Multi-Agent Reinforcement Learning of Karma Bidding Strategies

author names withheld

Under Review for NExT-Game 2026

Abstract

Capacity-constrained shared infrastructure systems require demand management mechanisms that balance efficiency and fairness. Karma mechanisms address this challenge using an artificial, non-tradable currency that enables decentralized allocation through repeated bidding, but computing equilibrium strategies in such settings is difficult due to unknown population dynamics, stochastic demand, and computation time in practice. Here, we study suitable regimes to prove convergence towards stochastic Nash equilibria when following a multi-agent reinforcement learning approach. Computational case studies demonstrate empirically that learned policies closely approximate equilibrium behavior, and further assess impact of learning algorithm and policy initialization on convergence speed. The work highlights the practical potential of the approach for real-world decision support in repeated access-allocation settings.

1. Introduction

Many shared infrastructure systems, such as transportation networks, communication bandwidth, and urban services, operate as capacity-constrained common resources [2]. Unregulated access to these resources often leads to over-consumption, inefficiencies, reduced utility, and negative societal externalities [23]. Demand management mechanisms can mitigate these effects by aligning individual incentives with system-wide objectives [37]. Common approaches include rationing, prioritization, queuing, reservation, and pricing, with pricing remaining the most widely used policy instrument [17, 41]. However, monetary pricing mechanisms often face limited public acceptance because they tie access rights to financial resources and may exacerbate inequality in societies with uneven wealth distributions [38].

To address these limitations, recent research has explored alternative demand management schemes that avoid direct monetary pricing, including tradable credits [35] and artificial currencies [19]. A prominent example is the class of *Karma mechanisms* [37]. Karma uses a non-tradable artificial currency that accounts for past consumption behavior and allocates scarce resources according to users' needs rather than financial means. Consuming resources decreases an agent's Karma balance, while abstaining increases it, encouraging balanced long-term participation [36]. Originally introduced for peer-to-peer file sharing [45], Karma mechanisms have since been studied across multiple domains [37].

A key challenge in Karma economies is strategic user behavior in repeated allocation settings. Game-theoretic analyses model Karma as a dynamic population game in which agents repeatedly bid for scarce resources based on urgency, current balance, and temporal preferences [13]. Prior work established the existence of a Stationary Nash Equilibrium (SNE) that characterizes optimal bidding behavior [14]. Behavioral experiments further show that human participants outperform random

© .

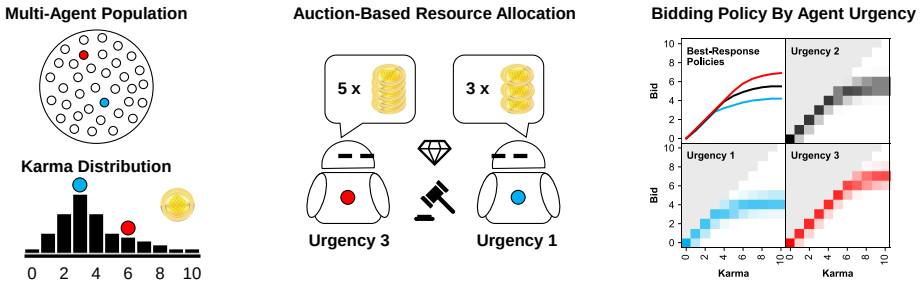

Figure 1: **Karma Game.** Randomly matched pairs from a multi-agent population, that differ by their urgency, repeatedly contest in a first-price auction for a resource, bidding within a capped Karma budget. The higher bidder wins and pays its bid. Agents learn optimal, stochastic bidding policies, leading to a stable population-level Karma distribution (Stationary Nash Equilibrium).

allocation but systematically deviate from equilibrium strategies [16]. These findings suggest that practical Karma deployments may require algorithmic decision support to guide users in repeated bidding environments.

Computing equilibrium policies in dynamic population games is generally intractable in realistic environments due to high-dimensional fixed-point problems [40]. In practice, equilibrium strategies depend on unknown population distributions and transition dynamics, making analytical solutions difficult. Recent work by Berriaud et al. [4] provides initial evidence that optimal bidding in repeated Karma auctions is learnable, but assumes adaptive pacing strategies and substantial knowledge of the population structure. Multi-agent reinforcement learning (MARL) offers a natural alternative for learning bidding strategies directly from repeated interactions under decentralized information [3, 9]. Recent studies have demonstrated the effectiveness of MARL in auctions, congestion games, and mechanism design settings [47]. These properties make MARL particularly suitable for Karma economies, where agents repeatedly adapt bids under stochastic demand and strategic competition.

In this work, we investigate MARL as a computational framework for learning bidding strategies in Karma economies and enabling algorithmic decision support for human users. We study conditions under which independent regularized Q-learning [42] converges toward SNE in finite-state Karma games, showing that MARL agents can efficiently learn stable bidding policies in dynamic population settings. Unlike mean-field game-theoretic approaches, our method operates under decentralized information and learns directly from local interaction. Finally, computational case studies analyze how learning algorithms and policy initialization affect convergence speed in Karma economies.

2. Proof: Independent Regularized Q-Learning Convergence in Finite-State Karma Games

In this section, we present the main result of this paper: the convergence of independent regularized Q-learning in finite-state Karma games. Our results show that independent MARL is a feasible approach for practical Karma economies supported by automated bidding agents.

The proof builds on standard terminology for Karma economies (Appendix B) and proceeds as follows. Section 2.1 introduces the finite-state Karma game and the independent Q-learning

update. Sections 2.2 and F.1 present the main equilibrium-learning results (Theorems 1 and 6; see Appendix D). These theorems show that independent regularized Q-learning converges to the regularized SNE in terms of Q-values, bidding policies, and empirical population distributions. The results further demonstrate that decentralized MARL recovers equilibrium bidding strategies without requiring explicit knowledge of the population distribution or transition kernel.

Section F.2 discusses a finite-population interpretation of the SNEs. Additional economic properties and extensions are deferred to Appendix C. Appendix D summarizes the assumptions and stability conditions under which IQRL is provably well behaved.

2.1. Independent Regularized Q-Learning in Karma Games

Finite-State Karma Game denotes a Karma economy, where urgency and Karma balance are discrete and bounded by an interval, with a constant amount of Karma in circulation (Appendix C.1). Each agent's local state is the pair $s = (u, k)$ where $u \in \mathcal{U}$ denotes urgency and $k \in \mathcal{K}$ denotes Karma. The limitation of Karma makes the learning problem tabular: agents learn one value for each urgency–karma–bid triple. Let $\mathcal{U} = \{u_1, \dots, u_m\}, \quad 0 < u_1 < \cdots < u_m$, be a finite ordered set of urgency levels, and let $\mathcal{K} = \{0, \dots, K\}$ be the finite Karma state space. Urgency evolves independently of current urgency according to a fixed distribution $\mu(u')$ over $\mathcal{U}$, implying $\Pr(u_{t+1} = u' \mid u_t) = \mu(u')$. Karma evolves according to a bounded transition $k \to k'$ (e.g. PBS, see Equation 1, or PBP see Equation 2, where $[x]_K = \min\{K, \max\{0, x\}\}$ denotes the bounded Karma balance between 0 and K, and σ denotes a population-level Karma surplus account that is used for equal redistribution of payments across all agents). Without loss of generality, we assume PBS payment rule $\mathcal{W}$ in the following. At state $s = (u, k)$, the feasible action space $\mathcal{A}(s)$ depends only on the Karma balance k, as it equals the set of possible bids $\mathcal{B}(k) = \{0, \dots, k\}$. The action $a \in \mathcal{A}$ in the Karma game is the agent's bid $b \in \mathcal{B}$. The stationary policy is denoted by $\pi(b \mid s)$. The population state is a distribution $d(s)$ over urgency-karma pairs. The induced population bid distribution is $\nu(b)$ (see Equation 3). The corresponding probability of winning with bid b is $w(b)$ (see Equation 4). If an agent loses the resource while having urgency u, it incurs cost u. Hence the expected one-period reward is $r(u, b)$ (see Equation 5).

$$k' = \begin{cases} [k - b + \sigma]_K & \text{if selected} \\ [k + \sigma]_K & \text{otherwise} \end{cases} \tag{1}$$

$$k' = \begin{cases} [k - b]_K & \text{if selected} \\ [k + b']_K & \text{otherwise} \end{cases} \tag{2}$$

$$\nu_{d,\pi}(b) = \sum_{u \in \mathcal{U}} \sum_{k=b}^{K} d((u, k))\pi(b \mid (u, k)) \tag{3}$$

$$w_{d,\pi}(b) = \sum_{b' < b} \nu_{d,\pi}(b') + \frac{1}{2}\nu_{d,\pi}(b) \tag{4}$$

$$r(u, b) = -u \times (1 - w(b)) \tag{5}$$

Independent Regularized Q-Learning. Each agent i maintains a local table $Q_t^i(s, b)$, observes only its own state (urgency, karma balance) and action (bid, realized reward, next state), and samples bids only from its own local policy $\pi_t^i(\cdot \mid s)$. For an exploration temperature $\gamma > 0$, let us define the soft value $V_\gamma^Q(s)$:

$$V_\gamma^Q(s) = \gamma \log \sum_{b \in \mathcal{A}(s)} \exp\left(\frac{Q(s,b)}{\gamma}\right), \tag{6}$$

and the softmax policy induced by a Q-table as $\Gamma_\gamma(Q)(b \mid s)$:

$$\Gamma_\gamma(Q)(b \mid s) = \frac{\exp\left(Q(s,b)/\gamma\right)}{\sum_{\tilde{b} \in \mathcal{A}(s)} \exp\left(Q(s,\tilde{b})/\gamma\right)}. \tag{7}$$

After observing $(s_t^i, b_t^i, r_t^i, s_{t+1}^i)$, agent i updates the visited entry by

$$Q_{t+1}^i(s_t^i, b_t^i) \leftarrow Q_t^i(s_t^i, b_t^i) + \eta_t \left[r_t^i + \alpha V_\gamma^{Q_t^i}(s_{t+1}^i) - Q_t^i(s_t^i, b_t^i)\right], \tag{8}$$

while all unvisited entries remain unchanged. The policy is updated by

$$\pi_{t+1}^i(\cdot \mid s) \leftarrow \pi_t^i(\cdot \mid s) + \beta_t \left[\Gamma_\gamma(Q_t^i)(\cdot \mid s) - \pi_t^i(\cdot \mid s)\right]. \tag{9}$$

η_t denotes the Q-learning rate, and β_t denotes the policy learning rate, where $\eta_t > \beta_t$ is assumed.

2.2. Convergence Result for Regularized SNE

Theorem 1 (Independent MARL Convergence to Regularized SNE) *Consider the diminishing-step-size version of independent regularized Q-learning in a finite-state Karma game with bounded rewards, sufficient feasible-action visits, two-timescale learning, the mean-field approximation, and a stable regularized response map Φ_γ (Appendix D). Then the independent regularized Q-learning process (IQLR) converges a.s. in the mean-field limit to the unique entropy-regularized stationary Nash equilibrium (SNE) $(d_\gamma^\star, \pi_\gamma^\star)$ induced by Φ_γ.*

Proof We first prove that, for fixed (d, π), the regularized Bellman equation has a unique solution. Let Q_1, Q_2 be two bounded Q-functions. For every state s, the soft value map is non-expansive in the sup norm $\left|V_\gamma^{Q_1}(s) - V_\gamma^{Q_2}(s)\right| \le \|Q_1 - Q_2\|_\infty$. Indeed, for any functions x and y on the finite set $\mathcal{A}(s)$,

$$\left|\gamma \log \sum_b \exp(x_b/\gamma) - \gamma \log \sum_b \exp(y_b/\gamma)\right| \le \max_b |x_b - y_b|. \tag{10}$$

Therefore, for the Bellman operator $T_{d,\pi}^\gamma$ we show it is a contraction [8] (since $\alpha < 1$):

$$\begin{aligned} \left\|T_{d,\pi}^\gamma Q_1 - T_{d,\pi}^\gamma Q_2\right\|_\infty &\le \alpha \max_{s,b} \sum_{s' \in \mathcal{S}} P_{d,\pi}(s' \mid s, b) \left|V_\gamma^{Q_1}(s') - V_\gamma^{Q_2}(s')\right| \\ &\le \alpha \|Q_1 - Q_2\|_\infty. \end{aligned} \tag{11}$$

By the Banach fixed-point theorem [22], the Bellman operator has a unique fixed point $Q_{d,\pi}^\gamma$.

Next consider the fast Q-learning recursion. Because the state-action space is finite, rewards are bounded, and every feasible state-action pair is visited infinitely often, the asynchronous stochastic approximation theorem for tabular Q-learning[21, 43, 46] applies for fixed (d, π). On the fast

timescale, the policy and population variables are quasi-static because $\beta_t/\eta_t \to 0$. Thus the Q-update tracks the moving fixed point:

$$\left\| Q_t^i - Q^\gamma_{d_t,\pi_t} \right\|_\infty \to 0 \qquad \text{a.s.} \tag{12}$$

This is the standard two-timescale stochastic approximation argument [5, 24]: the fast recursion sees (d_t, π_t) as frozen, while the slow recursion sees Q_t^i as equilibrated.

