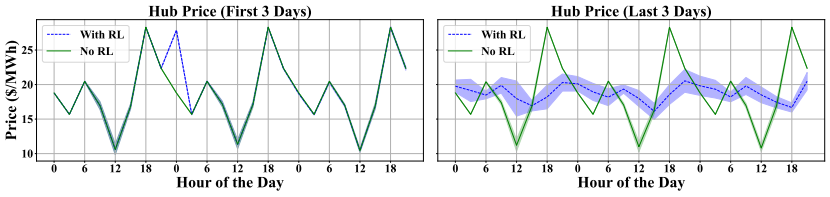


Figure 1: Comparison of nodal prices with and without learning. RL reduces price volatility and leads to more stable daily patterns.

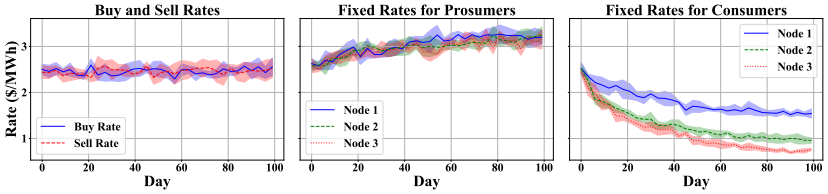


Figure 2: Learned buy/sell rates (left), fixed charges for prosumers (middle), and consumers (right).

## References


[1] S. Dempe and A. Zemkoho, "Bilevel optimization," in *Springer Optimization and Its Applications*. Springer, 2020, vol. 161.

[2] S. Zheng, A. Trott, S. Srinivasa, D. Parkes, and R. Socher, "The AI economist: Improving equality and productivity with AI-driven tax policies," *Science Advances*, vol. 8, no. 24, p. eabm1799, 2022.

[3] Y. Bai, C. Jin, H. Wang, and C. Xiong, "Sample-efficient learning of Stackelberg equilibria in general-sum games," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[4] T. Fiez, B. Chasnov, and L. J. Ratliff, "Convergence of learning dynamics in Stackelberg games," *arXiv preprint arXiv:1906.01217*, 2020.

[5] H. Zhong, Z. Yang, Z. Wang, and M. I. Jordan, "Can reinforcement learning find Stackelberg-Nash equilibria in general-sum markov games with myopically rational followers?" *Journal of Machine Learning Research*, vol. 24, no. 48, pp. 1–52, 2023.

[6] M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, ser. Wiley Series in Probability and Statistics. Wiley, 2014.

[7] A. Agarwal, N. Jiang, S. M. Kakade, and W. Sun, "Reinforcement learning: Theory and algorithms," *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, vol. 32, p. 96, 2019.

[8] B. Gao and L. Pavel, "On the properties of the softmax function with application in game theory and reinforcement learning," *arXiv preprint arXiv:1704.00805*, 2017.

[9] X. Guo, A. Hu, R. Xu, and J. Zhang, "A general framework for learning mean-field games," *Math. Oper. Res.*, vol. 48, no. 2, p. 656–686, May 2023.

[10] Y. Chen, A. L. Liu, M. Tanaka, and R. Takashima, "Optimal retail tariff design with prosumers: Pursuing equity at the expenses of economic efficiencies?" *IEEE Transactions on Energy Markets, Policy and Regulation*, vol. 1, no. 3, pp. 198–210, 2023.

[11] J. He and A. L. Liu, "Evaluating the impact of multiple DER aggregators on wholesale energy markets: A hybrid mean field approach," *arXiv preprint arXiv:2409.00107*, 2024.

[12] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[13] S. Banach, "Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales," *Fundamenta Mathematicae*, vol. 3, no. 1, pp. 133–181, 1922. [Online]. Available: http://eudml.org/doc/213289

[14] S. B. Nadler Jr, "Multi-valued contraction mappings," *Pacific Journal of Mathematics*, vol. 30, pp. 475–488, 1969.

[15] S. Cen, C. Cheng, Y. Chen, Y. Wei, and Y. Chi, "Fast global convergence of natural policy gradient methods with entropy regularization," *Operations Research*, vol. 70, no. 4, pp. 2563–2578, 2022.

[16] G. Neu, A. Jonsson, and V. Gómez, "A unified view of entropy-regularized Markov decision processes," *arXiv preprint arXiv:1705.07798*, 2017.

[17] S. Shalev-Shwartz, "Online learning: Theory, algorithms, and applications," *Ph.D. thesis, The Hebrew University of Jerusalem*, 08 2007.

[18] B. Anahtarci, C. D. Kariksiz, and N. Saldi, "Q-learning in regularized mean-field games," *Dynamic Games and Applications*, vol. 13, no. 1, pp. 89–117, 2023.

[19] M. Towers, A. Kwiatkowski, J. Terry, J. U. Balis, G. D. Cola, T. Deleu, M. Goulão, A. Kallinteris, M. Krimmel, A. KG, R. Perez-Vicente, A. Pierré, S. Schulhoff, J. J. Tai, H. Tan, and O. G. Younis, "Gymnasium: A standard interface for reinforcement learning environments," 2024. [Online]. Available: https://arxiv.org/abs/2407.17032

[20] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1364.html

[21] W. E. Hart, J.-P. Watson, and D. L. Woodruff, "Pyomo: modeling and solving mathematical programs in python," *Mathematical Programming Computation*, vol. 3, no. 3, pp. 219–260, 2011.

[22] M. L. Bynum, G. A. Hackebeil, W. E. Hart, C. D. Laird, B. L. Nicholson, J. D. Siirola, J.-P. Watson, and D. L. Woodruff, *Pyomo–optimization modeling in python*, 3rd ed. Springer Science & Business Media, 2021, vol. 67.

[23] A. Robinson, "Solar PV Analysis of Honolulu, United States," 2024. [Online]. Available: https://profilesolar.com/locations/United-States/Honolulu/

[24] Hawaiian Electric, "Power facts," 3 2024. [Online]. Available: https://www.hawaiianelectric.com/about-us/power-facts

[25] M. Roozbehani, M. A. Dahleh, and S. K. Mitter, "Volatility of power grids under real-time pricing," *IEEE Transactions on Power Systems*, vol. 27, no. 4, pp. 1926–1940, 2012.

[26] EIA, "Residential energy consumption survey 2015," https://www.eia.gov/consumption/residential/data/2020/.

# 4 Technical Appendices and Supplementary Material

## 4.1 Proof of Theorem 2.1 – Existence of an SSE

The proof relies on the well-known Banach fixed point theorem, for which we first restate the definition of a contraction mapping. In our work, we choose the distance function to be the $\ell_1$-norm.

**Definition 4.1.** Let $(\mathcal{X}, d)$ be a non-empty complete metric space, where $d$ is a metric on $\mathcal{X}$. A map $T : \mathcal{X} \mapsto \mathcal{X}$ is called a contraction mapping on $\mathcal{X}$ if for any $x, y \in \mathcal{X}$, there exists a constant $c \in [0, 1)$ such that $d(T(x), T(y)) \leq cd(x, y)$.

The Banach fixed point theorem [13] is stated as follows.

**Theorem 4.1.** *Let $(\mathcal{X}, d)$ be a non-empty complete metric space, and $T : \mathcal{X} \to \mathcal{X}$ be a contraction mapping. Then $T$ admits a unique fixed point $x^* \in \mathcal{X}$ such that $T(x^*) = x^*$.*

When the mapping is not single-valued but instead set-valued, the Banach fixed point theorem can be extended as follows [14].

