Graphon Mean-Field Subsampling for Cooperative Heterogeneous Multi-Agent Reinforcement Learning

Anonymous Authors

ABSTRACT

Coordinating large populations of interacting agents is a central challenge in multi-agent reinforcement learning (MARL), where the size of the joint state-action space scales exponentially with the number of agents. Mean-field methods alleviate this burden by aggregating agent interactions, but these approaches assume homogeneous interactions. Recent graphon-based frameworks capture heterogeneity, but are computationally expensive as the number of agents grows. Therefore, we introduce GMFS, a **G**raphon **M**ean-**F**ield **S**ubsampling framework for scalable cooperative MARL with heterogeneous agent interactions. By subsampling κ agents according to interaction strength, we approximate the graphon-weighted mean-field and learn a policy with sample complexity $\text{poly}(\kappa)$ and optimality gap $\widetilde{O}(1/\sqrt{\kappa})$. We verify our theory with numerical simulations in robotic coordination, showing that GMFS achieves near-optimal performance.

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent systems**.

KEYWORDS

Multi-agent Reinforcement Learning, Graphons, Non-uniform interactions, Sampling, Mean-Field Analyses, Large-Scale Systems

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) has emerged as a powerful framework for modeling complex, large-scale networked systems whose dynamics stem from the interactions between many individual agents. These systems are now ubiquitous, appearing in domains as varied as robotic swarms (54, 59), autonomous driving (21, 39, 40), ride-sharing (46), real-time bidding (36), stochastic games (13, 35), and efficient wireless networks (18, 78). In these settings, the objective is to derive optimal policies that maximize collective reward across the entire system. Drawing on the success of reinforcement learning (RL) in tasks like the game of Go (66) and robotic control (42), MARL extends these ideas to environments with multiple interacting agents.

However, quickly solving such systems exactly becomes infeasible as the number of agents grows, creating a barrier to large-scale deployment. The primary challenge in scaling MARL lies in the size of the joint state-action space $(|\mathcal{S}||\mathcal{A}|)^n$, which grows exponentially with the population size n and renders exhaustive sampling over all agents computationally intractable (6). To mitigate this "curse of dimensionality", previous works proposed using mean-field theory to approximate agent populations by replacing explicit interactions with population aggregates (10, 19, 29, 30, 47, 75). However, these frameworks typically require iterating over the entire population at each time step to compute aggregates. More recently, Anand et al. (4) mitigate this cost by sampling a subset of k agents to achieve polynomial complexity. However, these standard mean-field methods often fail at capturing the heterogeneous behaviors inherent in complex systems like traffic management (80). As noted by Carmona et al. (10), ignoring agent diversity can lead to systemic failures, including communication delays (61, 67), inefficient information dissemination (57), and unstable robotic control (76).

To model these non-uniformities, Caines and Huang (9), Gao and Caines (27), Lovász (53) represent network systems via graphons, where a limit function $W(x, y)$ defines interaction strengths between agents on a latent index $x \in [0, 1]$. Although Hu et al. (33) established an approximation error bound of $O(1/\sqrt{n})$ for such systems, no existing framework unifies the computational efficiency of neighborhood subsampling with the expressiveness of graphon-level heterogeneity. Therefore, we ask:

Can a mean-field MARL algorithm handle heterogeneity through non-uniform agent interactions, while reducing computation to a polylogarithmic dependence on the number of agents?

1.1 Contributions

Our key contributions are outlined below.

- GMFS **Algorithm (§3).** We introduce a subsampling framework for cooperative MARL with heterogeneous agents, where graphon mean-field interactions are approximated using $\kappa \boldsymbol{In}$ sampled neighbors.
- **Theoretical Guarantees (§4, §5).** We bound the optimality gap between the κ-sampled policy and the optimal graphon mean-field policy. For finite state-action spaces, we show that GMFS reduces the sample complexity from exponential in n, $(|\mathcal{S}|^n|\mathcal{A}|^n)$, to polynomial in κ, $(\kappa^{|\mathcal{S}||\mathcal{A}|})$.
- **Numerical Simulations (§A).** We observe monotonic improvements in average discounted return as κ increases, approaching the exhaustive mean-field limit in a heterogeneous robot-warehouse environment.

