Communication-Efficient Safety Coordination for Large Populations of LLM Web Agents

Anonymous Author(s)

Abstract

LLM-powered browser and web agents are increasingly used for scientific workflows, public-sector services, healthcare operations, education platforms, data collection, and large-scale interface automation. When many such agents act concurrently, they can overload shared browser workers, target websites, API limits, and compute resources, creating population-level failures invisible from a single-agent perspective. We study a controlled-fleet setting in which one operator deploys many similar web agents but wants to avoid per-agent centralized assignment. We formulate this problem as a constrained semi-Markov mean-field game with asynchronous macro-actions whose holding times and failure probabilities depend on aggregate action-specific congestion. We propose **Dual-MFG**, which broadcasts only a small congestion-price vector and a scalar safety price, allowing agents to adapt to population-level externalities. Under standard regularity, monotonicity, entropy-regularization, and feasibility assumptions, we prove existence of a regularized mean-field equilibrium and an $O(N^{-1/2})$ finite-population approximation bound. In a BrowserGym capacity-stress experiment with up to 128 parallel agents, Dual-MFG achieves the best decentralized success–violation tradeoff, improving task-benefit index, $\mathrm{TBI}_{\mathrm{excess}} = \mathrm{success} - \max(\mathrm{violation} - \rho, 0)$, from -0.170 for independent learning to -0.033 at $N = 128$, while the centralized oracle is the only method with positive TBI. A centralized capacity-aware oracle achieves higher safety-adjusted utility but requires per-agent assignment communication that grows with population size. Across $N \in \{32, 64, 128\}$, Dual-MFG recovers 41–51% of the centralized oracle's safety-adjusted improvement over independent learning while using 8.4×–31.4× less communication. Our evaluation uses a non-LLM macro-action abstraction that isolates this coordination signal; LLM-driven controllers are left to future work. These results support mean-field pricing as a communication-efficient mechanism for coordinating large web-agent fleets under shared resource contention.

Keywords

agentic AI, LLM web agents, browser agents, multi-agent coordination, mean-field games, safety, shared-resource congestion

1 Introduction

LLM-powered browser agents are increasingly deployed in batches rather than one at a time. A single operator may launch many agents to support scientific data collection, public-service workflows, healthcare operations, education platforms, interface testing, form checking, account workflows, or internal tool automation. Even when each agent is individually competent, the population can behave poorly as a system: simultaneous page loads, repeated clicks, and parallel tool calls compete for shared browser workers, target-site capacity, API quotas, and cloud resources. The resulting congestion can increase latency, induce timeouts, and raise the rate of operational safety violations. This makes scalable coordination an important systems problem for agentic AI intended to support scientific and societal applications.

This paper studies somehow a narrower and a more concrete problem than general LLM-agent safety. We focus on *controlled browser-agent fleets*: many similar agents are deployed by one operator, share a policy class, and interact through common resource bottlenecks. The operator wants to improve task performance while keeping the population-average violation rate below a specified threshold, but does not want to centrally observe and assign every action for every agent. Fully centralized control can require communication that scales with the population size, while independent learning ignores the externality each agent imposes on the others.

Our approach is to replace per-agent micromanagement with a small broadcast signal. At each round, the coordinator estimates aggregate congestion and violation statistics, then broadcasts a mean-field congestion price and a safety price to all agents. Each agent still acts using only its local state and the broadcast prices. The mechanism is analogous to traffic pricing: the signal does not specify which route every driver must take, but it makes individually chosen actions account for crowd-level congestion. We model this setting as a constrained semi-Markov mean-field game. The semi-Markov component captures browser-agent macro-actions with random holding times, such as clicking, waiting, or executing a browser option. The mean-field component captures the fact that each agent is affected by the aggregate population state rather than by the identity of every other agent. The constraint captures an operational safety requirement, such as keeping timeout, invalid-action, or rule-violation rates below a target. This framing turns a many-agent coordination problem into repeated single-agent learning against a slowly updated population summary.

This work focuses on a specific deployment regime: large managed fleets of web agents operating under shared resource contention. We show that mean-field pricing provides a principled and communication-efficient mechanism for coordinating population-level safety and congestion in this setting. Our contributions are:

- **Constrained semi-Markov formulation.** We formulate congested browser-agent deployment as a semi-Markov mean-field game with asynchronous macro-actions, aggregate latency, occupancy violation statistics, and a population-level safety constraint.
- **Dual-MFG algorithm.** We propose a primal-dual learning procedure that alternates between local policy improvement, aggregate mean-field updates, and a dual safety-price update. The coordinator broadcasts only aggregate prices rather than per-agent action assignments.
- **Idealized equilibrium analysis.** Under regularity, monotonicity, entropy-regularization, and feasibility assumptions, we prove

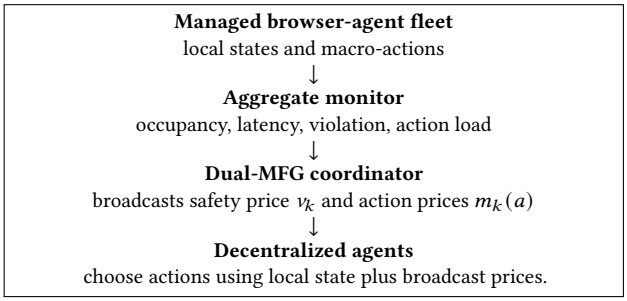

Figure 1: System overview. A coordinator observes only aggregate population statistics and broadcasts low-dimensional prices. It does not send per-agent action assignments; each browser agent acts locally using the broadcast prices.

existence of a regularized mean-field equilibrium, give a contraction condition for the fixed-point update, and obtain an $O(N^{-1/2})$ finite-population approximation bound.

- **BrowserGym capacity-stress evaluation.** We instantiate the method in a BrowserGym-based shared-resource stress test where macro-actions have finite action-specific capacities. The experiments isolate the setting in which mean-field feedback should help: safety-only control sees a scalar violation rate, whereas Dual-MFG receives action-specific load prices and can redistribute traffic while preserving constant-size communication.

Scope of the evaluation. Our experiments use lightweight PPO controllers over shared-capacity BrowserGym macro-actions. This isolates the broadcast coordination mechanism from perception and planning, while integration with LLM-driven controllers is left for future work.

2 Deployment Setting and Modeling Abstraction

We first make explicit the deployment regime considered in this work. The motivating setting is a platform or organization that controls a large fleet of browser agents running related tasks. Examples include automated user-interface testing, browser-based data collection, repetitive form workflows, or internal web-tool automation. The operator can broadcast a small global signal to all agents, but assigning actions centrally to every agent at every decision point is expensive and brittle. Figure 1 summarizes this coordination loop.

We abstract each browser agent as repeatedly choosing among macro-actions such as clicking elements, submitting forms, querying tools, waiting, or taking a no-op. The local browser observation is the state s; the distribution of active agents defines the mean field m; and shared infrastructure limits enter through aggregate occupancy, latency, violation, and action-capacity statistics. The semi-Markov structure captures variable macro-action durations such as page loads, API calls, browser waits, and reset overhead. This abstraction does not model every DOM-level transition. It isolates the population-level externality created when many agents concurrently use shared resources outside a single browser episode, such as worker pools, compute, target-site rate limits, API quotas, and server-side capacity. We summarize these effects through the mean-field aggregate.

The safety notion in this paper is operational rather than semantic. Here, safety refers to operational deployment failures such as timeouts, invalid actions, forbidden actions, excessive retries, or rule-threshold violations. Instead, the constraint represents measurable deployment failures such as timeouts, invalid actions, forbidden actions, excessive retry behavior, or rule-threshold violations. A population-level constraint is natural when the system operator must keep an aggregate error or violation rate below a service-level threshold.

The constraint is written as a population-average expected violation budget because many deployment resources are provisioned at the fleet level: a service can tolerate a small aggregate timeout or retry rate but not a large burst. This is also the regime in which a single broadcast dual price is meaningful. It is weaker than per-agent, worst-case, or high-probability safety. Those stronger notions would require individual risk state, confidence bounds, shielding, or robust control, and are outside the scope of the present mean-field pricing mechanism. We therefore report realized violation rates separately and treat per-agent or high-probability safety as future work.