We now analyze the slow variables. Let $z_t = (d_t, \pi_t)$ be the mean-field of the Karma game (population state, comprising Karma distribution d_t and policy π_t). After replacing Q_t^i by its fast-timescale limit $Q^\gamma_{d_t,\pi_t}$, the limiting mean update of the population state is

$$z_{t+1} = z_t + \beta_t \left[\Phi_\gamma(z_t) - z_t + M_{t+1} + \xi_t \right], \tag{13}$$

where $\Phi_\gamma(z_t)$ is a regularized mean-field response map (soft Bellman population update map, see Appendix C.5), M_{t+1} is a martingale-difference noise term [25, 26, 49] and $\xi_t \to 0$ a.s. is the tracking error from the fast recursion and the finite-population approximation. The associated ordinary differential equation is $\dot{z} = (d_z^\gamma, \pi_z^\gamma) - z = \Phi_\gamma(z) - z$. Since Φ_γ is a contraction (Lemma 5, see Appendix D, E), it has a unique fixed point $z_\gamma^\star = (d_\gamma^\star, \pi_\gamma^\star)$. Moreover, this fixed point is globally asymptotically stable for the ODE. To see this, let $e = z - z_\gamma^\star$. Since $\Phi_\gamma(z_\gamma^\star) = z_\gamma^\star$, the upper Dini derivative [10, 18] of the sup-norm distance satisfies:

$$\begin{aligned} D^+\|z - z_\gamma^\star\| &\le \|\Phi_\gamma(z) - \Phi_\gamma(z_\gamma^\star)\| - \|z - z_\gamma^\star\| \\ &\le -(1 - c_\gamma)\|z - z_\gamma^\star\|. \end{aligned} \tag{14}$$

Therefore all ODE trajectories converge to $z_\gamma^\star$. The stochastic approximation recursion tracks this globally stable ODE, and hence $z_t \to z_\gamma^\star$ a.s. Combining this with the fast-timescale tracking result gives $Q_t^i \to Q^\gamma_{d_\gamma^\star,\pi_\gamma^\star}$ a.s. $\forall\ i$.

It remains to verify that the fixed point is a regularized SNE. Since $z_\gamma^\star$ is a fixed point of Φ_γ,

$$\pi_\gamma^\star = \Gamma_\gamma(Q^\gamma_{d_\gamma^\star,\pi_\gamma^\star}), \tag{15}$$

and

$$d_\gamma^\star = d_\gamma^\star P_{d_\gamma^\star,\pi_\gamma^\star}. \tag{16}$$

Equation 15 states that, at every state, the representative agent plays the entropy-regularized best response to the stationary population state. Equation 16 states that the population distribution is stationary under the induced Markov chain. These are precisely the regularized SNE conditions (Appendix C.4). ■

3. Computational Case Study

To illustrate the proposed learning framework, we study a repeated Karma allocation setting in which agents compete pairwise for access to a scarce resource. Each agent observes only its local state, consisting of urgency and current Karma balance, and learns a decentralized bidding strategy without knowledge of population dynamics. Urgency $u \in \{1, 2, 3\}$ is sampled uniformly and determines the cost of not receiving priority access. Agents discount future rewards with factor $\alpha = 0.80$. The

population consists of $N = 100$ agents initialized with $k = 3$ Karma units, bounded by $K = 10$, and evenly distributed initial policies. The mechanism preserves total Karma, which is redistributed only through agent interactions. All experiments run for 1,000,000 iterations across 15 random seeds following RL evaluation best practices [34].

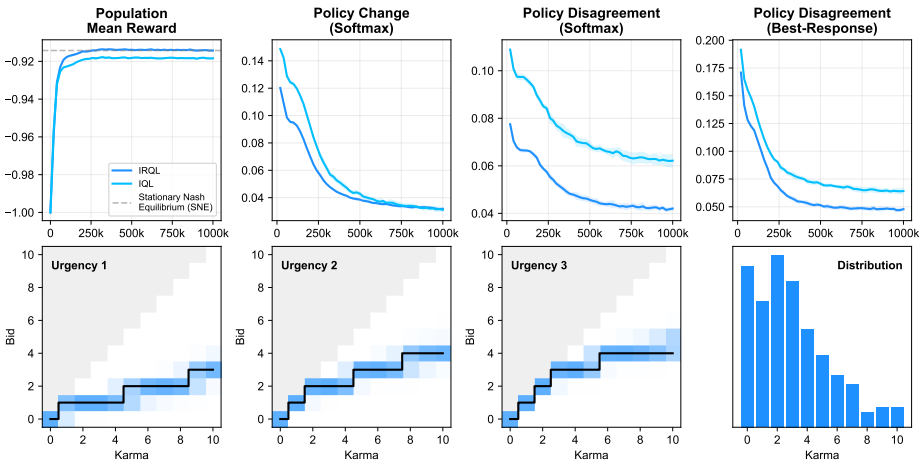

Figure 2: **Independent and Regularized Q-Learning in Karma Games.**

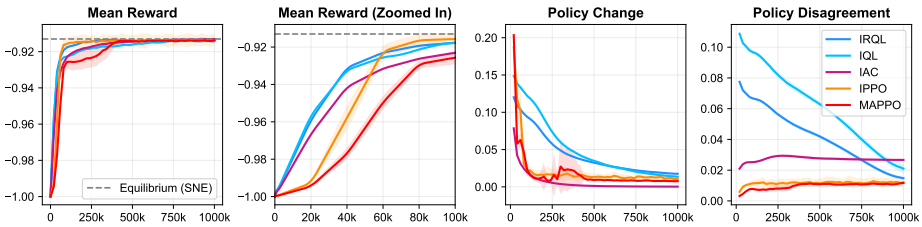

Figure 3: **MARL Benchmark in Karma Games.**

Figure 2 provides an overview of the learning process of IRQL and IQL. Despite their different exploration approaches, both algorithm's convergence trajectories resemble. After around 100k simulation iterations, both algorithms have almost achieved equilibrium-level rewards (top left), clear softmax and best-response policies (bottom), stable Karma distributions (bottom right). Going forward, policy change is decaying and agent policies homogenize (become more similar) (Appendix C.6). Figure 3 benchmarks five MARL algorithms that represent complementary learning paradigms. Independent Q-Learning (IQL) [42] and the derived Independent Regularized Q-Learning (IRQL) serve as the primary baseline due to its theoretical convergence properties in finite-state Karma games, while these do not count amongst the fastest. Independent Actor–Critic with entropy regularization (IAC) [28] provides a policy-gradient counterpart consistent with the regularized equilibrium formulation studied in the theoretical analysis, decays fastest in policy change, overall achieves lowest variance, but fails to homogenize policies amongst agents. In addition, we include IPPO [7], and MAPPO [48] as widely used centralized-training respectively as decentralized-execution baselines. IPPO converges continuously, fast, and with lower variance, while MAPPO experiences reactivation of exploration at 250k.

References

[1] Sachin Adlakha, Ramesh Johari, and Gabriel Y Weintraub. Equilibria of dynamic games with many players: Existence, approximation, and market structure. *Journal of Economic Theory*, 156:269–316, 2015. doi: 10.1016/j.jet.2013.07.002.

[2] Arun Agrawal. *Common Resources and Institutional Sustainability*, pages 41–85. National Academy Press, 2002. doi: 10.17226/10287.

[3] Stefano V. Albrecht, Filippos Christianos, and Lukas Schäfer. *Multi-Agent Reinforcement Learning: Foundations and Modern Approaches*. MIT Press, 2024. URL https://www.marl-book.com. ISBN: 9780262049375.

[4] Damien Berriaud, Ezzat Elokda, Devansh Jalota, Emilio Frazzoli, Marco Pavone, and Florian Dörfler. To spend or to gain: Online learning in repeated karma auctions. *AAMAS '25: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent System*, pages 289–297, 2025. doi: 10.5555/3709347.3743542.

[5] Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Springer, 2008. ISBN 978-93-86279-38-5. doi: 10.1007/978-981-99-8277-6.

[6] Robert Cogburn. The ergodic theory of Markov chains in random environments. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 66(1):109–128, 1984. doi: 10.1007/BF00532799.

[7] Christian Schroeder De Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the StarCraft multi-agent challenge? *arXiv Preprints*, 2020. doi: 10.48550/arXiv.2011.09533.

[8] Eric V Denardo. Contraction mappings in the theory underlying dynamic programming. *Siam Review*, 9(2):165–177, 1967. doi: 10.1137/1009030.

[9] Greg d'Eon, Neil Newman, and Kevin Leyton-Brown. Understanding iterative combinatorial auction designs via multi-agent reinforcement learning. *Proceedings of the 25th ACM Conference on Economics and Computation*, pages 1102–1130, 2024. doi: 10.1145/3670865.3673644.

[10] Ulisse Dini. Lezioni di analisi infinitesimale. *Calcolo Integrale*, 2:720, 1915. URL https://archive.org/details/lezionidianalisi22dini/page/n5/mode/2up.

[11] Roland L Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability & Its Applications*, 1(1):65–80, 1956. doi: 10.1137/1101006.

[12] Roland Lvovich Dobrushin. Gaussian and their subordinated self-similar random generalized fields. *The Annals of Probability*, 7:1–28, 1979. doi: 10.1214/aop/1176995145.

[13] Ezzat Elokda, Carlo Cenedese, Kenan Zhang, Andrea Censi, Saverio Bolognani, and Emilio Frazzoli. A dynamic population game model of non-monetary bottleneck congestion management under elastic demand using karma. *62nd IEEE Conference on Decision and Control (CDC)*, pages 120–125, 2023. doi: 10.1109/CDC49753.2023.10383388.

[14] Ezzat Elokda, Saverio Bolognani, Andrea Censi, Florian Dörfler, and Emilio Frazzoli. A self-contained karma economy for the dynamic allocation of common resources. *Dynamic Games and Applications*, 14(3):578–610, 2024. doi: 10.1007/s13235-023-00503-0.

[15] Ezzat Elokda, Saverio Bolognani, Andrea Censi, Florian Dörfler, and Emilio Frazzoli. Dynamic population games: A tractable intersection of mean-field games and population games. *IEEE Control Systems Letters*, 8:1072–1077, 2024. ISSN 2475-1456. doi: 10.1109/lcsys.2024.3406947.

[16] Ezzat Elokda, Saverio Bolognani, Florian Dörfler, and Heinrich H. Nax. Dynamic resource allocation with karma: An experimental study. *arXiv Preprints*, 2024. doi: 10.48550/arXiv.2404.02687.

[17] Michal Feldman, David Kempe, Brendan Lucier, and Renato Paes Leme. Pricing public goods for private sale. *Proceedings of the Fourteenth ACM Conference on Electronic Commerce*, pages 417–434, 2013. doi: 10.1145/2492002.2482594.

[18] Giorgio Giorgi and Sándor Komlósi. Dini derivatives in optimization—part i. *Rivista di matematica per le scienze economiche e sociali*, 15(1):3–30, 1992. doi: 10.1007/BF02086523.

[19] Artur Gorokh, Siddhartha Banerjee, and Krishnamurthy Iyer. Near-efficient allocation using artificial currency in repeated settings. *Available at SSRN 2852895*, 2016. doi: 10.2139/ssrn.2852895.

[20] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. A general framework for learning mean-field games. *Mathematics of Operations Research*, 48(2):656–686, 2023. doi: 10.1287/moor.2022.1274.

[21] Tommi Jaakkola, Michael Jordan, and Satinder Singh. Convergence of stochastic iterative dynamic programming algorithms. *Advances in Neural Information Processing Systems*, 6, 1993. URL https://proceedings.neurips.cc/paper/1993/hash/5807a685d1a9ab3b599035bc566ce2b9-Abstract.html.

[22] Jacek Jachymski, Izabela Jóźwik, and Małgorzata Terepeta. The banach fixed point theorem: selected topics from its hundred-year history. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A. Matemáticas*, 118(4):140, 2024. doi: 10.1007/s13398-024-01636-6.

[23] Inge Kaul, Isabelle Grunberg, and Marc Stern. Global public goods concepts, policies and strategies. *Global Public Goods: International Cooperation in the 21st Century*, 450, 1999. doi: 10.1093/0195130529.003.0023.

[24] Vijay R. Konda and John N. Tsitsiklis. Actor-critic algorithms. *Advances in Neural Information Processing Systems*, 12, 1999. URL https://proceedings.neurips.cc/paper/1999/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

[25] Harold J Kushner and G George Yin. Convergence with probability one: Martingale difference noise. In *Stochastic Approximation Algorithms and Applications*, pages 85–133. Springer, 1997. doi: 10.1007/978-1-4899-2696-8_5.

[26] Shuze Daniel Liu, Shuhang Chen, and Shangtong Zhang. The ode method for stochastic approximation and reinforcement learning with Markovian noise. *Journal of Machine Learning Research*, 26(24):1–76, 2025. URL https://www.jmlr.org/papers/v26/24-0100.html.

[27] Imanuel Marx and George Piranian. Lipschitz functions of continuous functions. *Pacific Journal of Mathematics*, 3:447–459, 1953. doi: 10.2140/pjm.1953.3.447.

[28] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. *International Conference on Machine Learning (ICML)*, pages 6820–6829, 2020. URL https://proceedings.mlr.press/v119/mei20b.html.

[29] A Yu Mitrophanov. Sensitivity and convergence of uniformly ergodic Markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005. doi: 10.1239/jap/1134587812.

[30] Farrukh Mukhamedov and Ahmed Al-Rawashdeh. Generalized dobrushin ergodicity coefficient and ergodicities of non-homogeneous Markov chains. *Banach Journal of Mathematical Analysis*, 16(1):18, 2022. doi: 10.1007/s43037-021-00173-3.

[31] Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv Preprints*, 2017. doi: 10.48550/arXiv.1705.07798.

[32] Anthony G Pakes. Some conditions for ergodicity and recurrence of Markov chains. *Operations Research*, 17(6):1058–1061, 1969. doi: 10.1287/opre.17.6.1058.

