OPTIMIZING MECHANISM PARAMETERS FOR MARKOV GAMES UNDER POLICY-SPACE EQUILIBRIUM CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Mechanism parameter optimization seeks to align the strategic behavior of players with global system objectives. While this problem is tractable in games with closed-form equilibrium solutions, extending it to Markov games remains challenging. The presence of stochastic transitions and high-dimensional state spaces precludes closed-form equilibrium characterization, necessitating computationally intensive multi-agent reinforcement learning (MARL) methods for equilibrium computation. Embedding such methods within an outer black-box optimization loop over mechanism parameters incurs substantial computational cost. We propose a gradient-based framework for optimizing mechanism parameters in two-player Markov games under quantal response equilibrium constraints. Our approach addresses the lack of differentiability in MARL-based equilibrium computation through two key components. First, we iteratively construct and evolve a manageable set of representative policies as mechanism parameters vary, mitigating the intractability of the full policy space. Second, we apply implicit differentiation to the equilibrium conditions over these representative policies, yielding gradient signals that enable efficient parameter optimization. We theoretically show that the surrogate gradient computed from the representative policy set approximates the full-space gradient, and analyze the convergence of gradient descent driven by this surrogate gradient. We evaluate our framework on two mechanism optimization settings: transition dynamics modification and reward function revision. Specifically, we consider minimizing unfairness in Kuhn Poker and balancing productivity and equality in the AI Economist environment. Experiments show that our framework outperforms the baselines in solution quality and search efficiency.

1 INTRODUCTION

Mechanism design aims to align the behavior of players in strategic environments with a global objective. Mechanism parameter optimization can be formulated as a bilevel problem. The upper-level problem selects mechanism parameters to optimize a designer-specified objective. The lower-level problem seeks an equilibrium among players that characterizes their strategic interactions under the given mechanism parameters. Existing studies often focus on parameter optimization in games that admit closed-form equilibrium solutions, such as reserve price optimization in one-shot first-price auctions (Riley & Samuelson, 1981). In these settings, the mapping from mechanism parameters to player behavior is analytically tractable, which enables straightforward optimization. However, many strategic interactions in practice, such as continuous double auctions and complex board games, are inherently Markov games with multi-round decisions, high-dimensional state spaces, and stochastic transitions. In such settings, equilibrium strategies typically do not admit closed-form expressions and must be approximated using computational approaches, such as multi-agent reinforcement learning methods (Lanctot et al., 2019). Nesting this equilibrium approximator inside an outer optimization loop makes parameter optimization computationally expensive.

In this work, we propose a framework for efficiently optimizing mechanism parameters in two-player Markov games under Quantal Response Equilibrium (QRE) constraints (McKelvey & Palfrey, 1995). QRE captures bounded rationality by modeling players as stochastic decision makers who may select suboptimal policies. Rather than treating the lower-level equilibrium solver as a black box without

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

analytical gradients, our framework tackles the bilevel problem via a gradient-based optimization pipeline through the following components. **Dynamic construction of restricted policy sets:** The policy space of a Markov game is typically unmanageably large. We leverage the Policy Space Response Oracles (PSRO) algorithm to iteratively construct a restricted set of representative player policies. As the mechanism parameters change, this policy set evolves to remain aligned with the changing strategic environment. This reduces equilibrium analysis to a manageable restricted policy space. **Implicit differentiation of equilibrium conditions:** To quantify how changes in mechanism parameters affect equilibrium behavior, we apply the Implicit Function Theorem to the QRE conditions defined on the restricted policy space. This yields the sensitivity of the equilibrium with respect to the mechanism parameters, which is then used to guide the search in parameter space. We provide rigorous theoretical analysis of our framework. We prove that (i) the surrogate gradient computed from the representative policy set can approximate the gradient of the full-space objective, and (ii) gradient descent using the surrogate gradient converges to a neighborhood of a stationary point of the full-space objective.

Our contributions are twofold. First, from a **methodological perspective**, we introduce a general gradient-based pipeline for optimizing mechanism parameters in two-player Markov games. By differentiating through the equilibrium over a restricted policy space, our framework generates a gradient signal that facilitates highly efficient optimization of mechanism parameters. Second, from an **empirical perspective**, we demonstrate the framework's efficacy through two case studies representing the two settings of mechanism optimization: transition dynamics modification and reward function revision. Specifically, we consider (1) minimizing unfairness between players in Kuhn Poker by optimizing card-dealing probabilities, and (2) balancing productivity and equality in the AI Economist environment via tax policy optimization. Our framework offers a scalable approach for designing mechanisms in real-world stochastic strategic environments.

2 Related Work

Mechanism Parameter Optimization Conventionally, mechanism parameter optimization has been restricted to static environments where the equilibrium behavior of strategic players can be characterized in closed form (Yuan et al., 2014). When the induced equilibrium does not admit a closed-form characterization, techniques such as penalty-based methods (Shen et al., 2025) and implicit differentiation (Ling et al., 2018) have been employed to enable gradient-based updates of mechanism parameters. As environments generalize to Markov games, obtaining analytical gradients becomes challenging. Consequently, many recent studies have turned to gradient-free approaches for mechanism parameter optimization, including reinforcement learning (Zheng et al., 2020), LLM-guided search over the parameter space (Karten et al., 2025), and black-box evolutionary methods such as CMA-ES (Pendurkar et al., 2023). Wang et al. (2023) focused on reward function revision in zero-sum Markov games by differentiating the equilibrium at the level of parameterized policies, requiring derivatives with respect to high-dimensional policy parameters. In contrast, our framework is applicable to both transition dynamics modification and reward function revision. It applies implicit differentiation to the equilibrium over a restricted policy space, substantially reducing the dimensionality of equilibrium differentiation while providing informative gradient signals for efficient mechanism parameter optimization.

Approximating Full Policy Spaces via Restricted Policy Sets To mitigate the intractable computation of exact equilibria in complex Markov games, one approach is to approximate the full policy space by a dynamically expanding restricted set of policies and compute equilibria over this restricted set. Building on Fictitious Play (Heinrich et al., 2015) and the Double Oracle method (Tang et al., 2023), PSRO (Lanctot et al., 2017) combines iterative equilibrium computation with reinforcement learning-based best-response generation to produce new policies against the current meta-strategy equilibrium. Recent literature validates that PSRO effectively constructs robust and reliable representations of the full policy space (Bighashdel et al., 2024). Integrating a meta-strategy solver such as α-Rank enables PSRO to capture globally stable dynamics in general-sum games (Omidshafiei et al., 2019; Muller et al., 2020). While prior work treats PSRO as a tool for equilibrium approximation, our framework exploits the resulting restricted policy set to construct informative gradients that guide the optimization of upper-level mechanism parameters.

3 PROBLEM SETTING

Parameterized Partially Observable Markov Game We consider a setting in which a designer chooses the parameters of a mechanism and two players interact in the induced partially observable Markov game. Let a continuous vector $\theta \in \Theta$ represent the mechanism parameters chosen by the designer, and denote the resulting two-player game by $\langle \mathcal{N}, \mathcal{Q}, \mathcal{A}, \mathcal{O}, P_\theta, \{R_{i,\theta}\}_{i\in\mathcal{N}}, \{\Omega_i\}_{i\in\mathcal{N}}, \gamma\rangle$, where $\mathcal{N} = \{1, 2\}$ is the player set, $\mathcal{Q}$ is the state space, $\mathcal{A}$ is the joint action space, and $\mathcal{O}$ is the joint observation space (Yao et al., 2023). The vector θ affects the state transition function P_θ, and the reward function $R_{i,\theta}$ of each player $i \in \mathcal{N}$. We study the case where P_θ and $R_{i,\theta}$ are differentiable with respect to θ. The observation function Ω_i specifies the observations available to player i.

Let $\boldsymbol{\pi} = (\pi_1, \pi_2)$ denote the players' policy profile. The expected discounted cumulative utility of player i with discount factor γ is defined as $U_i(\boldsymbol{\pi}; \theta) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{i,\theta}(s_t, \boldsymbol{a}_t)\right]$. The expectation is taken over trajectories induced by the state transition function, the observation functions, and the policy profile $\boldsymbol{\pi}$. Let $\boldsymbol{\Pi} = \Pi_1 \times \Pi_2$ denote a finite but potentially large set of feasible policy profiles. Let $\boldsymbol{\sigma} = (\sigma_1, \sigma_2) \in \Delta(\Pi_1) \times \Delta(\Pi_2)$ denote the distribution over these policies. We define the expected utility of player i under distribution $\boldsymbol{\sigma}$ as $J_i(\boldsymbol{\sigma}; \theta) \triangleq \mathbb{E}_{\boldsymbol{\pi}\sim\boldsymbol{\sigma}}\left[U_i(\boldsymbol{\pi}; \theta)\right]$.

Quantal Response Equilibrium Players may be boundedly rational and unable to choose the exact utility-maximizing policy. Let $\sigma_{i,j}$ denote the probability that player i adopts its j-th policy $\pi_i^j \in \Pi_i$, and let σ_{-i} denote the distribution profile of the other player. Each player i adopts a distribution σ_i by maximizing an entropy-regularized objective:

$$\max_{\sigma_i \in \Delta(\Pi_i)} \quad J_i\left(\sigma_i, \sigma_{-i}; \theta\right) - \frac{1}{\lambda} \sum_{j=1}^{|\Pi_i|} \sigma_{i,j} \log \sigma_{i,j}, \tag{1}$$

where $\lambda > 0$ controls the rationality level. As λ approaches infinity, the entropy term vanishes and (1) reduces to a utility maximization. A Quantal Response Equilibrium (QRE) under the parameter θ is a fixed point $\boldsymbol{\sigma}^*(\theta) = \left(\sigma_1^*(\theta), \sigma_2^*(\theta)\right)$ where each $\sigma_i^*(\theta)$ solves (1) given $\sigma_{-i}^*(\theta)$ (McKelvey & Palfrey, 1995). In two-player zero-sum normal-form games, the QRE is unique (Ling et al., 2018; Sokota et al., 2023). For two-player general-sum normal-form games, the QRE is unique when λ is below a threshold (McKelvey & Palfrey, 1995; Hofbauer & Sandholm, 2002). A formal statement and proof are provided in Appendix A. When λ exceeds this threshold, we adopt a consistent equilibrium selection criterion to isolate a unique equilibrium.

Mechanism Parameter Optimization under Equilibrium Constraints Let $f(\boldsymbol{\pi}, \theta)$ denote the designer's objective under the players' policy profile $\boldsymbol{\pi}$ and the design choice θ. We focus on the case where $f(\boldsymbol{\pi}, \theta)$ is differentiable with respect to θ. This holds in many applications, such as generalized Kuhn Poker game design and tax policy design, as will be demonstrated in our case studies. Our goal is to minimize the designer's expected objective subject to the players adopting a QRE:

$$\min_{\boldsymbol{\sigma}, \theta} \; F_{\boldsymbol{\Pi}}(\boldsymbol{\sigma}, \theta) \triangleq \mathbb{E}_{\boldsymbol{\pi}\sim\boldsymbol{\sigma}}\left[f(\boldsymbol{\pi}, \theta)\right] \quad \text{s.t. } \sigma_i \text{ solves Problem (1)}, \; \forall i. \tag{2}$$

4 EQUILIBRIUM-CONSTRAINED MECHANISM PARAMETER OPTIMIZATION

4.1 QUANTAL RESPONSE EQUILIBRIUM APPROXIMATION VIA PSRO

The first step in solving (2) is to derive the QRE of the underlying Markov game for a given mechanism parameter vector θ. For complex Markov games, the full policy space $\boldsymbol{\Pi}$ is enormous, making it computationally infeasible to compute the QRE over the entire policy space. To address this, we employ the Policy Space Response Oracles (PSRO) (Lanctot et al., 2017). Instead of solving the intractable full game directly, PSRO reduces it to a tractable empirical game defined over a restricted set of policies, and computes the QRE of the empirical game to approximate the true QRE.

The PSRO algorithm proceeds as follows. For each player i, we initialize a policy pool $\mathcal{S}_i^0$, where the initial policies can be generated randomly (Lian et al., 2025). At iteration k, we perform three steps. **(i) Empirical utility matrix construction:** We construct an empirical normal-form meta-game over the current policy pools. Let $\mathbb{U}_i^k(\theta)$ denote the utility matrix of player i in the meta-game. We

obtain its (m, n)-entry $\mathbb{U}_i^k(m, n; \theta)$ by evaluating the joint policy (π_1^m, π_2^n) via T_{simu} independent simulated rollouts, where π_1^m is the m-th policy in $\mathcal{S}_1^k$ and π_2^n is the n-th policy in $\mathcal{S}_2^k$. The value of $\mathbb{U}_i^k(m, n; \theta)$ is the average empirical utility of player i across these T_{simu} rollouts. **(ii) QRE computation:** We compute the QRE of the normal-form meta-game based on $\mathbb{U}_i^k(\theta)$ (Slumbers et al., 2023; Li et al., 2026). Under the QRE, each σ_i^k solves (1) given σ_{-i}^k. According to the first-order optimality conditions, this induces a system of nonlinear equations, which can be solved via an iterative fixed-point method (Rong et al., 2015; Goeree et al., 2020). **(iii) Policy pool expansion:** For each player i, we obtain a new policy π_i^{new} by training a best-response policy via reinforcement learning against the opponent's strategy σ_{-i}^k. Although π_i^{new} is trained using reinforcement learning, we assume a finite discretization of the candidate policy space in our theoretical analysis. We expand the policy pools for the next iteration, i.e., $\mathcal{S}_i^{k+1} = \mathcal{S}_i^k \cup \{\pi_i^{\text{new}}\}$ for all i. The PSRO algorithm terminates when every newly generated policy receives negligible probability mass under the QRE at iteration $k + 1$. This indicates that the current policy space is effectively saturated.

4.2 Objective Gradient Computation under Equilibrium Constraints

Let $\boldsymbol{\sigma}_{\boldsymbol{\Pi}}^*(\theta)$ be the QRE under the mechanism parameter vector θ and full policy space $\boldsymbol{\Pi}$. Next, we analyze the gradient $\nabla_\theta F_{\boldsymbol{\Pi}}(\boldsymbol{\sigma}_{\boldsymbol{\Pi}}^*(\theta), \theta)$, which will be used in our gradient method to solve (2).

We can expand the function $F_{\boldsymbol{\Pi}}(\boldsymbol{\sigma}_{\boldsymbol{\Pi}}^*(\theta), \theta)$ as follows:

$$F_{\boldsymbol{\Pi}}(\boldsymbol{\sigma}_{\boldsymbol{\Pi}}^*(\theta), \theta) = \sum_{m=1}^{|\Pi_1|} \sum_{n=1}^{|\Pi_2|} \sigma_{1,m}^*(\theta)\, \sigma_{2,n}^*(\theta)\, f(\pi_1^m, \pi_2^n, \theta). \tag{3}$$

Based on the chain rule, computing $\nabla_\theta F_{\boldsymbol{\Pi}}(\boldsymbol{\sigma}_{\boldsymbol{\Pi}}^*(\theta), \theta)$ requires $\nabla_\theta \sigma_{1,m}^*(\theta)$, $\nabla_\theta \sigma_{2,n}^*(\theta)$, and $\nabla_\theta f(\pi_1^m, \pi_2^n, \theta)$. Given the explicit form of $f(\pi_1^m, \pi_2^n, \theta)$, its gradient is straightforward to compute. The main challenge lies in computing the gradients $\nabla_\theta \sigma_{1,m}^*(\theta)$ and $\nabla_\theta \sigma_{2,n}^*(\theta)$, which capture how changes in the mechanism parameter vector θ affect the players' equilibrium strategies. Next, we investigate the computation of $\nabla_\theta \sigma_{1,m}^*(\theta)$. The computation of $\nabla_\theta \sigma_{2,n}^*(\theta)$ proceeds analogously.

Decomposition of Gradient $\nabla_\theta \sigma_{1,m}^*(\theta)$ Let $\mathbb{U}_i(\theta)$ denote the utility matrix of player i, where each entry $\mathbb{U}_i(u, v; \theta)$ denotes the utility under the policy profile (π_1^u, π_2^v), with $\pi_1^u \in \Pi_1$ and $\pi_2^v \in \Pi_2$. The $\sigma_{1,m}^*(\theta)$ (i.e., the probability that player 1 chooses its m-th policy under the QRE) is fully determined by the utility matrices $\mathbb{U}_1(\theta)$ and $\mathbb{U}_2(\theta)$. We can expand the gradient $\nabla_\theta \sigma_{1,m}^*(\theta)$ as

$$\nabla_\theta \sigma_{1,m}^*(\theta) = \sum_{i \in \{1,2\}} \sum_{u=1}^{|\Pi_1|} \sum_{v=1}^{|\Pi_2|} \frac{\partial \sigma_{1,m}^*(\theta)}{\partial \mathbb{U}_i(u, v; \theta)} \nabla_\theta \mathbb{U}_i(u, v; \theta). \tag{4}$$

Based on the score-function identity, the gradient $\nabla_\theta \mathbb{U}_i(u, v; \theta)$ can be expressed as

$$\nabla_\theta \mathbb{U}_i(u, v; \theta) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta R_{i,\theta}(s_t, \boldsymbol{a}_t) + \left(\sum_{t=0}^{\infty} \gamma^t R_{i,\theta}(s_t, \boldsymbol{a}_t) \right) \sum_{t=0}^{\infty} \nabla_\theta \log P_\theta(s_{t+1} \mid s_t, \boldsymbol{a}_t) \right].$$

To compute $\nabla_\theta \sigma_{1,m}^*(\theta)$ in (4), it remains to evaluate $\frac{\partial \sigma_{1,m}^*(\theta)}{\partial \mathbb{U}_i(u,v;\theta)}$, which we will address next.