A crucial challenge in heterogeneous systems is the position-dependent variance that makes uniform sampling ineffective. We show how to resolve this issue via graphon-weighted subsampling, where neighbors are sampled directly according to normalized graphon weights. By showing that the resulting sampled Bellman operator remains a γ-contraction, we prove that the algorithm converges to an approximately optimal policy. We further show that the optimality gap concentrates at an $O(1/\sqrt{\kappa})$ rate (Theorem C.1), unifying uniform mean-field and exhaustive population methods.

Proc. of the Adaptive and Learning Agents Workshop (ALA 2026), Aydeniz, Delgrange, Mohammedalamen, Yang (eds.), May 25 – 26, 2026, Paphos, Cyprus, https://alaworkshop2026.github.io/. 2026.

1.2 Related Literature

MARL has a rich history, beginning with early work on Markov games for multi-agent decision-making (52, 68), which can be viewed as multi-agent extensions of Markov Decision Processes (MDPs). Since then, MARL has been studied across a wide range of settings (79). It is closely related to "succinctly described" MDPs (6), where the joint state-action space is a product of individual agent spaces and agents optimize a collective objective. A promising recent line of work constrains such problems to sparse networked instances to enforce local interactions (51, 55, 60).In this formulation, agents correspond to graph vertices and interact only with immediate neighbors. By exploiting correlation decay (26), these methods mitigate the curse of dimensionality by searching over policies defined on truncated graphs. Yet, as networks become dense, agent neighborhoods grow, rendering graph-truncation approaches computationally intractable.

Mean-Field RL. To address large agent neighborhoods, mean-field theory replaces a finite set of local agents with an empirical distribution over agent states (43, 74, 75). This simplifies multi-agent interactions into a two-agent formulation, where each agent interacts with a representative "mean agent" that evolves according to the empirical distribution of all other agents (29). For a detailed overview of learning methodologies in mean-field games, we refer the reader to Laurière et al. (45). Pásztor et al. (58) explore efficient model-based approaches to further improve sample efficiency in these settings. By sampling a subset of the total population, Anand et al. (4) build on this mean-field abstraction to achieve a sample complexity that is polynomial in the number of subsampled agents. However, these approaches operate under the assumption that interactions between agents are uniform. Our GMFS algorithm utilizes graphon functions to model heterogeneous interactions in dense graphs, better representing realistic networked systems (23) while maintaining provable performance guarantees.

Graphon Mean-Field MARL. Building on classical mean-field MARL, a growing body of recent work relaxes homogeneity assumptions by modeling dense, heterogeneous interactions in large populations. Caines and Huang (8) introduced Graphon Mean-Field Games (GMFG), formalizing graphons as limiting objects that encode heterogeneous structures in infinite networks (7). Subsequent work expands GMFG concepts to broader classes of control and learning problems, including dynamical sequential games (20, 23), learning algorithms for realistic sparse graphs (24), and graphon estimation from sampled agents (22, 77). Moreover, even when the graphon is unknown, it can be efficiently estimated (1, 11). In the cooperative MARL setting, Hu et al. (33) adapts the GMFG framework to Graphon Mean-Field Control, where agents optimize a joint Q-function driven by graphon-weighted aggregates. While these works demonstrate that graphons provide a principled mechanism for capturing heterogeneity, existing methods typically require full population aggregation at each time step. In contrast, GMFS approximates these aggregates using only $\kappa < n$ sampled neighbors.

Other Related Work. Beyond the areas highlighted above, our work contributes to the literature on Centralized Training with Decentralized Execution (81), as we learn a provably near-optimal policy using centralized information while executing decisions based only on local observations. In distributed settings, V-learning (35) reduces the exponential dependence on the joint action space to an additive one. In contrast, our approach further reduces the complexity of the joint state space, which has not been previously achieved. Finally, while linear function approximation can be used to reduce Q-table complexity (35), bounding the resulting performance loss is generally intractable without strong assumptions such as Linear Bellman completeness (28) or low Bellman-Eluder dimension (34). While our work primarily considers the finite tabular setting, we also provide extensions to the non-tabular case under Linear Bellman completeness.