Why a mean-field model? In a large homogeneous fleet, the identity of any single other agent is less important than the aggregate traffic it contributes. Mean-field modeling exploits this structure by training a representative agent against a population summary rather than against all other agents individually. This approximation is most appropriate for large controlled or partially controlled fleets with repeated operation and shared bottlenecks. It is less appropriate for highly heterogeneous open-web ecosystems in which agents have unrelated objectives and no common broadcast mechanism; we return to this limitation in Section 7 and the discussion.

3 Related Work

Our work lies at the intersection of mean-field games, reinforcement learning, and web-agent benchmarks.

Mean-field games (MFGs) provide tractable models for large-population interactions, with existence and uniqueness results well established in the continuous-time literature [Huang et al. 2006; Lasry and Lions 2007]. Recent work has extended MFG methods into reinforcement learning and control [Anahtarci et al. 2023; Cui and Koeppl 2021; Guo et al. 2019], including entropy-regularized fixed-point and deep-RL methods that motivate our use of regularization for stable policy responses. In parallel, multi-agent reinforcement learning (MARL) has explored population-level coordination via centralized training or opponent modeling [Foerster et al. 2018; Lowe et al. 2017].

Mean-field MARL methods such as MF-Q and MF-Actor-Critic approximate large populations through average actions or population statistics [Anahtarci et al. 2023; Cui and Koeppl 2021; Guo et al. 2019; Yang et al. 2018]. They are natural related baselines, but standard versions do not directly include our two key coordination signals: a population safety dual and action-specific capacity prices. Our ablations isolate these mechanisms through Safety-only, Congestion-only, and their combination in Dual-MFG. Constrained

MF-Q/MF-AC variants with the same capacity wrapper are important future baselines.

Browser-based environments such as WebArena and BrowserGym [Le Sellier De Chezelles et al. 2024; Zhou et al. 2023] enable realistic agent evaluation, but congestion and large-population interactions remain underexplored. We combine MFG-style coordination with a practical large-population congestion evaluation.

Most standard MFGs couple agents through reward/cost terms. Our setting includes **population-level safety constraints**, connecting it to generalized, constrained, and primal-dual MFGs [Gomes and Voskanyan 2016; Guo et al. 2025; Jusup et al. 2024; Mondal et al. 2024]. Our focus is different: we use a Lagrangian safety price and action-specific congestion prices as low-bandwidth coordination signals for discrete semi-Markov web-agent populations. The analysis below is restricted to an idealized regularized setting.

4 Problem Formulation and Notation

We formulate the above problem as a **semi-Markov Decision Process (SMDP)** defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \mathcal{P})$. Here, $\mathcal{S}$ is a compact subset of Euclidean space, and $\mathcal{A}$ is the action space of options. $\mathcal{T}$ denotes the **holding time distribution**, where the duration of an action depends on the current congestion level $\tau \sim \mathcal{T}(\cdot|s, a, m)$. Let $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$ be the **congestion feature** (e.g., bandwidth usage) and $v : \mathcal{S} \times \mathcal{A} \to \{0, 1\}$ be the **binary safety violation indicator**. We equip the space of probability measures $\mathcal{P}(\mathcal{S})$ with the **Wasserstein-1 distance** W_1, and denote the space of aggregates by $\mathcal{M} \subset \mathbb{R}^d$, assuming it is compact and convex.

4.1 Finite Population Formulation

Consider a finite population of N indistinguishable agents and let $\mathcal{W}$ be the finite set of websites. The state space is $\mathcal{S} = \mathcal{W} \times [0, 1] \times \mathbb{R}_{\geq 0} \times \mathcal{Z}$, where $s = (w, y, b, z)$ represents the current website w, the **goal progress** $y \in [0, 1]$ (where $y = 1$ denotes completion), the remaining budget b, and observation features z.

Agents interact exclusively through the population aggregate $m \in \mathcal{M}$, defined as the tuple of occupancy, latency, and violation rate per website: $m = (\omega, \bar{\ell}, \bar{v}) \in \mathbb{R}^{3|\mathcal{W}|}$, where ω denotes occupancy. Formally, let $\mu \in \mathcal{P}(\mathcal{S})$ denote a probability measure over the state space. For a finite population, we define the **empirical measure** at time t as $\mu_t^N = \frac{1}{N} \sum_{i=1}^N \delta_{s_t^i}$. The aggregate vector is obtained via a projection map $G : \mathcal{P}(\mathcal{S}) \to \mathcal{M}$, denoted by $m_t = G(\mu_t^N)$, which computes the relevant moments (occupancy, latency, violation) from the distribution. At each decision epoch, an agent chooses an option a with random holding time $\tau \sim \mathcal{T}(\cdot|s, a, m)$, so action duration can depend on congestion. The agent receives a reward: $r(s, a, m) = r_{\text{task}}(s, a) - C_{\text{time}}(s, a) - \alpha\phi(s, a, m)$, where $r_{\text{task}}(s, a)$ denotes the intrinsic task reward (e.g., success or progress), $C_{\text{time}}(s, a)$ represents the cost of time (proportional to the holding time τ), and $\phi \geq 0$ captures the congestion penalty. Note that safety violations $v(s, a)$ are tracked separately for the constraint in Section 4.2. We define a **policy** $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ as a measurable mapping from the state space to the probability simplex over actions. We denote the set of all such stationary Markov policies by Π. At each decision epoch t, agent i observes its local state s_t^i and independently samples an action $a_t^i \sim \pi(\cdot|s_t^i)$. We remark that given the asynchronous nature of semi-Markov decision epochs, we model the agent's interaction against a stationary aggregate m, consistent with the Mean-Field Equilibrium analysis in Section 4.3.

4.2 Constrained Formulation

We seek to maximize the average expected return across the population while ensuring the **population-average** violation rate stays below a target ρ_{safe}. The finite-N constrained objective is:

$$\max_{\pi_1,\ldots,\pi_N} \frac{1}{N} \sum_{i=1}^N \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r(S_t^i, A_t^i, m_t)\right] \tag{1}$$

$$\text{s.t.} \quad \frac{1}{N} \sum_{i=1}^N \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} v(S_t^i, A_t^i)\right] \leq \rho_{\text{safe}}, \quad m_t = G(\mu_t^N).$$

Population-average safety permits occasional individual violations as long as the fleet-level expected violation rate stays below the target. This matches service-level deployment constraints and keeps the dual update scalar; per-agent or tail-risk safety would require individual risk budgets or high-probability control. We use a Lagrangian relaxation, justified by the following standard strong-duality condition for constrained RL.

Proposition 4.1 (Strong Duality [Paternain et al. 2019])**.** *Consider the agent's Best Response problem against a fixed aggregate m. Let $g(\pi \mid m) \triangleq \mathbb{E}\left[\frac{1}{T} \sum_{t=0}^{T-1} v(S_t, A_t)\right]$ be the expected violation rate, and define the Lagrangian as $\mathcal{L}(\pi, \nu \mid m) = J(\pi \mid m) - \nu(g(\pi \mid m) - \rho_{\text{safe}})$. Under Slater's condition (existence of a strictly safe policy), this subproblem satisfies strong duality. Consequently, the optimal constrained policy can be recovered by solving $\max_\pi \min_{\nu \geq 0} \mathcal{L}(\pi, \nu \mid m)$.*

4.3 Mean-Field Equilibrium

Definition 4.2 (Mean-field equilibrium (MFE))**.** A pair $(\pi^\star, \{\mu_t^\star\}_{t=0}^T)$ is a finite-horizon MFE if for all t,

$$\pi^\star \in \arg\max_\pi \mathbb{E}\Big[\sum_{k=t}^{T-1} \gamma^{k-t} r(S_k, A_k, m_k^\star)\Big]$$
$$\text{given } m_t^\star = G(\mu_t^\star), \qquad \mu_{t+1}^\star = \Phi(\mu_t^\star; \pi^\star). \tag{2}$$

Remark 4.3 (Relationship to GNE). Strictly speaking, the coupled safety constraint induces a Generalized Nash Equilibrium (GNE) problem, as the feasible policy space of one agent depends on the collective violation rate. However, by solving the Lagrangian relaxation, the coupled constraint is represented through a common shadow price $\nu^\star$. Conditional on this price, agents solve decoupled local problems. This does not remove the population-level nature of the original constraint; rather, it gives an equivalent priced formulation under strong duality and feasibility assumptions.