Implicit Differentiation of QRE For notational simplicity, let the column vectors $x \triangleq \sigma_{\boldsymbol{\Pi},1}^*(\theta)$ and $y \triangleq \sigma_{\boldsymbol{\Pi},2}^*(\theta)$ denote the QRE strategies. We define $\Phi_{\boldsymbol{\Pi},1}(y, \theta) \triangleq \text{softmax}(\lambda \mathbb{U}_1(\theta) y)$, which maps player 2's mixed strategy y to player 1's logit best-response strategy. Similarly, we define $\Phi_{\boldsymbol{\Pi},2}(x, \theta) \triangleq \text{softmax}\left(\lambda \mathbb{U}_2(\theta)^\top x\right)$. The optimality conditions yield the following result, and we leave the detailed proof in Appendix B.

Proposition 1 The solution to Problem (1) satisfies the following fixed-point system:

$$x = \Phi_{\boldsymbol{\Pi},1}(y, \theta), \quad y = \Phi_{\boldsymbol{\Pi},2}(x, \theta). \tag{5}$$

Both x and y are implicitly determined by the nonlinear system (5). We define matrices $\mathbb{D}_x \triangleq \text{diag}(x) - xx^\top$ and $\mathbb{D}_y \triangleq \text{diag}(y) - yy^\top$. Taking the total differential of both sides of the fixed-point system (5) yields the following result, and we leave the detailed proof in Appendix C.

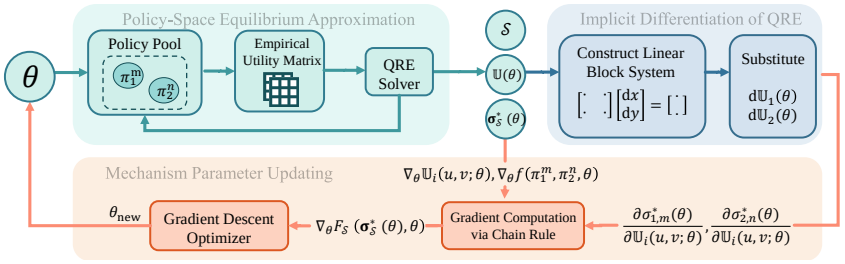


Figure 1: Overview of **ECMPO**.

Proposition 2 The differentials $\mathrm{d}x$ and $\mathrm{d}y$ satisfy

$$\mathrm{d}x = \lambda \mathbb{D}_x(\mathrm{d}\mathbb{U}_1(\theta)\cdot y + \mathbb{U}_1(\theta)\cdot \mathrm{d}y), \quad \mathrm{d}y = \lambda \mathbb{D}_y(\mathrm{d}\left(\mathbb{U}_2(\theta)^\top\right)\cdot x + \mathbb{U}_2(\theta)^\top \cdot \mathrm{d}x). \tag{6}$$

Let $M \triangleq |\Pi_1|$ and $N \triangleq |\Pi_2|$, and use $\mathbb{I}_M$ and $\mathbb{I}_N$ to denote the corresponding identity matrices. We can rewrite (6) as the following linear block system:

$$\begin{bmatrix} \mathbb{I}_M & -\lambda \mathbb{D}_x \mathbb{U}_1(\theta) \\ -\lambda \mathbb{D}_y \mathbb{U}_2(\theta)^\top & \mathbb{I}_N \end{bmatrix} \begin{bmatrix} \mathrm{d}x \\ \mathrm{d}y \end{bmatrix} = \begin{bmatrix} \lambda \mathbb{D}_x \mathrm{d}\mathbb{U}_1(\theta)\cdot y \\ \lambda \mathbb{D}_y \mathrm{d}\left(\mathbb{U}_2(\theta)^\top\right)\cdot x \end{bmatrix}. \tag{7}$$

To compute the partial derivative of x with respect to a specific utility entry $\mathbb{U}_1(u,v;\theta)$, we perturb only this entry while holding all other entries of $\mathbb{U}_1(\theta)$ and $\mathbb{U}_2(\theta)$ fixed. Let e_u and e_v denote the standard basis vectors with a 1 in the u-th and v-th entries, respectively, and zeros elsewhere. We substitute $\mathrm{d}\mathbb{U}_1(\theta) = e_u e_v^\top$ and $\mathrm{d}\mathbb{U}_2(\theta) = 0$ into (7). The partial derivative $\frac{\partial x}{\partial \mathbb{U}_1(u,v;\theta)}$ is then given by the solution vector $\mathrm{d}x$ to the linear system (7). Similarly, substituting $\mathrm{d}\mathbb{U}_1(\theta) = 0$ and $\mathrm{d}\mathbb{U}_2(\theta) = e_u e_v^\top$ into (7), we solve for $\mathrm{d}x$ and obtain $\frac{\partial x}{\partial \mathbb{U}_2(u,v;\theta)}$. The above procedure provides $\frac{\partial \sigma^*_{1,m}(\theta)}{\partial \mathbb{U}_i(u,v;\theta)}$ for $i = 1, 2$.

4.3 Equilibrium-Constrained Mechanism Parameter Optimization Framework

In Section 4.2, we derive the full-space gradient $\nabla_\theta F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta)$. In practice, the complete policy space $\mathbf{\Pi}$ is enormous, making direct gradient descent over the full QRE infeasible. Therefore, at each update of θ, we run PSRO to construct a restricted policy set $\mathcal{S} \triangleq \mathcal{S}_1 \times \mathcal{S}_2$ and compute the gradient of the restricted objective $F_{\mathcal{S}}(\boldsymbol{\sigma}^*_{\mathcal{S}}(\theta), \theta)$, which follows the same form as (3) but with the summations restricted to $\mathcal{S}_1$ and $\mathcal{S}_2$. This restricted gradient is used as a surrogate for the full-space gradient.

As illustrated in Figure 1, our Equilibrium-Constrained Mechanism Parameter Optimization (**ECMPO**) framework iterates through three components. **(i) Policy-Space Equilibrium Approximation:** It applies the PSRO algorithm to approximate the QRE under the current value of θ. As introduced in Section 4.1, this component consists of three steps and outputs a restricted policy space $\mathcal{S}$, the corresponding utility matrices $\mathbb{U}_1(\theta)$ and $\mathbb{U}_2(\theta)$, and the resulting QRE $\boldsymbol{\sigma}^*_{\mathcal{S}}(\theta)$. **(ii) Implicit Differentiation of QRE:** Using implicit differentiation, it constructs the linear block system (7) based on $\mathbb{U}_1(\theta)$ and $\mathbb{U}_2(\theta)$. By substituting different values of $\mathrm{d}\mathbb{U}_1(\theta)$ and $\mathrm{d}\mathbb{U}_2(\theta)$, it solves the system to obtain the partial derivatives $\frac{\partial \sigma^*_{1,m}(\theta)}{\partial \mathbb{U}_i(u,v;\theta)}$ and $\frac{\partial \sigma^*_{2,n}(\theta)}{\partial \mathbb{U}_i(u,v;\theta)}$. **(iii) Mechanism Parameter Updating:** It aggregates the gradient components to compute $\nabla_\theta F_{\mathcal{S}}(\boldsymbol{\sigma}^*_{\mathcal{S}}(\theta), \theta)$ and updates θ via gradient descent.

4.4 Gradient Approximation Error and Convergence Analysis

The full-space gradient in Section 4.2 is defined over the complete policy space $\mathbf{\Pi}$, whereas **ECMPO** computes the gradient from the restricted policy set $\mathcal{S}$ generated by PSRO. We show that this restricted gradient is a reliable surrogate: *its error decays exponentially with the quality of $\mathcal{S}$, and the resulting optimization update converges to a neighborhood of a stationary point of the full-space objective*.

Recall that $\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta)$ denotes the QRE over the full policy space $\mathbf{\Pi}$. Let $\mathcal{R}_i = \Pi_i \setminus \mathcal{S}_i$ for $i = 1, 2$. We use $\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)$ to denote the QRE over the restricted policy space $\mathcal{S}$, extended by assigning zero

probability to all policies in $\mathcal{R} \triangleq \mathcal{R}_1 \times \mathcal{R}_2$. This extension ensures that $\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta)$ and $\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)$ have the same dimension, thereby facilitating their comparison.

For each player i, we define the following restricted logit best-response mapping $\bar{\Phi}_{\mathcal{S},i}(\sigma_{-i}, \theta)$, where the coordinate corresponding to policy a is given by

$$\bar{\Phi}_{\mathcal{S},i}(\sigma_{-i}, \theta)_a \triangleq \begin{cases} \dfrac{\exp\{\lambda J_i(a, \sigma_{-i}; \theta)\}}{\sum_{c \in \mathcal{S}_i} \exp\{\lambda J_i(c, \sigma_{-i}; \theta)\}}, & \text{for } a \in \mathcal{S}_i, \\ 0, & \text{for } a \in \mathcal{R}_i. \end{cases} \tag{8}$$

The QRE $\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)$ satisfies $\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta) = \bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)$. Next, we make several assumptions on the mixed strategy profiles in a neighborhood of this QRE, defined as $B_r(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)) \triangleq \{\boldsymbol{\sigma} : \|\boldsymbol{\sigma} - \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\|_1 \le r\}$.

Assumption 1. *There exists a constant $\Delta > 0$ such that, for any $\boldsymbol{\sigma} \in B_r(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta))$,*

$$J_i(a, \boldsymbol{\sigma}_{-i}; \theta) \le \max_{c \in \mathcal{S}_i} J_i(c, \boldsymbol{\sigma}_{-i}; \theta) - \Delta, \quad \forall a \in \mathcal{R}_i, i \in \{1, 2\}. \tag{9}$$

Here, Δ is the *policy gap*, which characterizes the minimum payoff separation between omitted policies in $\mathcal{R}$ and the best policy within the restricted policy space $\mathcal{S}$. Let $\mathrm{Jac}_{\boldsymbol{\sigma}} \Phi$ and $\mathrm{Jac}_{\theta} \Phi$ denote the Jacobians of $\Phi(\boldsymbol{\sigma}, \theta)$ with respect to $\boldsymbol{\sigma}$ and θ, respectively.

Assumption 2. *There exists a constant $L < 1$ such that, for any $\boldsymbol{\sigma} \in B_r(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta))$,*

$$\|\mathrm{Jac}_{\boldsymbol{\sigma}} \Phi_{\boldsymbol{\Pi}}(\boldsymbol{\sigma}, \theta)\|_1 \le L, \qquad \|\mathrm{Jac}_{\boldsymbol{\sigma}} \bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma}, \theta)\|_1 \le L. \tag{10}$$

Assumption 3. *The functions $J_i(\boldsymbol{\sigma}, \theta)$, $F_{\boldsymbol{\Pi}}(\boldsymbol{\sigma}, \theta)$, and $F_{\mathcal{S}}(\boldsymbol{\sigma}, \theta)$ are twice continuously differentiable on $B_r(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta))$, with uniformly bounded values and first- and second-order derivatives under the induced 1-norm. The composite full-space objective $\theta \mapsto F_{\boldsymbol{\Pi}}(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta), \theta)$ is lower bounded by a constant $F^*_{\boldsymbol{\Pi}}$ and is L_{up}-smooth, i.e., for any θ, θ',*

$$\|\nabla_\theta F_{\boldsymbol{\Pi}}(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta), \theta) - \nabla_\theta F_{\boldsymbol{\Pi}}(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta'), \theta')\|_2 \le L_{\mathrm{up}} \|\theta - \theta'\|_2. \tag{11}$$

Assumption 4. *The omitted policy set satisfies the condition $2\left(|\mathcal{R}_1| + |\mathcal{R}_2|\right) e^{-\lambda\Delta} \le (1 - L)r$.*

Under these assumptions, we can prove that the surrogate gradient approximates the full-space gradient with an error that decreases exponentially with the policy gap Δ.

Theorem 1 (Gradient Approximation Error): There exists a constant $C_{\mathrm{loc}} > 0$ such that

$$\|\nabla_\theta F_{\boldsymbol{\Pi}}(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta), \theta) - \nabla_\theta F_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\|_1 \le C_{\mathrm{loc}} e^{-\lambda\Delta}. \tag{12}$$

We further establish the convergence guarantee of gradient descent using the surrogate gradient.

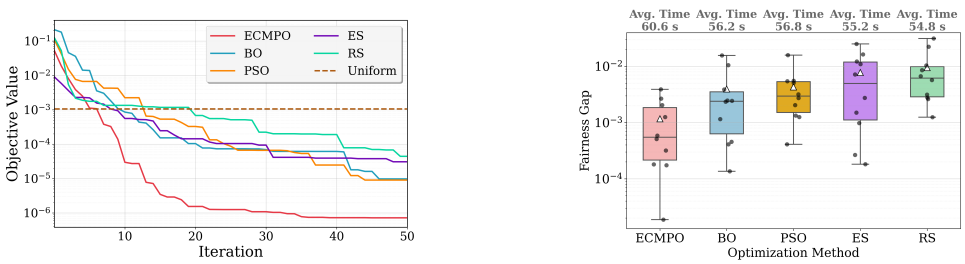
Theorem 2 (Convergence of the Surrogate-Gradient Method): Consider the update $\theta_{t+1} = \theta_t - \eta \nabla_\theta F_{\mathcal{S}_t}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}_t}(\theta_t), \theta_t)$ with $0 < \eta \le \frac{1}{L_{\mathrm{up}}}$. For T iterations, let $\Delta_{\min} = \min_{0 \le t \le T-1} \Delta_t$, where Δ_t is the policy gap associated with $\mathcal{S}_t$. Then

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla_\theta F_{\boldsymbol{\Pi}}(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta_t), \theta_t)\|_2^2 \le \frac{2\left(F_{\boldsymbol{\Pi}}(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta_0), \theta_0) - F^*_{\boldsymbol{\Pi}}\right)}{\eta T} + C^2_{\mathrm{loc}} e^{-2\lambda\Delta_{\min}}. \tag{13}$$

The left-hand side of (13) is the average squared norm of the full-space gradient along the optimization trajectory $\{\theta_0, \theta_1, \ldots, \theta_{T-1}\}$ generated by gradient descent using the surrogate gradient. As T increases, the right-hand side of (13) approaches $C^2_{\mathrm{loc}} e^{-2\lambda\Delta_{\min}}$, which decreases exponentially with the policy gap. The proofs of the above theorems are provided in Appendices D and E.

5 Case Study I: Optimizing Dealing Probabilities in Kuhn Poker

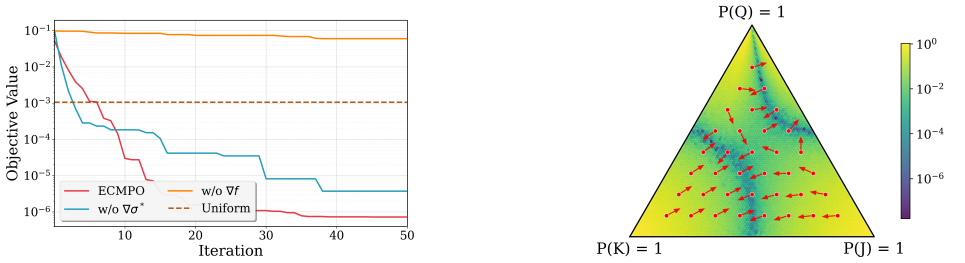
Experimental Setup In this section, we consider a setting in which we modify the transition dynamics in Markov games. Kuhn Poker is a two-player zero-sum game, where each player is dealt one private card from the deck $\{\mathrm{J}, \mathrm{Q}, \mathrm{K}\}$ and takes sequential actions from $\{\mathrm{bet}, \mathrm{pass}\}$ to compete



(a) Best Objective Value versus Search Iteration.

(b) Fairness Gap and Execution Time per Iteration.

Figure 2: Optimization Performance in Kuhn Poker (Across 10 Independent Runs).



(a)

(b)

Figure 3: (a) Ablation Study. (b) 2-D Heatmap of the Mechanism Parameter Space.

for the pot (Kuhn, 2016; Liu et al., 2026). In the standard setting, cards are drawn uniformly without replacement, so each card is dealt with probability $1/3$. This creates an unfair outcome, as the second player obtains a positive expected equilibrium utility (Hoehn et al., 2005; Lanctot, 2014). We study the problem of designing a generalized Kuhn Poker game that ensures fairness between the two players. Specifically, the cards are drawn according to a weighted sampling without replacement rule (Ben-Hamou et al., 2018), and the initial dealing probabilities of the three cards are parameterized by a mechanism parameter vector θ. Since the three dealing probabilities must sum to 1, θ is inherently two-dimensional.