**Theorem 4.2.** *Let $(\mathcal{X}, d)$ be a non-empty complete metric space, and let $T : \mathcal{X} \to CB(\mathcal{X})$ be a set-valued contraction mapping where $CB(\mathcal{X}) := \{\mathcal{Y} : \mathcal{Y}$ is a non-empty closed and bounded subset of $\mathcal{X}\}$ is endowed with the Hausdorff metric induced by $d$. Then $T$ has a fixed point $x^* \in \mathcal{X}$ such that $x^* \in T(x^*)$.*

*Proof of Theorem 2.1.* Fix $s_L, s_F$. For any $\pi_L, \pi'_L \in \mathcal{P}(\mathcal{S}_L)$, let $\pi^*_F \in \text{BR}_F(s_F, s_L, \pi_L)$ and $\pi^{*'}_F \in \text{BR}_F(s_F, s_L, \pi'_L)$, then

$$\begin{aligned} &D_H(\text{BR}_L(s_L, s_F, \pi^*_F), \text{BR}_L(s_L, s_F, \pi^{*'}_F)) \leq d_L \|\pi^*_F - \pi^{*'}_F\|_1 \\ &\leq d_L D_H(\text{BR}_F(s_F, s_L, \pi_L), \text{BR}_F(s_F, s_L, \pi'_L)) \leq d_L d_F \|\pi_L - \pi'_L\|_1. \end{aligned}$$

By (set-valued) Banach fixed-point theorem, with $0 \leq d_L d_F < 1$, there exists a fixed point of $\text{BR}_L$. As a result, there exists an SSE to the game $\mathcal{G}_S$. □

## 4.2 Proof of Theorem 2.2 – Error Bound for Projected Boltzmann Policy

*Proof.* Fix $s_F, s_L$. For notation simplicity, We drop the two state arguments in the two $\text{BR}_i$ mappings. We let the updates be (i) $\hat{\pi}^k_F = \text{proj}_\varepsilon(\tilde{\pi}_F)$, and (ii) $\hat{\pi}^{k+1}_L = \text{proj}_\varepsilon(\tilde{\pi}_L)$ where $\tilde{\pi}_i = \text{softmax}_{\alpha_i}(\hat{Q}^{*,\pi_{-i}})$. Then, at each step $k$, the following holds:

$$\begin{aligned} &\|\hat{\pi}^{k+1}_L - \pi^{\text{SE}}_L\|_1 = D_H(\{\hat{\pi}^{k+1}_L\}, \{\pi^{\text{SE}}_L\}) \\ &\leq D_H(\{\hat{\pi}^{k+1}_L\}, \{\tilde{\pi}^{k+1}_L\}) + D_H(\{\tilde{\pi}^{k+1}_L\}, \text{BR}_L(\hat{\pi}^k_F)) + D_H(\text{BR}_L(\hat{\pi}^k_F), \{\pi^{\text{SE}}_L\}) \\ &\leq \varepsilon + 2|\mathcal{A}_L| e^{-\alpha_L \phi(\varepsilon)} + d_L \|\hat{\pi}^k_F - \pi^{\text{SE}}_F\|_1, \end{aligned}$$

where we used the closedness of softmax and argmax [9], and the last term can be similarly bounded as follows:

$$\|\hat{\pi}^{k+1}_F - \pi^{\text{SSE}}_F\|_1 \leq \varepsilon + 2|\mathcal{A}_F| e^{-\alpha_F \phi(\varepsilon)} + d_F \|\hat{\pi}^k_L - \pi^{\text{SSE}}_L\|_1.$$

Combining the two recursive inequalities, we obtain:

$$\begin{aligned} \|\hat{\pi}^{k+1}_L - \pi^{\text{SSE}}_L\|_1 &\leq \varepsilon + 2|\mathcal{A}_L| e^{-\alpha_L \phi(\varepsilon)} + d_L \left(\varepsilon + 2|\mathcal{A}_F| e^{-\alpha_F \phi(\varepsilon)} + d_F \|\hat{\pi}^k_L - \pi^{\text{SSE}}_L\|_1\right) \\ &= (1 + d_L)\, \varepsilon + 2|\mathcal{A}_L| e^{-\alpha_L \phi(\varepsilon)} + 2d_L |\mathcal{A}_F| e^{-\alpha_F \phi(\varepsilon)} + d_L d_F \|\hat{\pi}^k_L - \pi^{\text{SSE}}_L\|_1. \end{aligned}$$

Unfolding the recursion over $k$ yields:

$$\begin{aligned} \|\hat{\pi}^{k+1}_L - \pi^{\text{SSE}}_L\|_1 \leq &\left((1 + d_L)\varepsilon + 2|\mathcal{A}_L| e^{-\alpha_L \phi(\varepsilon)} + 2d_L |\mathcal{A}_F| e^{-\alpha_F \phi(\varepsilon)}\right) \sum_{\kappa=0}^{k} (d_L d_F)^\kappa \\ &+ (d_L d_F)^{k+1} \|\hat{\pi}^0_L - \pi^{\text{SSE}}_L\|_1. \end{aligned}$$

Assuming $d_L d_F < 1$, and setting $\alpha_L = \alpha_F = \frac{\log(1/\varepsilon)}{\phi(\varepsilon)}$. At $K$, the bound is:

$$\begin{aligned}\|\hat{\pi}_L^K - \pi_L^{\text{SSE}}\|_1 &\leq \frac{(1+d_L)\varepsilon + 2|\mathcal{A}_L|e^{-\alpha_L\phi(\varepsilon)} + 2d_L|\mathcal{A}_F|e^{-\alpha_F\phi(\varepsilon)}}{1-d_L d_F} + (d_L d_F)^K \|\hat{\pi}_L^0 - \pi_L^{\text{SSE}}\|_1 \\ &\leq \frac{(1+d_L+2|\mathcal{A}_L|+2d_L|\mathcal{A}_F|)\varepsilon}{1-d_L d_F} + 2(d_L d_F)^K,\end{aligned}$$

where we used the fact that the $\ell_1$-norm between two distributions over a finite set is bounded by 2. To achieve $2(d_L d_F)^K \leq \varepsilon$, we need $K \geq \log_{d_L d_F}\left(\frac{\varepsilon}{2}\right)$, which bounds the error to $\|\hat{\pi}_L^K - \pi_L^{\text{SSE}}\|_1 \leq \left(\frac{1+d_L+2|\mathcal{A}_L|+2d_L|\mathcal{A}_F|}{1-d_L d_F} + 1\right)\varepsilon = O(\varepsilon)$. □

### 4.3 Regularization

Regularization is widely used in RL to promote stability, enhance exploration, and improve convergence rates [15, 16]. To facilitate the analysis, we define the diameter of the (finite) action space as the maximum distance between any two actions: $\text{diam}(\mathcal{A}_i) := \max_{a_i, a_i' \in \mathcal{A}_i} \|a_i - a_i'\|_1$. Without loss of generality, we normalize the action space so that $\text{diam}(\mathcal{A}_i) = 1$ for all $i \in \mathcal{I}$.

**Assumption 4.1** (Lipschitz Reward and Transition Kernel)**.** For each agent $i \in \mathcal{I}$, for any states $s_i \in \mathcal{S}_i, s_{-i} \in \mathcal{S}_{-i}$, and for any actions $a_i, a_i' \in \mathcal{A}_i, a_{-i}, a_{-i}' \in \mathcal{A}_{-i}$, the reward function $r_i$ and the transition kernel $P_i$ satisfy the condition that, there exists $d_r, d_P \geq 0$ such that

$$\begin{aligned}&|r_i(s_i, s_{-i}, a_i, a_{-i}) - r_i(s_i, s_{-i}, a_i', a_{-i}')| \\ &\quad\leq d_r(\|s_i - s_i'\|_1 + \|s_{-i} - s_{-i}'\|_1 + \|a_i - a_i'\|_1 + \|a_{-i} - a_{-i}'\|_1), \qquad (1) \\ &|P_i(s_i, a_i, a_{-i}) - r_i(s_i, a_i', a_{-i}')| \\ &\quad\leq d_P(\|s_i - s_i'\|_1 + \|s_{-i} - s_{-i}'\|_1 + \|a_i - a_i'\|_1 + \|a_{-i} - a_{-i}'\|_1), \qquad (2)\end{aligned}$$

and in addition, we assume $\gamma_i d_P/2 \in [0, 1]$.