4.4 Forward Operators

Let $P^{\pi,m}$ denote the **induced transition kernel** given by $P^{\pi,m}(s'|s) = \sum_a P(s'|s, a, m)\pi(a|s)$. Recall the aggregate map $m = G(\mu)$. The **population dynamics** are governed by the operator $\Phi(\mu; \pi)(B) = \int P^{\pi,G(\mu)}(B|s)\mu(\mathrm{d}s)$, which describes the physical evolution $\mu_{t+1} = \Phi(\mu_t; \pi)$. However, exact tracking of the time-varying measure μ_t is intractable. To establish the existence of a stable operating point, we focus on the **Stationary Mean-Field Equilibrium** and construct the **aggregate consistency operator** $\Psi(m, \pi)$ directly. We define

Ψ via the **discounted occupancy measure** induced by playing π against a fixed m:

$$d_\gamma^{\pi,m}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr^{\pi,m}(S_t = s \mid S_0 \sim \mu_0).$$

Defining the occupancy vector $\rho^{\pi,m}(w) = \sum_s d_\gamma^{\pi,m}(s)\mathbf{1}_{\text{site}(s)=w}$, the latency and violation aggregates are derived as regularized ratio statistics:

$$(\bar{\ell}, \text{viol})^{\pi,m}(w) = \frac{(1-\gamma)\,\mathbb{E}^{\pi,m}[\sum_{t\geq 0}\gamma^t \mathbf{1}_{W_t=w}(\tau_t, v_t)]}{\rho^{\pi,m}(w)+\varepsilon},$$

where the holding time is drawn from the congestion-dependent distribution $\tau_t \sim \mathcal{T}(\cdot|S_t, A_t, m)$ and v_t is the violation indicator. This defines the map $\Psi(m,\pi)$. The equilibrium is then characterized as the fixed point $m^* = T(m^*)$, where $T(m) \triangleq \Psi(m, \text{BR}(m))$.

Remark 4.4 (Stationary approximation and finite-horizon effects). A fully finite-horizon formulation would allow time-dependent aggregates m_t and policies π_t. We instead analyze a stationary discounted approximation: a representative agent responds to a fixed aggregate m, and the induced occupancy measure defines the consistency map Ψ. This matches long-running deployments focused on stable operating behavior, but it is not an exact finite-horizon Nash characterization near initial or terminal boundaries. Turnpike results [Cardaliaguet et al. 2012; Porretta 2018] motivate such approximations away from boundary layers; our experiments test the resulting steady-state coordination heuristic.

Scope of the analysis. The theory is an idealized justification for the update structure, not a formal certificate for every BrowserGym implementation: the implementation uses noisy finite samples and approximate PPO best responses.

Assumptions. We impose the following standard assumptions on the Semi-Markov Game structure. Let the aggregate space be $\mathcal{M} = \Delta(|\mathcal{W}|) \times [0,L]^{|\mathcal{W}|} \times [0,1]^{|\mathcal{W}|}$.

- **(A1) Discounted SMDP Contraction:** The reward function r is bounded. The congestion-dependent holding times $\tau(s,a,m) \sim \mathcal{T}(\cdot|s,a,m)$ satisfy $\sup_{s,a,m}\mathbb{E}[\gamma^{\tau(s,a,m)}] \leq \bar{\gamma} < 1$, ensuring the Bellman operator remains a contraction despite random action durations.
- **(A2) Regularity of Coupling:** The reward function $r(s,a,m)$, transition kernel $P(\cdot|s,a,m)$, and holding time distribution $\mathcal{T}(\cdot|s,a,m)$ are Lipschitz continuous in the aggregate vector m, with maximal constant L_P. Furthermore, the projection map $G : \mathcal{P}(\mathcal{S}) \to \mathcal{M}$ is Lipschitz continuous with respect to the **Wasserstein-1 metric** W_1.
- **(A3) Compact Best-Response:** The set of admissible policies Π is compact, and the objective $J(\pi|m)$ is continuous. Hence, best-response set $\arg\max_\pi J(\pi \mid m)$ is non-empty for any m.
- **(A4) Monotonicity (Lasry–Lions):** The reward function satisfies the strict monotonicity condition: for any two aggregates $m \neq m'$, $\langle J(\cdot|m) - J(\cdot|m'), m - m'\rangle < 0$. Intuitively, higher traffic strictly reduces the utility at that site, guaranteeing equilibrium uniqueness.
- **(A5) Feasibility (Slater's Condition):** There exists at least one strictly safe policy π_{safe} such that the violation rate $g(\pi_{safe} \mid m) < \rho_{\text{safe}}$ for all m. This guarantees strong duality for the constrained optimization.

Interpretation of assumptions. Assumption **(A1)** gives Bellman contraction under random action durations; **(A2)** smooths resource contention in expectation; **(A4)** encodes monotone congestion; and **(A5)** requires a strictly feasible conservative policy. Entropy regularization and stochastic holding times smooth the aggregate map G, supporting **(A2)** in expectation. Proofs are in the Appendix.

5 Existence, uniqueness, and ε-Nash equilibria

We now present the main results by first establishing existence in the mean-field limit ($N \to \infty$). To that end, we first note that while Definition 4.2 poses the standard maximization problem, direct optimization in mean-field settings can yield discontinuous responses to congestion prices. To guarantee learning stability and the existence of a smooth best-response map, we analyze the *regularized* objective, consistent with the maximum-entropy RL formulation used in our experiments (Section 7).

Lemma 5.1 (Lipschitz Continuity of Best Response)**.** *Let $J(\pi|m)$ denote the expected return of policy π under aggregate m. Consider the regularized objective $f(\pi,m) = J(\pi|m) + \lambda\mathcal{H}(\pi)$, where $\mathcal{H}(\pi)$ is a σ-**strongly concave** regularizer (e.g., Shannon entropy). Following standard sensitivity analysis results [Dafermos 1980; Fiacco 1983], since the regularization ensures the optimizer lies in the relative interior of the policy space Π, the best-response map $\text{BR}(m) = \arg\max_{\pi\in\Pi} f(\pi,m)$ is Lipschitz continuous. Specifically, if the policy gradient $\nabla_\pi J(\pi|m)$ is Lipschitz in m with constant $L_{\pi m}$, we have: $\|\text{BR}(m) - \text{BR}(m')\| \leq \frac{L_{\pi m}}{\lambda\sigma}\|m - m'\|, \quad \forall m, m' \in \mathcal{M}$.*

From Local Sensitivity to Global Equilibrium. While Lemma 5.1 recalls standard local stability guarantees for the optimizer, the next proposition shows how these properties can be combined with a weak-coupling condition on the aggregate dynamics. The resulting statement is a sufficient condition for contraction of the mean-field update operator $T(m)$; it should not be read as claiming that arbitrary browser environments automatically satisfy contraction.

Proposition 5.2 (Existence of Regularized MFE)**.** *Assume **(A1)–(A3)**. Consider the λ-entropy regularized game where agents play $\pi^* = \text{BR}_\lambda(m)$. Define the operator $T : \mathcal{M} \to \mathcal{M}$ by $T(m) = \Psi(m; \text{BR}_\lambda(m))$. Then T is single-valued and continuous, mapping the compact convex set $\mathcal{M}$ into itself. By Schauder's fixed-point theorem, there exists a regularized equilibrium (π^*, m^*).*

Existence alone does not rule out multiple fixed points. We next state sufficient monotonicity and weak-coupling conditions under which the regularized stationary model has a unique stable operating point.