We define the designer's objective as $(\mathbb{E}[U_1(\boldsymbol{\pi}, \theta)])^2$, which drives player 1's expected equilibrium utility in this zero-sum game to 0. Its gradient is $2\mathbb{E}[U_1(\boldsymbol{\pi}, \theta)]\nabla_\theta \mathbb{E}[U_1(\boldsymbol{\pi}, \theta)]$, where $\nabla_\theta \mathbb{E}[U_1(\boldsymbol{\pi}, \theta)]$ is obtained through our framework. We leave details of the game to Appendix F.

We compare **ECMPO** against several optimization methods: Bayesian Optimization (**BO**) (Snoek et al., 2012), Particle Swarm Optimization (**PSO**) (Kennedy & Eberhart, 1995), Evolution Strategies (**ES**) (Salimans et al., 2017), and Random Search (**RS**). While searching over mechanism parameters, these methods employ PSRO to compute the QRE for each parameter vector.

Experimental Results As illustrated in Figure 2a, **ECMPO** consistently reduces the objective value over successive iterations. While gradient-free methods also manage to decrease the objective initially, they soon plateau at suboptimal values. The brown dashed line represents the objective value under the uniform dealing distribution (i.e., the standard Kuhn Poker setting). We further compare the methods in reducing the *fairness gap*, defined as $|\mathbb{E}\left[U_1(\boldsymbol{\pi}, \theta) - U_2(\boldsymbol{\pi}, \theta)\right]|$ and measured in chips. Even a small deviation from zero can accumulate over repeated games, yielding a meaningful profit advantage for one player. Figure 2b shows the distribution of the fairness gap achieved by the best solution found by each method across 10 independent trials, as well as the average execution time per iteration. Our **ECMPO** achieves the lowest fairness gap without incurring any noticeable computational overhead.

Next, we conduct an ablation study. When computing the gradient of the objective function in (3), our **ECMPO** explicitly accounts for $\nabla_\theta \sigma^*_{1,m}(\theta)$, $\nabla_\theta \sigma^*_{2,n}(\theta)$, and $\nabla_\theta f(\pi_1^m, \pi_2^n, \theta)$. For comparison, we evaluate variants that omit either the gradient of the QRE $\boldsymbol{\sigma}^*$ or the gradient of the function f.



(a) Time = 0 (b) Time = 250 (c) Time = 500

Figure 4: Illustration of Agent Activities in the Spatial Grid World of the AI Economist.

The comparison in Figure 3a shows that **ECMPO** achieves the best performance, highlighting the importance of our design (e.g., computing the gradient of the QRE via implicit differentiation).

Figure 3b visualizes the objective value over the probability simplex of θ, where the vertices correspond to degenerate distributions assigning probability 1 to J, Q, and K, respectively. The dark blue regions correspond to low objective values. The red arrows represent the negative gradients computed by our **ECMPO**. These directions consistently guide the optimization along the steepest descent toward nearby valleys. Therefore, compared to gradient-free exploration, **ECMPO** effectively leverages gradient information to achieve more efficient search.

Due to space limitations, we include additional experimental results in the appendices. In Appendix G, we further investigate card-dealing probability optimization in Kuhn Poker under a different design objective. In Appendix H, we study card-dealing probability optimization in Leduc Poker.

6 Case Study II: Optimizing Tax Vectors in AI Economist

Experimental Setup In this section, we consider a setting in which we modify the reward functions in Markov games. The *AI Economist* provides an environment for simulating economic interactions among strategic agents, which can be modeled as a non-zero-sum Markov game (Zheng et al., 2020). As illustrated in Figure 4, two agents move in an 11×11 grid world, gather wood and stone, and build houses. Each action incurs a labor cost, and agents earn coins by building houses. A simulation episode lasts 500 time steps and is divided into 10 tax periods. At the beginning of the episode, the taxation mechanism designer announces bracketed marginal tax rates, parameterized by a **7-dimensional** vector θ. Taxes are collected at the end of each tax year, and the total amount collected is evenly redistributed to the agents.

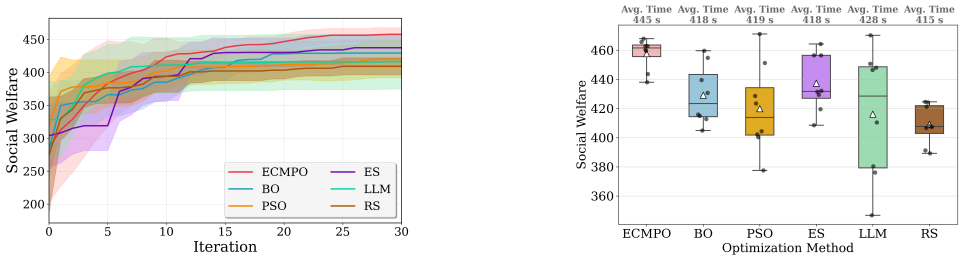
Let $x_i(\boldsymbol{\pi}, \theta)$ denote the final coin holdings of agent i at the end of the episode. Each agent optimizes its policy π_i to maximize its expected utility $U_i(\boldsymbol{\pi}, \theta)$, which strictly increases with $x_i(\boldsymbol{\pi}, \theta)$ and decreases with its cumulative labor cost. To balance total productivity and wealth equality, we define the objective function $f(\boldsymbol{\pi}, \theta)$ (to be minimized) as the negative of the social welfare function defined in Zheng et al. (2020):

$$\mathrm{swf}(\boldsymbol{\pi}, \theta) \triangleq (x_1(\boldsymbol{\pi}, \theta) + x_2(\boldsymbol{\pi}, \theta)) - |x_1(\boldsymbol{\pi}, \theta) - x_2(\boldsymbol{\pi}, \theta)|. \tag{14}$$

Our goal is to maximize $\mathbb{E}_{\boldsymbol{\pi} \sim \boldsymbol{\sigma}} [\mathrm{swf}(\boldsymbol{\pi}, \theta)]$ over θ. Further details regarding the environment configurations are deferred to Appendix I.

In addition to the baseline methods evaluated in Section 5, we compare our **ECMPO** framework against the LLM-guided search method (Karten et al., 2025). It prompts a large language model with the interaction history and uses a phase-based search strategy that shifts from broad exploration to fine-grained exploitation as context accumulates.

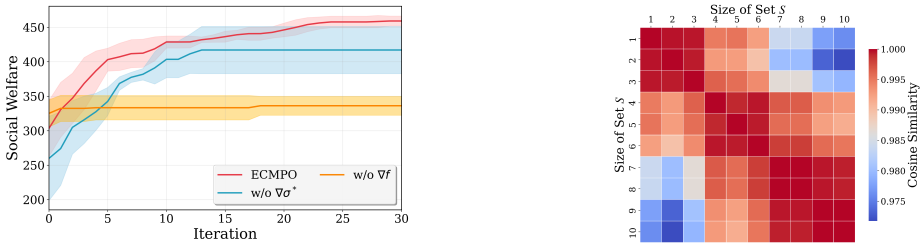
Experimental Results Figure 5a shows the maximum social welfare as a function of search iterations for different methods, with shaded regions indicating the standard deviation over 8 runs. While **BO** performs well when optimizing the low-dimensional parameter vector in the Kuhn Poker task, it suffers from the curse of dimensionality when scaling to the 7-dimensional parameter space in the AI Economist environment. In contrast, **ES** mitigates this issue and outperforms **BO** by



(a) Maximum Social Welfare versus Search Iteration.

(b) Social Welfare and Time per Iteration.

Figure 5: Optimization Performance in AI Economist (Across 8 Independent Runs).



(a)

(b)

Figure 6: (a) Ablation Study. (b) Pairwise Cosine Similarity Heatmap of Gradients.

leveraging gradient estimates obtained through stochastic perturbations, highlighting the scalability of search methods based on gradient estimates. Our **ECMPO** further improves upon **ES**, achieving higher mean social welfare by utilizing analytically derived gradients rather than the high-variance gradient estimates induced by stochastic zeroth-order perturbations in **ES**, thereby enabling more stable parameter updates. Although the LLM-based method initially improves rapidly by exploiting pre-trained economic priors, it quickly stagnates, as it lacks the mathematical structure necessary for fine-grained continuous optimization.

Figure 5b details the final social welfare distribution across 8 independent runs. **ECMPO** achieves the highest median and mean scores with the lowest variance, demonstrating robustness across random seeds. The figure also reports the average execution time per iteration for each method. **ECMPO** improves optimization performance and stability without sacrificing much computational efficiency.

Similar to Figure 3a, Figure 6a compares **ECMPO** with variants that omit either the gradient of the QRE $\boldsymbol{\sigma}^*$ or the gradient of the function f in the gradient computation. The results show that **ECMPO** achieves the best performance. Figure 6b illustrates the pairwise cosine similarity of the gradients $\nabla_\theta F$ as the restricted policy set size increases. Specifically, we expand the restricted policy set from size 1 to 10. For each size, we compute the corresponding 7-dimensional gradient $\nabla_\theta F$, and then evaluate the pairwise cosine similarity between these gradients. We observe that early additions to the policy set (from size 1 to 6) lead to substantial changes in the gradient vector. Beyond this point, the gradients stabilize, reaching a plateau once the set contains around 7 representative policies. This suggests that the gradient converges as the size of the restricted policy set increases.

7 Conclusion

We proposed the **ECMPO** framework for optimizing mechanism parameters in complex stochastic strategic environments. We provided a theoretical analysis of its convergence and evaluated its performance in two case studies: transition dynamics modification (Kuhn Poker) and reward function revision (the AI Economist). The main limitation of **ECMPO** is its focus on settings where state transition probabilities and reward functions are differentiable with respect to the mechanism parameter vector θ. This opens promising directions for future research on efficient mechanism parameter optimization in complex Markov games with discrete parameter spaces.

AI USE STATEMENT

In this work, we used generative AI tools for providing critical ingredients for proving mathematical claims and assisting in the writing of proofs. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, and interpreting results. Generating synthetic datasets, cleaning and reformatting datasets, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for searching for information, editing this research paper to improve readability, and improving the visual presentation of Figure 4. We have reviewed all AI-assisted work. We carefully checked and independently verified the AI-assisted portions of the proofs. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the detailed derivations supporting our theoretical results in Appendices A, B, C, D, and E. Detailed descriptions of the experimental environments and task-specific gradient computations are provided in Appendices F, G, H, and I. The main experimental setups and evaluation protocols are described in Sections 5 and 6, and their hyperparameters are specified in Appendices. All reported experimental results are aggregated over at least 8 independent runs.

REFERENCES

Anna Ben-Hamou, Yuval Peres, and Justin Salez. Weighted sampling without replacement. *Brazilian Journal of Probability and Statistics*, 32(3):657–669, 2018.

Ariyan Bighashdel, Yongzhao Wang, Stephen McAleer, Rahul Savani, and Frans A. Oliehoek. Policy space response oracles: A survey. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 7951–7961, 2024.

Jacob K. Goeree, Charles A. Holt, and Thomas R. Palfrey. Stochastic game theory for social science: A primer on quantal response equilibrium. In *Handbook of Experimental Game Theory*, pp. 8–47. Edward Elgar Publishing, 2020.

Johannes Heinrich, Marc Lanctot, and David Silver. Fictitious self-play in extensive-form games. In *Proceedings of the International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 805–813, 2015.

Bret Hoehn, Finnegan Southey, Robert C. Holte, and Valeriy Bulitko. Effective short-term opponent exploitation in simplified poker. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2005.

Josef Hofbauer and William H. Sandholm. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6):2265–2294, 2002.

Seth Karten, Wenzhe Li, Zihan Ding, Samuel Kleiner, Yu Bai, and Chi Jin. LLM economist: Large population models and mechanism design in multi-agent generative simulacra. *arXiv preprint arXiv:2507.15815*, 2025.

J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of the International Conference on Neural Networks*, volume 4, pp. 1942–1948, 1995.

Harold W. Kuhn. A simplified two-person poker. In *Contributions to the Theory of Games*, volume 1, pp. 97–104. Princeton University Press, 2016.

Marc Lanctot. Further developments of extensive-form replicator dynamics using the sequence-form representation. In *Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems*, pp. 1257–1264, 2014.

Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, 2017.

Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

Zun Li, John Schultz, Daniel Hennes, and Marc Lanctot. Discovering multiagent learning algorithms with large language models. *arXiv preprint arXiv:2602.16928*, 2026.

Jiesong Lian, Yucong Huang, Chengdong Ma, Mingzhi Wang, Ying Wen, Long Hu, and Yixue Hao. Fusion-PSRO: Nash policy fusion for policy space response oracles. In *Proceedings of the European Conference on Artificial Intelligence*, pp. 2562–2569, 2025.

Chun Kai Ling, Fei Fang, and J. Zico Kolter. What game are we playing? End-to-end learning in normal and extensive form games. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 396–402, 2018.

Bo Liu, Simon Yu, Zichen Liu, Leon Guertler, Penghui Qi, Daniel Balcells, Mickel Liu, Cheston Tan, Weiyan Shi, Min Lin, Wee Sun Lee, and Natasha Jaques. SPIRAL: Self-play on zero-sum games incentivizes reasoning via multi-agent multi-turn reinforcement learning. In *Proceedings of the International Conference on Learning Representations*, 2026.

Richard D. McKelvey and Thomas R. Palfrey. Quantal response equilibria for normal form games. *Games and Economic Behavior*, 10(1):6–38, 1995.

Paul Muller, Shayegan Omidshafiei, Mark Rowland, Karl Tuyls, Julien Perolat, Siqi Liu, Daniel Hennes, Luke Marris, Marc Lanctot, Edward Hughes, Zhe Wang, Guy Lever, Nicolas Heess, Thore Graepel, and Remi Munos. A generalized training approach for multiagent learning. In *Proceedings of the International Conference on Learning Representations*, 2020.

Shayegan Omidshafiei, Christos Papadimitriou, Georgios Piliouras, Karl Tuyls, Mark Rowland, Jean-Baptiste Lespiau, Wojciech M. Czarnecki, Marc Lanctot, Julien Perolat, and Remi Munos. Alpha-rank: Multi-agent evaluation by evolution. *Scientific Reports*, 9(1):9937, 2019.

Sumedh Pendurkar, Chris Chow, Luo Jie, and Guni Sharon. Bilevel entropy based mechanism design for balancing meta in video games. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 2134–2142, 2023.

John G. Riley and William F. Samuelson. Optimal auctions. *The American Economic Review*, 71(3): 381–392, 1981.

Jiang Rong, Tao Qin, and Bo An. Computing quantal response equilibrium for sponsored search auctions. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, 2015.

J. Ben Rosen. Existence and uniqueness of equilibrium points for concave n-person games. *Econometrica*, 33(3):520–534, 1965.

Tim Salimans, Jonathan Ho, Xi Chen, Szymon Sidor, and Ilya Sutskever. Evolution strategies as a scalable alternative to reinforcement learning. *arXiv preprint arXiv:1703.03864*, 2017.

Terence Conrad Schauenberg. Opponent modelling and search in poker. Master's thesis, University of Alberta, 2006.

Han Shen, Zhuoran Yang, and Tianyi Chen. Principled penalty-based methods for bilevel reinforcement learning and RLHF. *Journal of Machine Learning Research*, 26(114):1–49, 2025.

Oliver Slumbers, David Henry Mguni, Stefano B. Blumberg, Stephen Marcus McAleer, Yaodong Yang, and Jun Wang. A game-theoretic framework for managing risk in multi-agent systems. In *Proceedings of the International Conference on Machine Learning*, pp. 32059–32087, 2023.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 2951–2959, 2012.

Samuel Sokota, Ryan D'Orazio, J. Zico Kolter, Nicolas Loizou, Marc Lanctot, Ioannis Mitliagkas, Noam Brown, and Christian Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *Proceedings of the International Conference on Learning Representations*, 2023.

Xiaohang Tang, Le Cong Dinh, Stephen Marcus McAleer, and Yaodong Yang. Regret-minimizing double oracle for extensive-form games. In *Proceedings of the International Conference on Machine Learning*, volume 202, pp. 33599–33615, 2023.

Jing Wang, Meichen Song, Feng Gao, Boyi Liu, Zhaoran Wang, and Yi Wu. Differentiable arbitrating in zero-sum Markov games. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 1034—1043, 2023.

Jian Yao, Weiming Liu, Haobo Fu, Yaodong Yang, Stephen McAleer, Qiang Fu, and Wei Yang. Policy space diversity for non-transitive games. In *Advances in Neural Information Processing Systems*, volume 36, pp. 67771–67793, 2023.

Shuai Yuan, Jun Wang, Bowei Chen, Peter Mason, and Sam Seljan. An empirical study of reserve price optimisation in real-time bidding. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1897–1906, 2014.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, Nikhil Naik, Melvin Gruesbeck, David C. Parkes, and Richard Socher. The AI Economist: Improving equality and productivity with AI-driven tax policies. *arXiv preprint arXiv:2004.13332*, 2020.