[33] Ling Pan, Tabish Rashid, Bei Peng, Longbo Huang, and Shimon Whiteson. Regularized softmax deep multi-agent q-learning. *Advances in Neural Information Processing Systems*, 34:1365–1377, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/0a113ef6b61820daa5611c870ed8d5ee-Abstract.html.

[34] Andrew Patterson, Samuel Neumann, Martha White, and Adam White. Empirical design in reinforcement learning. *Journal of Machine Learning Research*, 25(318):1–63, 2024. URL https://www.jmlr.org/papers/v25/23-0183.html.

[35] Jesper Provoost, Oded Cats, and Serge Hoogendoorn. Design and classification of tradable mobility credit schemes. *Transport Policy*, 136:59–69, 2023. doi: 10.1016/j.tranpol.2023.03.010.

[36] Kevin Riehl, Anastasios Kouvelas, and Michail A. Makridis. Karma economies for sustainable urban mobility – a fair approach to public good value pricing. *npj Sustainable Mobility and Transport*, 1:14, 2024. doi: 10.1038/s44333-024-00014-4.

[37] Kevin Riehl, Anastasios Kouvelas, and Michail A. Makridis. Resource allocation with karma mechanisms—a review. *Economies*, 12(8), 2024. ISSN 2227-7099. doi: 10.3390/economies12080211.

[38] Kevin Riehl, Anastasios Kouvelas, and Michail A. Makridis. Quantitative fairness—a framework for the design of equitable cybernetic societies. *Computers in Human Behavior: Artificial Humans*, 6:100236, 2025. ISSN 2949-8821. doi: 10.1016/j.chbah.2025.100236.

[39] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407, 1951. doi: 10.1214/aoms/1177729586.

[40] William H Sandholm. *Population Games and Evolutionary Dynamics*. MIT press, 2010. ISBN 978-0-262-19587-4.

[41] Bruce A Scherr and Emerson M Babb. Pricing public goods: An experiment with two proposed pricing systems. *Public Choice*, pages 35–48, 1975. doi: 10.1007/BF01718088.

[42] Ming Tan et al. Multi-agent reinforcement learning: Independent vs. cooperative agents. *Proceedings of the Tenth International Conference on Machine Learning*, pages 330–337, 1993. URL https://web.media.mit.edu/~cynthiab/Readings/tan-MAS-reinfLearn.pdf.

[43] John N. Tsitsiklis. Asynchronous stochastic approximation and q-learning. *Machine Learning*, 16(3):185–202, 1994. doi: 10.1007/BF00993306.

[44] Richard L Tweedie. Sufficient conditions for ergodicity and recurrence of Markov chains on a general state space. *Stochastic Processes and Their Applications*, 3(4):385–403, 1975. doi: 10.1016/0304-4149(75)90033-2.

[45] Vivek Vishnumurthy, Sangeeth Chandrakumar, and Emin Gun Sirer. Karma: A secure economic framework for peer-to-peer resource sharing. *Workshop on Economics of Peer-to-peer Systems*, 35(6), 2003. URL https://netecon.seas.harvard.edu/P2PEcon03.html/Papers/Vishnumurthy_03.pdf.

[46] Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3):279–292, 1992. doi: 10.1007/bf00992698.

[47] Hongsheng Xu, Qinran Hu, Qiuwei Wu, Ke Wang, Feng Wu, and Jinyu Wen. Deep multi-task multi-agent reinforcement learning based joint bidding and pricing strategy of price-maker load serving entity. *IEEE Transactions on Power Systems*, 40(1):505–517, 2024. doi: 10.1109/TPWRS.2024.3403715.

[48] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35:24611–24624, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9c1535a02f0ce079433344e14d910597-Abstract-Datasets_and_Benchmarks.html.

[49] Ingvar Ziemann, Stephen Tu, George J Pappas, and Nikolai Matni. The noise level in linear regression with dependent data. *Advances in Neural Information Processing Systems*, 36:74903–74920, 2023. URL https://proceedings.neurips.cc/paper/2023/hash/ecffd829f90b0a4b6aa017b6df15904f-Abstract-Conference.html.

Appendix A. Appendix: Going Beyond - Contributions of This Work

Contributions The paper develops a MARL framework to learn bidding strategies in Karma economies from decentralized interaction. It proves convergence of independent regularized Q-learning to an entropy-regularized stationary Nash equilibrium in finite-state Karma games. Doing so, it bridges theory and practice by showing that equilibrium strategies can be approximated without explicit knowledge of population dynamics. It demonstrates that decentralized agents can learn stable policies using only local information, without access to global state distributions. Furthermore, the provided empirical evidence implies that learned policies closely match equilibrium behavior in simulated resource allocation scenarios, which highlights the feasibility of using MARL as decision-support for repeated allocation mechanisms with strategic agents.

Limitations The theoretical results are restricted to finite-state Karma games with bounded Karma and discrete urgency levels, limiting generalizability. The analysis relies on the assumptions of ergodicity, mean-field approximation, infinite visitation, and two-timescale learning. Convergence further depends on stability and uniqueness of the regularized equilibrium, which may not hold in more complex environments. The external validity to real-world deployments and human acceptance remains uncertain.

Practical Applications Karma economies could be applied to any, repeated real-world resource allocation context that is fairness-sensitive, complicated ethically, requires social acceptance, where monetary mechanisms are undesirable (e.g., organ transplant lists), or any form of coordination is required. The framework could support both human and agentic smart grid management in power electricity networks, congestion pricing in transportation networks, bandwidth allocation in communication networks with scarce capacity, medical services in social networks, urban services such as parking, energy usage, or shared mobility systems. It provides decision-support tools that guide users' bidding strategies in repeated allocation settings, and offers a practical alternative when computing equilibria analytically is infeasible in dynamic environments.

Appendix B. Appendix: Preliminaries on Karma Games

Dynamic Resource Allocation Mechanisms with Karma. We consider a population $\mathcal{N}$ of N agents, where each agent is endowed with a non-negative integer counter balance $k \in \mathbb{N}$, called Karma, which is private to the agent, similar to as introduced in [14, 36]. Furthermore, each agent has an urgency u that determines the cost for not accessing a resource of interest, and a temporal preference α that determines how much the agent prioritizes present vs. future consumption. At discrete global time instants, two random agents from $\mathcal{N}$ compete for a scarce, indivisible resource in a repeated setting. During such a resource competition, each competing agent submits a sealed non-negative integer bid b which is bounded by its Karma balance. The *resource allocation rule* $\mathcal{V}$ and the *payment rule* $\mathcal{W}$ determine the outcome of the interaction. $\mathcal{V} : \mathbb{N}^2 \to k \in \mathbb{N}$ determines which agent is selected to receive the contended resource (e.g., *highest bid* with fair coin toss as tie breaker). $\mathcal{W} : \mathbb{N}^2 \to \mathbb{Z}^2$ determines the non-negative Karma payments of the two competing agents (e.g., *pay-bid-to-peer* (PBP), and *pay-bid-to-society* (PBS)).

Dynamic Population Game. Dynamic Population Games (DPGs) are a class of discrete-time, finite-state-and-action, stationary mean-field game, which enables the application of equilibrium analysis tool-set from (static) population games on a subset of dynamic mean-field games [15]. Let $\mathcal{G}$ be a DPG, comprising a large population $\mathcal{N}$ of agents with types $\tau \in \mathcal{T} = \{1, \cdots, N_\tau\}$, and states $s \in \mathcal{S} = \{1, \cdots, N_s\}$. The distribution of types is the exogenous parameter $g \in \Delta(\mathcal{T})$. The state distribution of type τ is given by $d_\tau \in \mathcal{D}_\tau = g_\tau \Delta(\mathcal{S})$, and the joint type-state distribution is the concatenation of state distributions $d = (d_1, \cdots, d_{N_\tau}) \in \mathcal{D} = \Pi_{\tau \in \mathcal{T}} \mathcal{D}_\tau$. At discrete time steps, agents participate in an interaction in which each agent of type-state $[\tau, s]$ plays one of a finite number of actions $a \in \mathcal{A}_\tau[s]$. Type τ agents pick their action according to the stationary homogeneous policy $\pi_\tau : \mathcal{S} \to \Delta(\mathcal{A}_\tau[s])$, which maps the state to a probability distribution over the actions. The probability to play action a when in state s is denoted by $\pi_\tau[a|s] \in [0, 1]$. The set of policies of type τ is denoted by Π_τ, and the concatenation of the policies is denoted by $\pi = (\pi_1, \cdots, \pi_{N_\tau}) \in \Pi = \Pi_{\tau \in \mathcal{T}} \Pi_\tau$. The pair $(d, \pi) \in \mathcal{D} \times \Pi$ is referred to as the mean-field (in mean-field games) or social state (in population games). The discrete time state dynamics of type τ agents are governed by the *state transition function* $p_\tau[s^+|s, a](d, \pi)$, which is the probability of the state to transition from s to s^+ after playing a, as a function of the mean-field (d, π). The agent receives an *immediate payoff* for playing a when in state s, which is also a function of (d, π) given by reward $r_\tau[s, a](d, \pi)$. A DPG $\mathcal{G}$ is therefore specified by the pair of state transition p and immediate payoff functions r. The Karma resource allocation mechanism, for a sufficiently large N can be considered a DPG [15, 40], where types τ refer to temporal preferences α and states s refer to the tuple of urgency u and Karma k, and actions a correspond to the bids b.

Stationary Nash Equilibrium of Karma Economies. A natural solution concept for DPGs is the SNE (Stationary Nash Equilibrium) [1], which is the stationary mean-field equilibrium [20] in which microscopically, each agent follows an optimal policy, and macroscopically, the state distribution is stationary. This solution concept can help to calculate the payoff-optimal behavior that agents will follow in such DPGs (Karma economies), and therefore to predict agent behavior [36]. It is shown that there exists (at least) one SNE for every DPG $\mathcal{G}$ (under some assumptions) [15]. Therefore, a user equilibrium is guaranteed to exist in Karma Economies (DPGs that model Karma resource allocation mechanisms) [36].

The SNE of a DPG $\mathcal{G}$ is a mean-field $(d^*, \pi^*) \in \mathcal{D} \times \Pi$ which satisfies, for all $[\tau, s] \in \mathcal{T} \times \mathcal{S}$, following two conditions. Equation 17 implies that the agents have no incentive to unilaterally deviate from the optimal policy π^*. Equation 18 implies that the type-state distribution d^* is stationary under the stochastic process characterized by $P_\tau(d^*, \pi^*)$. $B_\tau[s](d, \pi)$ denotes the *state-dependent best response correspondence*, and $P_\tau[s^+|s](d, \pi)$ denotes the *state transition matrix*; more details can be found in [15].

$$\pi^*_\tau[\cdot|s] \in B_\tau[s](d^*, \pi^*), \tag{17}$$

$$d^*_\tau[s] = \sum_{s^- \in \mathcal{S}} d^*_\tau[s^-] P_\tau[s|s^-](d^*, \pi^*), \tag{18}$$

Appendix C. Appendix: Discussion of Finite-State Karma Game Terminology

C.1. Finite-State Karma Game

We prove convergence of independent regularized Q-learning for finite-state-space Karma games (Karma balances are capped at a finite level, and urgency takes values in a finite set). The bounds for Karma balances should not be considered a restriction, but rather be understood as desired property, as capped Karma balances are shown to be important to avoid hoarding behavior and guarantee liquidity in Karma economies [36]. This ensures that the induced learning dynamics define a finite, irreducible, ergodic, and aperiodic Markov process with entropy-regularized best responses, allowing convergence guarantees toward a regularized SNE, as we show in this work. Here, regularization refers to entropy-regularized best responses, implemented through softmax policies (instead of best-response policy), that encourages continuous exploration of agents, and which ensures continuity and uniqueness of the equilibrium mapping [28, 31, 33]. Ergodicity follows from the finiteness of the state space together with strictly positive softmax exploration and the independent urgency process, which ensure that the induced Markov chain over (u, k) is irreducible and aperiodic and therefore admits a unique stationary distribution [29].

C.2. Economic Properties (Karma)

Property 1 (Independent Urgency Process) *Urgency is drawn independently of current urgency from a full-support distribution μ:* $\Pr(u_{t+1} = u' \mid u_t) = \mu(u')$.

Role in the proof. *This separates current urgency from the continuation urgency process, which is what gives the Q-function increasing differences in urgency and bid.*

Implications in practice. *Urgency is considered something that cannot be planned or anticipated as [13]. Therefore, it is reasonable to assume that urgency originates from a random process, where current urgency is independent of previous urgency [14].*

Property 2 (Agent Homogeneity) *All agents share the same discount factor $\alpha \in (0, 1)$, urgency distribution μ, reward function, transition dynamics, and feasible action sets.*

Role in the proof. *Homogeneity lets us analyze a representative learner facing the population bid distribution, rather than tracking separate value functions for different agent types.*

Implications in practice. *Assuming a population-level distribution for the urgency process, and assuming it to be homogeneous across all agents, is like risk modeling in the context of insurance or finance [19], rather a tool than a restriction.*

Without loss of generality. *The proof can be easily extended to heterogeneous discount factors α, or heterogeneous urgency-processes g_τ by integrating them in the state $s = (u, k, \alpha, g_\tau)$. For clarity, in the proof we restrict the formulations to $s = (u, k)$.*

Property 3 (Bidding Monotonicity in Karma) *For fixed urgency u, a greater Karma balance can only lead to a greater or equal bid* $k_2 > k_1 \Rightarrow \pi(\cdot \mid u, k_2) \geq_{\text{FOSD}} \pi(\cdot \mid u, k_1)$.