Proposition 5.3 (Uniqueness and Contraction)**.** *Assume **(A1)–(A4)**. First, under the monotonicity condition **(A4)**, the MFE aggregate m^* is unique. Second, regarding algorithmic convergence: Let L_P be the Lipschitz constant of the environment dynamics (from **A2**) and $\bar{\gamma}$ be the contraction factor (from **A1**). We define the **environment stability constant** as $K_{dyn} \triangleq \frac{L_P}{1-\bar{\gamma}}$. If the environment is weakly coupled ($K_{dyn} < 1$) and the regularization is sufficient such that: $\lambda > \frac{L_\Psi^\pi L_{\pi m}}{\sigma(1-K_{dyn})}$, then the fixed-point operator $T(m)$ is a strict contraction on $\mathcal{M}$, giving linear convergence of the idealized exact best-response iteration.*

Finally, we bridge the gap between this theoretical infinite population limit and our practical finite agent setting.

Proposition 5.4 (Finite-N ε-Nash)**.** *Let (π^*, m^*) be a consistent mean-field pair where $m^* = \Psi(m^*; \pi^*)$ and π^* is the constrained best response. Under Assumptions **(A1)** and **(A2)**, the value function J and constraint function g are Lipschitz continuous in the aggregate m. Then, for the finite population game with N agents, the exploitability is bounded by a finite-size error of order $O(N^{-1/2})$. Specifically, for any unilateral deviation π', the gain is bounded by:*

$$\mathbb{E}[J(\pi', m^{N'})] - \mathbb{E}[J(\pi^*, m^N)] \leq \frac{C}{\sqrt{N}}$$

, and the constraint violation satisfies $\mathbb{E}[g(\pi^*|m^N)] \leq \rho_{\text{safe}} + \frac{C}{\sqrt{N}}$. Thus, (π^*, m^*) constitutes an asymptotic Nash equilibrium for the idealized finite-population game.

What the finite-N result does and does not imply. Proposition 5.4 controls the loss from replacing the empirical population by its mean-field limit under Lipschitz coupling and concentration. It does not by itself prove that PPO finds the exact constrained best response, nor that a finite experimental run exactly satisfies the safety target. Those algorithmic and finite-sample effects are handled empirically in Section 7.

6 Algorithm and Convergence

The preceding section studies an exact regularized mean-field model. The implemented algorithm is necessarily approximate: population statistics are estimated from finite rollouts, and the policy update is performed by PPO rather than by an exact dynamic-programming best response. We therefore separate the idealized convergence statement from the practical implementation.

Let $\pi_k \approx \widehat{\mathrm{BR}}_{\nu_k}(m_k)$ denote the approximate best response learned by the agents given the current safety price ν_k. We model the noisy population update $\widehat{\Psi}$ and violation metric $\widehat{g}$ as

$$\widehat{\Psi}(m_k; \pi_k) = \Psi(m_k; \pi_k) + \xi_k^{\Psi} + b_k^{\Psi}$$
$$\widehat{g}(\pi_k \mid m_k) = \mathrm{viol}(m_k) + \xi_k^{g} + b_k^{g},$$

where $\xi_k^{(\cdot)}$ are martingale-difference noise terms ($\mathbb{E}[\xi_k|\mathcal{F}_k] = 0$) arising from finite-sampling variance, and $b_k^{(\cdot)}$ are approximation or estimation bias terms. The convergence theorem below assumes these bias terms vanish asymptotically; the finite BrowserGym experiments test whether the same update structure is useful when this assumption only holds approximately.

6.1 Safe Mean-Field Learning Algorithm

We propose **Dual-MFG** (Algorithm 1), which alternates between policy optimization, mean-field aggregation, and safety dual updates. To enforce the safety constraint, we utilize the Lagrangian $\mathcal{L}(\pi, \nu \mid m) = J(\pi \mid m) - \nu(g(\pi \mid m) - \rho_{\text{safe}})$. The dual variable ν acts as an integrator: $\nu_{k+1} = \Big[\nu_k + \eta_k(\hat{g}_k - \rho_{\text{safe}})\Big]_+$. It increases the penalty price when violations exceed the target ρ_{safe} and decays when the system is safe.

Algorithm 1 Safe Mean-Field Learning (Dual-MFG)

Input: Initial policy θ_0, initial aggregate m_0, safety target ρ_{safe}.
Hyperparameters: Dual step size η, Mean-Field update rate β, PPO epochs E.
Initialize: Dual variable $\nu_0 = 0$.
for iteration $k = 0, 1, 2, \ldots$ **do**
  // **Phase 1: Interaction & Data Collection**
  Broadcast aggregate m_k and safety price ν_k to all N agents.
  Agents execute π_{θ_k} to collect trajectories $\mathcal{D}_k = \{(s_t^{(i)}, a_t^{(i)}, r_t^{(i)}, v_t^{(i)})\}_{i=1}^N$.
  // **Phase 2: Primal Update (Approx. Best Response)**
  Construct Lagrangian reward: $r_t^{\nu} = r(s_t, a_t, m_k) - \nu_k v_t$.
  Update θ_{k+1} via PPO on $\mathcal{D}_k$ maximizing r^{ν} (run E epochs).
  // **Phase 3: Dual Update (Safety)**
  Estimate empirical violation rate: $\hat{g}_k = \frac{1}{|\mathcal{D}_k|} \sum_{i,t} v_t^{(i)}$.
  Update dual variable: $\nu_{k+1} \leftarrow [\nu_k + \eta_k(\hat{g}_k - \rho_{\text{safe}})]_+$.
  // **Phase 4: Mean-Field Update (Congestion)**
  Compute empirical aggregate $\hat{m}_k$ from batch $\mathcal{D}_k$.
  Smooth update: $m_{k+1} \leftarrow (1 - \beta_k)m_k + \beta_k \hat{m}_k$.
end for

6.2 Convergence Analysis

We analyze the coupled stochastic approximation assuming the RL subroutine tracks the regularized best response sufficiently accurately at each step. This is analytical rather than empirical: since PPO is an approximate optimizer, the theorem is a guarantee for the idealized two-timescale update, not a certificate that each finite run computes an exact Nash equilibrium.

Theorem 6.1 (Two-timescale Convergence)**.** *Consider the joint updates with estimators $\widehat{\Psi}, \widehat{g}$ defined previously. Recall that $g(\pi \mid m)$ is the expected violation rate defined in Proposition 4.1. The update rules are given by:*

$$m_{k+1} = (1 - \beta_k)m_k + \beta_k \widehat{\Psi}\big(m_k; \widehat{\mathrm{BR}}_{\nu_k}(m_k)\big)$$
$$\nu_{k+1} = \Big[\, \nu_k + \eta_k\big(\widehat{g}(\pi_k \mid m_k) - \rho_{\text{safe}}\big) \Big]_+.$$

*Assume the step sizes satisfy the standard Robbins-Monro conditions ($\sum \beta_k = \infty, \sum \beta_k^2 < \infty$, etc.). Assume a **timescale separation** $\eta_k/\beta_k \to 0$, implying the dual update (safety) is slower than the aggregate update (congestion). Define the **equilibrium violation map** $\Lambda(\nu)$ as: $\Lambda(\nu) \triangleq g(\mathrm{BR}_\nu(m^*(\nu)) \mid m^*(\nu))$. Assume $\Lambda(\nu)$ is strictly monotonically decreasing near the optimal dual ν^*. Then, in the idealized two-timescale limit, the sequence (m_k, ν_k) converges almost surely to the constrained regularized mean-field equilibrium (m^*, ν^*).*

Role of entropy regularization. Entropy regularization serves two roles. Theoretically, it smooths the best-response map and makes the sensitivity bound in Lemma 5.1 meaningful. Algorithmically, it maintains exploration over browser options and prevents early collapse to a high-load action. The regularized equilibrium is therefore not identical to the unregularized constrained optimum: too little entropy can make fixed-point learning unstable, while too much entropy can bias the policy toward unnecessary randomization. In

experiments we use a small entropy temperature and report raw success and violation metrics rather than regularized returns.