TABLE OF CONTENTS

A Theoretical Analysis of QRE Uniqueness **14**
A.1 QRE Uniqueness and the Diagonally Strictly Concave Condition 14
A.2 Negative Definiteness of the Jacobian 14
A.3 Upper Bound on λ for QRE Uniqueness 15

B Proof of Proposition 1 (The QRE Fixed-Point System) **16**

C Proof of Proposition 2 (Differential of linear system) **17**

D Proof of Theorem 1 (Gradient Approximation Error) **17**
D.1 Preliminary Lemmas 17
D.2 Proof of Theorem 1 23

E Proof of Theorem 2 (Convergence of the Surrogate-Gradient Method) **24**
E.1 Preliminary Lemma 24
E.2 Proof of Theorem 2 25

F Two-Player Kuhn Poker **26**
F.1 Game Rules 26
F.2 Differentiating the Objective 27
F.3 Hyperparameter Settings 27

G Optimizing Fairness with Entropy Regularization in Kuhn Poker **28**
G.1 Gradient of the Regularized Objective 28
G.2 Experimental Results 28

H Fairness Optimization in Leduc Poker **29**
H.1 Game Rules 29
H.2 Leduc-Specific Gradient 30
H.3 Experimental Results 30

I Social Welfare Optimization in AI Economist **31**
I.1 Environment 31
I.2 Direct Gradient of the Social Welfare Objective 32
I.3 Gradient of the QRE Mixture 32
I.4 Hyperparameter Settings 33

A THEORETICAL ANALYSIS OF QRE UNIQUENESS

This section details the theoretical proofs for QRE uniqueness. Adapting the analysis in Rosen (1965) to two-player games, Subsection A.1 proves that the Diagonally Strictly Concave (DSC) condition guarantees uniqueness, and Subsection A.2 shows that a strictly negative definite Jacobian satisfies this condition. Building on this, Subsection A.3 derives an explicit sufficient upper bound on λ that guarantees QRE uniqueness.

A.1 QRE UNIQUENESS AND THE DIAGONALLY STRICTLY CONCAVE CONDITION

In a two-player game, each player iteratively maximizes its respective regularized expected utility function (1). The gradients of the players' objective functions jointly define the pseudo-gradient vector field of the game:

$$g(\sigma_1, \sigma_2) \triangleq \begin{bmatrix} \nabla_{\sigma_1} u_1 \\ \nabla_{\sigma_2} u_2 \end{bmatrix}.$$

Proposition 3 For any two distinct strategy profiles $\boldsymbol{\sigma}^A$ and $\boldsymbol{\sigma}^B$, if the function $g(\cdot)$ satisfies the following Diagonally Strictly Concave (DSC) condition, defined as

$$(\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B)^\top (g(\boldsymbol{\sigma}^A) - g(\boldsymbol{\sigma}^B)) < 0, \tag{15}$$

then the QRE is unique.

Proof. We prove the proposition by contradiction. We assume that there exist multiple QREs, and consider two distinct equilibrium points $\boldsymbol{\sigma}^A$ and $\boldsymbol{\sigma}^B$. For each player i, the strategy $\boldsymbol{\sigma}_i$ is subject to two constraints:

- **Positivity Constraint:** Due to the infinite marginal utility of the entropy regularization term as $\sigma_{ij} \to 0$, the non-negativity constraint $\sigma_{ij} \geq 0$ is strictly inactive. Consequently, the corresponding Lagrange multipliers are zero and can be omitted from the stationary equations.
- **Simplex Constraint:** The probabilities must sum to 1, defined as $h(\boldsymbol{\sigma}_i) \triangleq 1 - \sum_j \sigma_{ij} = 0$.

From the Karush-Kuhn-Tucker (KKT) conditions, for all $i \in \{1, 2\}$, we have:

$$\nabla_{\sigma_i} u_i(\boldsymbol{\sigma}_i^A) + \gamma_i^A \nabla_{\sigma_i} h(\boldsymbol{\sigma}_i^A) = 0, \qquad \nabla_{\sigma_i} u_i(\boldsymbol{\sigma}_i^B) + \gamma_i^B \nabla_{\sigma_i} h(\boldsymbol{\sigma}_i^B) = 0.$$

Here, γ_i^A, γ_i^B are the Lagrange multipliers associated with player i's simplex constraint. Multiplying both equations by $(\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B)^\top$ and subtracting the second equation from the first over all players yields:

$$(\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B)^\top \left(g(\boldsymbol{\sigma}^A) - g(\boldsymbol{\sigma}^B)\right) + \sum_i (\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B)^\top \left(\gamma_i^A \nabla_{\sigma_i} h(\boldsymbol{\sigma}_i^A) - \gamma_i^B \nabla_{\sigma_i} h(\boldsymbol{\sigma}_i^B)\right) = 0.$$

Given the DSC condition, the first term evaluates to a strictly negative value, implying that the remaining summation must be strictly positive. However, since both $\boldsymbol{\sigma}^A$ and $\boldsymbol{\sigma}^B$ reside on the probability simplex, the sum of the elements in the deviation vector $\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B$ is exactly 0. Furthermore, the gradient of the simplex constraint is $\nabla_{\boldsymbol{\sigma}_i} h(\boldsymbol{\sigma}_i) = -\mathbf{1}$. Thus, the inner product evaluates to 0, producing a contradiction. Hence, the QRE must be unique. □

A.2 NEGATIVE DEFINITENESS OF THE JACOBIAN

Proposition 4 Let $G(\boldsymbol{\sigma})$ denote the Jacobian matrix of $g(\boldsymbol{\sigma})$. If $G(\boldsymbol{\sigma}) + G^\top(\boldsymbol{\sigma})$ is strictly negative definite on all feasible tangent directions, then $g(\boldsymbol{\sigma})$ strictly satisfies the DSC condition.

Proof. For any two distinct strategy profiles $\boldsymbol{\sigma}^A$ and $\boldsymbol{\sigma}^B$, we parameterize the line segment connecting them as $\boldsymbol{\sigma}^x(\theta) = \theta\boldsymbol{\sigma}^A + (1-\theta)\boldsymbol{\sigma}^B$, where $\theta \in [0, 1]$. Applying the chain rule, we obtain the derivative of $g(\boldsymbol{\sigma}^x(\theta))$ with respect to θ:

$$\frac{dg(\boldsymbol{\sigma}^x(\theta))}{d\theta} = G(\boldsymbol{\sigma}^x(\theta)) \frac{d\boldsymbol{\sigma}^x(\theta)}{d\theta} = G(\boldsymbol{\sigma}^x(\theta))(\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B).$$

Integrating this derivative from $\theta = 0$ to $\theta = 1$ yields:

$$g(\boldsymbol{\sigma}^A) - g(\boldsymbol{\sigma}^B) = \int_0^1 G(\boldsymbol{\sigma}^x(\theta))(\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B)d\theta.$$

Left-multiplying both sides by $(\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B)^\top$, we obtain:

$$\begin{aligned}
&(\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B)^\top \left(g(\boldsymbol{\sigma}^A) - g(\boldsymbol{\sigma}^B)\right) \\
=&\int_0^1 (\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B)^\top G(\boldsymbol{\sigma}^x(\theta))(\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B)d\theta \\
=&\frac{1}{2}\int_0^1 (\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B)^\top \left[G(\boldsymbol{\sigma}^x(\theta)) + G^\top(\boldsymbol{\sigma}^x(\theta))\right](\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B)d\theta.
\end{aligned}$$

By assumption, the symmetric matrix $G(\boldsymbol{\sigma}) + G^\top(\boldsymbol{\sigma})$ is strictly negative definite. Consequently, the quadratic form within the integral is strictly negative for any non-zero vector $(\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B)$, causing the entire expression to evaluate to a strictly negative value. Rearranging the resulting inequality, $(\boldsymbol{\sigma}^A - \boldsymbol{\sigma}^B)^\top \left(g(\boldsymbol{\sigma}^A) - g(\boldsymbol{\sigma}^B)\right) < 0$, directly yields the exact form of the DSC condition. Thus, the condition is strictly satisfied. □

A.3 Upper Bound on λ for QRE Uniqueness

Proposition 5 If the regularization parameter satisfies $\lambda < \frac{2}{\|P_m(A+B)P_n\|_{\max}}$, where $P_m = \mathbb{I}_m - \frac{1}{m}\mathbf{1}\mathbf{1}^\top$, $P_n = \mathbb{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ are the projection matrices, then $G(\boldsymbol{\sigma}) + G^\top(\boldsymbol{\sigma})$ is strictly negative definite on the tangent space of $\Delta(\Pi_1) \times \Delta(\Pi_2)$.

Proof. Since any strategy profile $\boldsymbol{\sigma}$ and its perturbed state must reside on the probability simplex, their elements must sum to 1. Consequently, any valid perturbation vector $\mathbf{v} = [x^\top, y^\top]^\top$ must lie in the tangent space of the simplex, meaning $1_m^\top x = 0$ and $1_n^\top y = 0$. We introduce the projection matrices P_m and P_n, which project vectors onto the corresponding simplex tangent spaces, ensuring $P_m x = x$ and $P_n y = y$.

The Jacobian matrix $G(\boldsymbol{\sigma})$ and its symmetric part are given by:

$$G(\boldsymbol{\sigma}) + G^\top(\boldsymbol{\sigma}) = \begin{bmatrix} -\frac{2}{\lambda}\text{diag}(\boldsymbol{\sigma}_1)^{-1} & A + B \\ A^\top + B^\top & -\frac{2}{\lambda}\text{diag}(\boldsymbol{\sigma}_2)^{-1} \end{bmatrix}.$$

To ensure strict negative definiteness, the quadratic form $\mathbf{v}^\top(G + G^\top)\mathbf{v}$ must be strictly negative for any non-zero $\mathbf{v}$:

$$-\frac{2}{\lambda}x^\top \text{diag}(\boldsymbol{\sigma}_1)^{-1}x - \frac{2}{\lambda}y^\top \text{diag}(\boldsymbol{\sigma}_2)^{-1}y + 2x^\top(A + B)y < 0.$$

Using the projection property $x = P_m x$ and $y = P_n y$, and the fact that P_m and P_n are symmetric, we can rewrite the cross-interaction term as $x^\top(A + B)y = x^\top P_m(A + B)P_n y$. Rearranging the inequality, we aim to satisfy:

$$\frac{2}{\lambda}\left(x^\top \text{diag}(\boldsymbol{\sigma}_1)^{-1}x + y^\top \text{diag}(\boldsymbol{\sigma}_2)^{-1}y\right) > 2x^\top P_m(A + B)P_n y.$$

Step 1: Lower Bounding the Left Side. By applying the Cauchy-Schwarz inequality, $\left(\sum_i x_i^2/\boldsymbol{\sigma}_{1,i}\right)\left(\sum_i \boldsymbol{\sigma}_{1,i}\right) \geq \left(\sum_i |x_i|\right)^2$. Since $\sum_i \boldsymbol{\sigma}_{1,i} = 1$, this simplifies to $x^\top \text{diag}(\boldsymbol{\sigma}_1)^{-1}x \geq \|x\|_1^2$. Applying this to both players, the left side is lower-bounded by $\frac{2}{\lambda}(\|x\|_1^2 + \|y\|_1^2)$.

Step 2: Upper Bounding the Right Side. Let $C = P_m(A + B)P_n$. The bilinear term is upper-bounded by extracting its maximum absolute element $\|C\|_{\max} = \max_{i,j}|C_{ij}|$:

$$2x^\top Cy \leq 2\|C\|_{\max}\|x\|_1\|y\|_1 \leq \|C\|_{\max}(\|x\|_1^2 + \|y\|_1^2).$$

Combining both bounds, the negative definite condition on the tangent space of $\Delta(\Pi_1) \times \Delta(\Pi_2)$ is strictly satisfied if:

$$\frac{2}{\lambda}(\|x\|_1^2 + \|y\|_1^2) > \|P_m(A+B)P_n\|_{\max}(\|x\|_1^2 + \|y\|_1^2).$$

Dividing by the strictly positive term $(\|x\|_1^2 + \|y\|_1^2)$, we obtain the threshold $\lambda < \frac{2}{\|P_m(A+B)P_n\|_{\max}}$. □

According to Propositions 3, 4, and 5, $\lambda < \frac{2}{\|P_m(A+B)P_n\|_{\max}}$ is a sufficient condition for the QRE uniqueness. In two-player zero-sum games, we have $B = -A$. In this case, $\|P_m(A+B)P_n\|_{\max} = 0$, implying that the QRE uniqueness unconditionally holds for any $\lambda > 0$.

B Proof of Proposition 1 (The QRE Fixed-Point System)

We derive the logit fixed-point characterization of the quantal response equilibrium used in (5).

Proof. Let $M = |\Pi_1|$ and $N = |\Pi_2|$. Denote the mixed strategies of players 1 and 2 by $x \in \Delta_M$ and $y \in \Delta_N$, respectively. Given the mixed strategy y of player 2, player 1 solves the entropy-regularized best-response problem:

$$\max_{x \in \Delta_M} \; x^\top U_1(\theta) y - \frac{1}{\lambda} \sum_{m=1}^{M} x_m \log x_m. \tag{16}$$

The entropy regularization term $-\sum_m x_m \log x_m$ is strictly concave on the interior of the probability simplex. Moreover, its marginal derivative tends to positive infinity as any coordinate x_m approaches zero.

Introducing a Lagrange multiplier μ_1 for the simplex constraint $\mathbf{1}^\top x = 1$, the Lagrangian is

$$\mathcal{L}_1(x, \mu_1) = x^\top U_1(\theta) y - \frac{1}{\lambda} \sum_{m=1}^{M} x_m \log x_m + \mu_1(1 - \mathbf{1}^\top x).$$

The first-order condition for each $m \in \{1, \ldots, M\}$ is

$$[U_1(\theta) y]_m - \frac{1}{\lambda}(1 + \log x_m) - \mu_1 = 0.$$

Rearranging it gives

$$\log x_m = \lambda [U_1(\theta) y]_m - \lambda \mu_1 - 1,$$

and hence

$$x_m = \exp(-\lambda \mu_1 - 1) \exp(\lambda [U_1(\theta) y]_m).$$

The factor $\exp(-\lambda \mu_1 - 1)$ is shared by all coordinates. Using $\sum_{m=1}^{M} x_m = 1$, we obtain

$$x_m = \frac{\exp(\lambda [U_1(\theta) y]_m)}{\sum_{u=1}^{M} \exp(\lambda [U_1(\theta) y]_u)}.$$

Therefore,

$$x = \mathrm{softmax}(\lambda U_1(\theta) y).$$

Similarly,

$$y = \mathrm{softmax}\big(\lambda U_2(\theta)^\top x\big).$$

At a QRE, each player simultaneously adopts its unique entropy-regularized best response to the other player's mixed strategy. Thus, the equilibrium satisfies (5). □

C Proof of Proposition 2 (Differential of linear system)

Proof. Let $s(z) = \mathrm{softmax}(z)$. For any vector $z \in \mathbb{R}^d$, the j-th component of $s(z)$ is

$$s_j(z) = \frac{\exp(z_j)}{\sum_{k=1}^{d} \exp(z_k)}.$$

Its partial derivative with respect to z_k is

$$\frac{\partial s_j(z)}{\partial z_k} = s_j(z)(\mathbb{I}\{j = k\} - s_k(z)).$$

Therefore, the Jacobian of the softmax mapping is $\mathrm{diag}(s(z)) - s(z)s(z)^\top$.

We first consider player 1. At the QRE,

$$x = \mathrm{softmax}(\lambda U_1(\theta) y)\,.$$

Applying the total differential, we obtain

$$dx = \big(\mathrm{diag}(x) - xx^\top\big)\, d(\lambda U_1(\theta) y)\,.$$

By the product rule,

$$d(U_1(\theta) y) = dU_1(\theta) y + U_1(\theta) dy.$$

Defining $\mathbb{D}_x = \mathrm{diag}(x) - xx^\top$ therefore gives

$$dx = \lambda \mathbb{D}_x \left(dU_1(\theta) y + U_1(\theta) dy\right). \tag{17}$$

Similarly, the QRE condition for player 2 is

$$y = \mathrm{softmax}\big(\lambda U_2(\theta)^\top x\big)\,.$$

Applying the total differential yields

$$dy = \big(\mathrm{diag}(y) - yy^\top\big)\, d\big(\lambda U_2(\theta)^\top x\big)\,.$$

Using the product rule again,

$$d\big(U_2(\theta)^\top x\big) = d\big(U_2(\theta)^\top\big)\, x + U_2(\theta)^\top dx.$$

Thus, with $\mathbb{D}_y = \mathrm{diag}(y) - yy^\top$, we have

$$dy = \lambda \mathbb{D}_y \left(d\big(U_2(\theta)^\top\big)\, x + U_2(\theta)^\top dx\right). \tag{18}$$

Equations (17) and (18) establish (6). □

D Proof of Theorem 1 (Gradient Approximation Error)

D.1 Preliminary Lemmas

To prove Theorem 1, we first prove several lemmas.