Now we define the (regularized) best response mapping for each $i \in \mathcal{I}$ as $\text{BR}_i : \mathcal{S}_i \times \mathcal{S}_{-i} \times \mathcal{P}(\mathcal{A}_{-i}) \mapsto \mathcal{P}(\mathcal{A}_i)$. That is, follows:

$$\text{BR}_i^{\text{reg}}(s_i, s_{-i}, \pi_{-i}) := \text{argmax}_{\pi_i} V_i^{\text{reg}}(s_i, s_{-i}, \pi_i, \pi_{-i}). \qquad (3)$$

Then, the Lipschitz continuity condition can be established:

**Theorem 4.3** (Lipschitz Regularized Best Response)**.** *Under Assumptions 2.1 and 4.1, the best response mapping $BR_i^{reg}$ for each agent $i \in \mathcal{I}$ to $\mathcal{G}_S$ with regularized reward is Lipschitz with respect to the other agent's policy $\pi_{-i}$; that is, for any $\pi_L, \pi_L' \in \mathcal{P}(\mathcal{A}_L)$ and $\pi_F, \pi_F' \in \mathcal{P}(\mathcal{A}_F)$, there exist constants $d_L^{reg}, d_F^{reg} \geq 0$ such that,*

$$\|BR_L^{reg}(s_L, s_F, \pi_F) - BR_L^{reg}(s_L, s_F, \pi_F')\| \leq d_L^{reg}\|\pi_F - \pi_F'\|, \qquad (4)$$

$$\|BR_F^{reg}(s_F, s_L, \pi_L) - BR_F^{reg}(s_F, s_L, \pi_L')\| \leq d_F^{reg}\|\pi_L - \pi_L'\|, \qquad (5)$$

*where the constants are defined symmetrically in the form of:*

$$d_i^{reg} = \frac{d_r}{\rho}\left(1 + \frac{\gamma_i}{(1-\gamma_i)(1-\gamma_i d_P/2)} + \frac{\gamma_i d_P/2}{1-\gamma_i d_P/2}\right), \forall i \in \mathcal{I}. \qquad (6)$$

We first prove that adding a strongly concave regularization term to the value function can ensure the uniqueness as well as the continuity of the argmax operator. As the proof is symmetric for both agents, we drop the agent's index $i$ for simplicity and use superscript $^\dagger$ to denote the opponent's components. With a slight abuse of notations, in this proof only, we use $\mathbf{s} = (s, s^\dagger), \mathbf{a} = (a, a^\dagger)$ to represent the joint states and actions, and when necessary, we unpack the argument list. We use $\pi, \pi^\dagger$ to indicate the policies to the agent and its opponent, respectively. The regularized reward and value functions can be re-written concisely as follows:

$$r^{\text{reg}}(\mathbf{s}, \mathbf{a}) = r(\mathbf{s}, \mathbf{a}) + H(\pi), \qquad (7)$$

$$V^{\text{reg}}(\mathbf{s}, \pi, \pi^\dagger) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^{\text{reg}}(\mathbf{s}_t, \mathbf{a}_t) \Big| \mathbf{s}_0 = \mathbf{s}\right], \qquad (8)$$

where $H$ is a $\rho$-strongly concave function. The following lemma is first needed:

**Lemma 4.1.** *The $\text{argmax}_\pi V^{reg}$ admits a unique solution.*

*Proof.* We first argue that the expected reward $\mathbb{E}[r(\mathbf{s}, \mathbf{a})]$ is linear w.r.t. $\pi^\dagger$. In fact, the linearity is a direct consequence of the Lebesgue measure by viewing the distribution $\pi^\dagger$ as the measure function. Then the sum of a linear function and a $\rho$-strongly concave function preserves the $\rho$-strong concavity. Thus, the $\text{argmax}_\pi V^{\text{reg}}$ admits a unique solution. □

To proceed with our analysis, we state the following properties of the Fenchel conjugate, established in Lemma 15 of [17].

**Lemma 4.2** (Fenchel Conjugate Properties [17])**.** *Let $E = \mathbb{R}^m, m \geq 1$ with inner product $\langle \cdot, \cdot \rangle$. Let function $g : E \mapsto \mathbb{R}^+$ be a differentiable and $\rho$-strongly convex function with respect to some norm $\| \cdot \|$, where $\mathbb{R}^+ = \mathbb{R} \cup \{-\infty, \infty\}$. Let $X$ be its domain. The Fenchel conjugate $g^\star$ is defined as $g^\star(y) = \max_{x \in X} \langle x, y \rangle - g(x)$. Then 3 properties hold: (i) $g^\star$ is differentiable on $E$, (ii) $\nabla g^\star(y) = \text{argmax}_{x \in X} \langle x, y \rangle - g(x)$, and (iii) $g^\star$ is $\frac{1}{\rho}$-smooth with respect to $\| \cdot \|_\star$, the dual norm of $\| \cdot \|$. That is, for any $y_1, y_2 \in E$, $\|\nabla g^\star(y_1) - \nabla g^\star(y_2)\| \leq \frac{1}{\rho}\|y_1 - y_2\|_\star$.*

We need the property of $\ell_1$-norm of distributions on the set of probability distribution over finite sets.

**Lemma 4.3.** *Suppose that there exists a real-valued function $f$ on a finite set $E$. For any two probability distributions $\psi_1, \psi_2$, we have*

$$\Big| \sum_{x \in E} f(x)\psi_1(x) - \sum_{x \in E} f(x)\psi_2(x) \Big| \leq \frac{\max_{x \in E} f(x) - \min_{x \in E} f(x)}{2} \|\psi_1 - \psi_2\|_1, \tag{9}$$

*Proof.* We first have that $\sum_x (\psi_1(x) - \psi_2(x)) = 0$ and hence for any constant $c \in \mathbb{R}$, one has $\sum_x c(\psi_1(x) - \psi_2(x)) = 0$, then $\Big| \sum_{x \in E} f(x)\psi_1(x) - \sum_{x \in E} f(x)\psi_2(x) \Big| = \Big| \sum_{x \in E} (f(x) - c)(\psi_1(x) - \psi_2(x)) \Big| \leq \sum_{x \in E} |f(x) - c| \cdot |\psi_1(x) - \psi_2(x)| \leq \max_{x \in E} |f(x) - c| \cdot \sum_{x \in E} |\psi_1(x) - \psi_2(x)|$. By choosing $c = (\max_{x \in E} f(x) + \min_{x \in E} f(x))/2$, we get (9). □

The proof to Theorem 4.3 is adapted from [18] in which they proved the argmax operator is Lipschitz continuous with respect to the MF in their MF game setting. Our proof replaces the MF with opponent's policy, and we will show that our result matches theirs.