7 Experiments

7.1 Environment and Capacity-Stress Task

All experiments use the open-source BrowserGym platform [Le Sellier De Chezelles et al. 2024] with miniwob.click-button in headless Chromium. The task is semantically simple by design: it provides a controlled browser workload for studying concurrent macro-actions and shared-resource contention, not open-ended web reasoning. To isolate the population externality, we add an *action-specific shared-capacity wrapper*. At each iteration, N agents run in parallel and choose among five macro-actions: `click_any_button`, `click_top2`, `click_top3`, `click_biggest`, and `noop`. Each macro-action is treated as a finite-capacity shared resource. If too many agents choose the same action in one batch, excess agents receive a capacity-induced violation and, in the main experiment, forced task failure; `noop` is effectively unconstrained. This creates a controlled "tragedy of the commons," where an action can be useful when chosen by few agents but harmful when overused. The wrapper isolates the shared-resource externality that arises when many browser agents contend for workers, target-site capacity, or API quotas. This setting separates scalar safety feedback from action-specific congestion feedback. A scalar safety price reacts to high global violations, but cannot identify which action bucket is overloaded. A mean-field congestion price broadcasts action-specific load information, allowing agents to shift away from overloaded resources without centralized assignment.

7.2 Learning Objective and Broadcast Signals

Agents target a population-level safety constraint $\mathbb{E}[V] \leq \rho_{\text{safe}}$ with $\rho_{\text{safe}} = 0.10$. For an agent i selecting option a_i, Dual-MFG optimizes the shaped return $G_i^{\text{mfg}} = R_i - \nu_k[V_i - \rho_{\text{safe}}]_+ - m_k(a_i)\phi_i(a_i)$, where ν_k is the scalar safety price, $m_k(a)$ is the broadcast action-specific congestion price, and $\phi_i(a_i)$ is the observed load feature for the selected option. The dual update uses an exponential moving average of the violation rate to stabilize finite-sample learning. The mean-field update maintains a smoothed per-action load estimate and broadcasts the resulting price vector in the next round.

7.3 Compared Methods

We compare four variants: **Independent PPO** uses no prices; **Safety-only** uses only the scalar dual price ν_k; **Congestion-only** uses only the action-specific price vector m_k; and **Dual-MFG** uses both.

This ablation tests whether action-specific aggregate feedback improves over scalar safety control. The variants share an *identical* capacity wrapper, isolating the marginal value of each broadcast signal; constrained MF-Q / MF-Actor-Critic variants with the same wrapper are the key external comparison for future work.

As an upper-reference, we evaluate a centralized capacity-aware oracle. At each iteration, the oracle observes population-level action scores and known capacities, then sends an explicit macro-action assignment to each agent. This allows it to eliminate capacity-induced violations in the wrapper, but its communication payload scales with N; it is therefore an upper-reference for explicit coordination, not a low-bandwidth deployment policy.

The underlying BrowserGym task is miniwob.click-button; the shared-capacity wrapper adds the population externality. We keep the population controlled so that component effects are interpretable; richer multi-step BrowserGym/WebArena tasks and mixed controlled/uncontrolled populations are natural extensions.

7.4 Metrics and Protocol

We report final-window averages over the last quarter of training, averaged over three random seeds. We use population sizes $N \in \{32, 64, 128\}$. The main outcome metrics are success rate, violation rate, capacity-induced violation rate, and $\text{TBI}_{\text{excess}} = \text{Success} - \max(0, \text{Violation} - \rho_{\text{safe}})$. The total violation rate is the logical OR of the original BrowserGym/environment violation and the capacity-induced violation. We separately log capacity violations to verify that the shared-resource wrapper is the mechanism driving the difference between variants. Communication is measured as serialized bytes per iteration. Independent PPO sends an empty payload, Safety-only sends ν_k, Congestion-only sends m_k, and Dual-MFG sends both. The centralized oracle sends an action assignment for each agent, so its communication grows with N. The capacity-stress benchmark isolates the population-level mechanism studied in the theory: many agents choosing among shared finite-capacity action channels. This lets us test broadcast safety and congestion prices without confounding the result with perception, planning, or page-state errors.

7.5 Main Results: Capacity-Stress and Centralized Reference

Table 1 reports final-window means over three seeds. Dual-MFG is the strongest decentralized method at every population size: it has the highest success rate, lowest total violation rate, and highest TBI among Independent PPO, Safety-only, and Congestion-only. The centralized capacity oracle obtains the best overall TBI by directly assigning actions to avoid overload, but its communication grows with the number of agents. The comparison quantifies the mean-field tradeoff. Relative to Independent PPO, Dual-MFG recovers 41.4%, 50.9%, and 48.1% of the centralized oracle's TBI improvement at $N = 32, 64, 128$, respectively. At the same time, it uses 8.4×, 15.9×, and 31.4× less communication than centralized assignment. Thus, Dual-MFG trades off optimal centralized assignment for a fixed-size broadcast that recovers a substantial fraction of the centralized benefit. These are *relative* gains: under this severe wrapper all decentralized methods (including Dual-MFG) retain negative absolute $\text{TBI}_{\text{excess}}$ (-0.033 vs. the oracle's $+0.115$ at N=128), so the recovered fraction is a communication–utility tradeoff, not net-positive utility.

Figure 2 shows the same pattern visually: Dual-MFG gives the best decentralized tradeoff, while the centralized oracle is the high-communication upper-reference. The gap between Congestion-only and Dual-MFG shows that reducing overload alone is not equivalent to optimizing the full success–violation objective.

Mechanism. The component ablation separates the roles of the two prices. Safety-only improves over Independent PPO because it reacts to high violation rates, but it cannot distinguish which

Table 1: Capacity-stress comparison with a centralized capacity-aware oracle. Values are final-window means over three seeds. Boldface marks the best decentralized method; the centralized oracle is an upper-reference that uses per-agent assignment communication. Communication is serialized broadcast or assignment bytes per iteration.

N	Method	Success ↑	Viol. ↓	Cap. viol. ↓	$\mathrm{TBI}_{\mathrm{excess}}$ ↑	Comm (B) ↓
32	Independent PPO	0.236	0.568	0.217	−0.232	2.0
32	Safety-only	0.249	0.526	0.177	−0.177	14.0
32	Congestion-only	0.272	0.492	**0.063**	−0.120	53.5
32	**Dual-MFG**	**0.286**	**0.490**	0.139	**−0.103**	66.5
32	Centralized oracle	0.332	0.353	0.000	0.079	559.5
64	Independent PPO	0.248	0.544	0.188	−0.196	2.0
64	Safety-only	0.268	0.509	0.163	−0.141	14.0
64	Congestion-only	0.294	0.467	**0.058**	−0.073	53.5
64	**Dual-MFG**	**0.298**	**0.462**	0.120	**−0.065**	66.5
64	Centralized oracle	0.328	0.366	0.000	0.062	1056.4
128	Independent PPO	0.260	0.530	0.184	−0.170	2.0
128	Safety-only	0.281	0.491	0.153	−0.111	14.0
128	Congestion-only	0.303	0.453	**0.050**	−0.050	53.4
128	**Dual-MFG**	**0.310**	**0.443**	0.101	**−0.033**	66.5
128	Centralized oracle	0.354	0.340	0.000	0.115	2087.0