2 PRELIMINARIES

We formally introduce the problem, state motivating examples for our setting, and provide technical details about the mean-field model and graphon-based techniques used.

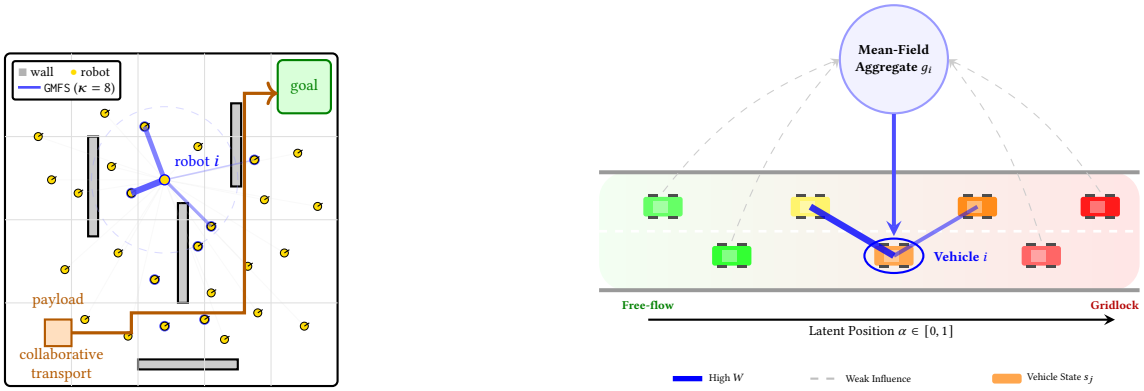
Notation. For $n, \kappa \in \mathbb{N}$ with $\kappa \le n$, let $\binom{[n]}{\kappa}$ denote the set of all κ-sized subsets of $[n] = \{1, \dots, n\}$. For any vector $z \in \mathbb{R}^d$, let $\|z\|_1$ and $\|z\|_\infty$ denote the standard ℓ_1 and ℓ_∞ norms, respectively. Given a collection of variables $s_1, \dots, s_n$ and $\Delta \subseteq [n]$, the shorthand s_Δ denotes the set $\{s_i : i \in \Delta\}$. We use $\widetilde{O}(\cdot)$ notation to suppress polylogarithmic factors in all problem parameters except n. For a discrete measurable space $(\mathcal{X}, \mathcal{F})$, the total variation distance between probability measures ρ_1 and ρ_2 is defined as $\mathrm{TV}(\rho_1, \rho_2) = \frac{1}{2} \sum_{x \in \mathcal{X}} |\rho_1(x) - \rho_2(x)|$. We write $x \sim \mathcal{D}$ to indicate that x is sampled from the distribution $\mathcal{D}$, and $x \sim \mathcal{U}(\Omega)$ to denote sampling from the uniform distribution over a finite set Ω.

Graphons. A graphon is a bounded, measurable, symmetric function $W : \mathcal{I}^2 \to \mathcal{I}$, where $\mathcal{I} = [0, 1]$, that encodes interaction weights in dense graphs. Given deterministic latent points $\{\alpha_i\}_{i=1}^n \subset [0, 1]$, the graphon induces a complete weighted interaction matrix with entries $w_{ij} \coloneqq W(\alpha_i, \alpha_j)$ (and $w_{ii} \coloneqq 0$). Graphons arise as limit objects for dense graphs, and any finite weighted graph G admits an associated underlying graphon W_G constructed from its adjacency matrix (53).