Lemma 1 For any $\boldsymbol{\sigma} = (x, y) \in B_r(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta))$, the response-map approximation error is upper-bounded as

$$\left\|\Phi_{\Pi}(\boldsymbol{\sigma}, \theta) - \bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma}, \theta)\right\|_1 \le 2(|\mathcal{R}_1| + |\mathcal{R}_2|)e^{-\lambda\Delta}. \tag{19}$$

Proof. Given $y \in \Delta(\Pi_2)$, define

$$D^{(1)}_{\Pi}(y) \triangleq \sum_{c \in \Pi_1} \exp \lambda J_1(c, y; \theta),$$
$$D^{(1)}_{\mathcal{S}}(y) \triangleq \sum_{c \in \mathcal{S}_1} \exp \lambda J_1(c, y; \theta),$$
$$D^{(1)}_{\mathcal{R}}(y) \triangleq \sum_{c \in \mathcal{R}_1} \exp \lambda J_1(c, y; \theta).$$

Since $\mathcal{R}_1 = \Pi_1 \setminus \mathcal{S}_1$, we have $D^{(1)}_{\Pi}(y) = D^{(1)}_{\mathcal{S}}(y) + D^{(1)}_{\mathcal{R}}(y)$. For $\boldsymbol{\sigma} = (x, y) \in B_r(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta))$, define $V_1^{\mathcal{S}}(y;\theta) \triangleq \max_{c\in\mathcal{S}_1} J_1(c, y;\theta)$. By Assumption 1, for any $a \in \mathcal{R}_1$, $J_1(a, y;\theta) \le V_1^{\mathcal{S}}(y;\theta) - \Delta$. Therefore,

$$\exp\{\lambda J_1(a, y;\theta)\} \le \exp\{\lambda(V_1^{\mathcal{S}}(y;\theta) - \Delta)\} = e^{-\lambda\Delta} \exp\{\lambda V_1^{\mathcal{S}}(y;\theta)\}.$$

We now bound the ratio between the omitted-policy mass and the retained-policy mass:

$$\frac{D^{(1)}_{\mathcal{R}}(y)}{D^{(1)}_{\mathcal{S}}(y)} = \frac{\sum_{a\in\mathcal{R}_1} \exp\{\lambda J_1(a, y;\theta)\}}{\sum_{c\in\mathcal{S}_1} \exp\{\lambda J_1(c, y;\theta)\}}.$$

The numerator satisfies

$$\sum_{a\in\mathcal{R}_1} \exp\{\lambda J_1(a, y;\theta)\} \le |\mathcal{R}_1| e^{-\lambda\Delta} \exp\{\lambda V_1^{\mathcal{S}}(y;\theta)\}.$$

Since the maximum $V_1^{\mathcal{S}}(y;\theta)$ is attained by at least one retained policy, the denominator satisfies

$$\sum_{c\in\mathcal{S}_1} \exp\{\lambda J_1(c, y;\theta)\} \ge \exp\{\lambda V_1^{\mathcal{S}}(y;\theta)\}.$$

It follows that

$$\sum_{a\in\mathcal{R}_1} \Phi_{\Pi,1}(y,\theta)_a = \frac{D^{(1)}_{\mathcal{R}}(y)}{D^{(1)}_{\Pi}(y)} = \frac{D^{(1)}_{\mathcal{R}}(y)}{D^{(1)}_{\mathcal{R}}(y) + D^{(1)}_{\mathcal{S}}(y)} \le \frac{D^{(1)}_{\mathcal{R}}(y)}{D^{(1)}_{\mathcal{S}}(y)} \le |\mathcal{R}_1| e^{-\lambda\Delta}.$$

The same argument applies to player 2 and yields

$$\sum_{b\in\mathcal{R}_2} \Phi_{\Pi,2}(x,\theta)_b \le |\mathcal{R}_2| e^{-\lambda\Delta}.$$

We define $\eta_1(y) \triangleq \frac{D^{(1)}_{\mathcal{R}}(y)}{D^{(1)}_{\mathcal{S}}(y)}$, and for any $a \in \mathcal{S}_1$, we have

$$\Phi_{\Pi,1}(y,\theta)_a = \frac{\exp\{\lambda J_1(a, y;\theta)\}}{D^{(1)}_{\mathcal{S}}(y) + D^{(1)}_{\mathcal{R}}(y)} = \frac{\bar{\Phi}_{\mathcal{S},1}(y,\theta)_a}{1 + \eta_1(y)}.$$

Therefore,

$$\left|\Phi_{\Pi,1}(y,\theta)_a - \bar{\Phi}_{\mathcal{S},1}(y,\theta)_a\right| = \bar{\Phi}_{\mathcal{S},1}(y,\theta)_a \frac{\eta_1(y)}{1 + \eta_1(y)} \le \bar{\Phi}_{\mathcal{S},1}(y,\theta)_a \eta_1(y).$$

Summing over $a \in \mathcal{S}_1$ gives

$$\sum_{a\in\mathcal{S}_1} \left|\Phi_{\Pi,1}(y,\theta)_a - \bar{\Phi}_{\mathcal{S},1}(y,\theta)_a\right| \le \eta_1(y) \le |\mathcal{R}_1| e^{-\lambda\Delta}.$$

For $a \in \mathcal{R}_1$, we have $\bar{\Phi}_{\mathcal{S},1}(y,\theta)_a = 0$, and hence

$$\sum_{a\in\mathcal{R}_1} \left|\Phi_{\Pi,1}(y,\theta)_a - \bar{\Phi}_{\mathcal{S},1}(y,\theta)_a\right| = \sum_{a\in\mathcal{R}_1} \Phi_{\Pi,1}(y,\theta)_a \le |\mathcal{R}_1| e^{-\lambda\Delta}.$$

Combining the retained and omitted parts, we obtain

$$\|\Phi_{\Pi,1}(y,\theta) - \bar{\Phi}_{\mathcal{S},1}(y,\theta)\|_1 \le 2|\mathcal{R}_1| e^{-\lambda\Delta},$$
$$\|\Phi_{\Pi,2}(x,\theta) - \bar{\Phi}_{\mathcal{S},2}(x,\theta)\|_1 \le 2|\mathcal{R}_2| e^{-\lambda\Delta}.$$

Therefore,

$$\|\Phi_{\Pi}(\boldsymbol{\sigma},\theta) - \bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma},\theta)\|_1 \le 2(|\mathcal{R}_1| + |\mathcal{R}_2|) e^{-\lambda\Delta}.$$

□

Lemma 2 For any $\boldsymbol{\sigma} \in B_r(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta))$, the full response map satisfies

$$\Phi_\Pi(\boldsymbol{\sigma}, \theta) \in B_r(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)). \tag{20}$$

Proof. Since $\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)$ is the extended restricted QRE, it satisfies

$$\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta) = \bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta).$$

Therefore, we need to prove

$$\left\|\Phi_\Pi(\boldsymbol{\sigma}, \theta) - \bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\right\|_1 \le r.$$

Adding and subtracting $\bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma}, \theta)$, we obtain

$$\|\Phi_\Pi(\boldsymbol{\sigma}, \theta) - \bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}, \theta)\|_1 \le \|\Phi_\Pi(\boldsymbol{\sigma}, \theta) - \bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma}, \theta)\|_1 + \|\bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma}, \theta) - \bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}_{\mathcal{S}}(\theta)^*, \theta)\|_1.$$

By Lemma 1,

$$\left\|\Phi_\Pi(\boldsymbol{\sigma}, \theta) - \bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma}, \theta)\right\|_1 \le 2(|\mathcal{R}_1| + |\mathcal{R}_2|)e^{-\lambda\Delta}.$$

By Assumption 2,

$$\left\|\bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma}, \theta) - \bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\right\|_1 \le L \left\|\boldsymbol{\sigma} - \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\right\|_1.$$

Since $\boldsymbol{\sigma} \in B_r(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta))$, we have

$$\left\|\boldsymbol{\sigma} - \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\right\|_1 \le r.$$

Thus,

$$\left\|\bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma}, \theta) - \bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\right\|_1 \le Lr.$$

Combining the above inequalities gives

$$\left\|\Phi_\Pi(\boldsymbol{\sigma}, \theta) - \bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\right\|_1 \le 2(|\mathcal{R}_1| + |\mathcal{R}_2|)e^{-\lambda\Delta} + Lr.$$

By Assumption 4,

$$2(|\mathcal{R}_1| + |\mathcal{R}_2|)e^{-\lambda\Delta} + Lr \le r.$$

Therefore,

$$\left\|\Phi_\Pi(\boldsymbol{\sigma}, \theta) - \bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\right\|_1 \le r,$$

which proves that $\Phi_\Pi(\boldsymbol{\sigma}, \theta) \in B_r(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta))$. □

Together with Assumption 2, this self-mapping property implies that Φ_Π is a contraction from the local neighborhood $B_r(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta))$ into itself. By the Banach fixed-point theorem, Φ_Π admits a unique fixed point in this neighborhood, and the fixed-point iteration initialized from any point in the neighborhood converges to this fixed point. Since the full-space QRE branch is unique, this local fixed point coincides with the full-space QRE $\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta)$. Therefore,

$$\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta) \in B_r(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)).$$

Lemma 3 The distance between the full-space QRE and the extended restricted QRE satisfies

$$\|\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta) - \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\|_1 \le r. \tag{21}$$

Proof. Since both QREs are fixed points of their corresponding response maps, we have

$$\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta) = \Phi_\Pi(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta), \theta), \qquad \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta) = \bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta).$$

Therefore,

$$\begin{aligned}\|\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta) - \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\|_1 &= \|\Phi_\Pi(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta), \theta) - \bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\|_1 \\ &\le \|\Phi_\Pi(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta), \theta) - \bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta), \theta)\|_1 + \|\bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta), \theta) - \bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\|_1.\end{aligned}$$

By Lemma 2, the full-space QRE lies in the local neighborhood $B_r(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta))$. Hence, Lemma 1 gives

$$\|\Phi_\Pi(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta), \theta) - \bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta), \theta)\|_1 \le 2(|\mathcal{R}_1| + |\mathcal{R}_2|)e^{-\lambda\Delta}.$$

Moreover, by Assumption 2,

$$\|\bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta),\theta) - \bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta),\theta)\|_1 \le L\|\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta) - \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\|_1.$$

Combining the above inequalities yields

$$\|\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta) - \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\|_1 \le 2(|\mathcal{R}_1| + |\mathcal{R}_2|)e^{-\lambda\Delta} + L\|\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta) - \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\|_1.$$

Rearranging terms gives

$$(1-L)\|\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta) - \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\|_1 \le 2(|\mathcal{R}_1| + |\mathcal{R}_2|)e^{-\lambda\Delta}.$$

Since $L < 1$, Assumption 4 yields

$$\|\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta) - \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\|_1 \le \frac{2(|\mathcal{R}_1| + |\mathcal{R}_2|)}{1-L}e^{-\lambda\Delta} \le r.$$

□

Lemma 4 The first-order derivatives of the full response map and the zero-padded restricted response map are exponentially close. Specifically, there exist constants $C_\sigma > 0$ and $C_\theta > 0$, independent of Δ, such that for any $\boldsymbol{\sigma} \in B_r(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta))$,

$$\left\|\mathrm{Jac}_{\boldsymbol{\sigma}}\,\Phi_\Pi(\boldsymbol{\sigma},\theta) - \mathrm{Jac}_{\boldsymbol{\sigma}}\,\bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma},\theta)\right\|_1 \le C_\sigma e^{-\lambda\Delta}, \tag{22}$$

and

$$\left\|\mathrm{Jac}_{\theta}\,\Phi_\Pi(\boldsymbol{\sigma},\theta) - \mathrm{Jac}_{\theta}\,\bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma},\theta)\right\|_1 \le C_\theta e^{-\lambda\Delta}. \tag{23}$$

Proof. We prove the bound for the derivative with respect to the mixed strategy. The proof for the derivative with respect to θ follows from the same argument and the smoothness assumption. Consider player 1 and fix $y \in \Delta(\Pi_2)$. Let $p = \Phi_{\Pi,1}(y,\theta)$ and $q = \bar{\Phi}_{\mathcal{S},1}(y,\theta)$. By Lemma 1, we have

$$\|p - q\|_1 \le 2|\mathcal{R}_1|e^{-\lambda\Delta}.$$

For $a \in \Pi_1$, write

$$p_a = \frac{\exp\{\lambda J_1(a,y;\theta)\}}{\sum_{c\in\Pi_1}\exp\{\lambda J_1(c,y;\theta)\}}, \qquad J_1(c,y;\theta) = \sum_{d\in\Pi_2} y_d J_1(c,d;\theta).$$

By the chain rule,

$$\frac{\partial p_a}{\partial y_d} = \sum_{c\in\Pi_1}\frac{\partial p_a}{\partial J_1(c,y;\theta)}\frac{\partial J_1(c,y;\theta)}{\partial y_d}.$$

The softmax derivative satisfies

$$\frac{\partial p_a}{\partial J_1(c,y;\theta)} = \begin{cases} \lambda p_a(1-p_c), & c = a, \\ -\lambda p_a p_c, & c \ne a. \end{cases}$$

Since $\frac{\partial J_1(c,y;\theta)}{\partial y_d} = J_1(c,d;\theta)$, we obtain

$$\frac{\partial p_a}{\partial y_d} = \lambda p_a\left[J_1(a,d;\theta) - \sum_{c\in\Pi_1} p_c J_1(c,d;\theta)\right].$$

Similarly, for $a \in \mathcal{S}_1$,

$$\frac{\partial q_a}{\partial y_d} = \lambda q_a\left[J_1(a,d;\theta) - \sum_{c\in\mathcal{S}_1} q_c J_1(c,d;\theta)\right],$$

while for $a \in \mathcal{R}_1$, $q_a = 0$, and hence its derivative is also zero.

We next show that the difference between the two derivatives is exponentially small. By Assumption 3, there exists a constant $B_J > 0$ such that $|J_i(c,d;\theta)| \le B_J$ for all relevant policies and all θ in the local neighborhood.

First consider $a \in \mathcal{R}_1$. Since $\frac{\partial q_a}{\partial y_d} = 0$, we have

$$\left|\frac{\partial p_a}{\partial y_d} - \frac{\partial q_a}{\partial y_d}\right| = \left|\frac{\partial p_a}{\partial y_d}\right|.$$

Using the derivative formula above,

$$\left|\frac{\partial p_a}{\partial y_d}\right| = \lambda p_a \left| J_1(a,d;\theta) - \sum_{c\in\Pi_1} p_c J_1(c,d;\theta)\right| \le 2\lambda B_J p_a.$$

Therefore,

$$\sum_{a\in\mathcal{R}_1}\left|\frac{\partial p_a}{\partial y_d} - \frac{\partial q_a}{\partial y_d}\right| \le 2\lambda B_J \sum_{a\in\mathcal{R}_1} p_a.$$

By the tail bound in Lemma 1,

$$\sum_{a\in\mathcal{R}_1} p_a \le |\mathcal{R}_1| e^{-\lambda\Delta}.$$

Thus,

$$\sum_{a\in\mathcal{R}_1}\left|\frac{\partial p_a}{\partial y_d} - \frac{\partial q_a}{\partial y_d}\right| \le 2\lambda B_J |\mathcal{R}_1| e^{-\lambda\Delta}.$$

Now consider $a \in \mathcal{S}_1$. For simplicity, define

$$h_c \triangleq J_1(c,d;\theta), \qquad m_p \triangleq \sum_{c\in\Pi_1} p_c h_c, \qquad m_q \triangleq \sum_{c\in\Pi_1} q_c h_c.$$

Since $q_c = 0$ for $c \in \mathcal{R}_1$, this definition of m_q is equivalent to summing over $\mathcal{S}_1$. For $a \in \mathcal{S}_1$,

$$\frac{\partial p_a}{\partial y_d} - \frac{\partial q_a}{\partial y_d} = \lambda\left[p_a(h_a - m_p) - q_a(h_a - m_q)\right].$$

Adding and subtracting $q_a m_p$ gives

$$p_a(h_a - m_p) - q_a(h_a - m_q) = (p_a - q_a)(h_a - m_p) + q_a(m_q - m_p).$$

Therefore,

$$\left|\frac{\partial p_a}{\partial y_d} - \frac{\partial q_a}{\partial y_d}\right| \le \lambda\left[|p_a - q_a||h_a - m_p| + q_a|m_q - m_p|\right].$$

Since $|h_a| \le B_J$ and $|m_p| \le B_J$, we have $|h_a - m_p| \le 2B_J$. Moreover,

$$|m_q - m_p| = \left|\sum_{c\in\Pi_1}(q_c - p_c)h_c\right| \le B_J \sum_{c\in\Pi_1}|q_c - p_c| = B_J\|q - p\|_1 \le 2B_J|\mathcal{R}_1|e^{-\lambda\Delta}.$$

Summing over $a \in \mathcal{S}_1$ yields

$$\begin{aligned}
&\sum_{a\in\mathcal{S}_1}\left|\frac{\partial p_a}{\partial y_d} - \frac{\partial q_a}{\partial y_d}\right| \\
\le&\lambda\left[2B_J\sum_{a\in\mathcal{S}_1}|p_a - q_a| + 2B_J|\mathcal{R}_1|e^{-\lambda\Delta}\sum_{a\in\mathcal{S}_1} q_a\right] \\
\le&4\lambda B_J|\mathcal{R}_1|e^{-\lambda\Delta}.
\end{aligned}$$

Combining the bounds over $\mathcal{S}_1$ and $\mathcal{R}_1$, for every $d \in \Pi_2$ we obtain

$$\sum_{a\in\Pi_1}\left|\frac{\partial p_a}{\partial y_d} - \frac{\partial q_a}{\partial y_d}\right| \le 6\lambda B_J|\mathcal{R}_1|e^{-\lambda\Delta}.$$

Thus, each column of $\frac{\partial\Phi_{\Pi,1}}{\partial y} - \frac{\partial\bar{\Phi}_{\mathcal{S},1}}{\partial y}$ has an exponentially small ℓ_1 norm. The same argument for player 2 gives an analogous bound for $\frac{\partial\Phi_{\Pi,2}}{\partial x} - \frac{\partial\bar{\Phi}_{\mathcal{S},2}}{\partial x}$.