Property 4 (Bidding Monotonicity in Urgency) *For fixed Karma k, a greater urgency can only lead to a greater or equal bid* $u_2 > u_1 \Rightarrow \pi(\cdot \mid u_2, k) \geq_{\text{FOSD}} \pi(\cdot \mid u_1, k)$.

Property 5 (Value Function Concavity in Karma) *For every urgency level u, the value function is discrete concave in Karma:* $V(u, k+1) - V(u, k) \geq V(u, k+2) - V(u, k+1)$.

Role in the proof. *Concavity formalizes diminishing marginal value of additional Karma and is the key condition behind increasing differences in Karma and bid.*

C.3. Independent Regularized Q-Learning (IRQL)

The original IQL [42] uses a hard Bellman target,

$$V_\gamma^Q(s) = V^Q(s) = \max_{b \in \mathcal{A}(s)} Q_t(s_{t+1}, b), \tag{19}$$

while for IRQLs the Boltzmann/softmax is only the behavior policy used for exploration

$$V_\gamma^Q(s) = \gamma \log \sum_{b \in \mathcal{A}(s)} \exp\left(\frac{Q(s,b)}{\gamma}\right). \tag{20}$$

This makes the induced softmax policy part of the optimality criterion rather than merely an exploration rule. Thus, fixed-temperature IQL does not converge, even in principle, to the entropy-regularized best response used in the proof. IRQL changes the Bellman target itself by replacing the hard maximum with the soft value, which enables a smooth, single-valued best-response map, persistent full-support exploration, and a plausible stochastic-approximation proof.

The independent regularized Q-learning (IRQL) algorithm is a soft Bellman analogue of independent Q-learning (IQL) [42]. As in IQL, this is an *independent* MARL rule: each agent updates only from its own local transition $(s_t^i, b_t^i, r_t^i, s_t^{i+})$ and treats all other agents as part of the environment, without exchanging Q-values, experiences, policies, gradients, or population statistics. The key difference lies in the Bellman recursion. IQL learns standard Q-values using a hard maximum over next actions, while a softmax (Boltzmann) distribution is used only as a behavior policy for exploration. In contrast, IRQL replaces the hard maximum in the Bellman target by the soft value V_γ^Q, so that the induced softmax policy is no longer merely exploratory, but defines the optimality criterion itself (Appendix C.3). This modification corresponds to *entropy regularization* as the agent does not optimize expected return alone, but instead maximizes a trade-off between value and policy entropy. Equivalently, the soft value V_γ^Q can be written as the expected Q-value under the induced policy plus a temperature-weighted Shannon entropy term. As a result, IRQL favors stochastic policies that keep multiple actions active rather than collapsing immediately to a deterministic maximizer. The limiting policy is therefore an entropy-regularized best response, which approaches the standard greedy optimum of IQL as the temperature $\gamma \downarrow 0$.

C.4. Regularized Stationary Nash Equilibrium

In the SNE definition, regularization changes only the microscopic optimality condition, not the stationarity condition. The ordinary SNE condition requires

$$\pi^\star(\cdot \mid s) \in \mathrm{BR}_s(d^\star, \pi^\star), \tag{21}$$

meaning that the policy places probability only on exact best-response bids. The entropy-regularized SNE replaces this by

$$\pi_\gamma^\star(\cdot \mid s) \in \mathrm{BR}_{\gamma,s}(d_\gamma^\star, \pi_\gamma^\star), \tag{22}$$

where the regularized best response solves

$$\max_{\pi(\cdot|s) \in \Delta(\mathcal{A}(s))} \left\{ \sum_{b \in \mathcal{A}(s)} \pi(b \mid s) Q(s,b) + \gamma H(\pi(\cdot \mid s)) \right\}, \tag{23}$$

where $H(\pi_i(\cdot \mid s))$ is the Shannon entropy over feasible bids. The optimizer is the softmax policy. The population stationarity condition remains

$$d^\star_\gamma = d^\star_\gamma P_{d^\star_\gamma, \pi^\star_\gamma}. \tag{24}$$

Thus, regularization means that equilibrium agents are not perfectly greedy; they are entropy-regularized or logit-rational best responders. This smooths the best-response map, avoids discontinuities from ties, keeps all feasible bids visited, and enables convergence analysis.

C.5. Regularized Mean-Field Response Map

Φ_γ represents the *regularized mean-field response map* or *soft Bellman population update map*. It maps the current social state $z = (d, \pi)$ to the population-policy pair (Equation 28) obtained by three operations: (i) solve the regularized Bellman equation at (d, π) (Equation 25); (ii) compute the induced soft best-response policy (Equation 26); (iii) update the population distribution under that policy (Equation 27).

$$Q^\gamma_z = T^\gamma_{d,\pi} Q^\gamma_z \tag{25}$$

$$\pi^\gamma_z = \Gamma_\gamma(Q^\gamma_z) \tag{26}$$

$$d^\gamma_z(s') = \sum_{s \in \mathcal{S}} d(s) P_{d,\pi^\gamma_z}(s' \mid s) \tag{27}$$

$$\Phi_\gamma(z) = (d^\gamma_z, \pi^\gamma_z) \tag{28}$$

Consecutively, the slow ODE $\dot{z} = \Phi_\gamma(z) - z$ has a simple interpretation: the social state moves toward the state obtained by applying one regularized best-response and population-update step. A fixed point of this map is exactly a regularized SNE.

C.6. Empirical Convergence Assessment: Reward, Policy Change & Policy Disagreement

The empirical convergence plots use checkpoint-level diagnostics computed from the saved per-agent policy tensors. Let t_c denote checkpoint c, let $\pi^i_c(b \mid u, k)$ be the policy of agent i at that checkpoint, and let $M = \sum_{k=0}^{K}(k+1)$ be the number of feasible Karma–bid entries per urgency level. All policy diagnostics exclude infeasible bids $b > k$. The checkpoint reward is the interval-average population reward since the previous checkpoint,

$$R_c = \frac{1}{N|I_c|} \sum_{t \in I_c} \sum_{i=1}^{N} r^i_t, \qquad I_c = \{t_{c-1} + 1, \ldots, t_c\}. \tag{29}$$

This is the quantity stored as `mean_reward`; the corresponding `std_reward` is the standard deviation across agents of their interval-average rewards.

Policy change measures temporal stability of the learned bidding tables. For $c \geq 1$, it is the normalized entrywise ℓ_1 distance, implemented as a mean absolute difference over feasible policy probabilities:

$$C_c(\pi) = \frac{1}{N|\mathcal{U}|M} \sum_{i=1}^{N} \sum_{u \in \mathcal{U}} \sum_{k=0}^{K} \sum_{b=0}^{k} \left| \pi^i_c(b \mid u, k) - \pi^i_{c-1}(b \mid u, k) \right|, \tag{30}$$

with $C_0(\pi) = 0$ by convention. Thus, policy change is not a KL divergence or a distance between logits; it is the average absolute probability movement between two consecutive policy checkpoints.

Policy disagreement measures cross-agent heterogeneity at a fixed checkpoint. Let the population-average policy be

$$\pi_c^{\mathrm{pop}}(b \mid u, k) = \frac{1}{N} \sum_{j=1}^{N} \pi_c^j(b \mid u, k). \tag{31}$$

The disagreement diagnostic is

$$D_c(\pi) = \frac{1}{N|\mathcal{U}|M} \sum_{i=1}^{N} \sum_{u \in \mathcal{U}} \sum_{k=0}^{K} \sum_{b=0}^{k} \left| \pi_c^i(b \mid u, k) - \pi_c^{\mathrm{pop}}(b \mid u, k) \right|. \tag{32}$$

Lower D_c therefore means greater policy agreement across agents.

The same definitions are applied to both the stochastic policy and its greedy best-response projection. For each checkpoint, the plotted best-response policy is

$$G(\pi_c^i)(b \mid u, k) = \mathbf{1}\{b = b_{i,c}^\star(u, k)\}, \qquad b_{i,c}^\star(u, k) \in \arg\max_{a \in \{0,\ldots,k\}} \pi_c^i(a \mid u, k), \tag{33}$$

with ties resolved by the smallest maximizing bid in the implementation. Substituting $G(\pi)$ into C_c and D_c gives the best-response policy-change and best-response disagreement panels, while using π directly gives the softmax-policy panels. For multi-seed summaries, runs are aligned at common checkpoints. The main convergence panels report the cross-seed mean diagnostic with one-standard-deviation bands; the population-size variability analysis reports cross-seed standard deviations of the same checkpoint diagnostics. The first two aligned checkpoint values are omitted when enough checkpoints are available so that initialization transients do not dominate the scale.

These diagnostics are empirical convergence indicators rather than separate equilibrium conditions. A reward plateau indicates that realized welfare has stabilized, $C_c(\pi)$ close to zero indicates that the learned policy tables are no longer moving substantially, and $D_c(\pi)$ close to zero indicates that homogeneous agents have learned essentially the same state-dependent bidding rule. These convergence trends were observed consistently for best-response and probabilistic best-response policies. Under the uniqueness and stability regime used in the convergence argument, such temporal stability and cross-agent agreement are the finite-population signatures expected as the learned policy profile approaches a stationary, no-profitable-deviation equilibrium policy.

Appendix D. Appendix: Necessary Conditions

We use the following assumptions.

Assumption 1 (Finite State and Bounded Rewards) *The state space $\mathcal{S}$ and all feasible bid sets $\mathcal{A}(s)$ are finite. Rewards are uniformly bounded, and the discount factor satisfies $\alpha \in (0, 1)$.*

Role in the proof. *Finiteness makes the Bellman equation tabular and compactness arguments available, while bounded rewards imply bounded Q-values and bounded stochastic-approximation noise.*

Assumption 2 (Softmax Exploration and Sufficient Visits) *For every feasible state-action pair (s, b) with $b \in \mathcal{B}(k)$, the policy is not containing deprecated feasible bids $\pi_t(b \mid s) > 0$, and the induced learning process visits each feasible state-action pair infinitely often.*

Role in the proof. *Positive exploration prevents unobserved feasible bids, while sufficient visits are the standard sampling condition needed for tabular Q-learning convergence.*

Implications in practice. *We need to select a maximum Karma amount that is not too far of from the average Karma per agent in the population, to keep the Karma distribution sufficiently high across the state-space. Throughout our computational experiments, we could show, even in cases where the space was not sufficiently visited, the best-response-policy still emerged quite fast, just the probability distribution over the action space took more time to concentrate around the best-response-policy on the right tail of the Karma balances.*

Assumption 3 (Two-Timescale Learning) *The Q-learning and policy-learning step-sizes η_t and β_t satisfy the Robbins–Monro conditions [39]:*

$$\sum_t \eta_t = \sum_t \beta_t = \infty, \qquad \sum_t \eta_t^2 < \infty, \qquad \sum_t \beta_t^2 < \infty, \tag{34}$$

and the policy update is slower than the Q-update: $\frac{\beta_t}{\eta_t} \to 0$.

Role in the proof. *This lets the Q-values equilibrate on the fast timescale while the policy changes slowly enough to be analyzed by a limiting ordinary differential equation.*

Implications in practice. *It is reasonable to assume a two-timescale learning, as the frequency with which agents change their actions (e.g. in the form of software updates or recalculations) lags behind the perception for the need of an update (through updated value perceptions).*

Assumption 4 (Mean-Field Limit) *The population is large enough that the effect of a single agent on the empirical population distribution is negligible. Equivalently, the representative learner faces the deterministic population state d_t generated by the population policy, while the finite-population sampling error is a martingale noise term [25] whose magnitude vanishes as $N \to \infty$.*

Role in the proof. As discussed in the finite-population interpretation (Section F.2), this assumption is an integral, approximate infinite-population corollary for the proof.

Implications in practice. It is reasonable to assume a sufficiently large yet finite population, and as our computational experiments show, $N = 100$ is already producing stable, contractual convergence dynamics (Section 3).

Assumption 5 (Ergodicity) *For every full-support policy π, the finite Markov chain over $\mathcal{S}$ induced by $P_{d,\pi}$ is irreducible and aperiodic. It therefore admits a unique stationary distribution.*

This assumption is standard in the literature on Markov chain ergodicity and reinforcement learning [6, 32, 44].

Assumption 6 (Unique Regularized Equilibrium) *The entropy-regularized bounded Karma game has a unique stationary equilibrium* $(d^\star, \pi^\star)$.

Role in the proof. *Uniqueness turns convergence to the set of regularized fixed points into convergence to the single equilibrium* $(d^\star, \pi^\star)$. *Existence and uniqueness of stationary equilibria in finite dynamic population games such as the SNE follows under standard continuity and compactness conditions; see, for example, [1, 15].*

Appendix E. Appendix: Regularized Response Map Stability

This appendix defines conditions and identifies regimes, under which the regularized mean-field response map Φ_γ is a contraction. It follows when the Karma game is uniformly mixing, the dependence on the population state is weak enough, and the entropy temperature is large enough to smooth the best-response map.