*Proof.* We first define the opponent-policy averaged reward and transition as follows:

$$\bar{r}^{\text{REG}^\dagger}(\mathbf{s}, a, \pi^\dagger) := \mathbb{E}_{a^\dagger \sim \pi^\dagger}[r^{\text{reg}}(\mathbf{s}, a, a^\dagger)], \text{ and } \bar{P}^\dagger(\mathbf{s}, a, \pi^\dagger) := \mathbb{E}_{a^\dagger \sim \pi^\dagger}[P(\mathbf{s}, a, a^\dagger)].$$

It is easy to show that both $\bar{r}^{\text{REG}^\dagger}$ and $\bar{P}^\dagger$ are Lipschitz continuous in $\pi^\dagger$ under Assumption 4.1 with the same constants $d_r, d_P \geq 0$ respectively. For any $\pi^\dagger, \pi^{\dagger\prime} \in \mathcal{P}(\mathcal{A}^\dagger)$,

$$\begin{aligned}
&|\bar{r}^{\text{REG}^\dagger}(\mathbf{s}, a, \pi^\dagger) - \bar{r}^{\text{REG}^\dagger}(\mathbf{s}, a, \pi^{\dagger\prime})| = \Big| \mathbb{E}_{a^\dagger \sim \pi^\dagger}[r^{\text{reg}}(\mathbf{s}, a, a^\dagger)] - \mathbb{E}_{a^{\dagger\prime} \sim \pi^{\dagger\prime}}[r^{\text{reg}}(\mathbf{s}, a, a^{\dagger\prime}) \Big| \\
&= \mathbb{E}_{(a^\dagger, a^{\dagger\prime}) \sim \text{Coupling}(\pi^\dagger, \pi^{\dagger\prime})} \Big| r(\mathbf{s}, a, a^\dagger) - r(\mathbf{s}, a, a^{\dagger\prime}) \Big| \leq \mathbb{E}_{(a^\dagger, a^{\dagger\prime}) \sim \text{Coupling}(\pi^\dagger, \pi^{\dagger\prime})} d_r \|a^\dagger - a^{\dagger\prime}\|_1
\end{aligned}$$

As this works for any coupling of $(\pi^\dagger, \pi^{\dagger\prime})$, we can pick the optimal coupling that achieves the $\ell_1$-Wasserstein distance, defined as $W_1(\pi^\dagger, \pi^{\dagger\prime}) = \inf_{\nu \in \text{Coupling}(\pi^\dagger, \pi^{\dagger\prime})} \int_{\mathcal{A}^\dagger \times \mathcal{A}^\dagger} \|a^\dagger - a^{\dagger\prime}\|_1 d\nu(a^\dagger, a^{\dagger\prime})$, in which the infimum can be replaced by minimum when the coupling space is compact. Indeed, when the action space is discrete and finite in our case, the compactness is guaranteed. Then,

$$|\bar{r}^{\text{REG}^\dagger}(\mathbf{s}, a, \pi^\dagger) - \bar{r}^{\text{REG}^\dagger}(\mathbf{s}, a, \pi^{\dagger\prime})| \leq d_r \mathbb{E}_{a \sim \pi}[W_1(\pi^\dagger, \pi^{\dagger\prime})] = d_r W_1(\pi^\dagger, \pi^{\dagger\prime}) \leq d_r \|\pi^\dagger - \pi^{\dagger\prime}\|_1.$$

The last inequality can be established by noticing that for any optimal coupling $\nu^{\text{TV}}$ that attains the minimum of the total variance distance, which is defined as $d_{\text{TV}}(\pi^\dagger, \pi^\dagger) = \nu^{\text{TV}}(a^\dagger \neq a^{\dagger\prime}) :=$

$\inf_{\nu \in \text{Coupling}(\pi^\dagger, \pi^{\dagger\prime})} \nu(a^\dagger \neq a^{\dagger\prime}) = \frac{1}{2}\|\pi^\dagger - \pi^{\dagger\prime}\|_1$. The following condition must be satisfied with the assumption that $\text{diam}(\mathcal{A}^\dagger) = 1$ has been normalized:

$$
\begin{aligned}
&W_1(\pi^\dagger, \pi^{\dagger\prime}) \leq \mathbb{E}_{(a^\dagger, a^{\dagger\prime}) \sim \nu^{\text{TV}}}[\|a^\dagger - a^{\dagger\prime}\|_1] \\
&= \nu^{\text{TV}}(a^\dagger = a^{\dagger\prime}) \mathbb{E}[\|a^\dagger - a^{\dagger\prime}\|_1 \mid a^\dagger = a^{\dagger\prime}] + \nu^{\text{TV}}(a^\dagger \neq a^{\dagger\prime}) \mathbb{E}[\|a^\dagger - a^{\dagger\prime}\|_1 \mid a^\dagger \neq a^{\dagger\prime}] \\
&\leq \text{diam}(\mathcal{A}^\dagger) \nu^{\text{TV}}(a^\dagger \neq a^{\dagger\prime}) = \frac{1}{2}\|\pi^\dagger - \pi^{\dagger\prime}\|_1 \leq \|\pi^\dagger - \pi^{\dagger\prime}\|_1.
\end{aligned}
$$

We immediately have that $|\bar{r}^{\text{REG}^\dagger}(\mathbf{s}, a, \pi^\dagger) - \bar{r}^{\text{REG}^\dagger}(\mathbf{s}, a, \pi^{\dagger\prime})| \leq d_r \|\pi^\dagger - \pi^{\dagger\prime}\|_1$. The proof to $\bar{P}^\dagger$ being $d_P$-Lipschitz with respect to $\pi^\dagger$ is symmetric. Now we can look at the learning problem. Since at different rounds, we solve a different RL problem, we are essentially dealing with different $Q$-functions. we define

$$
Q^{\pi^\dagger}(\mathbf{s}, a) = \bar{r}^{\text{REG}^\dagger}(\mathbf{s}, a, \pi^\dagger) + \gamma \sum_{\mathbf{s}' \in \mathcal{S}} Q^{*, \pi^\dagger}(\mathbf{s}') \bar{P}^\dagger(\mathbf{s}' | \mathbf{s}, a, \pi^\dagger), \tag{10}
$$

where $Q^{*, \pi^\dagger}(\mathbf{s}) = \max_{a \in \mathcal{A}} Q^{\pi^\dagger}(\mathbf{s}, a)$ for all $\mathbf{s}$. The next is to prove that $Q^{*, \pi^\dagger}$ is $d_Q$-Lipschitz continuous with respect to the states $\mathbf{s}$, where $d_Q = \frac{d_r}{1 - \gamma d_P/2}$. Define $T_{\pi^\dagger}$ as the Bellman operator for the problem with $\pi^\dagger$. We can rewrite $Q^{*, \pi^\dagger}$ in the form of $T_{\pi^\dagger}$ as follows

$$
Q^{*, \pi^\dagger}(\mathbf{s}) = \max_{a \in \mathcal{A}} \left\{ \bar{r}^{\text{REG}^\dagger}(\mathbf{s}, a, \pi^\dagger) + \gamma \sum_{\mathbf{s}' \in \mathcal{S}} Q^{*, \pi^\dagger}(\mathbf{s}') \bar{P}^\dagger(\mathbf{s}' | \mathbf{s}, a, \pi^\dagger) \right\} = T_{\pi^\dagger} Q^{*, \pi^\dagger}(\mathbf{s}), \tag{11}
$$

which is the Bellman optimality condition. It is known that the operator forms a $\gamma$-contraction mapping. Start with any $Q$, and apply $T_{\pi^\dagger}$, by Banach fixed point theorem, $\lim_{n \to \infty} T^n_{\pi^\dagger} Q \to Q^{*, \pi^\dagger}$. Choose the initial $Q$ to be $d_K$-Lipschitz where $d_K < d_r$, then $Q/d_K$ is 1-Lipschitz. For any $\mathbf{s}_1, \mathbf{s}_2$, the following holds

$$
\begin{aligned}
|T_{\pi^\dagger} Q(\mathbf{s}_1) - T_{\pi^\dagger} Q(\mathbf{s}_2)| &\leq \max_{a \in \mathcal{A}} \Big\{ |\bar{r}^{\text{REG}^\dagger}(\mathbf{s}_1, a, \pi^\dagger) - \bar{r}^{\text{REG}^\dagger}(\mathbf{s}_2, a, \pi^\dagger)| \\
&\qquad + \gamma \Big| \sum_{s' \in \mathcal{S}} Q(\mathbf{s}') \bar{P}^\dagger(\mathbf{s}' | \mathbf{s}_1, a, \pi^\dagger) - \sum_{s' \in \mathcal{S}} Q(\mathbf{s}') \bar{P}^\dagger(\mathbf{s}' | \mathbf{s}_2, a, \pi^\dagger) \Big| \Big\} \\
&\leq \max_{a \in \mathcal{A}} \Big\{ d_r \|\mathbf{s}_1 - \mathbf{s}_2\|_1 + \gamma d_K \Big| \sum_{\mathbf{s}' \in \mathcal{S}} \frac{Q(\mathbf{s}')}{d_K} \bar{P}^\dagger(\mathbf{s}' | \mathbf{s}_1, a, \pi^\dagger) - \sum_{\mathbf{s}' \in \mathcal{S}} \frac{Q(\mathbf{s}')}{d_K} \bar{P}^\dagger(\mathbf{s}' | \mathbf{s}_2, a, \pi^\dagger) \Big| \Big\} \\
&\leq \left( d_r + \gamma \frac{d_K d_P}{2} \right) \|\mathbf{s}_1 - \mathbf{s}_2\|_1.
\end{aligned}
$$