The full response map has the block Jacobian

$$\mathrm{Jac}_{\boldsymbol{\sigma}}\,\Phi_\Pi(\boldsymbol{\sigma},\theta)=\begin{bmatrix}0 & \frac{\partial\Phi_{\Pi,1}}{\partial y}\\ \frac{\partial\Phi_{\Pi,2}}{\partial x} & 0\end{bmatrix},$$

and similarly

$$\mathrm{Jac}_{\boldsymbol{\sigma}}\,\bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma},\theta)=\begin{bmatrix}0 & \frac{\partial\bar{\Phi}_{\mathcal{S},1}}{\partial y}\\ \frac{\partial\bar{\Phi}_{\mathcal{S},2}}{\partial x} & 0\end{bmatrix}.$$

Therefore, there exists a constant $C_\sigma > 0$, independent of Δ, such that

$$\|\,\mathrm{Jac}_{\boldsymbol{\sigma}}\,\Phi_\Pi(\boldsymbol{\sigma},\theta)-\mathrm{Jac}_{\boldsymbol{\sigma}}\,\bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma},\theta)\|_1\le C_\sigma e^{-\lambda\Delta}.$$

It remains to control the derivative with respect to θ, with $\boldsymbol{\sigma}=(x,y)$ held fixed. We again consider player 1 and write $p=\Phi_{\Pi,1}(y,\theta)$ and $q=\bar{\Phi}_{\mathcal{S},1}(y,\theta)$. By the softmax chain rule,

$$\nabla_\theta p_a=\lambda p_a\left[\nabla_\theta J_1(a,y;\theta)-\sum_{c\in\Pi_1}p_c\nabla_\theta J_1(c,y;\theta)\right],$$

and the same formula holds for q_a with the summation restricted to $\mathcal{S}_1$, while $\nabla_\theta q_a=0$ for $a\in\mathcal{R}_1$. This is identical in form to the derivative with respect to y_d, except that the term $J_1(c,d;\theta)$ is replaced by $\nabla_\theta J_1(c,y;\theta)$. Since Assumption 3 gives a uniform bound on these parameter gradients, the retained-policy and omitted-policy bounds above apply unchanged. Therefore, the same argument for both players yields a constant $C_\theta > 0$, independent of Δ, such that

$$\left\|\mathrm{Jac}_\theta\,\Phi_\Pi(\boldsymbol{\sigma},\theta)-\mathrm{Jac}_\theta\,\bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma},\theta)\right\|_1\le C_\theta e^{-\lambda\Delta}.$$

□

Lemma 5 The derivatives of the full-space QRE and the extended restricted QRE are exponentially close. Specifically, there exists a constant $C_d > 0$, independent of Δ, such that

$$\|\nabla_\theta\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta)-\nabla_\theta\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\|_1\le C_d e^{-\lambda\Delta}. \tag{24}$$

Proof. The full-space QRE satisfies

$$\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta)=\Phi_\Pi(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta),\theta).$$

Differentiating both sides with respect to θ gives

$$\nabla_\theta\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta)=\mathrm{Jac}_{\boldsymbol{\sigma}}\,\Phi_\Pi(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta),\theta)\nabla_\theta\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta)+\mathrm{Jac}_\theta\,\Phi_\Pi(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta),\theta).$$

Define

$$A_\Pi\triangleq\mathrm{Jac}_{\boldsymbol{\sigma}}\,\Phi_\Pi(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta),\theta),\qquad b_\Pi\triangleq\mathrm{Jac}_\theta\,\Phi_\Pi(\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta),\theta).$$

Then

$$(I-A_\Pi)\nabla_\theta\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta)=b_\Pi.$$

By Assumption 2, $\|A_\Pi\|_1\le L<1$, and hence $I-A_\Pi$ is invertible with

$$\|(I-A_\Pi)^{-1}\|_1\le\sum_{k=0}^{\infty}\|A_\Pi\|_1^k\le\frac{1}{1-L}.$$

Therefore,

$$\nabla_\theta\boldsymbol{\sigma}^*_{\boldsymbol{\Pi}}(\theta)=(I-A_\Pi)^{-1}b_\Pi.$$

Similarly, the extended restricted QRE satisfies

$$\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)=\bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta),\theta).$$

Define

$$A_{\mathcal{S}}\triangleq\mathrm{Jac}_{\boldsymbol{\sigma}}\,\bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta),\theta),\qquad b_{\mathcal{S}}\triangleq\mathrm{Jac}_\theta\,\bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta),\theta).$$

Then

$$\nabla_\theta \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta) = (I - A_{\mathcal{S}})^{-1} b_{\mathcal{S}}, \qquad \|(I - A_{\mathcal{S}})^{-1}\|_1 \le \frac{1}{1-L}.$$

We next compare the two implicit systems. By adding and subtracting intermediate terms,

$$\begin{aligned}\|A_\Pi - A_{\mathcal{S}}\|_1 \le &\left\|\mathrm{Jac}_{\boldsymbol{\sigma}}\, \Phi_\Pi(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta) - \mathrm{Jac}_{\boldsymbol{\sigma}}\, \bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta)\right\|_1 \\ &+ \left\|\mathrm{Jac}_{\boldsymbol{\sigma}}\, \bar{\Phi}_{\mathcal{S}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta) - \mathrm{Jac}_{\boldsymbol{\sigma}}\, \bar{\Phi}_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\right\|_1.\end{aligned}$$

The first term is bounded by Lemma 4. The second term is bounded by the Lipschitz continuity of $\mathrm{Jac}_{\boldsymbol{\sigma}}\, \bar{\Phi}_{\mathcal{S}}$ and Lemma 3. Thus, there exists a constant $C_A > 0$, independent of Δ, such that

$$\|A_\Pi - A_{\mathcal{S}}\|_1 \le C_A e^{-\lambda\Delta}.$$

The same argument, using the Lipschitz continuity of $\mathrm{Jac}_\theta\, \bar{\Phi}_{\mathcal{S}}$ and the θ-derivative bound in Lemma 4, gives a constant $C_b > 0$ such that

$$\|b_\Pi - b_{\mathcal{S}}\|_1 \le C_b e^{-\lambda\Delta}.$$

Now, we compare the two QRE derivatives:

$$\begin{aligned}&\|\nabla_\theta \boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta) - \nabla_\theta \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\|_1 \\ =&\left\|(I - A_\Pi)^{-1} b_\Pi - (I - A_{\mathcal{S}})^{-1} b_{\mathcal{S}}\right\|_1 \\ \le&\|(I - A_\Pi)^{-1}\|_1 \|b_\Pi - b_{\mathcal{S}}\|_1 + \|(I - A_\Pi)^{-1} - (I - A_{\mathcal{S}})^{-1}\|_1 \|b_{\mathcal{S}}\|_1.\end{aligned}$$

Using the matrix inverse identity

$$(I - A_\Pi)^{-1} - (I - A_{\mathcal{S}})^{-1} = (I - A_\Pi)^{-1}(A_\Pi - A_{\mathcal{S}})(I - A_{\mathcal{S}})^{-1},$$

we obtain

$$\|(I - A_\Pi)^{-1} - (I - A_{\mathcal{S}})^{-1}\|_1 \le \frac{C_A}{(1-L)^2} e^{-\lambda\Delta}.$$

Moreover, $b_{\mathcal{S}}$ is uniformly bounded by Assumption 3. Combining the bounds above yields a constant $C_d > 0$, such that

$$\|\nabla_\theta \boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta) - \nabla_\theta \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\|_1 \le C_d e^{-\lambda\Delta}.$$

□

D.2 Proof of Theorem 1

Now, we are ready to prove Theorem 1, i.e., there exists a constant $C_{\mathrm{loc}} > 0$, such that

$$\|\nabla_\theta F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta) - \nabla_\theta F_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\|_1 \le C_{\mathrm{loc}} e^{-\lambda\Delta}.$$

Proof. By the chain rule,

$$\nabla_\theta F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta) = \mathrm{Jac}_{\boldsymbol{\sigma}}\, F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta) \nabla_\theta \boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta) + \partial_\theta F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta),$$

and

$$\nabla_\theta F_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta) = \mathrm{Jac}_{\boldsymbol{\sigma}}\, F_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta) \nabla_\theta \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta) + \partial_\theta F_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta).$$

We first control the direct θ-derivative term. Since $F_{\mathcal{S}}$ is the restriction of $F_{\mathbf{\Pi}}$ to the zero-padded restricted policy space, we have

$$\partial_\theta F_{\mathbf{\Pi}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta) = \partial_\theta F_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta).$$

Therefore,

$$\|\partial_\theta F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta) - \partial_\theta F_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\|_1 = \|\partial_\theta F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta) - \partial_\theta F_{\mathbf{\Pi}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\|_1.$$

By Assumption 3, $\partial_\theta F_{\mathbf{\Pi}}$ is Lipschitz in $\boldsymbol{\sigma}$ on the local neighborhood. Hence, there exists $L_\theta > 0$ such that

$$\|\partial_\theta F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta) - \partial_\theta F_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\|_1 \le L_\theta \|\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta) - \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\|_1.$$

Next, consider the equilibrium-response term. Adding and subtracting $\mathrm{Jac}_{\boldsymbol{\sigma}} F_{\mathbf{\Pi}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\nabla_\theta \boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta)$ gives

$$\begin{aligned}
&\mathrm{Jac}_{\boldsymbol{\sigma}} F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta)\nabla_\theta \boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta) - \mathrm{Jac}_{\boldsymbol{\sigma}} F_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\nabla_\theta \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta) \\
=& \left[\mathrm{Jac}_{\boldsymbol{\sigma}} F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta) - \mathrm{Jac}_{\boldsymbol{\sigma}} F_{\mathbf{\Pi}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\right] \nabla_\theta \boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta) \\
&+ \mathrm{Jac}_{\boldsymbol{\sigma}} F_{\mathbf{\Pi}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta) \left[\nabla_\theta \boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta) - \nabla_\theta \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\right].
\end{aligned}$$

In the last line, we used the zero-padded definition of $F_{\mathcal{S}}$: on the restricted coordinates, its derivative agrees with that of $F_{\mathbf{\Pi}}$, while the omitted coordinates of $\nabla_\theta \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)$ are zero.

Taking the matrix 1-norm and using submultiplicativity,

$$\begin{aligned}
&\left\|\mathrm{Jac}_{\boldsymbol{\sigma}} F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta)\nabla_\theta \boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta) - \mathrm{Jac}_{\boldsymbol{\sigma}} F_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\nabla_\theta \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\right\|_1 \\
\le& \left\|\mathrm{Jac}_{\boldsymbol{\sigma}} F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta) - \mathrm{Jac}_{\boldsymbol{\sigma}} F_{\mathbf{\Pi}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\right\|_1 \left\|\nabla_\theta \boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta)\right\|_1 \\
&+ \left\|\mathrm{Jac}_{\boldsymbol{\sigma}} F_{\mathbf{\Pi}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\right\|_1 \left\|\nabla_\theta \boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta) - \nabla_\theta \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\right\|_1.
\end{aligned}$$

By Assumption 3, there exist constants $L_\sigma > 0$ and $B_F > 0$ such that

$$\left\|\mathrm{Jac}_{\boldsymbol{\sigma}} F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta) - \mathrm{Jac}_{\boldsymbol{\sigma}} F_{\mathbf{\Pi}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\right\|_1 \le L_\sigma \left\|\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta) - \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\right\|_1,$$

and

$$\left\|\mathrm{Jac}_{\boldsymbol{\sigma}} F_{\mathbf{\Pi}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\right\|_1 \le B_F.$$

By Lemma 3, there exists $C_q > 0$ such that

$$\left\|\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta) - \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\right\|_1 \le C_q e^{-\lambda \Delta}.$$

Moreover, the implicit derivative formula, Assumption 2, and Assumption 3 imply that there exists $B_d > 0$ such that

$$\left\|\nabla_\theta \boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta)\right\|_1 \le B_d.$$

Finally, by Lemma 5,

$$\left\|\nabla_\theta \boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta) - \nabla_\theta \bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta)\right\|_1 \le C_d e^{-\lambda \Delta}.$$

Combining the above bounds gives

$$\left\|\nabla_\theta F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta) - \nabla_\theta F_{\mathcal{S}}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}}(\theta), \theta)\right\|_1 \le \left(L_\theta C_q + L_\sigma C_q B_d + B_F C_d\right) e^{-\lambda \Delta}.$$

Setting

$$C_{\mathrm{loc}} = L_\theta C_q + L_\sigma C_q B_d + B_F C_d$$

proves the claim. □

E Proof of Theorem 2 (Convergence of the Surrogate-Gradient Method)

For simplicity, define the true full-space gradient and the restricted gradient at iteration t as

$$g^*_t \triangleq \nabla_\theta F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta_t), \theta_t), \qquad \hat{g}_t \triangleq \nabla_\theta F_{\mathcal{S}_t}(\bar{\boldsymbol{\sigma}}^*_{\mathcal{S}_t}(\theta_t), \theta_t).$$

The update rule is therefore

$$\theta_{t+1} = \theta_t - \eta \hat{g}_t.$$

For notational clarity, let

$$G(\theta) = F_{\mathbf{\Pi}}(\boldsymbol{\sigma}^*_{\mathbf{\Pi}}(\theta), \theta).$$

E.1 Preliminary Lemma

We first introduce the following lemma.

Lemma 6 For any θ_t and θ_{t+1},

$$G(\theta_{t+1}) \le G(\theta_t) + \langle \nabla_\theta G(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{L_{\mathrm{up}}}{2} \|\theta_{t+1} - \theta_t\|_2^2. \tag{25}$$

Proof. Define the one-dimensional path between θ_t and θ_{t+1} by

$$\theta(s) \triangleq \theta_t + s(\theta_{t+1} - \theta_t), \qquad s \in [0, 1].$$

By the fundamental theorem of calculus and chain rule,

$$G(\theta_{t+1}) - G(\theta_t) = \int_0^1 \langle \nabla_\theta G(\theta(s)), \theta_{t+1} - \theta_t \rangle \, ds.$$

Adding and subtracting $\nabla_\theta G(\theta_t)$ inside the integral gives

$$G(\theta_{t+1}) - G(\theta_t) = \langle \nabla_\theta G(\theta_t), \theta_{t+1} - \theta_t \rangle + \int_0^1 \langle \nabla_\theta G(\theta(s)) - \nabla_\theta G(\theta_t), \theta_{t+1} - \theta_t \rangle \, ds.$$

By the Cauchy–Schwarz inequality and L_{up}-smoothness,

$$\langle \nabla_\theta G(\theta(s)) - \nabla_\theta G(\theta_t), \theta_{t+1} - \theta_t \rangle \le \|\nabla_\theta G(\theta(s)) - \nabla_\theta G(\theta_t)\|_2 \, \|\theta_{t+1} - \theta_t\|_2 \le L_{\mathrm{up}} \, \|\theta(s) - \theta_t\|_2 \, \|\theta_{t+1} - \theta_t\|_2 .$$

Since

$$\theta(s) - \theta_t = s(\theta_{t+1} - \theta_t),$$

we have

$$\|\theta(s) - \theta_t\|_2 = s \, \|\theta_{t+1} - \theta_t\|_2 .$$

Therefore,

$$\int_0^1 \langle \nabla_\theta G(\theta(s)) - \nabla_\theta G(\theta_t), \theta_{t+1} - \theta_t \rangle \, ds \le \int_0^1 L_{\mathrm{up}} s \, \|\theta_{t+1} - \theta_t\|_2^2 \, ds = \frac{L_{\mathrm{up}}}{2} \, \|\theta_{t+1} - \theta_t\|_2^2 .$$

Combining the above inequalities yields

$$G(\theta_{t+1}) \le G(\theta_t) + \langle \nabla_\theta G(\theta_t), \theta_{t+1} - \theta_t \rangle + \frac{L_{\mathrm{up}}}{2} \, \|\theta_{t+1} - \theta_t\|_2^2 .$$

□

E.2 Proof of Theorem 2

Proof. Substituting $\theta_{t+1} - \theta_t = -\eta \hat{g}_t$ into (25) gives

$$G(\theta_{t+1}) \le G(\theta_t) - \eta \langle g_t^*, \hat{g}_t \rangle + \frac{L_{\mathrm{up}} \eta^2}{2} \|\hat{g}_t\|_2^2.$$

Let $e_t \triangleq \hat{g}_t - g_t^*$. Then

$$-\langle g_t^*, \hat{g}_t \rangle = \frac{1}{2}\|e_t\|_2^2 - \frac{1}{2}\|g_t^*\|_2^2 - \frac{1}{2}\|\hat{g}_t\|_2^2.$$