Notation. Let $z = (d, \pi)$ denote the social state and define the norm

$$\|z - z'\| = \|d - d'\|_1 + \max_{s \in \mathcal{S}} \|\pi(\cdot \mid s) - \pi'(\cdot \mid s)\|_1. \tag{35}$$

Let

$$A_{\max} = \max_{s \in \mathcal{S}} |\mathcal{A}(s)|. \tag{36}$$

Assume rewards are bounded by $R_{\max} < \infty$:

$$|r_{d,\pi}(s, b)| \le R_{\max} \qquad \forall (d, \pi), s, b. \tag{37}$$

For fixed $z = (d, \pi)$, write Q_z^γ for the fixed point of the regularized Bellman operator, and write

$$\pi_z^\gamma = \Gamma_\gamma(Q_z^\gamma). \tag{38}$$

Let $K_{d,\varrho}$ be the policy-averaged transition kernel induced by population distribution d and (generic) policy (argument) ϱ:

$$K_{d,\varrho}(s' \mid s) = \sum_{b \in \mathcal{A}(s)} \varrho(b \mid s) P_{d,\varrho}(s' \mid s, b). \tag{39}$$

Then the population component of the response map is

$$d_z^\gamma = d K_{d,\pi_z^\gamma}. \tag{40}$$

Assumption 7 (Primitive Lipschitz Conditions) *There exist constants $L_r, L_P, L_d, L_\varrho < \infty$ such that, for all $z = (d, \pi)$ and $z' = (d', \pi')$,*

$$\max_{s,b} |r_{d,\pi}(s, b) - r_{d',\pi'}(s, b)| \le L_r \|z - z'\|, \tag{41}$$

$$\max_{s,b} \|P_{d,\pi}(\cdot \mid s, b) - P_{d',\pi'}(\cdot \mid s, b)\|_1 \le L_P \|z - z'\|, \tag{42}$$

and, for all policies ϱ, ϱ',

$$\max_s \|K_{d,\varrho}(\cdot \mid s) - K_{d',\varrho'}(\cdot \mid s)\|_1 \le L_d \|d - d'\|_1 + L_\varrho \max_s \|\varrho(\cdot \mid s) - \varrho'(\cdot \mid s)\|_1. \tag{43}$$

Lemma 2 (Primitive Lipschitz Conditions for the Implemented Pairwise Karma Game) *For the implemented pairwise Karma game with bounded Karma K, finite urgency set $\mathcal{U}$, bounded losing cost c_{loss}, and either settlement rule $\mathcal{W}$, Assumption 7 holds with finite constants. One may take $R_{\max} = u_{\max} c_{\text{loss}}$, $L_r = u_{\max} c_{\text{loss}}$, $L_P = 1$, $L_d = 1$, and $L_\varrho = 2$ under the norm above.*

Proof Let $z = (d, \varrho)$ and define the opponent state-action mixture

$$m_z(\tilde{s}, \tilde{b}) = d(\tilde{s})\varrho(\tilde{b} \mid \tilde{s}), \qquad \tilde{b} \in \mathcal{A}(\tilde{s}). \tag{44}$$

For two social states $z = (d, \varrho)$ and $z' = (d', \varrho')$,

$$\begin{aligned} \|m_z - m_{z'}\|_1 &\le \sum_{\tilde{s}} |d(\tilde{s}) - d'(\tilde{s})| + \sum_{\tilde{s}} d'(\tilde{s}) \|\varrho(\cdot \mid \tilde{s}) - \varrho'(\cdot \mid \tilde{s})\|_1 \\ &\le \|d - d'\|_1 + \max_{\tilde{s}} \|\varrho(\cdot \mid \tilde{s}) - \varrho'(\cdot \mid \tilde{s})\|_1 \le \|z - z'\|. \end{aligned} \tag{45}$$

For a focal state $s = (u, k)$ and bid b, the probability of losing the pairwise auction is

$$\ell_z(b) = \sum_{\tilde{s}, \tilde{b}} m_z(\tilde{s}, \tilde{b}) \left[\mathbf{1}\{\tilde{b} > b\} + \frac{1}{2} \mathbf{1}\{\tilde{b} = b\} \right]. \tag{46}$$

The implemented reward is $r_z((u, k), b) = -u c_{\text{loss}} \ell_z(b)$, hence

$$|r_z(s, b) - r_{z'}(s, b)| \le u_{\max} c_{\text{loss}} \|m_z - m_{z'}\|_1 \le u_{\max} c_{\text{loss}} \|z - z'\|. \tag{47}$$

This proves the reward bound and the reward Lipschitz bound.

For the transition kernel, fix the focal state-action pair (s, b) and the opponent pair $(\tilde{s}, \tilde{b})$. The implemented auction resolution, tie-breaking, clipping at the Karma cap, and the chosen settlement rule induce a probability kernel

$$H(k^+ \mid s, b, \tilde{s}, \tilde{b}) \tag{48}$$

over the focal next Karma balance. This kernel is independent of the social state z. Since next urgency is sampled from the fixed distribution μ, the mean-field transition kernel can be written as

$$P_z((u^+, k^+) \mid s, b) = \mu(u^+) \sum_{\tilde{s}, \tilde{b}} m_z(\tilde{s}, \tilde{b}) H(k^+ \mid s, b, \tilde{s}, \tilde{b}). \tag{49}$$

Therefore,

$$\begin{aligned} \|P_z(\cdot \mid s, b) - P_{z'}(\cdot \mid s, b)\|_1 &\le \sum_{\tilde{s}, \tilde{b}} |m_z(\tilde{s}, \tilde{b}) - m_{z'}(\tilde{s}, \tilde{b})| \\ &= \|m_z - m_{z'}\|_1 \le \|z - z'\|. \end{aligned} \tag{50}$$

Thus $L_P = 1$ is sufficient.

Finally, the policy-averaged kernel satisfies

$$K_{d,\varrho}(\cdot \mid s) = \sum_{b \in \mathcal{A}(s)} \varrho(b \mid s) P_{d,\varrho}(\cdot \mid s, b). \tag{51}$$

For any s,

$$\begin{aligned} \|K_{d,\varrho}(\cdot \mid s) - K_{d',\varrho'}(\cdot \mid s)\|_1 &\le \|\varrho(\cdot \mid s) - \varrho'(\cdot \mid s)\|_1 + \max_{s,b} \|P_{d,\varrho}(\cdot \mid s, b) - P_{d',\varrho'}(\cdot \mid s, b)\|_1 \\ &\le \|d - d'\|_1 + 2 \max_{s} \|\varrho(\cdot \mid s) - \varrho'(\cdot \mid s)\|_1. \end{aligned} \tag{52}$$

This gives $L_d = 1$ and $L_\varrho = 2$. ■

Assumption 8 (Uniform Mixing) *There exists $\kappa < 1$ such that every kernel $K_{d,\varrho}$ induced by a full-support policy has Dobrushin coefficient [11, 12, 30] at most κ:*

$$\sup_{d_1 \neq d_2} \frac{\|d_1 K_{d,\varrho} - d_2 K_{varrho}\|_1}{\|d_1 - d_2\|_1} \leq \kappa. \tag{53}$$

Assumption 9 (Small Coupling or Sufficient Regularization) *Let*

$$B_\gamma = \frac{R_{\max} + \gamma \log A_{\max}}{1 - \alpha}. \tag{54}$$

Define

$$L_Q^\gamma = \frac{L_r + \alpha B_\gamma L_P}{1 - \alpha}, \qquad L_\Gamma^\gamma = \frac{A_{\max}}{\gamma}, \qquad L_\pi^\gamma = L_\Gamma^\gamma L_Q^\gamma. \tag{55}$$

The primitive constants satisfy

$$c_\gamma = \kappa + L_d + (1 + L_\varrho) L_\pi^\gamma < 1. \tag{56}$$

Lemma 3 (Lipschitz Continuity of the Regularized Q-Fixed Point) *Under Assumption 7,*

$$\|Q_z^\gamma - Q_{z'}^\gamma\|_\infty \leq L_Q^\gamma \|z - z'\|. \tag{57}$$

Proof For any fixed Q, the soft value satisfies

$$|V_\gamma^Q(s)| \leq \|Q\|_\infty + \gamma \log A_{\max}. \tag{58}$$

Since rewards are bounded, the fixed point satisfies

$$\|Q_z^\gamma\|_\infty \leq \frac{R_{\max} + \alpha\gamma \log A_{\max}}{1 - \alpha}, \tag{59}$$

and hence

$$\|V_\gamma^{Q_z^\gamma}\|_\infty \leq B_\gamma. \tag{60}$$

Using the contraction of the regularized Bellman operator in Q,

$$\begin{aligned} \|Q_z^\gamma - Q_{z'}^\gamma\|_\infty &\leq \frac{1}{1 - \alpha} \|T_z^\gamma Q_{z'}^\gamma - T_{z'}^\gamma Q_{z'}^\gamma\|_\infty \\ &\leq \frac{1}{1 - \alpha} \left(L_r \|z - z'\| + \alpha B_\gamma L_P \|z - z'\| \right). \end{aligned} \tag{61}$$

This proves the claim. ■

Lemma 4 (Lipschitz Continuity of the Soft Best Response) *Under Assumption 7,*

$$\max_s \|\pi_z^\gamma(\cdot \mid s) - \pi_{z'}^\gamma(\cdot \mid s)\|_1 \leq L_\pi^\gamma \|z - z'\|. \tag{62}$$

Proof On each finite feasible action set, the softmax map is Lipschitz:

$$\|\Gamma_\gamma(Q)(\cdot \mid s) - \Gamma_\gamma(Q')(\cdot \mid s)\|_1 \le \frac{A_{\max}}{\gamma}\|Q - Q'\|_\infty. \tag{63}$$

Combining this bound with Lemma 3 gives

$$\max_s \|\pi_z^\gamma(\cdot \mid s) - \pi_{z'}^\gamma(\cdot \mid s)\|_1 \le L_\Gamma^\gamma L_Q^\gamma \|z - z'\|. \tag{64}$$

■

Lemma 5 (Stability of Regularized Best-Response Map) *Under Assumptions 7–9, the regularized mean-field response map Φ_γ is a contraction:*

$$\|\Phi_\gamma(z) - \Phi_\gamma(z')\| \le c_\gamma \|z - z'\|, \qquad c_\gamma < 1. \tag{65}$$

Proof The policy component of Φ_γ satisfies, by Lemma 4,

$$\|\pi_z^\gamma - \pi_{z'}^\gamma\| \le L_\pi^\gamma \|z - z'\|. \tag{66}$$

For the population component,

$$\begin{aligned} \|d_z^\gamma - d_{z'}^\gamma\|_1 &= \|dK_{d,\pi_z^\gamma} - d'K_{d',\pi_{z'}^\gamma}\|_1 \\ &\le \|dK_{d,\pi_z^\gamma} - d'K_{d,\pi_z^\gamma}\|_1 + \|d'K_{d,\pi_z^\gamma} - d'K_{d',\pi_{z'}^\gamma}\|_1. \end{aligned} \tag{67}$$

The first term is bounded by uniform mixing:

$$\|dK_{d,\pi_z^\gamma} - d'K_{d,\pi_z^\gamma}\|_1 \le \kappa \|d - d'\|_1. \tag{68}$$

The second term is bounded by the Lipschitz continuity of the transition kernel:

$$\|d'K_{d,\pi_z^\gamma} - d'K_{d',\pi_{z'}^\gamma}\|_1 \le L_d \|d - d'\|_1 + L_\varrho \|\pi_z^\gamma - \pi_{z'}^\gamma\|. \tag{69}$$

Therefore,

$$\|d_z^\gamma - d_{z'}^\gamma\|_1 \le (\kappa + L_d)\|d - d'\|_1 + L_\varrho L_\pi^\gamma \|z - z'\|. \tag{70}$$

Adding the policy and population components gives

$$\begin{aligned} \|\Phi_\gamma(z) - \Phi_\gamma(z')\| &= \|d_z^\gamma - d_{z'}^\gamma\|_1 + \|\pi_z^\gamma - \pi_{z'}^\gamma\|_1 \\ &\le [\kappa + L_d + (1 + L_\varrho)L_\pi^\gamma]\, \|z - z'\| \\ &= c_\gamma \|z - z'\|. \end{aligned} \tag{71}$$

By Assumption 9, $c_\gamma < 1$. Thus Φ_γ is a contraction. ■

Interpretation. The condition $c_\gamma < 1$ has a direct economic meaning. The term κ measures how quickly the Karma state distribution mixes under a fixed policy. The terms L_d and L_ϱ measure how strongly the transition kernel reacts to changes in the population distribution and policy. The term L_π^γ measures how sharply the regularized best response reacts to changes in Q-values. Larger entropy temperature γ makes the softmax response smoother, while weaker population coupling makes the mean-field interaction less destabilizing. The contraction proof therefore applies to Karma games with sufficiently strong mixing, sufficiently weak strategic coupling, and sufficiently smooth regularized best responses.

Appendix F. Appendix: Further Extending Proofs

F.1. Zero-Temperature Limits Towards Unregularized SNE

Theorem 6 (Zero-Temperature Limit of Regularized SNE.) *If $\gamma_n \downarrow 0$ and every limit point of $(d^\star_{\gamma_n}, \pi^\star_{\gamma_n})$ coincides with the same equilibrium, then the limiting equilibrium is an ordinary stationary Nash equilibrium of the unregularized finite-state Karma game.*

Proof Finally, let us discuss a decline in exploration temperature $\gamma_n \downarrow 0$. For every Q-function Q and every state s, the soft value satisfies

$$\max_{b\in\mathcal{A}(s)} Q(s,b) \le V^Q_{\gamma_n}(s) \le \max_{b\in\mathcal{A}(s)} Q(s,b) + \gamma_n \log|\mathcal{A}(s)|. \tag{72}$$

Thus the regularized Bellman equation converges to the ordinary Bellman optimality equation. Furthermore, if b is not a maximizer of $Q(s,\cdot)$ and the gap is

$$\Delta_Q(s,b) = \max_{\tilde{b}\in\mathcal{A}(s)} Q(s,\tilde{b}) - Q(s,b) > 0, \tag{73}$$

then the softmax probability of b satisfies

$$\Gamma_{\gamma_n}(Q)(b \mid s) \le \exp\left(-\frac{\Delta_Q(s,b)}{\gamma_n}\right). \tag{74}$$

Therefore, as $\gamma_n \downarrow 0$, all probability mass is placed on greedy best responses BR_s. Any limit point $(\bar{d}, \bar{\pi})$ of $(d^\star_{\gamma_n}, \pi^\star_{\gamma_n})$ therefore satisfies

$$\bar{\pi}(\cdot \mid s) \in \mathrm{BR}_s(\bar{d}, \bar{\pi}) \qquad \forall s \in \mathcal{S}, \tag{75}$$

and

$$\bar{d} = \bar{d} P_{\bar{d},\bar{\pi}}. \tag{76}$$

Hence $(\bar{d}, \bar{\pi})$ is an ordinary SNE of the unregularized finite-state Karma game. ■

F.2. Finite-Population Interpretation

The theorems are stated in the mean-field limit because the SNE of the Karma dynamic population game is a population equilibrium concept. For a finite population of N agents, the same argument gives an approximate equilibrium. If payoffs and transitions are Lipschitz continuous [27] in the empirical population distribution, and if the empirical distribution differs from its mean-field limit by δ_N, then a unilateral deviator can change the population state by at most order $1/N$.