Inductively, we have for all $n \geq 1$, it holds that, $|T^n_{\pi^\dagger} Q(\mathbf{s}_1) - T^n_{\pi^\dagger} Q(\mathbf{s}_2)| \leq \left( d_r \sum_{k=0}^{n-1} \left( \frac{\gamma d_P}{2} \right)^k + d_K \left( \frac{\gamma d_P}{2} \right)^n \right) \|\mathbf{s}_1 - \mathbf{s}_2\|_1 \leq d_r \sum_{k=0}^{n} \left( \frac{\gamma d_P}{2} \right)^k \|\mathbf{s}_1 - \mathbf{s}_2\|_1 \leq \frac{d_r}{1 - \gamma d_P/2} \|\mathbf{s}_1 - \mathbf{s}_2\|_1$, where the second inequality is a result of $d_K < d_r$, and the third inequality uses the fact that $\gamma d_P / 2 \in [0, 1]$ with which the geometric series is bounded above. Hence, $T^n_{\pi^\dagger}$ is $\frac{d_r}{1 - \gamma d_P/2}$-continuous for all $n$, which holds true when $n \to \infty$, where $T^n_{\pi^\dagger} Q \to Q^{*, \pi^\dagger}(\mathbf{s})$. We then set $d_Q = \frac{d_r}{1 - \gamma d_P/2}$ for notation easiness. We now claim that $Q^{*, \pi^\dagger}$ is $d_0$-Lipschitz continuous with respect to $\mathbf{s}$, where

$d_0 = \frac{1}{1-\gamma}\Big(d_r + \gamma\frac{d_P d_Q}{2}\Big)$. For any $\pi_1^\dagger, \pi_2^\dagger \in \mathcal{P}(\mathcal{A}^\dagger)$, we have

$$\begin{aligned}
\|Q^\star_{\pi_1^\dagger} - Q^\star_{\pi_2^\dagger}\|_\infty &= \max_{s,a}\Big|\bar{r}^{\text{REG}^\dagger}(\mathbf{s},a,\pi_1^\dagger) + \gamma\sum_{s'\in\mathcal{S}} Q^\star_{\pi_1^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_1^\dagger) \\
&\qquad - \bar{r}^{\text{REG}^\dagger}(\mathbf{s},a,\pi_2^\dagger) - \gamma\sum_{s'\in\mathcal{S}} Q^\star_{\pi_2^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_2^\dagger)\Big| \\
&\le |\bar{r}^{\text{REG}^\dagger}(\mathbf{s},a,\pi_1^\dagger) - \bar{r}^{\text{REG}^\dagger}(\mathbf{s},a,\pi_2^\dagger)| \\
&\qquad + \gamma\Big|\sum_{s'\in\mathcal{S}} Q^\star_{\pi_1^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_1^\dagger) - \sum_{s'\in\mathcal{S}} Q^\star_{\pi_1^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_2^\dagger)\Big| \\
&\qquad + \gamma\Big|\sum_{s'\in\mathcal{S}} Q^\star_{\pi_1^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_2^\dagger) - \sum_{s'\in\mathcal{S}} Q^\star_{\pi_2^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_2^\dagger)\Big| \\
&\le d_r\|\pi_1^\dagger - \pi_2^\dagger\|_1 + \gamma\frac{d_P d_Q}{2}\|\pi_1^\dagger - \pi_2^\dagger\|_1 + \gamma\|Q^\star_{\pi_1^\dagger} - Q^\star_{\pi_2^\dagger}\|_\infty,
\end{aligned}$$

where the first term follows the Lipschitz assumption on the reward and the last term uses the fact that $\bar{P}^\dagger$ is probability. The second term can be bounded as follows. Notice that for any $\pi^\dagger$, $Q^{*,\pi^\dagger}$ is $d_Q$-Lipschitz continuous implies $Q^{*,\pi^\dagger}/d_Q$ is 1-Lipschitz continuous with respect to $\mathbf{s}$. Then,

$$\begin{aligned}
&\Big|\sum_{s'\in\mathcal{S}} Q^\star_{\pi_1^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_1^\dagger) - \sum_{s'\in\mathcal{S}} Q^\star_{\pi_1^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_2^\dagger)\Big| \\
&\quad = d_Q\Big|\sum_{s'\in\mathcal{S}} \frac{Q^\star_{\pi_1^\dagger}(\mathbf{s}')}{d_Q}\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_1^\dagger) - \sum_{s'\in\mathcal{S}} \frac{Q^\star_{\pi_1^\dagger}(\mathbf{s}')}{d_Q}\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_2^\dagger)\Big| \\
&\quad \le \frac{d_Q}{2}\|\bar{P}^\dagger(\mathbf{s},a,\pi_1^\dagger) - \bar{P}^\dagger(\mathbf{s},a,\pi_2^\dagger)\|_1 \le \frac{d_P d_Q}{2}\|\pi_1^\dagger - \pi_2^\dagger\|_1,
\end{aligned}$$

where we use equation (9) and Lipschitz continuity on the transition kernel. Then by rearranging the terms, we obtain that $\|Q^\star_{\pi_1^\dagger} - Q^\star_{\pi_2^\dagger}\|_\infty \le d_0\|\pi_1^\dagger - \pi_2^\dagger\|_1$ where $d_0 = \frac{1}{1-\gamma}\Big(d_r + \gamma\frac{d_P d_Q}{2}\Big)$. Equation (10) can be rewritten as follows:

$$Q^{\pi^\dagger}(\mathbf{s},a) = \bar{r}^{\text{REG}^\dagger}(\mathbf{s},a,\pi^\dagger) + \gamma\sum_{s'\in\mathcal{S}} Q^{*,\pi^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi^\dagger) - H(\pi) = \langle q_{\pi^\dagger,\mathbf{s}}, a\rangle - H(\pi), \quad (12)$$

where $q_{\pi^\dagger,\mathbf{s}} = \bar{r}^{\text{REG}^\dagger}(\mathbf{s},\cdot,\pi^\dagger) + \gamma\sum_{s'\in\mathcal{S}} Q^{*,\pi^\dagger}(\mathbf{s}')P(\mathbf{s}'|\mathbf{s},\cdot,\pi^\dagger)$ for any $\mathbf{s}$. We now prove that is $\Big(d_r + \gamma d_0 + \gamma\frac{d_P d_Q}{2}\Big)$-Lipschtiz continuous with respect to $\pi^\dagger$. Indeed, one has