2.1 Problem Formulation

We consider a system of n cooperative agents indexed by $[n] = \{1, \dots, n\}$. Let $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces of each agent, respectively (Assumption 3.2). At each time t, the agents have a joint state $s(t) = (s_1(t), \dots, s_n(t)) \in \mathcal{S}^n$ and select a joint action $a(t) = (a_1(t), \dots, a_n(t)) \in \mathcal{A}^n$. The interactions between the agents can be written as a weighted graph $\mathbb{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V} = [n]$, where edge weights are determined by a measurable, bounded, symmetric graphon $W : \mathcal{I}^2 \to \mathcal{I}$ for $\mathcal{I} = [0, 1]$. Each agent i is assigned a latent coordinate $\alpha_i = i/n \in \mathcal{I}$, enabling the graphon to represent non-uniform interaction structures within the network.

Definition 2.1 (Graphon-weighted neighborhood state-action feature for agent i). *Let $W : [0, 1]^2 \to [0, 1]$ be a graphon. We fix deterministic latent coordinates $\{\alpha_i\}_{i=1}^n \subset [0, 1]$. The graphon induces a complete weighted interaction graph $w_{ij} \coloneqq W(\alpha_i, \alpha_j)$ where $w_{ii} \coloneqq 0$, and normalized influence weights for each i, $\bar{w}_{ij} \coloneqq \frac{w_{ij}}{\sum_{m \neq i} w_{im}}$, where $\bar{w}_{ii} \coloneqq 0$. Then, for a joint state/action pair $(\mathbf{s}, \mathbf{a}) \in \mathcal{S}^n \times \mathcal{A}^n$, agent i's*



(a) Warehouse robots **(b) Traffic coordination**

Figure 1: Graphon mean-field systems with distance-decay interactions. (a) Warehouse robots collaborate to transport a payload, where robot i uses $\kappa = 8$ subsampled neighbors with interaction strength $W(x_i, x_j)$ indicated by line thickness. (b) Traffic vehicles coordinate using graphon-weighted aggregates g_i of the population, where interaction strength decays with distance in latent position space $\alpha \in [0, 1]$.

graphon-weighted neighborhood state/action feature is a probability mass function $z_i \in \mathcal{Z} \subset \mathcal{P}(\mathcal{S} \times \mathcal{A})$, defined for all $(x, u) \in \mathcal{S} \times \mathcal{A}$,

$$z_i(x, u) \coloneqq \sum_{j \neq i} \bar{w}_{ij} \mathbb{1}\{\mathbf{s}_j = x, \mathbf{a}_j = u\},$$

where its state marginal is $g_{z_i}(x) = \sum_{u \in \mathcal{A}} z_i(x, u) \in \mathcal{G}$ and action marginal is $h_{z_i}(u) = \sum_{x \in \mathcal{S}} z_i(x, u) \in \mathcal{H}$.

Agent dynamics. Let $\mathfrak{G}(\mathcal{S})$ denote the space of neighborhood state features, the probability simplex over $\mathcal{S}$. At time t, agent i observes its local state $s_i(t)$ and its neighborhood feature. It selects an action $a_i(t) \in \mathcal{A}$, and the next state evolves according to the transition kernel $P : \mathcal{S} \times \mathcal{A} \times \mathfrak{G}(\mathcal{S}) \to \mathcal{P}(\mathcal{S})$, such that $s_i(t+1) \sim P(\cdot \mid s_i(t), a_i(t), g_i(t))$ for all $i \in [n]$, where $g_i(t) \in \mathfrak{G}(\mathcal{S})$ is agent i's graphon-weighted neighborhood state feature, i.e., the agent transitions depend only on empirical neighborhood features $g \in \mathfrak{G}(\mathcal{S})$.