Let J_i be the expected discounted return of agent i under a policy profile:

$$J_i(\pi_i, \pi_{-i}; d_0) = \mathbb{E}_{\pi_i,\pi_{-i}}\left[\sum_{t=0}^{\infty} \alpha^t r_i(s_t, b_t) \,\middle|\, d_0\right]. \tag{77}$$

Consequently, for some constant $C < \infty$ independent of N, the approximate Nash statement can be bounded:

$$\sup_{\pi_i'} \left[J_i(\pi_i', \pi^\star_{-i}; d_0) - J_i(\pi^\star_i, \pi^\star_{-i}; d_0)\right] \le C\left(\delta_N + \frac{1}{N} + \gamma \log A_{\max}\right), \tag{78}$$

where $A_{\max} = \max_s |\mathcal{A}(s)|$, and J_i denotes the expected discounted return under a policy profile. Thus, for large N and small γ, the learned independent policy profile is an approximate Nash equilibrium of the finite-agent game, and finite-population approximation of the mean-field SNE.

Appendix G. Appendix: Further Computational Experiments

Figure 4 assesses how different initial policy configurations affect convergence speed toward equilibrium bidding strategies (for Urgency 2). Specifically, we compare four structured initializations: a *bid-nothing* policy, in which agents initially refrain from bidding; a *bid-all-you-have* policy, in which agents initially bid their full Karma balance and do not economize; *random-even* policy, which starts with evenly distributed probabilities across the action space; and a *random-monotone* policy, which samples admissible monotone bidding strategies consistent with the structural assumptions of the model. The results highlight that a stable best-response policy emerges (for all initializations) around 50,000 iterations, but not yet a fully "crystallized" softmax policy, where *random-even* and *bid-nothing* achieve high rewards the fastest.

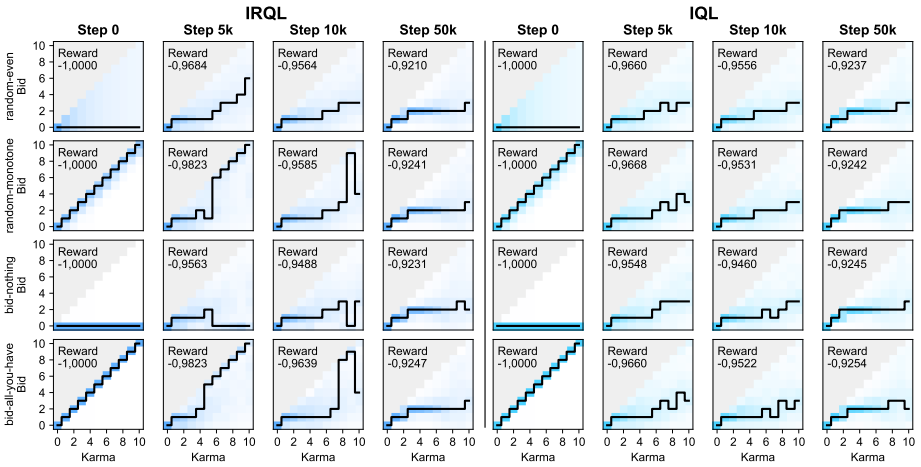

Figure 4: **Policy Emergence for Different Policy Initializations.**

Figure 5 demonstrates how population size affects learning dynamics and convergence behavior by evaluating the framework for $N \in \{50, 100, 150, 200\}$. Since the mean-field limit ($N \to \infty$) is a core assumption of the proof, and Karma allocation mechanisms rely on repeated interactions within large populations, varying the number of agents changes the stability of aggregate bid distributions and thus influences convergence toward stationary equilibrium policies. The major observation from this analysis: slightly faster convergence for larger populations, and higher variance for smaller populations.

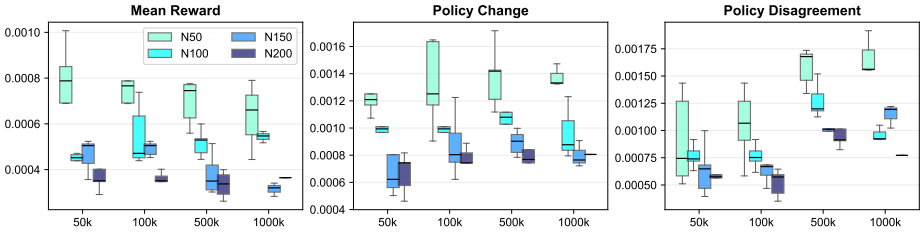

Figure 5: **Cross-Seed Standard Deviation Analysis for Different Population Sizes.**

Appendix H. Appendix: Implemented MARL Algorithms

This appendix describes the tabular MARL algorithms used in the computational experiments. All algorithms operate on the local Karma state $s_i = (u_i, k_i)$ where u_i is urgency and $k_i \in \{0, \dots, K\}$ is the current Karma balance. The feasible action set is the set of admissible bids $\mathcal{A}(k_i) = \{0, \dots, k_i\}$. After bids are submitted, agents are randomly paired in a two-agent auction, receive an immediate reward r_i, transition to a new state $s_i^+ = (u_i^+, k_i^+)$, and update their local learning tables.

The implemented MARL algorithms span three standard families of decentralized learning methods: value-based learners (IQL, IRQL), actor–critic policy-gradient learners (IAC, IPPO), and centralized-critic actor–critic learners (MAPPO). They differ primarily in how policies are represented, whether entropy regularization enters the Bellman recursion or only the policy update, and whether cross-agent information is used during training. Table 1 summarizes their structural differences.

Table 1: **Comparison of Benchmarked MARL Algorithms.**

Algorithm	Learner Type	Policy Representation	Bellman Target	Entropy Role	Cross-Agent Information
IQL	Value-Based	Softmax(Q) Implicit	Hard Max	Exploration Only	None
IRQL	Value-Based Regularized	Explicit Policy Table	Soft Value Log-Sum-Exp	Objective Regularization	None
IAC	Actor–Critic	Softmax Logits	Expected Soft Value	Policy-Gradient Regularization	None
IPPO	Actor–Critic Proximal	Parametric Policy Net	PPO Advantage	Entropy Bonus	None
MAPPO	Centralized-Critic & Actor–Critic	Parametric Policy Net	Centralized Advantage	Entropy Bonus	Centralized Critic Only

The computational experiments are practical, finite-time approximations, not literal instantiations of the theorem, as the Robbins-Monro assumption uses diminishing step sizes, while here we use (if not specified differently) constant, non-decaying learning rates.

H.1. [IQL] Independent Q-Learning

The implementation denoted IQL follows the independent Q-learning structure of Tan et al. [42]. Each agent i maintains its own table $Q_i(s, b) = Q_i((u, k), b)$ and treats the other learning agents as part of the environment. The bidding policy is not represented by an independently learned policy

table. Instead, it is induced directly by the current Q-values through a Boltzmann distribution,

$$\pi_i(b \mid s) = \pi_i(b \mid u, k) = \frac{\exp(Q_i((u,k),b)/T_{\mathrm{IQL}})}{\sum_{b' \in \mathcal{A}(k)} \exp(Q_i((u,k),b')/T_{\mathrm{IQL}})}, \qquad b \in \mathcal{A}(k), \tag{79}$$

where $T_{\mathrm{IQL}} > 0$ is the exploration temperature. At each step, the agent samples its bid from this distribution. After observing (s_i, b_i, r_i, s_i^+), only the visited table entry is updated:

$$Q_i(s_i, b_i) \leftarrow Q_i(s_i, b_i) \; + \; \eta_{\mathrm{IQL}} \left[r_i \; + \; \alpha \max_{b' \in \mathcal{A}(k_i^+)} Q_i(s_i^+, b') - Q_i(s_i, b_i) \right]. \tag{80}$$

Here $\eta_{\mathrm{IQL}} > 0$ is configured by `marl_iql_q_learning_rate`, and T_{IQL} is configured by `marl_iql_temperature`. For plotting and check-pointing, the implementation stores the policy induced by the Boltzmann equation above, but this policy is a deterministic function of the Q-table rather than a separate learner.

This algorithm assumes that the local state $s = (u, k)$ is the agent's only decision-relevant private state, that the environment is finite because Karma and urgency are bounded, and that persistent Boltzmann exploration visits feasible bids sufficiently often. The usual single-agent convergence guarantees do not directly apply in the MARL setting because all agents learn simultaneously, making the environment non-stationary from the perspective of each individual learner.

H.2. [IRQL] Independent Regularized Q-Learning

The implementation denoted IRQL is the entropy-regularized analogue of IQL. Each agent i again maintains a private state-action table $Q_i(s, b) = Q_i((u, k), b)$ and updates only from its own local transition (s_i, b_i, r_i, s_i^+). However, in contrast to IQL, IRQL keeps an explicit policy table $\pi_i(\cdot \mid s)$ and uses a soft Bellman continuation value in the Q-learning target. For a temperature $T_{\mathrm{IRQL}} > 0$, define the soft value

$$V_i^{T_{\mathrm{IRQL}}}(s) = T_{\mathrm{IRQL}} \log \sum_{b' \in \mathcal{A}(k)} \exp\left(\frac{Q_i(s, b')}{T_{\mathrm{IRQL}}} \right), \qquad s = (u, k). \tag{81}$$

The corresponding regularized best-response map is the softmax distribution

$$\Gamma_{T_{\mathrm{IRQL}}}(Q_i)(b \mid s) = \frac{\exp(Q_i(s,b)/T_{\mathrm{IRQL}})}{\sum_{b' \in \mathcal{A}(k)} \exp(Q_i(s,b')/T_{\mathrm{IRQL}})}, \qquad b \in \mathcal{A}(k). \tag{82}$$

At each step, the agent samples its bid from the current policy table $\pi_i(\cdot \mid s_i)$. After observing (s_i, b_i, r_i, s_i^+), only the visited Q-entry is updated:

$$Q_i(s_i, b_i) \leftarrow Q_i(s_i, b_i) \; + \; \eta_{\mathrm{IRQL}} \left[r_i \; + \; \alpha V_i^{T_{\mathrm{IRQL}}}(s_i^+) - Q_i(s_i, b_i) \right]. \tag{83}$$

The policy is then moved toward the softmax policy induced by the current Q-table:

$$\pi_i(\cdot \mid s_i) \leftarrow \pi_i(\cdot \mid s_i) \; + \; \beta_{\mathrm{IRQL}} \left[\Gamma_{T_{\mathrm{IRQL}}}(Q_i)(\cdot \mid s_i) - \pi_i(\cdot \mid s_i) \right]. \tag{84}$$

All probabilities are restricted to the feasible bid set $\mathcal{A}(k_i) = \{0, \ldots, k_i\}$, and infeasible bids receive probability zero. In the implementation, η_{IRQL} is configured by `marl_irql_q_learning_rate`, β_{IRQL} by `marl_irql_policy_learning_rate`, and T_{IRQL} by `marl_irql_temperature`.

The distinction between IQL and IRQL is therefore not the use of a softmax formula alone. Both algorithms use a softmax-type distribution over feasible bids at finite temperature. The difference is the role of that distribution in the Bellman recursion. In IQL, the Boltzmann distribution is the behavior policy used to sample bids, while the continuation value remains the hard optimality value

$$\max_{b' \in \mathcal{A}(k_i^+)} Q_i(s_i^+, b'). \tag{85}$$

Thus IQL learns ordinary, unregularized Q-values and the stored policy is a deterministic function of the Q-table. At fixed $T_{\mathrm{IQL}} > 0$, this policy should be interpreted as a finite-temperature exploration or perturbed best-response policy.

By contrast, IRQL replaces the hard maximum in the Q-target by the entropy-regularized value $V_i^{T_{\mathrm{IRQL}}}$. The induced softmax is therefore not only an exploration device; it is the regularized best-response map associated with the objective optimized by the soft Bellman update. Equivalently, for $\pi_i = \Gamma_{T_{\mathrm{IRQL}}}(Q_i)$,

$$V_i^{T_{\mathrm{IRQL}}}(s) = \sum_{b \in \mathcal{A}(k)} \pi_i(b \mid s) Q_i(s, b) \; + \; T_{\mathrm{IRQL}} H(\pi_i(\cdot \mid s)), \tag{86}$$

where $H(\pi_i(\cdot \mid s))$ is the Shannon entropy over feasible bids. This entropy term is what makes the limiting best response regularized. As $T_{\mathrm{IRQL}} \downarrow 0$, the soft value approaches the hard maximum and IRQL approaches the ordinary IQL Bellman target, up to the separate policy-learning timescale.