$$\begin{aligned}
\|q_{\pi_1^\dagger,\mathbf{s}} - q_{\pi_2^\dagger,\mathbf{s}}\|_\infty &= \max_{a\in\mathcal{A}}\Big|\bar{r}^{\text{REG}^\dagger}(\mathbf{s},a,\pi_1^\dagger) + \gamma\sum_{s'\in\mathcal{S}} Q^{*,\pi^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_1^\dagger) \\
&\qquad - \bar{r}^{\text{REG}^\dagger}(\mathbf{s},a,\pi_2^\dagger) - \gamma\sum_{s'\in\mathcal{S}} Q^{*,\pi^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_2^\dagger)\Big| \\
&\le d_r\|\pi_1^\dagger - \pi_2^\dagger\|_1 + \gamma\max_{a\in\mathcal{A}}\Big|\sum_{s'\in\mathcal{S}} Q^\star_{\pi_1^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_1^\dagger) - \sum_{s'\in\mathcal{S}} Q^\star_{\pi_1^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_2^\dagger)\Big| \\
&\qquad + \gamma\max_{a\in\mathcal{A}}\Big|\sum_{s'\in\mathcal{S}} Q^\star_{\pi_1^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_2^\dagger) - \sum_{s'\in\mathcal{S}} Q^\star_{\pi_2^\dagger}(\mathbf{s}')\bar{P}^\dagger(\mathbf{s}'|\mathbf{s},a,\pi_2^\dagger)\Big| \\
&\le d_r\|\pi_1^\dagger - \pi_2^\dagger\|_1 + \gamma\|Q^\star_{\pi_1^\dagger} - Q^\star_{\pi_2^\dagger}\|_\infty + \gamma\frac{d_P d_Q}{2}\|\pi_1^\dagger - \pi_2^\dagger\|_1 \\
&= \Big(d_r + \gamma d_0 + \gamma\frac{d_P d_Q}{2}\Big)\|\pi_1^\dagger - \pi_2^\dagger\|_1.
\end{aligned}$$

We now apply Lemma 4.2. For any $s\in\mathcal{S}$, we write $\text{BR}^{\text{reg}}(\mathbf{s},\pi^\dagger) = \nabla H^\star(q_{\pi^\dagger,\mathbf{s}})$ where $H^\star$ is the Fenchel conjugate of $H$. Then,

$$\|\text{BR}^{\text{reg}}(\mathbf{s},\pi_1^\dagger) - \text{BR}^{\text{reg}}(\mathbf{s},\pi_2^\dagger)\|_1 \le \frac{1}{\rho}\|q_{\pi_1^\dagger,\mathbf{s}} - q_{\pi_2^\dagger,\mathbf{s}}\|_\infty = \frac{d_r + \gamma d_0 + \gamma d_P d_Q/2}{\rho}\|\pi_1^\dagger - \pi_2^\dagger\|_1.$$

The argmax is therefore Lipschitz with constant $\frac{d_r+\gamma d_0+\gamma d_P d_Q/2}{\rho}$. Then by substituting $d_0, d_Q$ and bringing back the agent index $i$, we get

$$d_i^{\text{reg}} = \frac{d_r}{\rho}\left(1 + \frac{\gamma_i}{(1-\gamma_i)(1-\gamma_i d_P/2)} + \frac{\gamma_i d_P/2}{1-\gamma_i d_P/2}\right), \forall i \in \mathcal{I}. \tag{13}$$

□

### 4.4 Stationary Stackelberg Markov Equilibrium with Mean-Field Followers

In this section, we present the proof of Theorem 2.3 and introduce an algorithm that iteratively computes an SS-MFE.

#### 4.4.1 Assumptions of Theorem 2.3

To establish the existence and uniqueness of an SS-MFE, we adopt the following assumption:

**Assumption 4.2** (Uniqueness of Best Response and MF Update)**.** For each agent $i \in \mathcal{I}$, for any $s_i \in \mathcal{S}_i, s_{-i} \in \mathcal{S}_{-i}, \pi_{-i} \in \mathcal{P}(\mathcal{A}_{-i})$, and for any follower's MF $\mu_F \in \mathcal{P}(\mathcal{S}_F \times \mathcal{A}_F)$, agent $i$'s best response function $\text{BR}_i(s_i, s_{-i}, \pi_{-i}, \mu_F)$ admits a unique solution. In addition, the MF update map $\Gamma(\mu_F, \pi_L, \pi_F)$ also returns a unique solution.

**Assumption 4.3** (Lipschitz Best Responses)**.** There exist constants $d_L^F, d_L^\mu, d_F^L, d_F^\mu, d_\mu^L, d_\mu^F, d_\mu^\mu \geq 0$ such that for any admissible leader's policies $\pi_L, \pi_L' \in \mathcal{P}(\mathcal{A}_L)$, the follower's policies $\pi_F, \pi_F' \in \mathcal{P}(\mathcal{A}_F)$, and follower's MF $\mu_F, \mu_F' \in \mathcal{P}(\mathcal{S}_F \times \mathcal{A}_F)$:

$$\sup_{s_F, s_L} \|\text{BR}_F(s_F, s_L, \pi_L, \mu_F) - \text{BR}_F(s_F, s_L, \pi_L', \mu_F')\|_1 \leq d_F^L \|\pi_L - \pi_L'\|_1 + d_F^\mu \|\mu - \mu'\|_1, \tag{14}$$

$$\|\Gamma(\mu_F, \pi_L, \pi_F) - \Gamma(\mu_F', \pi_L', \pi_F')\|_1 \leq d_\mu^\mu \|\mu_F - \mu_F'\|_1 + d_\mu^L \|\pi_L - \pi_L'\|_1 + d_\mu^F \|\pi_F - \pi_F'\|_1, \tag{15}$$

$$\sup_{s_L, s_F} \|\text{BR}_L(s_L, s_F, \pi_F, \mu_F) - \text{BR}_L(s_L, s_F, \pi_F', \mu_F')\|_1 \leq d_L^F \|\pi_F - \pi_F'\|_1 + d_L^\mu \|\mu - \mu'\|_1. \tag{16}$$

#### 4.4.2 Proof of Theorem 2.3 – Existence and Uniqueness of SS-MFE under Regularization

*Proof.* We define the map $\text{BR}_{F\mu} : \mathcal{S}_F \times \mathcal{S}_L \times \mathcal{P}(\mathcal{A}_L) \mapsto \mathcal{P}(\mathcal{A}_F) \times \mathcal{P}(\mathcal{S}_F \times \mathcal{A}_F)$, which is simply a composite update map from $BR_F$ and $\Gamma$; that is, at the outer iteration $k$, given current states $s_F, s_L$ and leader's policy $\pi_L^k$, the inner iteration returns $\text{BR}_{F\mu}(s_F, s_L, \pi_L^k) = (\pi_F^{k*}, \mu_F^{k*})$. Fix a leader policy $\pi_L \in \mathcal{P}(\mathcal{A}_L)$ and states $s_L \in \mathcal{S}_L, s_F \in \mathcal{S}_F$. We first show that the mapping $\text{BR}_{F\mu}$ returns a unique solution by showing that $\Gamma$ is contractive, then we show that $\text{BR}_{F\mu}$ forms a contractive mapping. Consider any pair of follower policies $\pi_F, \pi_F' \in \mathcal{P}(\mathcal{A}_F)$ and mean-field distributions $\mu_F, \mu_F' \in \mathcal{P}(\mathcal{S}_F \times \mathcal{A}_F)$ that satisfy $\pi_F = \text{BR}_F(s_F, s_L, \pi_L, \mu_F)$ and $\pi_F' = \text{BR}_F(s_F, s_L, \pi_L, \mu_F')$, we have:

$$\begin{aligned}
&\|\Gamma(\mu_F, \pi_L, \pi_F) - \Gamma(\mu_F', \pi_L', \pi_F')\|_1 \leq d_\mu^\mu \|\mu_F - \mu_F'\|_1 + d_\mu^F \|\pi_F - \pi_F'\|_1 \\
&\leq d_\mu^\mu \|\mu_F - \mu_F'\|_1 + d_\mu^F \|\text{BR}_F(s_F, s_L, \pi_L, \mu_F) - \text{BR}_F(s_F, s_L, \pi_L, \mu_F')\|_1 \\
&\leq (d_\mu^\mu + d_\mu^F d_F^\mu) \|\mu_F - \mu_F'\|_1.
\end{aligned}$$