Therefore,

$$G(\theta_{t+1}) \le G(\theta_t) - \frac{\eta}{2}\|g_t^*\|_2^2 + \frac{\eta}{2}\|e_t\|_2^2 - \frac{\eta}{2}(1 - L_{\mathrm{up}}\eta)\|\hat{g}_t\|_2^2.$$

Because $\eta \le 1/L_{\mathrm{up}}$, the last term is non-positive. Hence,

$$G(\theta_{t+1}) \le G(\theta_t) - \frac{\eta}{2}\|g_t^*\|_2^2 + \frac{\eta}{2}\|e_t\|_2^2.$$

Rearranging gives

$$\frac{\eta}{2}\|g_t^*\|_2^2 \le G(\theta_t) - G(\theta_{t+1}) + \frac{\eta}{2}\|e_t\|_2^2.$$

Summing over $t = 0, \ldots, T-1$ yields

$$\frac{\eta}{2} \sum_{t=0}^{T-1} \|g_t^*\|_2^2 \leq G(\theta_0) - G(\theta_T) + \frac{\eta}{2} \sum_{t=0}^{T-1} \|e_t\|_2^2.$$

Using the lower bound $G(\theta_T) \geq F_{\mathbf{\Pi}}^*$ and dividing both sides by $\eta T/2$, we obtain

$$\frac{1}{T} \sum_{t=0}^{T-1} \|g_t^*\|_2^2 \leq \frac{2\left(G(\theta_0) - F_{\mathbf{\Pi}}^*\right)}{\eta T} + \frac{1}{T} \sum_{t=0}^{T-1} \|e_t\|_2^2.$$

By Theorem 1,

$$\|e_t\|_2^2 \leq \|e_t\|_1^2 \leq C_{\mathrm{loc}}^2 e^{-2\lambda\Delta_t} \leq C_{\mathrm{loc}}^2 e^{-2\lambda\Delta_{\min}},$$

we have

$$\frac{1}{T} \sum_{t=0}^{T-1} \|e_t\|_2^2 \leq C_{\mathrm{loc}}^2 e^{-2\lambda\Delta_{\min}}.$$

Substituting this bound gives

$$\frac{1}{T} \sum_{t=0}^{T-1} \|\nabla_\theta G(\theta_t)\|_2^2 \leq \frac{2\left(G(\theta_0) - F_{\mathbf{\Pi}}^*\right)}{\eta T} + C_{\mathrm{loc}}^2 e^{-2\lambda\Delta_{\min}}.$$

□

F Two-Player Kuhn Poker

F.1 Game Rules

In two-player Kuhn Poker, the deck consists of three cards labeled J, Q, and K with ranking J $<$ Q $<$ K. Cards are dealt according to a mechanism parameter $\theta = (\theta_J, \theta_Q, \theta_K)$, with $\theta_J, \theta_Q, \theta_K > 0$ and $\theta_J + \theta_Q + \theta_K = 1$. Under θ, the two players are dealt one private card each using weighted sampling without replacement (Ben-Hamou et al., 2018). The remaining card is not dealt and is unobserved by both players.

At the beginning of each game, both players contribute one chip as an ante, forming an initial pot of two chips. Player 1 acts first, and players subsequently act according to the game history. To match the notation used in our implementation, we represent the action space by two symbols $\{P, B\}$. When no bet has been made, P denotes check and B denotes bet. When facing a bet, P denotes fold and B denotes call.

The Kuhn Poker zero-sum payoff is determined as follows:

- Both players check: the player with the higher-ranked card wins the pot of two chips, resulting in a net payoff of $+1$ for the winner and -1 for the loser.
- A bet is called: the player with the higher-ranked card wins the pot of four chips, resulting in a net payoff of $+2$ for the winner and -2 for the loser.
- One player bets and the other player folds: the betting player wins the pot immediately, resulting in a net payoff of $+1$ for the bettor and -1 for the folder.

We treat θ_{J} and θ_{Q} as the independent parameters and set $\theta_{\mathrm{K}} = 1 - \theta_{\mathrm{J}} - \theta_{\mathrm{Q}}$. For an ordered card type $\mathcal{C}_k = (a, b)$, $a, b \in \{\mathrm{J}, \mathrm{Q}, \mathrm{K}\}$, $a \neq b$, where player 1 receives card a and player 2 receives card b, the probability of this card type is

$$P_{ab}(\theta) = \theta_a \frac{\theta_b}{1 - \theta_a}.$$

For $l \in \{\mathrm{J}, \mathrm{Q}\}$, applying the chain rule gives

$$\frac{\partial P_{ab}(\theta)}{\partial \theta_l} = \frac{\partial \theta_a}{\partial \theta_l} \frac{\theta_b}{(1 - \theta_a)^2} + \frac{\partial \theta_b}{\partial \theta_l} \frac{\theta_a}{1 - \theta_a},$$

where

$$\frac{\partial \theta_a}{\partial \theta_l} = \begin{cases} 1, & \text{if } a = l, \\ -1, & \text{if } a = \mathrm{K}, \\ 0, & \text{otherwise}, \end{cases} \qquad \frac{\partial \theta_b}{\partial \theta_l} = \begin{cases} 1, & \text{if } b = l, \\ -1, & \text{if } b = \mathrm{K}, \\ 0, & \text{otherwise}. \end{cases}$$

The negative derivatives in the cases $a = \mathrm{K}$ or $b = \mathrm{K}$ account explicitly for the dependence $\theta_{\mathrm{K}} = 1 - \theta_{\mathrm{J}} - \theta_{\mathrm{Q}}$.

F.2 Differentiating the Objective

F.2.1 Gradient of the fairness objective

As defined in Section 5, the designer minimizes $(\mathbb{E}[U_1(\boldsymbol{\pi}, \theta)])^2$. To compute its gradient, **ECMPO** first evaluates $\nabla_\theta \mathbb{E}[U_1(\boldsymbol{\pi}, \theta)]$ by treating $U_1(\boldsymbol{\pi}, \theta)$ as $f(\boldsymbol{\pi}, \theta)$. For a fixed policy profile $\boldsymbol{\pi}$, player 1's expected utility can be decomposed over the ordered card types:

$$U_1(\boldsymbol{\pi}, \theta) = \sum_{k=1}^{|\mathcal{C}|} U_1(\boldsymbol{\pi} \mid \mathcal{C}_k) P_{\mathcal{C}_k}(\theta).$$

Here, $U_1(\boldsymbol{\pi} \mid \mathcal{C}_k)$ denotes player 1's conditional expected utility given the ordered card type $\mathcal{C}_k$. Once both the policy profile and the card type are fixed, the conditional utility does not directly depend on the dealing probability parameter θ. Therefore,

$$\nabla_\theta U_1(\boldsymbol{\pi}, \theta) = \sum_{k=1}^{|\mathcal{C}|} U_1(\boldsymbol{\pi} \mid \mathcal{C}_k) \nabla_\theta P_{\mathcal{C}_k}(\theta).$$

The conditional utilities $U_1(\boldsymbol{\pi} \mid \mathcal{C}_k)$ are estimated through simulation, while $\nabla_\theta P_{\mathcal{C}_k}(\theta)$ is computed analytically from the weighted sampling without replacement rule (Ben-Hamou et al., 2018). The expected utility $U_1(\boldsymbol{\pi}, \theta)$ is then obtained by aggregating the conditional utilities according to the corresponding card-type probabilities.

F.2.2 Gradient of the QRE Mixture

According to (4), the partial derivative $\frac{\partial \sigma^*_{1,m}(\theta)}{\partial \mathbb{U}_i(u,v;\theta)}$ is obtained by implicitly differentiating the QRE fixed-point equations. The payoff gradient $\nabla_\theta \mathbb{U}_i(u, v; \theta)$ is computed from the conditional-utility decomposition above. The gradient of $\sigma^*_{2,n}(\theta)$ is obtained analogously.

Applying the chain rule to (3) combines the direct utility gradient $\nabla_\theta U_1(\boldsymbol{\pi}, \theta)$ with the QRE-mixture gradients $\nabla_\theta \sigma^*_{1,m}(\theta)$ and $\nabla_\theta \sigma^*_{2,n}(\theta)$, yielding $\nabla_\theta \mathbb{E}[U_1(\boldsymbol{\pi}, \theta)]$. The gradient of the fairness objective is consequently

$$\nabla_\theta \left(\mathbb{E}[U_1(\boldsymbol{\pi}, \theta)]\right)^2 = 2\mathbb{E}[U_1(\boldsymbol{\pi}, \theta)] \nabla_\theta \mathbb{E}[U_1(\boldsymbol{\pi}, \theta)].$$

This gradient is used by **ECMPO** to update the dealing probabilities in the Kuhn Poker experiment.

F.3 Hyperparameter Settings

Table 1: Shared PSRO hyperparameters for Kuhn Poker.

Hyperparameter	Value
PSRO negligible probability	10^{-6}
Payoff simulations per matrix entry	50
DQN hidden layers	4 layers with 128 units each
DQN batch size	32
DQN learning rate	5×10^{-3}
DQN target-network update period	1,000 steps
DQN learning frequency	Every 10 steps
DQN optimizer	SGD
QRE rationality parameter	$\lambda = 1$

Table 2: Hyperparameters of **ECMPO** and the gradient-free baselines in Kuhn Poker.

Method	Hyperparameters
ECMPO	Learning rate = 0.01 cosine decay over 50 outer iterations.
Objective ablation	Only the direct objective-gradient term $\nabla_\theta f$.
Strategy ablation	Only the QRE-mixture-gradient term $\nabla_\theta \boldsymbol{\sigma}^*$.
RS	Uniformly samples (p_J, p_Q) from the feasible simplex.
BO	UCB acquisition with $\kappa = 1$.
PSO	5 particles; inertia weight 0.7; cognitive coefficient 1.5; social coefficient 1.5.
ES	2 perturbation pairs; perturbation scale $\sigma = 0.1$; learning rate = 0.15.

G Optimizing Fairness with Entropy Regularization in Kuhn Poker

G.1 Gradient of the Regularized Objective

The Kuhn Poker experiment in Section 5 uses the fairness objective $(\mathbb{E}[U_1(\boldsymbol{\pi}, \theta)])^2$. This objective encourages the equilibrium utility of player 1 to approach zero. However, fairness alone may not uniquely identify a desirable dealing mechanism. In particular, a highly non-uniform dealing rule may also achieve a small fairness gap, while substantially deviating from the standard random dealing rule of Kuhn Poker. To incorporate the additional preference that the optimized mechanism should preserve randomness in card dealing, we add an entropy regularization term to the upper-level objective.

Let $u = (1/3, 1/3, 1/3)$ denote the uniform dealing distribution. We define the regularized objective as $(\mathbb{E}[U_1(\boldsymbol{\pi}, \theta)])^2 + \rho D_{\mathrm{KL}}(\theta \| u)$, where $\rho > 0$ controls the strength of the regularization and

$$D_{\mathrm{KL}}(\theta \| u) = \sum_{c \in \{\mathrm{J},\mathrm{Q},\mathrm{K}\}} \theta_c \log \frac{\theta_c}{u_c}.$$

Since minimizing $D_{\mathrm{KL}}(\theta \| u)$ is equivalent to encouraging a higher-entropy dealing distribution up to an additive constant, the regularized objective balances two goals: reducing the equilibrium fairness gap and preventing the dealing distribution from collapsing to an overly deterministic distribution.

The KL term acts only as a regularizer on the mechanism parameter and does not alter the lower-level game. Therefore, the QRE constraints and the implicit differentiation procedure for $\sigma_1^*(\theta)$ and $\sigma_2^*(\theta)$ remain unchanged. The only modification is that the explicit gradient of the designer objective with respect to θ receives an additional closed-form term.

Following Appendix F, the fairness part has gradient $\nabla_\theta \left(\mathbb{E}[U_1(\boldsymbol{\pi}, \theta)]\right)^2$. Under the simplex parameterization, the KL term satisfies

$$\frac{\partial D_{\mathrm{KL}}(\theta \| u)}{\partial \theta_J} = \log \frac{\theta_J}{\theta_K}, \qquad \frac{\partial D_{\mathrm{KL}}(\theta \| u)}{\partial \theta_Q} = \log \frac{\theta_Q}{\theta_K}.$$

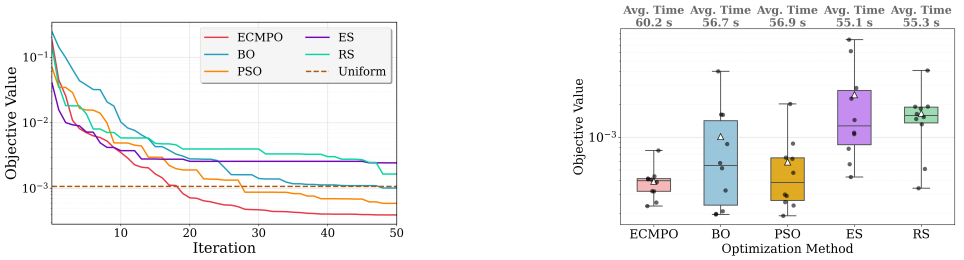
Thus, the explicit gradient of the regularized objective with respect to the two free variables is

$$\nabla_\theta \left(\mathbb{E}[U_1(\boldsymbol{\pi}, \theta)]\right)^2 + \rho \begin{pmatrix} \log(\theta_J/\theta_K) \\ \log(\theta_Q/\theta_K) \end{pmatrix}.$$

G.2 Experimental Results

We evaluate the regularized objective in Kuhn Poker using the same experimental protocol as in Section 5. All methods optimize the objective defined in Appendix G.1 with $\rho = 0.1$. The dashed line denotes the objective value under the uniform dealing rule.

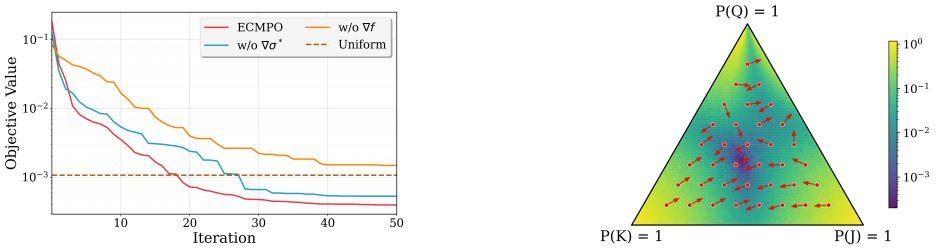
As shown in Figure 7a, **ECMPO** consistently decreases the regularized objective and reaches the lowest objective value among all methods. Although the gradient-free baselines also improve over the initial solution, they either converge more slowly or plateau at higher objective values. Figure 7b further shows that **ECMPO** achieves lower final objective values with less variability while maintaining comparable per-iteration running time.



(a) Best Objective Value Versus Search Iteration.

(b) Objective Value and Time per Iteration.

Figure 7: Optimization Performance in Kuhn Poker with Entropy Regularization Across 10 Independent Runs.



(a)

(b)

Figure 8: (a) Ablation Study. (b) 2-D Heatmap of the Mechanism Parameter Space.

We also conduct an ablation study to examine the contribution of different gradient components. Figure 8a shows that the full **ECMPO** update achieves the best optimization performance. Removing either the QRE sensitivity term or the explicit objective gradient degrades convergence, indicating that both components remain important after entropy regularization is introduced. Figure 8b visualizes the regularized objective over the probability simplex. The KL term penalizes boundary regions corresponding to highly biased dealing rules, while the red arrows show that **ECMPO** still provides descent directions toward low-objective regions.

H Fairness Optimization in Leduc Poker

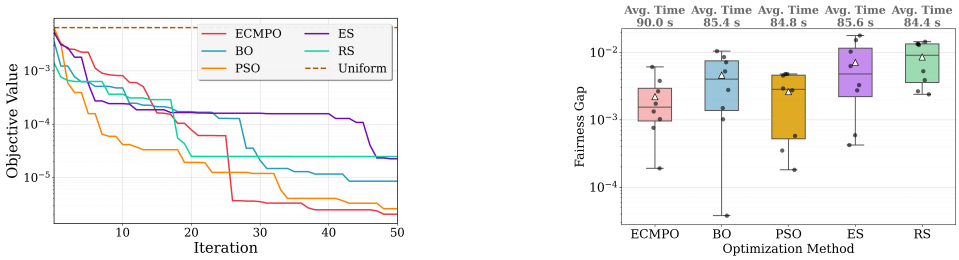
H.1 Game Rules

We further evaluate **ECMPO** on two-player Leduc Poker, a larger imperfect-information poker game than Kuhn Poker. The deck consists of six physical cards: $\mathrm{J}_1, \mathrm{J}_2, \mathrm{Q}_1, \mathrm{Q}_2, \mathrm{K}_1, \mathrm{K}_2$, where the ranks satisfy $\mathrm{J} < \mathrm{Q} < \mathrm{K}$.

At the beginning of the game, each player contributes an ante. Player 1 and player 2 are then dealt one private card each. After the first betting round, if neither player folds, one public card is dealt and a second betting round follows. At showdown, a player who forms a pair between the private card and the public card beats any non-pair hand. If neither player forms a pair, the player with the higher private-card rank wins. If one player folds before showdown, the other player wins the pot immediately. This default rule is not position-fair.