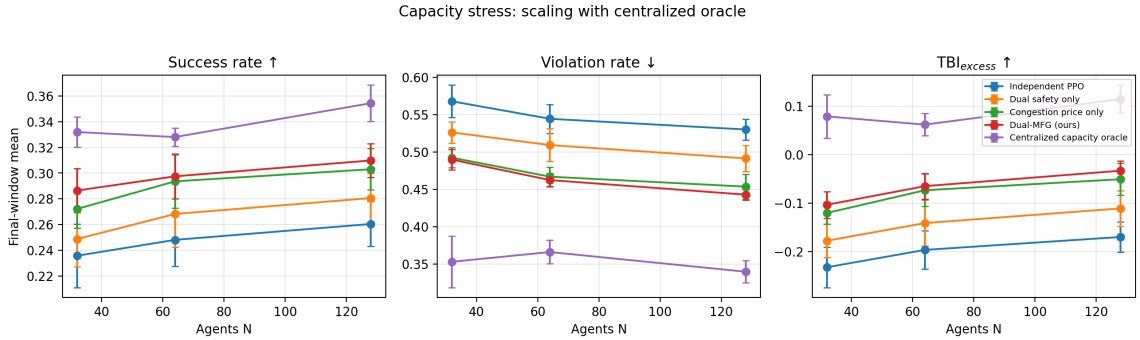

Figure 2: Capacity-stress performance across population size. Dual-MFG is the best decentralized method on success, total violation, and TBI; the centralized capacity oracle is stronger but requires per-agent assignment.

action bucket is overloaded. Congestion-only most strongly reduces capacity-induced violations, especially at $N = 128$ (0.050), but it lacks the safety dual and therefore does not optimize the full success–violation tradeoff. Dual-MFG balances both signals: it avoids high-capacity-cost actions while maintaining enough productive clicking to achieve the best decentralized TBI.

7.6 Communication–Performance Tradeoff

Figure 3 visualizes the systems tradeoff. The decentralized methods cluster at small, constant payload sizes: an empty header for Independent PPO, one scalar for Safety-only, one length-$|\mathcal{A}|$ vector for Congestion-only, and both signals for Dual-MFG. Dual-MFG is not the lowest-communication decentralized method; rather, it spends a small constant number of bytes to obtain the best decentralized safety-adjusted utility. The centralized oracle sends an action assignment for each agent, so its communication grows from 559.5 bytes at $N = 32$ to 2087.0 bytes at $N = 128$.

Runtime. Runtime is dominated by BrowserGym episode execution rather than by the price update or centralized assignment computation. At $N = 128$, final-window iteration times are 97.5 seconds for Independent PPO, 97.5 seconds for Safety-only, 94.8 seconds for Congestion-only, 96.0 seconds for Dual-MFG, and 98.6 seconds for the centralized oracle. Thus the practical distinction in this experiment is communication and coordination structure, not planner wall-clock overhead.

7.7 Discussion

Evidence for mean-field pricing. The ablation supports the central hypothesis: Dual-MFG helps when the population coupling is action-specific. A scalar safety controller improves violations but cannot route traffic across overloaded actions. The congestion price supplies that action-specific signal, while the dual safety price prevents congestion-only learning from optimizing load at the expense of task success. Thus Congestion-only attains the lowest capacity-induced violation, whereas Dual-MFG attains the best overall TBI.

Limits of the safety constraint. Dual-MFG reduces violations but does not satisfy the strict target $\rho_{\mathrm{safe}} = 0.10$ in the noisy BrowserGym setting; at $N = 128$, the final-window violation rate is 0.443 ± 0.007. We interpret the method as improving the safety–performance frontier, not as guaranteeing hard constraint satisfaction.

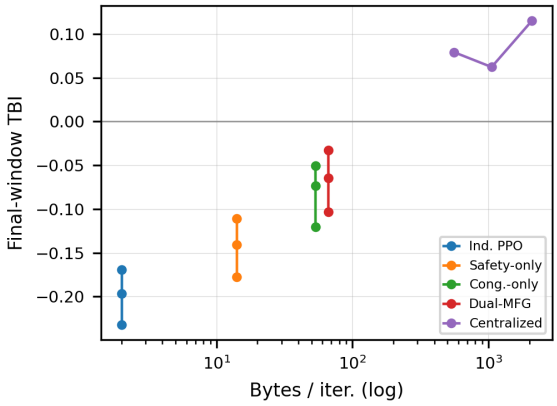

Figure 3: Communication–performance tradeoff under capacity stress. The x-axis uses a log scale. The centralized oracle obtains higher TBI but uses per-agent assignment communication that grows with N; Dual-MFG is the strongest decentralized method using a constant-size broadcast.

Efficiency at scale. The communication advantage is structural: centralized assignment sends per-agent actions and grows with N, whereas Dual-MFG broadcasts a scalar safety price and fixed-length action-price vector, remaining $O(|\mathcal{A}|)$ (about 66.5 bytes per iteration in our implementation).

What the experiment validates. The experiment tests whether action-specific prices improve coordination under action-specific overload; the theory justifies their form and constant $O(|\mathcal{A}|)$ communication scaling in an idealized stationary, regularized model.

Limitations and future work. The capacity wrapper is a controlled stress test, not a complete model of open-ended web-agent deployment. Natural extensions include richer BrowserGym/WebArena tasks, exogenous background agents, learned centralized planners, constrained MF-Q/MF-Actor-Critic baselines, LLM-based navigation, and stronger safety notions such as per-agent constraints, high-probability bounds, and robust shielding.

8 Conclusion

We studied communication-efficient safety coordination for controlled LLM/web-agent fleets under shared-resource contention. Dual-MFG broadcasts aggregate congestion and safety prices rather than per-agent assignments. Under idealized assumptions, the model admits equilibrium and finite-population approximation guarantees. In a BrowserGym capacity-stress experiment, Dual-MFG achieves the best decentralized success–violation tradeoff and recovers a substantial fraction of a centralized oracle's benefit using constant-size broadcast communication. The results support mean-field pricing as a practical coordination mechanism when the population externality is action-specific.

Reproducibility. We report final-window averages over three random seeds and release the scripts, result CSVs, and plotting code used to generate all tables and figures. Communication is measured as serialized JSON payload size per iteration: broadcast messages for decentralized methods and per-agent action assignments for the centralized oracle.

References

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. 2023. Q-Learning in Regularized Mean-field Games. *Dynamic Games and Applications* 13, 1 (2023), 89–117.

Vivek S. Borkar. 2008. *Stochastic Approximation: A Dynamical Systems Viewpoint.* Cambridge University Press, Cambridge, UK.

Pierre Cardaliaguet, Jean-Michel Lasry, Pierre-Louis Lions, and Alessio Porretta. 2012. Long Time Average of Mean Field Games. *Networks and Heterogeneous Media* 7, 2 (2012), 279–301. doi:10.3934/nhm.2012.7.279

Kai Cui and Heinz Koeppl. 2021. Approximately Solving Mean Field Games via Entropy-Regularized Deep Reinforcement Learning. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 130).* PMLR, 1909–1917. https://proceedings.mlr.press/v130/cui21a.html

Stella Dafermos. 1980. Traffic equilibrium and variational inequalities. *Transportation Science* 14, 1 (1980), 42–54.

Stella Dafermos. 1988. Sensitivity analysis in variational inequalities. *Mathematics of Operations Research* 13, 3 (1988), 421–434.

Anthony V. Fiacco. 1983. *Introduction to Sensitivity and Stability Analysis in Nonlinear Programming.* Mathematics in Science and Engineering, Vol. 165. Academic Press, New York.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual Multi-Agent Policy Gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence.* https://arxiv.org/abs/1705.08926

Diogo A. Gomes and Vardan K. Voskanyan. 2016. Extended Deterministic Mean-Field Games. *SIAM Journal on Control and Optimization* 54, 2 (2016), 1030–1055. doi:10.1137/130944503

Xin Guo, Anran Hu, Jiacheng Zhang, and Yufei Zhang. 2025. Continuous-Time Mean Field Games: A Primal-Dual Characterization. arXiv:2503.01042 [math.OC] https://arxiv.org/abs/2503.01042

Xin Guo, Renyuan Hu, Jiacheng Xu, and Junzi Zhang. 2019. Learning mean-field games. In *Advances in Neural Information Processing Systems (NeurIPS),* Vol. 32.

Minyi Huang, Peter E Caines, and Roland P Malhamé. 2006. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems* 6, 3 (2006), 221–252.