As with the other implemented independent MARL algorithms, IRQL treats the simultaneously learning population as part of the environment and does not exchange Q-values, policies, gradients, or population statistics across agents. The finite Karma cap and finite urgency set make the local state-action space tabular. The regularized convergence interpretation relies on persistent exploration, bounded rewards, and a policy-learning rate that is slower than the Q-learning rate, so that the Q-values track the soft Bellman fixed point while the policy evolves toward the regularized best response.

H.3. [IAC] Independent Actor–Critic with Entropy-Regularized Softmax Policy Gradient

The implementation denoted IAC is an independent actor–critic algorithm based on the tabular softmax policy-gradient and entropy-regularized policy-gradient formulation of Mei et al. [28]. Each agent maintains two tabular objects. The critic is a state-action table $Q_i(s, b)$ and the actor is a logit table $\theta_i(s, b)$. The policy is obtained by applying the softmax transform to the feasible bid logits:

$$\pi_i(b \mid s) = \frac{\exp(\theta_i(s, b))}{\sum_{b' \in \mathcal{A}(k)} \exp(\theta_i(s, b'))}, \qquad b \in \mathcal{A}(k). \tag{87}$$

This corresponds to the tabular softmax parametrization analyzed by Mei et al. [28].

The critic estimates the action-value function of the current policy. After observing (s_i, b_i, r_i, s_i^+), the implementation first computes a soft continuation value. Let $\gamma_{\mathrm{IAC}} \geq 0$ denote the entropy temperature. For a state $s = (u, k)$, define the soft action score

$$\widetilde{Q}_i(s, b) = Q_i(s, b) - \gamma_{\mathrm{IAC}} \log \pi_i(b \mid s), \tag{88}$$

and the corresponding soft state value

$$\widetilde{V}_i(s) = \sum_{b \in \mathcal{A}(k)} \pi_i(b \mid s) \widetilde{Q}_i(s, b). \tag{89}$$

When $\gamma_{\mathrm{IAC}} = 0$, these reduce to the ordinary expected Q-value under the current policy. The critic update is

$$Q_i(s_i, b_i) \leftarrow Q_i(s_i, b_i) \; + \; \eta^Q_{i,t} \left[r_i \; + \; \alpha \widetilde{V}_i(s_i^+) - Q_i(s_i, b_i) \right], \tag{90}$$

where the critic step-size is visit-count dependent,

$$\eta^Q_{i,t} = \frac{\eta^Q_{\mathrm{IAC}}}{n_i(s_i, b_i)^{\rho^Q_{\mathrm{IAC}}}}. \tag{91}$$

The parameters η^Q_{IAC} and ρ^Q_{IAC} are configured by `marl_iac_critic_learning_rate` and `marl_iac_critic_learning_decay`.

The actor update follows the softmax-gradient structure. For the current state $s = (u, k)$, define the soft advantage

$$\widetilde{A}_i(s, b) = \widetilde{Q}_i(s, b) - \widetilde{V}_i(s). \tag{92}$$

Using the tabular softmax gradient identity, the logit update is

$$\theta_i(s, b) \leftarrow \theta_i(s, b) \; + \; \eta^\theta_{i,t} \pi_i(b \mid s) \widetilde{A}_i(s, b), \qquad b \in \mathcal{A}(k). \tag{93}$$

This is the actor-critic analogue of the entropy-regularized softmax policy-gradient update: the true soft advantage is replaced by the critic estimate. The actor step size is

$$\eta^\theta_{i,t} = \frac{\eta^\theta_{\mathrm{IAC}}}{m_i(s)^{\rho^\theta_{\mathrm{IAC}}}}, \tag{94}$$

where $m_i(s)$ is the number of actor updates performed in state s. The parameters $\eta^\theta_{\mathrm{IAC}}$ and $\rho^\theta_{\mathrm{IAC}}$ are configured by `marl_iac_actor_learning_rate` and `marl_iac_actor_learning_decay`.

The entropy temperature is initialized by `marl_iac_entropy_temperature`, multiplied after each update by `marl_iac_entropy_temperature_decay`, and lower bounded by `marl_iac_min_entropy_t` For $\gamma_{\mathrm{IAC}} > 0$, the actor is optimized toward an entropy-regularized objective, which penalizes near-deterministic policies and keeps action probabilities away from zero. For $\gamma_{\mathrm{IAC}} = 0$, `iac` reduces to an unregularized softmax actor-critic update.

The implementation makes several simplifying assumptions. First, each agent is an independent learner and treats the changing policies of other agents as part of the environment. Second, the critic is assumed to provide a useful local estimate of the policy-gradient advantage even though updates are sampled from simultaneous multi-agent interaction rather than a stationary single-agent MDP. Third, the feasible action set changes with Karma, so the softmax and all gradients are restricted to bids $b \leq k$ and assign zero probability to infeasible bids. Fourth, the entropy-regularized convergence results of Mei et al. [28] apply to tabular softmax policy-gradient with exact gradients in finite MDPs; `IAC` should therefore be interpreted as a practical actor-critic approximation to that update, not as a direct instantiation of the paper's exact-gradient algorithm.

H.4. [IPPO] Independent Proximal Policy Optimization

The implementation denoted IPPO [48] applies PPO independently to the agents, while using shared neural-network parameters for compactness. The actor and critic both receive only the local observation of the acting agent. The local observation is encoded as normalized urgency, normalized Karma, and optionally a one-hot agent identifier. Thus the actor input for agent i is

$$x_i = \left(\frac{\iota(u_i)}{|\mathcal{U}| - 1}, \frac{k_i}{K}, e_i \right), \tag{95}$$

where $\iota(u_i)$ is the index of urgency u_i and e_i is included only when `marl_ippo_use_agent_id` is true. The actor is a multilayer perceptron that produces logits over the full bid set $\{0, \dots, K\}$. Before the categorical distribution is formed, infeasible bids are masked by assigning them very negative logits. The implemented policy is therefore

$$\pi_\theta(b \mid x_i) = \frac{\exp(g_\theta(x_i)_b)}{\sum_{b'=0}^{k_i} \exp(g_\theta(x_i)_{b'})}, \qquad b \in \{0, \dots, k_i\}, \tag{96}$$

with zero probability assigned to $b > k_i$. The critic is a separate multilayer perceptron $V_\phi(x_i)$ with the same local input. The cached table `policy` is not the learned object itself, but a tabular snapshot of π_θ used for aggregation, plotting, and check-pointing.

At each environment step, the implementation samples all agents' bids from the current actor and stores the transition tuple. The stored tuple contains local observations, actions, rewards, next local observations, old action log-probabilities, and a terminal indicator. Training is triggered when `marl_ippo_rollout_length` transitions have accumulated, or earlier if a terminal transition is observed (in this case $\zeta_t = 0$). For a rollout of length T, the implementation computes TD(λ) returns by the backward recursion

$$R^\lambda_{t,i} = r_{t,i} + \alpha(1 - \zeta_t) \left[(1 - \lambda) V_\phi(x_{t+1,i}) + \lambda R^\lambda_{t+1,i} \right], \tag{97}$$

where λ is configured by `marl_ippo_td_lambda`. The advantage estimate is

$$\widehat{A}_{t,i} = R^\lambda_{t,i} - V_\phi(x_{t,i}). \tag{98}$$

Advantages are normalized over the rollout when `marl_ippo_normalize_advantage` is true. Returns are optionally normalized when `marl_ippo_normalize_return` is true.

The rollout is flattened across time and agents before PPO minibatches are formed. If `marl_ippo_minibatch_size` is positive and smaller than the resulting batch size, the flattened rollout is split into randomly shuffled minibatches. Otherwise, the full flattened rollout is used as a single batch. For each sampled transition, the probability ratio is

$$\rho_{t,i}(\theta) = \exp\left(\log \pi_\theta(b_{t,i} \mid x_{t,i}) - \log \pi_{\theta_{\text{old}}}(b_{t,i} \mid x_{t,i}) \right). \tag{99}$$

The actor minimizes the negative clipped surrogate with an entropy bonus:

$$\mathcal{L}^{\text{IPPO}}_{\text{actor}}(\theta) = -\mathbb{E}\left[\min\left(\rho_{t,i}(\theta) \widehat{A}_{t,i}, \text{clip}(\rho_{t,i}(\theta), 1 - \epsilon, 1 + \epsilon) \widehat{A}_{t,i} \right) \right] - c_H \mathbb{E}\left[H(\pi_\theta(\cdot \mid x_{t,i})) \right]. \tag{100}$$

Here ϵ is configured by `marl_ippo_clip` and c_H is configured by `marl_ippo_entropy_coef`. The critic minimizes a mean-squared return loss:

$$\mathcal{L}_{\text{critic}}^{\text{IPPO}}(\phi) = c_V \mathbb{E}\left[\left(V_\phi(x_{t,i}) - R_{t,i}^{\lambda}\right)^2\right], \tag{101}$$

where c_V is configured by `marl_ippo_value_loss_coef`. Actor and critic parameters are optimized by the configured PyTorch optimizer, usually Adam, using separate learning rates. Gradients are clipped to `marl_ippo_max_grad_norm` when this value is positive. If `marl_ippo_target_kl` is positive, the implementation stops the PPO epoch loop early when the approximate KL estimate exceeds that threshold.

This implementation is independent in the information used for value estimation. Each value target for agent i is computed from that agent's own observation, reward, next observation, and current critic value. The only cross-agent coupling enters through the environment dynamics and through the optional agent-id feature in the shared networks. No centralized state, opponent action, opponent reward, or population statistic is passed to the IPPO critic.

H.5. [MAPPO] Multi-Agent Proximal Policy Optimization

The implementation denoted MAPPO reuses the decentralized actor architecture of IPPO and replaces the critic with a centralized critic [48]. The actor still conditions only on the local feature vector x_i, so execution remains decentralized. The policy used for bidding is the same masked categorical policy

$$\pi_\theta(b \mid x_i) = \frac{\exp(g_\theta(x_i)_b)}{\sum_{b'=0}^{k_i} \exp(g_\theta(x_i)_{b'})}, \qquad b \in \{0, \dots, k_i\}. \tag{102}$$

The centralized critic receives the joint observation of all agents. For a population of N agents, the joint feature vector is

$$y = \left(\frac{\iota(u_1)}{|\mathcal{U}| - 1}, \frac{k_1}{K}, \dots, \frac{\iota(u_N)}{|\mathcal{U}| - 1}, \frac{k_N}{K}\right). \tag{103}$$

When `marl_mappo_use_agent_id` is true, the queried agent's one-hot identifier is appended to this joint feature vector. The critic therefore estimates an agent-specific centralized value $V_\psi(y, i)$ rather than a purely local value $V_\phi(x_i)$.

The rollout buffer stores the same fields as IPPO. During training, the actor minibatch uses local features and feasible-action masks, while the critic minibatch uses centralized features. The implementation computes generalized advantage estimates using centralized values:

$$\delta_{t,i} = r_{t,i} + \alpha(1 - \zeta_t)V_\psi(y_{t+1}, i) - V_\psi(y_t, i), \tag{104}$$

$$\widehat{A}_{t,i}^{\text{GAE}} = \delta_{t,i} + \alpha\lambda(1 - \zeta_t)\widehat{A}_{t+1,i}^{\text{GAE}}. \tag{105}$$

The return target is

$$R_{t,i}^{\text{GAE}} = \widehat{A}_{t,i}^{\text{GAE}} + V_\psi(y_t, i). \tag{106}$$

The discount factor is configured by `marl_mappo_gamma` and λ is configured by `marl_mappo_gae_lambda`. Advantages and returns are normalized according to the corresponding MAPPO configuration flags.

The MAPPO actor update uses the same PPO clipped objective as IPPO, but the advantage comes from the centralized critic. With

$$\rho_{t,i}(\theta) = \exp\left(\log \pi_\theta(b_{t,i} \mid x_{t,i}) - \log \pi_{\theta_{\text{old}}}(b_{t,i} \mid x_{t,i})\right), \tag{107}$$

the actor loss is

$$\mathcal{L}_{\text{actor}}^{\text{MAPPO}}(\theta) = -\mathbb{E}\left[\min\left(\rho_{t,i}(\theta)\widehat{A}_{t,i}^{\text{GAE}}, \text{clip}(\rho_{t,i}(\theta), 1-\epsilon, 1+\epsilon)\widehat{A}_{t,i}^{\text{GAE}}\right)\right] - c_H\mathbb{E}\left[H(\pi_\theta(\cdot \mid x_{t,i}))\right]. \tag{108}$$

The centralized critic loss is

$$\mathcal{L}_{\text{critic}}^{\text{MAPPO}}(\psi) = c_V\mathbb{E}\left[\left(V_\psi(y_t, i) - R_{t,i}^{\text{GAE}}\right)^2\right]. \tag{109}$$

The implementation optimizes actor and critic parameters separately, clips gradients when configured, and can stop PPO epochs early when the approximate KL divergence exceeds `marl_mappo_target_kl`. The MAPPO configuration also supports fallback to the corresponding IPPO configuration fields when a MAPPO-specific field is absent.

This implementation follows the centralized-training and decentralized-execution pattern. During training, the critic can use all agents' urgency and Karma states to reduce variance and account for strategic coupling. During execution, each agent's bid is sampled from a policy that only sees its own urgency, its own Karma, and optionally its own identifier. Thus MAPPO differs from IPPO only in the information available to the critic and in the use of GAE returns rather than the TD(λ) return recursion used by IPPO.

H.6. Computing Infrastructure & Runtimes

Hardware. All computational experiments were conducted on our institution's compute nodes. The hardware and system configuration used in our experiments are summarized in Table 2.

Table 2: Hardware and system configuration used for the experiments.