Team objective. We define a local reward function $r_\ell : \mathcal{S} \times \mathcal{A} \times \mathfrak{G}(\mathcal{S}) \to \mathbb{R}$ for each agent, and the team stage reward defined on $(\mathbf{s}_{1:n}(t), \mathbf{a}_{1:n}(t), \mathbf{g}_{1:n}(t)) \in \mathcal{S}^n \times \mathcal{A}^n \times \mathcal{G}^n$ as r_t where

$$r_t \coloneqq r_t(\mathbf{s}(t), \mathbf{a}(t), \mathbf{g}(t)) \coloneqq \frac{1}{n} \sum_{i=1}^{n} r_\ell\big(\mathbf{s}_i(t), \mathbf{a}_i(t), \mathbf{g}_i(t)\big).$$

Fix a discounting factor $\gamma \in (0, 1)$. Then, the discounted team return with joint policy $\pi = (\pi_1, \ldots \pi_n)$ and initial state $\mathbf{s}_{1:n} \in \mathcal{S}^n$ is then given by

$$V^{\pi}_{\text{team}}(\mathbf{s}_{1:n}) \coloneqq \mathbb{E}_\pi \left[\sum_{t \geq 0} \gamma^t r_t \middle| \mathbf{s}_{1:n}(0) = \mathbf{s}_{1:n} \right].$$

Note that each agent i's state transition and reward depends on agents $[n] \setminus i$ only through the state marginal $g_i(x)$, but the evolution of $g_i(x)$ depends on how they choose actions based on their states, which is captured in $Z_i(x, u)$.

Definition 2.2 (ϵ-optimal policy). *A policy π is ϵ-optimal if for all $\mathbf{s}_{1:n} \in \mathcal{S}^n$, we have $V^{\pi}_{team}(\mathbf{s}) \geq \sup_{\pi^*} V^{\pi^*}_{team}(\mathbf{s}) - \epsilon$.*

Motivating Examples. Below we give examples of cooperative graphon MARL settings naturally captured by this formulation. Our empirical results show that the average cumulative discounted return of learned policies increases monotonically as the subsampling parameter κ approaches n, while providing a computational speedup over graphon mean-field Q-learning methods.[1]

- **Robot Coordination in Constrained Workspaces.** Consider a swarm of n mobile robots performing collaborative tasks (e.g., cargo transport in a warehouse). Since robots operate in constrained regions (aisles and loading zones), local congestion can significantly affect task completion times. If robot i is assigned a location $\alpha_i \in [0, 1]^2$, we can model their interactions with a graphon $W(\alpha_i, \alpha_j) = \mathbb{1}\{\|\alpha_i - \alpha_j\|_2 \leq r\}$ and a fixed interaction radius $r > 0$, reflecting concerns like collision avoidance, and its empirical neighborhood state-action distribution $\widehat{z}_i^{(\kappa)}$ and its state marginal $\widehat{g}_i^{(\kappa)}$ are used to compute its reward. For tractability, each agent constructs a κ-sampled approximation of its neighborhood aggregate via graphon-weighted subsampling, so that a robot's optimal policy is primarily influenced by nearby agents (e.g., those sharing an aisle) rather than further ones. These heterogeneous densities can be represented with a block-structured graphon where different subsections of the unit square correspond to workspace sectors with varying congestion levels (Figure 1a).
- **Autonomous Vehicle Coordination.** Consider n autonomous vehicles navigating a road network where efficient coordination is needed to minimize travel times and prevent gridlock. In classical mean-field settings, all agents are assumed to interact uniformly, but real-world traffic dynamics exhibit significant locality as a vehicle's congestion experience typically depends on the density of nearby traffic rather than the state of the

[1] We provide more details of our numerical experiments in section A and have released the code here.

entire network. This spatial structure is captured by a distance-decay graphon $W(\alpha_i, \alpha_j) = \exp(-\beta|\alpha_i - \alpha_j|)$, where vehicles that are close to one another exert a strong mutual influence while those further away have a negligible impact on local flow. To enable scalable learning in this environment, each vehicle could use graphon-weighted subsampling to approximate the local traffic density without having to process the full network information. This allows decentralized policies that respect the underlying locality of traffic dynamics, thereby allowing vehicles to make informed decisions to improve overall network throughput (Figure 1b).

Capturing heterogeneity in state/action spaces. Following Mondal et al. (55), we model heterogeneity in agent states and actions by incorporating agent types directly into the state. Specifically, each agent i is assigned a type $\varepsilon_i \in \mathcal{E}$, and we define the state space $\mathcal{S} = \mathcal{E} \times \mathcal{S}'$, where $\mathcal{E}$ indexes agent types and $\mathcal{S}'$ denotes the latent state space. The transition and reward functions may then depend explicitly on the agent's type.