Matej Jusup, Barna Pásztor, Tadeusz Janik, Kenan Zhang, Francesco Corman, Andreas Krause, and Ilija Bogunovic. 2024. Safe Model-Based Multi-Agent Mean-Field Reinforcement Learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems.* 973–982. https://www.ifaamas.org/Proceedings/aamas2024/pdfs/p973.pdf

Jean-Michel Lasry and Pierre-Louis Lions. 2007. Mean field games. *Japanese Journal of Mathematics* 2, 1 (2007), 229–260.

Thibault Le Sellier De Chezelles, Maxime Gasse, Alexandre Drouin, Massimo Caccia, Léo Boisvert, Megh Thakkar, Tom Marty, Rim Assouel, Sahar Omidi Shayegan, Lawrence Keunho Jang, Xing Han Lù, Ori Yoran, Dehan Kong, Frank F. Xu, Siva Reddy, Quentin Cappart, Graham Neubig, Ruslan Salakhutdinov, Nicolas Chapados, and Alexandre Lacoste. 2024. The BrowserGym Ecosystem for Web Agent Research. arXiv:2412.05467 [cs.AI]

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. 2017. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In *Advances in Neural Information Processing Systems (NeurIPS).* https://arxiv.org/abs/1706.02275

Washim Uddin Mondal, Vaneet Aggarwal, and Satish V. Ukkusuri. 2024. Mean-Field Approximation of Cooperative Constrained Multi-Agent Reinforcement Learning. *Journal of Machine Learning Research* 25, 280 (2024), 1–42. https://jmlr.org/papers/v25/22-0956.html

Santiago Paternain, Luiz F. O. Chamon, Miguel Calvo-Fullana, and Alejandro Ribeiro. 2019. *Constrained reinforcement learning has zero duality gap.* Curran Associates Inc., Red Hook, NY, USA.

Alessio Porretta. 2018. On the Turnpike Property for Mean Field Games. *Minimax Theory and its Applications* 3, 2 (2018), 285–312.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. 2018. Mean Field Multi-Agent Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML).* PMLR, 5571–5580. https://proceedings.mlr.press/v80/yang18d.html

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2023. WebArena: A Realistic Web Environment for Building Autonomous Agents. arXiv:2307.13854 [cs.AI] https://arxiv.org/abs/2307.13854

A Proofs

Proof of Lemma 5.1. The result follows from the Implicit Function Theorem applied to the stationarity condition $\nabla_\pi f(\pi, m) = 0$. Since f is $\lambda\sigma$-strongly concave (due to the regularizer $\lambda\mathcal{H}$), the inverse Hessian of the objective is bounded by $(\lambda\sigma)^{-1}$. Consequently, the sensitivity of the optimizer is bounded by the ratio of the gradient Lipschitz constant to the curvature: $\frac{L_{\pi m}}{\lambda\sigma}$. This corresponds to the condition number of the regularized objective, as established in classical sensitivity analysis for variational inequalities [Dafermos 1988]. □

Proof of Proposition 5.2. The proof relies on Schauder's Fixed Point Theorem applied to the aggregate fixed-point operator $T(m) = \Psi(m; \mathrm{BR}(m))$. First, observe that the aggregate space $\mathcal{M} = \Delta(|W|) \times [0, L]^{|W|} \times [0, 1]^{|W|}$ is compact and convex, being a finite product of compact convex sets. Second, we establish the continuity of the map T. By Assumption **(A2)**, the environment dynamics and rewards are Lipschitz continuous in m. Crucially, due to the entropy regularization utilized in **Lemma 5.1**, the best-response map $m \mapsto \mathrm{BR}(m)$ is single-valued and Lipschitz continuous. (Note: Assumption **(A3)** ensures the policy space is compact, making the underlying optimization well-posed, but the uniqueness and continuity strictly follow from the strong concavity of the regularizer). Since the aggregate operator Ψ is continuous with respect to the policy and the underlying kernel, the composition T is a continuous map from $\mathcal{M}$ into itself. Thus, by Schauder's Fixed Point Theorem, there exists a fixed point $m^* \in \mathcal{M}$ such that $T(m^*) = m^*$, which alongside $\pi^* = \mathrm{BR}(m^*)$ constitutes a Mean-Field Equilibrium. □

Proof of Proposition 5.3. While the proof follows the standard Banach contraction framework, it requires explicit derivation to account for the semi-Markov dynamics. Specifically, we must establish how the random holding times affect the Lipschitz properties of the forward operator Ψ via the discounted occupancy measure. We establish that the aggregate update operator $T(m) = \Psi(m; \mathrm{BR}(m))$ is a contraction on $\mathcal{M}$ under sufficient regularization. Let $m, m' \in \mathcal{M}$. By the triangle inequality:

$$\begin{aligned}\|T(m) - T(m')\| &= \|\Psi(m; \mathrm{BR}(m)) - \Psi(m'; \mathrm{BR}(m'))\| \\ &\le \underbrace{\|\Psi(m; \mathrm{BR}(m)) - \Psi(m'; \mathrm{BR}(m))\|}_{\text{Term A}} \\ &\quad + \underbrace{\|\Psi(m'; \mathrm{BR}(m)) - \Psi(m'; \mathrm{BR}(m'))\|}_{\text{Term B}}.\end{aligned}$$

Bounding Term A (Dynamic Coupling): We derive the bound directly using the Wasserstein contraction properties implied by **(A2)**. Let ν_m and $\nu_{m'}$ be the stationary distributions (occupancies) under aggregates m and m', i.e., $\nu_m P_m = \nu_m$. Using the triangle inequality for W_1:

$$\begin{aligned}\|\nu_m - \nu_{m'}\|_{W_1} &= \|\nu_m P_m - \nu_{m'} P_{m'}\|_{W_1} \\ &\le \|\nu_m P_m - \nu_{m'} P_m\|_{W_1} + \|\nu_{m'} P_m - \nu_{m'} P_{m'}\|_{W_1}.\end{aligned}$$

By Assumption **(A1)** (Discounted SMDP Contraction), the kernel P_m acts as a $\bar{\gamma}$-contraction on the space of measures: $\|\mu P_m - \mu' P_m\|_{W_1} \le \bar{\gamma}\|\mu - \mu'\|_{W_1}$. By Assumption **(A2)**, the kernel is Lipschitz in m with constant L_P: $\|\nu_{m'}(P_m - P_{m'})\|_{W_1} \le L_P\|m - m'\|$. Substituting these back:

$$\|\nu_m - \nu_{m'}\|_{W_1} \le \bar{\gamma}\|\nu_m - \nu_{m'}\|_{W_1} + L_P\|m - m'\|.$$

Rearranging terms to solve for $\|\nu_m - \nu_{m'}\|_{W_1}$:

$$\|\nu_m - \nu_{m'}\|_{W_1} \le \frac{L_P}{1 - \bar{\gamma}}\|m - m'\|.$$

Thus, we set $K_{\mathrm{dyn}} \triangleq \frac{L_P}{1-\bar{\gamma}}$.

Bounding Term B (Policy Sensitivity): This term captures the shift in occupancy due to the policy change. Let L_Ψ^π be the Lipschitz constant of the forward operator w.r.t. the policy. From Lemma 5.1, the Best Response map is Lipschitz with constant $L_{\mathrm{BR}} = \frac{L_{\pi m}}{\lambda\sigma}$. Therefore:

$$\|\text{Term B}\| \le L_\Psi^\pi \|\mathrm{BR}(m) - \mathrm{BR}(m')\| \le \frac{L_\Psi^\pi L_{\pi m}}{\lambda\sigma}\|m - m'\|.$$

Combining these, the total Lipschitz constant is $L_T = K_{\mathrm{dyn}} + \frac{L_\Psi^\pi L_{\pi m}}{\lambda\sigma}$. To guarantee a contraction ($L_T < 1$), we require:

$$\lambda > \frac{L_\Psi^\pi L_{\pi m}}{\sigma(1 - K_{\mathrm{dyn}})}.$$

Provided that the baseline dynamic coupling is stable ($K_{\mathrm{dyn}} < 1$), sufficient regularization λ guarantees uniqueness and convergence. □

Proof of Proposition 5.4. We analyze the exploitability gap by bounding the difference between the finite-population return and the mean-field limit. Let $J(\pi_i, m)$ be the expected return for agent i.