Component	Specification
CPU	AMD Ryzen Threadripper PRO 7995WX, 96 cores
GPU	NVIDIA GeForce RTX 5090, 32 GB memory
GPU driver	NVIDIA driver 590.48.01
Operating system	Ubuntu 24.04.1 LTS
Job scheduler	Slurm

Execution time. We report the runtime required to train each MARL algorithm for 1,000,000 iterations with 100 agents. The runtimes of the representative experiments are summarized in Table 3.

Table 3: Runtime for training each MARL algorithm for 1,000,000 iterations with 100 agents.

Algorithm	Runtime
IQL	$\sim$ 66 min
IRQL	$\sim$ 66 min
IAC	$\sim$ 80 min
IPPO	$\sim$ 172 min
MAPPO	$\sim$ 172 min

H.7. Finetuning & Parameters

We tuned the hyperparameters of all MARL algorithms using grid search. The final values selected for each method are reported in Table 4.

Table 4: **Computational Experiment Hyperparameter.**

Hyperparameter	**IQL**	**IRQL**	**IAC**	**MAPPO**	**IPPO**
Learning and exploration					
Learning rate(s)	Q: 0.04	Q: 0.04 policy: 0.02	actor: 0.05 critic: 0.5	actor: 1×10^{-4} critic: 5×10^{-4}	actor: 3×10^{-4} critic: 1×10^{-3}
Temperature	0.02	0.02	0.01	–	–
Learning-rate decay	–	–	actor: 1×10^{-3} critic: 0.6	–	–
Entropy-temperature decay	–	–	0.05	–	–
Minimum entropy temperature	–	–	0.0	–	–
Policy optimization					
Rollout length	–	–	–	512	128
Discount factor α	–	–	–	0.8	0.8
TD/GAE parameter λ	–	–	–	0.95	0.95
PPO clip parameter ϵ	–	–	–	0.2	0.2
Entropy coefficient	–	–	–	0.005	0.01
Value-loss coefficient	–	–	–	1.0	1.0
Optimization epochs	–	–	–	4	4
Minibatch size	–	–	–	4096	0
Max gradient norm	–	–	–	0.5	0.5
Target KL	–	–	–	0.02	0.0
Normalize advantage	–	–	–	True	True
Normalize return	–	–	–	False	False
Network architecture					
Use agent ID	–	–	–	True	True
Actor hidden dimension	–	–	–	64	64
Actor layers	–	–	–	2	2
Critic hidden dimension	–	–	–	256	64
Critic layers	–	–	–	3	2
Optimizer	–	–	–	Adam	Adam
Initialization					
Initial policy fit steps	–	–	–	0	0
Initial policy fit learning rate	–	–	–	1×10^{-2}	1×10^{-2}

Note: – indicates that the hyperparameter is not used by the corresponding method or was not applicable in the implementation. For IAC, the entropy-temperature decay is selected through our hyperparameter grid search and implements rapid annealing, so that entropy regularization mainly affects the initial updates.

Appendix I. Appendix: List of Symbols

Table 5 summarizes the main mathematical notation used in the manuscript. Symbols with algorithm-specific subscripts are local to the corresponding implemented MARL algorithm.

Table 5: List of symbols used in the manuscript.

Symbol	Meaning
Karma Game & Dynamic Population Game	
$\mathcal{N}$	Agent population.
N	Number of agents in a finite population.
i, j	Agent indices.
t	Discrete time or learning iteration index.
τ	Agent type in the dynamic population game; in Karma games, used for temporal-preference types.
$\mathcal{T}$	Finite type space.
N_τ	Number of types.
$g \in \Delta(\mathcal{T})$	Exogenous distribution over types.
g_τ	Population mass of type τ.
s	Local state. In the Karma game, $s = (u, k)$.
s^+, s'	Next state or successor state.
$\mathcal{S}$	Finite state space.
N_s	Number of states.
u	Urgency level.
u'	Next urgency realization.
$\mathcal{U}$	Finite urgency set.
μ	Full-support urgency distribution.
k	Karma balance.
k_i	Karma balance of agent i.
K	Maximum Karma balance or Karma cap.
$\mathbb{N}$	Non-negative integer set used for Karma balances and bids.
α	Discount factor or temporal preference, with $\alpha \in (0, 1)$.
a	Generic action in the dynamic population game.
$A_\tau[s]$	Feasible action set for type τ in state s.
$\mathcal{A}(s)$	Feasible bid/action set in state s.
$\mathcal{A}(k)$	Feasible bid set at Karma balance k, equal to $\{0, \ldots, k\}$.
$\mathcal{B}(k)$	Feasible bid set at Karma balance k, equal to $\{0, \ldots, k\}$.
$b, \tilde{b}, b'$	Bid variables.
b_i	Bid submitted by agent i.
$\mathcal{V}$	Resource allocation rule.
$\mathcal{W}$	Payment rule.
d_τ	State distribution of type τ.
$\mathcal{D}_\tau$	Feasible state-distribution set for type τ.

Continued on next page

Symbol		Meaning
d	$=$	Joint type-state distribution; in the representative Karma game, population state distribution.
$(d_1, \ldots, d_{N_\tau})$		
$\mathcal{D}$		Set of feasible joint type-state distributions.
d_t		Population state distribution at time t.
$d^\star$		Stationary equilibrium population distribution.
$d^\star_\gamma$		Entropy-regularized stationary equilibrium population distribution at temperature γ.
$\bar{d}$		Limit-point population distribution as $\gamma \downarrow 0$.
d_0		Initial population distribution.
δ_N		Finite-population deviation of the empirical distribution from its mean-field limit.
π_τ		Stationary homogeneous policy of type τ.
Π_τ		Policy set for type τ.
π		Policy or policy profile.
Π		Set of policy profiles.
π^i_t		Policy of agent i at learning time t.
π_i		Policy of agent i.
π_{-i}		Policy profile of all agents except i.
$\pi^\star$		Stationary equilibrium policy.
$\pi^\star_\gamma$		Entropy-regularized stationary equilibrium policy at temperature γ.
$\bar{\pi}$		Limit-point policy as $\gamma \downarrow 0$.
$\nu_{d,\pi}(b)$		Population probability mass bidding exactly b under (d, π).
$w_{d,\pi}(b)$		Probability of winning with bid b against the population bid distribution under (d, π).
$p_\tau[s^+$	$\mid$	Type-specific transition probability in the dynamic population game.
$s, a](d, \pi)$		
$P_\tau[s^+ \mid s](d, \pi)$		Type-specific state transition matrix under policy.
$P_{d,\pi}$		Transition kernel induced by population distribution d and policy π.
$P_{d,\pi}(s' \mid s, b)$		Transition probability from s to s' after bid b under (d, π).
$r_\tau[s, a](d, \pi)$		Immediate payoff for type τ in state s taking action a.
$r_i(s, b)$		Reward of agent i in state s after bid b.
$J_i(\pi_i, \pi_{-i}; d_0)$		Expected discounted return of agent i from initial distribution d_0.
$B_\tau[s](d, \pi)$		State-dependent best-response correspondence for type τ.
BR_s		Unregularized best-response set in state s.
$\mathrm{BR}_{\gamma,s}$		Entropy-regularized best-response map in state s.
$\Delta(\cdot)$		Probability simplex over a finite set.
$\geq_{\mathrm{FOSD}}$		First-order stochastic dominance relation.
Regularized Q-Learning & Convergence Proof		
Q, Q^i_t		Action-value function or Q-table; Q^i_t is agent i's Q-table at time t.
$Q^\gamma_{d,\pi}$		Fixed point of the regularized Bellman operator for frozen (d, π).
Q^γ_z		Fixed point of the regularized Bellman operator at social state z.

Continued on next page

Symbol	Meaning
V^Q	Hard optimality value induced by Q-values.
V_γ^Q	Entropy-regularized soft value induced by Q-values at temperature γ.
$T_{d,\pi}^\gamma$	Regularized Bellman operator under frozen (d, π).
$\Gamma_\gamma(Q)$	Softmax policy induced by Q-table Q at temperature γ.
γ	Entropy regularization or exploration temperature in the theoretical IRQL analysis.
γ_n	Sequence of entropy temperatures approaching zero.
$H(\pi(\cdot \mid s))$	Shannon entropy of the feasible-bid distribution in state s.
η_t	Q-learning step size.
β_t	Policy-learning step size.
$z = (d, \pi)$	Social state, consisting of population distribution and policy.
z_t	Social state at time t.
$z_\gamma^\star$	Fixed point of the regularized response map.
e	Deviation from the fixed point, $e = z - z_\gamma^\star$.
Φ_γ	Regularized mean-field response map.
d_z^γ	Population distribution component returned by $\Phi_\gamma(z)$.
π_z^γ	Policy component returned by $\Phi_\gamma(z)$.
M_{t+1}	Martingale-difference noise term.
ξ_t	Tracking error from the fast recursion and finite-population approximation.
D^+	Upper Dini derivative.
c_γ	Contraction constant of Φ_γ.
$\Delta_Q(s, b)$	Q-value gap between the best feasible bid and bid b in state s.
$A_{\max}$	Maximum feasible-action set size, $\max_s \lvert\mathcal{A}(s)\rvert$.
C	Finite constant in the approximate Nash bound.
Response-Map Stability (Appendix)	
$\lVert \cdot \rVert$	Norm on social states used in the contraction proof.
$\lVert \cdot \rVert_1$	ℓ_1 norm.
$\lVert \cdot \rVert_\infty$	Supremum norm.
$R_{\max}$	Uniform reward bound.
$K_{d,\varrho}$	Policy-averaged transition kernel induced by distribution d and generic policy ϱ.
ϱ	Generic policy argument in $K_{d,\varrho}$.
ϱ'	Alternative generic policy argument.
κ	Mixing/contraction coefficient for transition kernels.
L_r	Lipschitz constant for rewards.
L_P	Lipschitz constant for transition probabilities.
L_d	Lipschitz constant for kernel dependence on the population distribution.
L_ϱ	Lipschitz constant for kernel dependence on the policy argument.
B_γ	Bound on the regularized value.
L_Q^γ	Lipschitz constant for the regularized Q fixed point.

Continued on next page

Symbol	Meaning
L_Γ^γ	Lipschitz constant for the softmax map.
L_π^γ	Lipschitz constant for the induced policy response.
Implemented MARL Algorithms	
T_{IQL}	IQL Boltzmann exploration temperature.
η_{IQL}	IQL Q-learning rate.
T_{IRQL}	IRQL entropy temperature.
η_{IRQL}	IRQL Q-learning rate.
β_{IRQL}	IRQL policy-learning rate.
$V_i^{T_{\mathrm{IRQL}}}$	Agent i's IRQL soft value at temperature T_{IRQL}.
$\Gamma_{T_{\mathrm{IRQL}}}(Q_i)$	IRQL softmax policy induced by Q_i.
$\theta_i(s,b)$	IAC actor logit for agent i in state s and bid b.
$\widetilde{Q}_i(s,b)$	IAC entropy-adjusted action score.
$\widetilde{V}_i(s)$	IAC entropy-adjusted state value.
$\widetilde{A}_i(s,b)$	IAC entropy-adjusted advantage.
γ_{IAC}	IAC entropy temperature.
$\eta_{i,t}^Q$	IAC critic step size for agent i at time t.
η_{IAC}^Q	IAC base critic learning rate.
ρ_{IAC}^Q	IAC critic learning-rate decay exponent.
$n_i(s_i,b_i)$	Number of critic updates or visits for agent i at state-action pair (s_i,b_i).
$\eta_{i,t}^\theta$	IAC actor step size for agent i at time t.
$\eta_{\mathrm{IAC}}^\theta$	IAC base actor learning rate.
$\rho_{\mathrm{IAC}}^\theta$	IAC actor learning-rate decay exponent.
$m_i(s)$	Number of actor updates for agent i in state s.
x_i	Local neural-network observation for agent i.
e_i	Optional one-hot agent identifier.
$\iota(u_i)$	Index of urgency level u_i.
$\ell_\theta(x_i)_b$	Actor-network logit for bid b at local observation x_i.
π_θ	Neural actor policy parameterized by θ.
$V_\phi(x_i)$	IPPO critic value parameterized by ϕ for local observation x_i.
ϕ	IPPO critic parameters.
ζ_t	Terminal/continuation indicator used in rollout-return recursions.
T	Rollout length in PPO-style algorithms.
$R_{t,i}^\lambda$	IPPO TD(λ) return for agent i at time t.
λ	TD(λ) or GAE trace parameter.
$\widehat{A}_{t,i}$	Estimated advantage for agent i at time t.
$\rho_{t,i}(\theta)$	PPO probability ratio for agent i at time t.
θ_{old}	Previous actor parameters used to compute PPO importance ratios.
ϵ	PPO clipping parameter.
c_H	Entropy-bonus coefficient.
c_V	Value-loss coefficient.

Continued on next page

Symbol	Meaning
$\mathcal{L}_{\text{actor}}^{\text{IPPO}}$	IPPO actor loss.
$\mathcal{L}_{\text{critic}}^{\text{IPPO}}$	IPPO critic loss.
y	MAPPO centralized joint observation vector.
$V_\psi(y, i)$	MAPPO centralized critic value for agent i at joint observation y.
ψ	MAPPO centralized critic parameters.
$\delta_{t,i}$	MAPPO temporal-difference residual.
$\widehat{A}_{t,i}^{\text{GAE}}$	MAPPO generalized advantage estimate.
$R_{t,i}^{\text{GAE}}$	MAPPO GAE return target.
$\mathcal{L}_{\text{actor}}^{\text{MAPPO}}$	MAPPO actor loss.
$\mathcal{L}_{\text{critic}}^{\text{MAPPO}}$	MAPPO centralized critic loss.