As $d_\mu^\mu + d_\mu^F d_F^\mu \in [0, 1)$, $\Gamma$ forms a contractive mapping by Banach's fixed point theorem. And since $\text{BR}_F$ returns unique solution, we conclude that the follower's side converges to a unique fixed point. For any $\pi_L, \pi_L'$, let the corresponding follower's fixed points be $(\pi_F^*, \mu_F^*) = \text{BR}_{F\mu}(s_F, s_L, \pi_L)$, and $(\pi_F^{*'}, \mu_F^{*'}) = \text{BR}_{F\mu}(s_F, s_L, \pi_L')$. Then, the following holds:

$$\begin{aligned}
&\|\text{BR}_{F\mu}(s_F, s_L, \pi_L) - \text{BR}_{F\mu}(s_F, s_L, \pi_L')\| = \|\pi_F^* - \pi_F^{*'}\|_1 + \|\mu_F^* - \mu_F^{*'}\|_1 \\
&= \|\text{BR}_F(s_F, s_L, \pi_L, \mu_F^*) - \text{BR}_F(s_F, s_L, \pi_L', \mu_F^{*'})\|_1 + \|\Gamma(\mu_F^*, \pi_L, \pi_F^*) - \Gamma(\mu_F^{*'}, \pi_L', \pi_F^{*'})\|_1 \\
&\leq (d_F^L + d_\mu^L)\|\pi_L - \pi_L'\|_1 + (d_F^\mu + d_\mu^\mu)\|\mu_F^* - \mu_F^{*'}\|_1 + d_\mu^F \|\pi_F^* - \pi_F^{*'}\|_1 \\
&\leq (d_F^L + d_\mu^L)\|\pi_L - \pi_L'\|_1 + (d_F^\mu + d_\mu^\mu + d_\mu^F)\left(\|\pi_F^* - \pi_F^{*'}\|_1 + \|\mu_F^* - \mu_F^{*'}\|_1\right).
\end{aligned}$$

By rearranging the term, we get

$$\|\text{BR}_{F\mu}(s_F, s_L, \pi_L) - \text{BR}_{F\mu}(s_F, s_L, \pi'_L)\|_1 = \|\pi^*_F - \pi^{*'}_F\|_1 + \|\mu^*_F - \mu^{*'}_F\|_1$$
$$\leq \frac{d^L_F + d^L_\mu}{1 - (d^\mu_F + d^\mu_\mu + d^F_\mu)} \|\pi_L - \pi'_L\|_1.$$

Finally, the leader's best response satisfies the following:

$$\|\text{BR}_L(\cdot, \pi^*_F, \mu^*_F) - \text{BR}_L(\cdot, \pi^{*'}_F, \mu^{*'}_F)\|_1 \leq d^F_L \|\pi^*_F - \pi^{*'}_F\|_1 + d^\mu_L \|\mu^* - \mu^{*'}\|_1$$
$$\leq \frac{d^L_F + d^L_\mu}{1 - (d^\mu_F + d^\mu_\mu + d^F_\mu)} \max\{d^F_L, d^\mu_L\} \|\pi_L - \pi'_L\|_1.$$

Because $\frac{d^L_F + d^L_\mu}{1 - (d^\mu_F + d^\mu_\mu + d^F_\mu)} \max\{d^F_L, d^\mu_L\} \in [0, 1)$, $\text{BR}_L$ forms a contractive mapping by Banach's fixed point theorem. As a result, there exists a unqiue SS-MFE to $\mathcal{G}_{\text{MF}}$. □

#### 4.4.3 Algorithm 1 – RL-Framework for Finding an SS-MFE

We now present a RL procedure for computing an SS-MFE. At each *outer* iteration, the leader updates its policy based on the aggregate follower response, while the *inner* loop computes the consistent mean-field and best response for the followers. The complete procedure is outlined below.

**Algorithm 1:** An RL to Single-Leader-MF-Follower Stackelberg Games

**Input:** Initial states $s^0_L, s^0_F$, leader's policy $\pi^0_L$, initial follower's MF $\mu^0_F$, tolerance `tol`, RL algorithms $\texttt{Alg}_L, \texttt{Alg}_F$

**for** *Iteration* $k = 0, 1, 2, \cdots$ **do**

- Leader takes action $a^k_L \sim \pi^k_L(\cdot|s^k_L)$ ;
- Set $\mu^{k,0}_F = \mu^k_F$ ;
- **for** *Iteration* $\tau = 0, 1, 2, \cdots$ **do**
  - Follower learns its best response policy $\pi^{k,\tau}_F = \text{BR}_F(s^k_F, s^k_L, \pi^k_L, \mu^{k,\tau}_F)$ through $\texttt{Alg}_F$ ;
  - Follower's MF updates as $\mu^{k,\tau+1}_F = \Gamma(\mu^{k,\tau}_F, \pi^k_L, \pi^{k,\tau}_F)$ ;
  - If $\|\mu^{k,\tau+1}_F - \mu^{k,\tau}_F\|_1 \leq \texttt{tol}$, set $(\pi^k_F, \mu^k_F) = (\pi^{k,\tau}_F, \mu^{k,\tau}_F)$ and exit the inner loop.
- **end**
- Follower takes action $a^k_F \sim \pi^k_F(\cdot|s^k_F)$ ;
- Leader learns its best response policy $\pi^{k+1}_L = \text{BR}_L(s^k_L, s^k_F, \pi^k_F, \mu^k_F)$ through $\texttt{Alg}_L$ ;
- State transition $s^{k+1}_L \sim P_L(s^k_L, a^k_L, a^k_F, \mu^k_F)$, $s^{k+1}_F \sim P_F(s^k_F, a^k_F, a^k_L, \mu^k_F)$ ;
- If $\|\pi^{k+1}_L - \pi^k_L\|_1 \leq \texttt{tol}$, exit the outer loop.

**end**

Return $(\pi^{\text{SE}}_L, \pi^{\text{SE}}_F, \mu^{\text{SE}}_F) = (\pi^k_L, \pi^k_F, \mu^k_F)$ as the SS-MFE.

### 4.5 Numerical Experiment Specification and Results

#### 4.5.1 Input Data and Hyper-parameters

Our numerical simulation's data and code can be found at `https://anonymous.4open.science/r/StackelbergGame-B592` and also in the supplemental materials. To reproduce our results, we require Python 3.10.11. All necessary packages are included in the `requirement.txt` file. The main packages used are: Gymnasium (version 1.0.0, [19]) for environment setting; Stable-Baselines3 (version 2.3.2, [20]) for RL; and Pyomo (version 6.7.2, [21, 22]) for solving the economic dispatch linear programming problem. We use a stylized 3-bus power system as the test case. The input specifications for the bus nodes (including demographic data of prosumers and consumers), transmission lines, and grid-level generators are provided in Tables 1, 2, and 3, respectively. The numerical experiment uses PPO as the RL algorithm. The training specification is listed in Table 4.