In two-player Leduc Poker, the second player has a positive equilibrium advantage under the standard rules (Schauenberg, 2006). This positional imbalance indicates that the default dealing rule admits room for mechanism-level optimization. We therefore optimize the chance dynamics to reduce the equilibrium fairness gap.

As in Kuhn Poker, the mechanism parameter controls the chance dynamics. However, Leduc Poker contains two physical copies for each rank. We therefore optimize rank-level dealing probabilities $\theta = (\theta_J, \theta_Q, \theta_K)$, with $\theta_J, \theta_Q, \theta_K > 0$ and $\theta_J + \theta_Q + \theta_K = 1$. The two physical cards with the same rank share the same rank probability. Specifically, for a physical card c, let $r(c) \in \{\mathrm{J}, \mathrm{Q}, \mathrm{K}\}$



(a) Best Objective Value Versus Search Iteration.

(b) Fairness Gap and Execution Time per Iteration.

Figure 9: Optimization Performance in Leduc Poker Across 8 Independent Runs.

denote its rank. We assign the physical-card weight $w_c(\theta) = \frac{\theta_{r(c)}}{2}$. Thus, the total initial probability mass assigned to the two copies of each rank equals the corresponding rank probability. Cards are sampled without replacement according to these physical-card weights (Ben-Hamou et al., 2018).

H.2 Leduc-Specific Gradient

Unlike Kuhn Poker, Leduc Poker contains three ordered chance events: player 1's private card, player 2's private card, and the public card. We denote an ordered physical-card type by $\mathcal{L}_k = (a, b, c)$, where a, b, and c are distinct physical cards. There are $|\mathcal{L}| = 6 \times 5 \times 4 = 120$ possible ordered card types. The probability of $\mathcal{L}_k = (a, b, c)$ is

$$P_{\mathcal{L}_k}(\theta) = w_a \frac{w_b}{1 - w_a} \frac{w_c}{1 - w_a - w_b}.$$

For $l \in \{J, Q\}$, let $g_l(d) = \frac{\partial w_d}{\partial \theta_l}$. Since $\theta_K = 1 - \theta_J - \theta_Q$, these derivatives are

$$g_J(d) = \begin{cases} \frac{1}{2}, & r(d) = \mathrm{J}, \\ -\frac{1}{2}, & r(d) = \mathrm{K}, \\ 0, & r(d) = \mathrm{Q}, \end{cases} \qquad g_Q(d) = \begin{cases} \frac{1}{2}, & r(d) = \mathrm{Q}, \\ -\frac{1}{2}, & r(d) = \mathrm{K}, \\ 0, & r(d) = \mathrm{J}. \end{cases}$$

The ordered card-type probability satisfies

$$\frac{\partial P_{\mathcal{L}_k}(\theta)}{\partial \theta_l} = P_{\mathcal{L}_k}(\theta) \left[\frac{g_l(a)}{w_a} + \frac{g_l(b)}{w_b} + \frac{g_l(c)}{w_c} + \frac{g_l(a)}{1 - w_a} + \frac{g_l(a) + g_l(b)}{1 - w_a - w_b} \right].$$

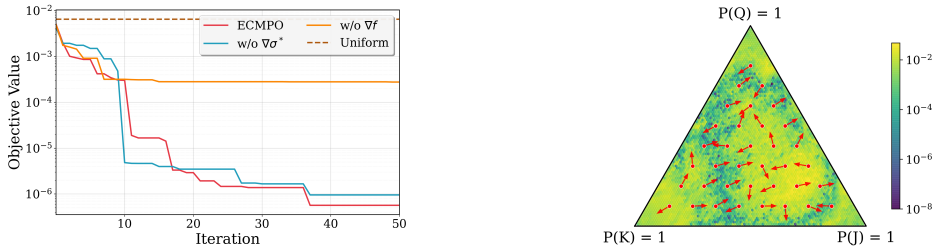
All remaining steps follow Appendix F. Specifically, **ECMPO** combines the direct utility gradients above with the implicit gradients of the QRE mixture to obtain $\nabla_\theta \mathbb{E}[U_1(\boldsymbol{\pi}, \theta)]$. The gradient of the fairness objective is then

$$\nabla_\theta \left(\mathbb{E}[U_1(\boldsymbol{\pi}, \theta)] \right)^2 = 2\mathbb{E}[U_1(\boldsymbol{\pi}, \theta)] \nabla_\theta \mathbb{E}[U_1(\boldsymbol{\pi}, \theta)].$$

H.3 Experimental Results

We use the same baselines as in Section 5. The dashed line denotes the fairness objective under uniform rank-level dealing probabilities. As shown in Figure 9, **ECMPO** reaches the lowest objective and fairness gap among the evaluated methods, while maintaining comparable per-iteration computational cost.

Figure 10a shows that removing either the QRE-mixture gradient or the direct utility gradient degrades optimization performance. Figure 10b further illustrates that the gradients produced by **ECMPO** provide effective descent directions over the rank-level probability simplex.



(a) (b)

Figure 10: (a) Ablation Study. (b) 2-D Heatmap of the Mechanism Parameter Space.

I Social Welfare Optimization in AI Economist

I.1 Environment

We conduct the AI Economist experiment in an 11×11 grid-world environment with two learning agents. At each time step, agents can move, gather resources, and build houses. The central water region and houses built by the other agent are impassable. Each agent's utility is defined as

$$u_i(x_{i,t}, l_{i,t}) \triangleq \frac{x_{i,t}^{1-\eta} - 1}{1-\eta} - l_{i,t},$$

where $x_{i,t}$ is the coin endowment of agent i at time step t, $l_{i,t}$ is its accumulated labor cost, and $\eta = 0.23$ in our experiments. The per-step reward is the incremental change in this utility. Gathering either wood or stone incurs a labor cost of 0.2, while building a house incurs a labor cost of 10. Building a house consumes one unit of wood and one unit of stone. To introduce heterogeneity, agent 1 receives 8 coins from building a house, while agent 2 receives 12 coins.

Each episode lasts $H = 500$ time steps. Taxes are collected every 50 time steps, leading to $P = 10$ tax periods. Let z_i^p denote the pre-tax income earned by agent i during tax period p. The tax function is parameterized by marginal tax rates $\theta = (\theta_0, \ldots, \theta_{B-1})$:

$$\mathrm{Tax}(z;\theta) = \sum_{b=0}^{B-1} \theta_b g_b(z),$$

where the tax brackets are $m = [0, 5, 10, 20, 40, 60, 100]$, and

$$g_b(z) = \begin{cases} 0, & \text{if } z \le m_b, \\ z - m_b, & \text{if } m_b < z \le m_{b+1},\ b < B-1, \\ m_{b+1} - m_b, & \text{if } z > m_{b+1},\ b < B-1, \\ z - m_b, & \text{if } z > m_b,\ b = B-1. \end{cases}$$

Thus, for each bracket b,

$$\frac{\partial\, \mathrm{Tax}(z;\theta)}{\partial \theta_b} = g_b(z).$$

After each tax period, the collected taxes are redistributed equally between the two agents. Therefore, the post-tax income increments are

$$\tilde{z}_1^p = z_1^p - \frac{\mathrm{Tax}(z_1^p;\theta) - \mathrm{Tax}(z_2^p;\theta)}{2}, \qquad \tilde{z}_2^p = z_2^p - \frac{\mathrm{Tax}(z_2^p;\theta) - \mathrm{Tax}(z_1^p;\theta)}{2}.$$

The spatial observation consists of a global feature tensor `world-map`, which represents spatial information such as resources, houses, and agent occupancy. The observation also contains `world-idx_map`, which provides discrete index information for entities located on the map.

Non-spatial information is represented by the vector `flat`, which contains scalar agent and environment features: labor, skill, and tax-period information. The observation additionally includes `time` to indicate the current progress of the episode. Each agent has a discrete action space with six actions. The corresponding `action_mask` indicates which actions are available at the current time step.

I.2 Direct Gradient of the Social Welfare Objective

For a policy profile $\boldsymbol{\pi}$, our objective $f(\boldsymbol{\pi}, \theta)$ is to minimize the negative of the social welfare $\mathrm{swf}(\boldsymbol{\pi}, \theta)$. In the AI Economist experiment, the $\mathrm{swf}(\boldsymbol{\pi}, \theta)$ is defined as the product of equality and productivity. In the two-agent case, it simplifies to

$$\begin{aligned}\mathrm{swf}(\boldsymbol{\pi}, \theta) &\triangleq \left(1 - \frac{|x_{1,H}(\boldsymbol{\pi}, \theta) - x_{2,H}(\boldsymbol{\pi}, \theta)|}{x_{1,H}(\boldsymbol{\pi}, \theta) + x_{2,H}(\boldsymbol{\pi}, \theta)}\right)(x_{1,H}(\boldsymbol{\pi}, \theta) + x_{2,H}(\boldsymbol{\pi}, \theta)) \\ &= x_{1,H}(\boldsymbol{\pi}, \theta) + x_{2,H}(\boldsymbol{\pi}, \theta) - |x_{1,H}(\boldsymbol{\pi}, \theta) - x_{2,H}(\boldsymbol{\pi}, \theta)|\,.\end{aligned}$$

For brevity, we write $x_{i,H}$ for the final coin holdings. Since taxes are only redistributed between the two agents, the total final coin holdings satisfy

$$\begin{aligned}x_{1,H} + x_{2,H} &= x_{1,0} + x_{2,0} + \sum_{p=1}^{P}(\tilde{z}_1^p + \tilde{z}_2^p) \\ &= x_{1,0} + x_{2,0} + \sum_{p=1}^{P}(z_1^p + z_2^p).\end{aligned}$$

Under the observation configuration in Appendix I.1, the tax-rate vector θ is not included in either agent's observation. Consequently, for a fixed policy pair, the agents' action probabilities, pre-tax incomes, labor costs, and action histories have zero direct derivative with respect to θ. Hence, in the gradient computation below, θ affects the final coin holdings only through tax collection and redistribution. Therefore, the total final coin holdings are independent of θ, and we have $\nabla_\theta(x_{1,H} + x_{2,H}) = 0$.

Let $s = \mathrm{sgn}(x_{1,H} - x_{2,H})$. The direct gradient of the equality term is

$$\begin{aligned}-\nabla_\theta|x_{1,H} - x_{2,H}| &= -s\nabla_\theta(x_{1,H} - x_{2,H}) \\ &= s\sum_{p=1}^{P}\left(\nabla_\theta \,\mathrm{Tax}(z_1^p; \theta) - \nabla_\theta \,\mathrm{Tax}(z_2^p; \theta)\right).\end{aligned}$$

Therefore,

$$\nabla_\theta \mathrm{swf}(\boldsymbol{\pi}, \theta) = s\sum_{p=1}^{P}\left(\nabla_\theta \,\mathrm{Tax}(z_1^p; \theta) - \nabla_\theta \,\mathrm{Tax}(z_2^p; \theta)\right),$$

where

$$\nabla_\theta \,\mathrm{Tax}(z; \theta) = (g_0(z), \ldots, g_{B-1}(z))^\top.$$

I.3 Gradient of the QRE Mixture

Let $\mathcal{S}_1$ and $\mathcal{S}_2$ denote the restricted policy sets used to form the empirical game. For a policy pair (π_1^u, π_2^v), let $\mathbb{U}_i(u, v; \theta)$ be the empirical payoff of player i. The derivative of the QRE mixture with respect to θ is obtained by combining QRE implicit differentiation with the derivative of the empirical payoff matrix:

$$\nabla_\theta \sigma_{1,m}^*(\theta) = \sum_{i\in\{1,2\}}\sum_{u=1}^{|\mathcal{S}_1|}\sum_{v=1}^{|\mathcal{S}_2|}\frac{\partial \sigma_{1,m}^*(\theta)}{\partial \mathbb{U}_i(u, v; \theta)}\nabla_\theta \mathbb{U}_i(u, v; \theta).$$

The same formula applies to $\nabla_\theta \sigma_{2,n}^*(\theta)$. The terms $\partial\sigma_{1,m}^*(\theta)/\partial\mathbb{U}_i(u, v; \theta)$ are computed by implicit differentiation of the QRE fixed point.

For T_{simu} sampled rollouts, the empirical payoff gradient is estimated as

$$\nabla_\theta \mathbb{U}_i(u, v; \theta) = \frac{1}{T_{\mathrm{simu}}}\sum_{n=1}^{T_{\mathrm{simu}}}\nabla_\theta u_i^{(n)}(\pi_1^u, \pi_2^v; \theta).$$

The utility gradient of agent 1 is

$$\begin{aligned}\nabla_\theta u_1(\pi_1^u, \pi_2^v; \theta) &= \nabla_\theta \left(\frac{x_{1,H}^{1-\eta} - 1}{1-\eta} - l_{1,H} \right) \\ &= x_{1,H}^{-\eta} \nabla_\theta x_{1,H} \\ &= -\frac{1}{2} x_{1,H}^{-\eta} \sum_{p=1}^{P} \left(\nabla_\theta \operatorname{Tax}(z_1^p; \theta) - \nabla_\theta \operatorname{Tax}(z_2^p; \theta) \right).\end{aligned}$$

Similarly, for agent 2,

$$\nabla_\theta u_2(\pi_1^u, \pi_2^v; \theta) = \frac{1}{2} x_{2,H}^{-\eta} \sum_{p=1}^{P} \left(\nabla_\theta \operatorname{Tax}(z_1^p; \theta) - \nabla_\theta \operatorname{Tax}(z_2^p; \theta) \right).$$

Since the upper-level objective is defined as the negative social welfare, its direct gradient is $\nabla_\theta f(\boldsymbol{\pi}, \theta) = -\nabla_\theta \mathrm{swf}(\boldsymbol{\pi}, \theta)$. Combining the direct objective gradient $\nabla_\theta f(\boldsymbol{\pi}, \theta)$ with the QRE-mixture gradients $\nabla_\theta \sigma_{1,m}^*(\theta)$ and $\nabla_\theta \sigma_{2,n}^*(\theta)$ gives the full **ECMPO** gradient used to update the tax rates in the AI Economist experiment.

I.4 Hyperparameter Settings

Table 3: Shared PSRO hyperparameters for the AI Economist.

Hyperparameter	Value
PPO train batch size	25,000
PPO minibatch size	2,500
PPO SGD epochs	16
PPO rollout fragment length	500
PPO entropy coefficient	0.05
Discount factor	$\gamma = 0.99$
PSRO negligible probability	0.01
PSRO policy budget	10 policies per player
QRE rationality parameter	$\lambda = 1$

Table 4: Hyperparameters of **ECMPO** and the gradient-free baselines in the AI Economist.

Method	Hyperparameters
ECMPO	Learning rate 0.01; rates clipped to $[0.01, 0.99]$.
Objective ablation	Only the direct objective-gradient term.
Strategy ablation	Only the QRE-mixture-gradient term.
RS	Each tax rate is sampled independently and uniformly from the unit interval.
BO	UCB acquisition with $\kappa = 2.5$.
PSO	10 particles; inertia weight 0.2; cognitive coefficient 3; social coefficient 1.
ES	3 perturbation pairs; perturbation scale $\sigma = 0.05$; learning rate = 0.01.

We use DeepSeek-v4-pro as the LLM Economist, and use the following prompt.

System prompt.

> You are an expert tax planner and mechanism designer. You are optimizing a fixed 7-bracket marginal tax vector in a two-agent AI Economist environment.
>
> **Objective.** Maximize the social welfare score $\mathrm{swf} = 2 \times \min(c_0, c_1)$, where c_0 and c_1 are the two agents' final coin holdings after equilibrium policy adaptation. This objective strongly rewards balanced final wealth. A high total output is useful only if both agents end with high coin holdings. Policies that produce one rich agent and one poor agent should be treated as risky even if total production is high.

Optimization method. Use in-context reinforcement learning as a tax-planner agent: infer the empirical response surface from historical tax-rate to SWF evaluations, then propose the next tax vector. Use only the historical evaluations provided in the user message. Do not assume any tax-shape prior such as monotonicity, progressivity, flatness, or smoothness unless the observed data supports it. Early evaluations should cover diverse regions of the 7-dimensional box. Later evaluations should exploit empirical regularities found in your own history. You may use economic reasoning to interpret trade-offs, but the final decision must be driven by observed SWF values. Do not copy the global best exactly. If exploiting, perturb a few coordinates to test a clear empirical hypothesis.

Decision procedure. Identify the best and worst historical rate structures; compare only the patterns that actually appear in the history; infer whether the current bottleneck appears to be insufficient equality, insufficient productivity, or excessive tax distortion based only on observed scores; choose a diverse exploratory point if the history is short or all scores are close; finally propose one new 7-dimensional absolute tax vector.

Output requirements. Respond only with a valid JSON object and do not include markdown. The JSON object must contain exactly two fields: `reasoning`, a brief explanation of the phase and empirical pattern used, and `tax_rates`, an array of seven floats.

User-message template.

Current phase: exploration or optimization.

Global best so far: rates equal to the current best 7-dimensional tax vector; SWF equal to the current best observed value.

Historical evaluations: a step-by-step list of all previously evaluated tax vectors and their SWF values.

Generate the next 7-dimensional absolute tax-rates array. Output strictly in JSON.