Step 1: Concentration of the Empirical Aggregate. The finite population interacts through the aggregate vector m^N. While m contains nonlinear ratio statistics (specifically latency $\bar{\ell}_w = \frac{\sum \tau}{\sum \mathrm{occ} + \varepsilon}$), it can be expressed as a smooth function of linear sample moments. Let $\mathbf{M}^N = \frac{1}{N}\sum_{j=1}^N \xi(s_j)$ be the vector of sufficient statistics, where the mapping $\xi : \mathcal{S} \to \mathbb{R}^k$ extracts the relevant moments (occupancy, total latency mass, violation counts). By the Central Limit Theorem, these linear averages concentrate around their mean-field limits $\mathbf{M}^*$ with rate $O(N^{-1/2})$:

$$\mathbb{E}\|\mathbf{M}^N - \mathbf{M}^*\| \le \frac{C}{\sqrt{N}}.$$

The final aggregate is given by $m^N = G(\mathbf{M}^N)$. Due to the regularization term $\varepsilon > 0$ in the denominator of the latency statistic, the gradient of the projection map G is uniformly bounded. Thus, G is Lipschitz continuous, and the concentration rate is preserved:

$$\mathbb{E}\|m^N - m^*\| \le L_G \mathbb{E}\|\mathbf{M}^N - \mathbf{M}^*\| \le O(N^{-1/2}).$$

Step 2: Decomposition of Exploitability. Consider a unilateral deviation by a single agent i to an arbitrary policy π', while the remaining $N - 1$ agents continue to play the equilibrium policy π^*. Let m^N denote the empirical aggregate when all N agents play π^*, and let $m^{N'}$ denote the empirical aggregate when agent i plays π' (while others play π^*). We define the finite-population exploitability gap Δ_N as the maximum expected gain an agent can achieve by deviating:

$$\Delta_N \triangleq \mathbb{E}\left[J(\pi', m^{N'})\right] - \mathbb{E}\left[J(\pi^*, m^N)\right].$$

To analyze this gap, we add and subtract the mean-field limit terms $J(\pi', m^*)$ and $J(\pi^*, m^*)$ to perform the following decomposition:

$$\begin{aligned}\Delta_N &= \mathbb{E}[J(\pi', m^{N'})] - \mathbb{E}[J(\pi^*, m^N)] \\ &= \underbrace{\mathbb{E}[J(\pi', m^{N'}) - J(\pi', m^*)]}_{\text{Term A}} + \underbrace{[J(\pi', m^*) - J(\pi^*, m^*)]}_{\text{Term B}} \\ &+ \underbrace{\mathbb{E}[J(\pi^*, m^*) - J(\pi^*, m^N)]}_{\text{Term C}}.\end{aligned}$$

Since (π^*, m^*) is a Mean-Field Equilibrium, π^* is a best response to m^*, implying Term B ≤ 0. Thus $\Delta_N \leq$ Term A + Term C. We now bound the remaining terms using the Lipschitz continuity of the value function $J(\pi|m)$. By Assumptions **(A1)** and **(A2)**, J is Lipschitz with respect to m with constant L_J. Note that Term C captures the error due to finite-population noise when all agents play the equilibrium policy. Using the Lipschitz property and the concentration result from Step 1:

$$\text{Term C} = \mathbb{E}[J(\pi^*, m^*) - J(\pi^*, m^N)] \leq L_J \mathbb{E}\|m^* - m^N\| \leq L_J \frac{C}{\sqrt{N}}.$$

Similarly, Term A captures the effect of the single agent's deviation on the aggregate. First, we apply the Lipschitz condition:

$$\text{Term A} = \mathbb{E}[J(\pi', m^{N'}) - J(\pi', m^*)] \leq L_J \mathbb{E}\|m^{N'} - m^*\|.$$

By the Triangle Inequality, $\|m^{N'} - m^*\| \leq \|m^{N'} - m^N\| + \|m^N - m^*\|$. The first part, $\|m^{N'} - m^N\|$, represents the shift in the aggregate caused by one agent changing their action. Since the aggregate is an average over N agents, this shift is bounded by $\frac{K}{N}$ (where K depends on the bounded diameter of the state space). Substituting this back:

$$\text{Term A} \leq L_J \left(\frac{K}{N} + \mathbb{E}\|m^N - m^*\| \right) \leq L_J \left(\frac{K}{N} + \frac{C}{\sqrt{N}} \right).$$

Summing the bounds for Term A and Term C:

$$\Delta_N \leq L_J \left(\frac{K}{N} + \frac{2C}{\sqrt{N}} \right).$$

Since $\frac{1}{N}$ decays faster than $\frac{1}{\sqrt{N}}$, the dominant term is of order $O(N^{-1/2})$. Thus, $\Delta_N \leq O(N^{-1/2})$.

Step 3: Constraint Violation. Finally, the constraint function $g(\pi|m)$ inherits the same Lipschitz properties. Thus:

$$\mathbb{E}[g(\pi^*|m^N)] \leq g(\pi^*|m^*) + L_g \mathbb{E}\|m^N - m^*\| \leq \rho_{\text{safe}} + O(N^{-1/2}).$$

We conclude that (π^*, m^*) is an $O(N^{-1/2})$-Nash equilibrium. □

Proof of Theorem 6.1. The proof relies on the timescale separation $\eta_k/\beta_k \to 0$ and the framework of Borkar [2008]. We decompose the dynamics into fast and slow processes.

1. Fast Timescale (m_k): Treating the slow variable $\nu_k \approx \nu$ as quasi-static, the recursion for m_k tracks the ODE:

$$\dot{m}(t) = \Psi(m(t); \text{BR}_\nu(m(t))) - m(t).$$

From Proposition 5.3, we established that under the weak coupling condition ($K_{\text{dyn}} < 1$) and sufficient regularization λ, the operator $T_\nu(m) = \Psi(m; \text{BR}_\nu(m))$ is a strict contraction. Thus, the fast process m_k converges almost surely to the unique equilibrium $m^*(\nu_k)$.

2. Slow Timescale (ν_k): The slow variable ν_k tracks the projected ODE driven by the equilibrated violation rate:

$$\dot{\nu}(t) = \Pi_{\geq 0}\left[\Lambda(\nu(t)) - \rho_{\text{safe}}\right].$$

Since agents rationally reduce violations as penalties increase, $\Lambda(\nu)$ is monotonically decreasing. To prove convergence to ν^*, consider the Lyapunov function $V(\nu) = \frac{1}{2}(\nu - \nu^*)^2$. Differentiating with respect to time gives $\dot{V}(\nu) = (\nu - \nu^*)\dot{\nu}$. We consider two cases:

(1) **Case 1 (Active Constraint):** If $\Lambda(0) > \rho_{\text{safe}}$, then $\nu^* > 0$ and $\Lambda(\nu^*) = \rho_{\text{safe}}$. Since Λ is strictly decreasing, if $\nu > \nu^*$, then $\Lambda(\nu) < \rho_{\text{safe}}$ and $\dot{\nu} < 0$. If $\nu < \nu^*$, then $\Lambda(\nu) > \rho_{\text{safe}}$ and $\dot{\nu} > 0$. In both sub-cases, $\dot{V} < 0$.

(2) **Case 2 (Inactive Constraint):** If $\Lambda(0) \leq \rho_{\text{safe}}$, then $\nu^* = 0$. Here, for any $\nu > 0$, we have $\Lambda(\nu) \leq \Lambda(0) \leq \rho_{\text{safe}}$, so $\dot{\nu} < 0$. Thus ν decreases to 0, ensuring $\dot{V} < 0$ for all $\nu \neq \nu^*$.

In all cases, (m_k, ν_k) converges almost surely to (m^*, ν^*). □