Table 1: Bus Node Data

| | Bus Node Name | | |
|---|---|---|---|
| **Parameter** | **b1** | **b2** | **b3** |
| P Load (kW) | 110.0 | 110.0 | 95.0 |
| Q Load (kVar) | 40.0 | 40.0 | 50.0 |
| Max Voltage (p.u.) | 1.1 | 1.1 | 1.1 |
| Min Voltage (p.u.) | 0.9 | 0.9 | 0.9 |
| Voltage Magnitude | 1.1 | 0.92617 | 0.9 |
| Voltage Angle | 0.0 | 7.25883 | -17.2671 |
| Base KV | 345 | 345 | 345 |
| Prosumer Population | 1,000 | 500 | 300 |
| Energy Storage Capacity (kWh) | 30 | 60 | 100 |
| Energy Storage One-way Efficiency | 0.8 | 0.8 | 0.8 |
| Prosumer Income/household (US$) | 25,000 | 45,000 | 65,000 |
| Consumer Population | 3,000 | 3,000 | 3,000 |
| Consumer Income/household (US$) | 15,000 | 15,000 | 15,000 |

Table 2: Transmission Line Data

| | Transmission Line Name | | |
|---|---|---|---|
| **Parameter** | **l1** | **l2** | **l3** |
| Source Bus | b1 | b3 | b1 |
| Target Bus | b3 | b2 | b2 |
| Reactance ($\Omega$) | 0.065 | 0.025 | 0.042 |
| Susceptance (S) | 0.62 | 0.75 | 0.9 |
| Normal Flow Limit (MW) | 100 | 100 | 100 |

Table 3: Grid-Level Generator Data

| | Grid-Level Generator Name | | | |
|---|---|---|---|---|
| **Parameter** | **g1** | **g2** | **solar** | **solar2** |
| Bus | b1 | b2 | b3 | b1 |
| Fuel Type | Oil | Oil | Solar | Solar |
| Cost Curve Coefficients* | [0.2, 5.0, 0.0] | [0.2, 4.0, 0.0] | Free | Free |
| Max Production (MW) | 2000.0 | 1500.0 | 30.0 | 30.0 |

* The cost curve is represented as an array, where the first entry is the quadratic coefficient, the second is the linear coefficient, and the third is the constant term. For an array $[a, b, c]$, the cost function is $C(p) = ap^2 + bp + c$, where $p$ is amount of energy consumption in MWh.

Table 4: Hyper-parameters for PPO Agents

| **Hyperparameter** | **Aggregators** | **Utility Company** |
|---|---|---|
| Learning Rate | 0.0003 | 0.0003 |
| Discount Factor ($\gamma$) | 0.9999 | 0.9999 |
| Entropy Coefficient | 0.01 | 0.01 |
| Batch Size | 128 | 128 |
| Number of Epochs | 10 | 10 |
| Steps per Update | 1200 | 1200 |
| Clip Range | 0.2 | 0.2 |
| Policy Network † | [18, 36] | [24, 36] |
| Value Network ° | [18, 36] | [24, 36] |
| Training length | 2000 | 2000 |

†,° All policy and value networks are fully connected neural networks. Each array lists the number of neurons at each hidden layer.

### 4.5.2 Input Data of Solar and Demand Shapes

We set each day to be of 8 time steps, each of which represents a 3-hour interval. Figure 3 shows the input solar capacity shape from [23] and energy demand shapes for both prosumers and consumers at each timestep adapted from [24]. Let $\Delta(a, b, c)$ denote a triangular distribution with lower limit $a$, upper limit $b$, and mode $c$. In our simulation, we assume each consumer/prosumer's demand and solar profile follow the average shapes shown in Figure 3, scaled by a random factor drawn from the triangular distribution $\Delta(0.8, 1.2, 1)$. This introduces variability across agents while preserving the overall profile shape. All data is scaled relative to the average individual storage capacity across all prosumers and consumers, computed using Table 1. We assume each consumer has a reference storage capacity of 10kWh. The demand input shows the energy consumed in each time step. The shapes show the total consumption for consumers, and net demand after subtracting their solar generation for prosumers.

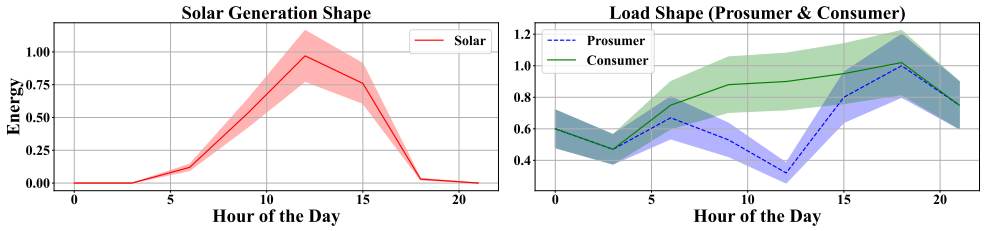


Figure 3: Input shapes for solar capacity shape (left, data adapted from [23]) and load demand shapes for prosumers and consumers (right, data adapted from [24]). All data is scaled to be with respect to the average storage capacity. The shadow areas indicate the noise bound of each shape.

### 4.5.3 Follower's Learning Result – Price Volatility and Prosumer Net Demand Shape

To better measure the impacts of energy storage coupled with the RL algorithms on locational marginal price (LMP) volatility, we adopt *incremental mean volatility* (IMV) from [25] as the metric. For a sequence of LMPs $\{\text{LMP}_t\}_{t=1}^{\infty}$, the IMV is defined as $\text{IMV} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \left|\text{LMP}_{t+1} - \text{LMP}_t\right|$. Figure 4 shows the IMV of the last 3 days between the two scenarios: with storage and RL, and without storage or RL. Results indicate that the scenario with storage and RL achieves a significant reduction in IMV, approximately 3 units lower, highlighting notably less volatile and more stable electricity prices.

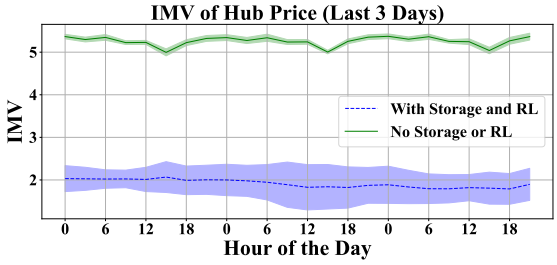


Figure 4: Comparison of IMV of the last 3 days between two scenarios: with storage and RL, and without storage or RL. Shadow areas show the 1-sigma error bounds across all simulations.

To further understand how RL influences consumption behavior, we examine the resulting net demand profiles and compare them to the original input demand. As shown in Figure 5, the RL-based strategy significantly reshapes prosumer net demand. It shifts a considerable portion of energy consumption (charging) toward midday, as a response to low electricity prices and abundant solar generation. The net demand turns negative during peak evening hours, indicating energy selling back to the grid when

prices are high. The curve after learning is less smooth due to the existence of cost-free grid-level solar generators, prosumers can increase their consumption without increasing the price too much.

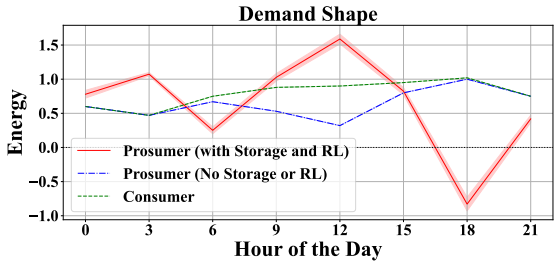


Figure 5: The vertical axis indicates the energy amount scaled down by a factor of the total storage level of the corresponding agent types (prosumers or consumers). The shaded areas indicate one standard deviation error bounds computed over all 10 days and all simulation runs.

### 4.5.4 Leader's Learning Result - Energy Expenditure Incidence (EEI)

The EEI for both prosumers and consumers is shown in Figure 6. Under our experimental setup, the utility company's optimal strategy reduces the EEI gap between prosumers and consumers from approximately 1% to about 0.7%, indicating improved equity across different income groups and customer types. We note that the EEI values are typically small since energy spending constitutes only a minor portion of total household income [26].

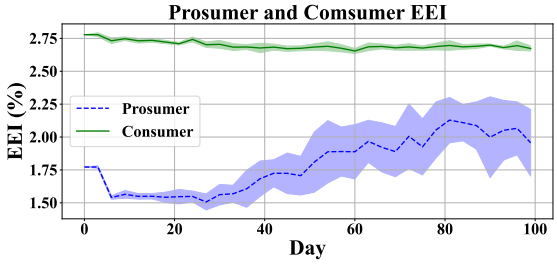


Figure 6: EEI over time for prosumers and consumers. The learned policy reduces the EEI gap between the two groups, indicating improved income-based equity. Shaded regions represent one standard deviation across simulation runs.