3 GRAPHON MEAN-FIELD SUBSAMPLING

We propose Graphon Mean-Field Subsampling (GMFS), a framework that handles agent heterogeneity through non-uniform interactions modeled by graphons. To ensure scalability in large-scale systems, we introduce a subsampled neighborhood approximation, where each agent constructs an empirical neighborhood histogram from a small subset of the population. Using the graphon aggregates from Definition 2.1, GMFS learns an approximately optimal policy that achieves near-optimal team rewards under a sampled mean-field limit. This approach extends local sampling used in graph neural networks (31) to the MARL setting.

Assumption 3.1 (Well-defined graphon weights)**.** *We assume the underlying graphon $W : [0,1]^2 \to [0,1]$ satisfies $\int_0^1 W(\alpha, \beta)d\beta > 0$ for all $\alpha \in [0,1]$, ensuring that all row-normalized graphon weights are well defined.*

Definition 3.1 (Graphon-weighted Subsampling)**.** *Fix $\kappa \geq 1$. For each agent $i \in [n]$, sample a multiset of κ neighbors $\Delta_i = (J_i^{(1)}, \cdots, J_i^{(\kappa)})$ where $J_i^{(m)} \sim \bar{w}_{i,\cdot}$ on $[n] \setminus \{i\}$. Across all the agents, this yields a collection $\{\Delta_i\}_{i=1}^n$.*

Definition 3.2 (Sampled neighborhood aggregates)**.** *For $\kappa \geq 1$, let $\mathcal{G}_\kappa := \mathcal{P}_\kappa(\mathcal{S}) \subset \mathfrak{G}(\mathcal{S})$ and $\mathcal{Z}_\kappa := \mathcal{P}_\kappa(\mathcal{S} \times \mathcal{A})$ be the sets of empirical histograms with denominator κ. Given Δ_i from Definition 3.1, define the empirical κ-sample neighborhood state-action distribution $\widehat{Z}_i^{(\kappa)} \in \mathcal{Z}_\kappa$, where*

$$\widehat{Z}_i^{(\kappa)}(s,a) := \frac{1}{\kappa} \sum_{m=1}^{\kappa} \mathbb{1}\{s_{J_i^{(m)}} = s,\ a_{J_i^{(m)}} = a\},$$

and let the sampled neighborhood state marginal $\widehat{g}_i^{(\kappa)} \in \mathcal{G}_\kappa$ be given by $\widehat{g}_i^{(\kappa)}(s) := \sum_{a \in \mathcal{A}} \widehat{Z}_i^{(\kappa)}(s,a)$.

Sampled local reward. Each agent evaluates its local reward through the sampled neighborhood feature $r_\ell(s_i, a_i, \widehat{g}_i^{(\kappa)})$ as an approximation of the true mean-field reward. The discrepancy between this sampled value and the ground truth is governed by the regularity conditions established in the following assumptions.

Assumption 3.2 (Finite state/action spaces)**.** *We assume that the state and action spaces of all the agents in the system are finite: $|\mathcal{S}|, |\mathcal{A}| < \infty$. Appendix G relaxes this assumption to the non-tabular setting with infinite continuous states.*

Assumption 3.3 (Bounded rewards)**.** *We assume that the components of the reward function are bounded. Specifically, we assume $\|r_\ell\|_\infty < \infty$.*

Assumption 3.4 (The local component of the reward function is $2\|r_\ell\|_\infty$-Lipschitz in the empirical distribution)**.** *We assume that for all $s \in \mathcal{S}, a \in \mathcal{A}$, and $g, g' \in \mathfrak{G}(\mathcal{S})$,*

$$|r_\ell(s,a,g) - r_\ell(s,a,g')| \leq 2\|r_\ell\|_\infty \cdot \mathrm{TV}(g, g').$$

Assumption 3.5 (Transitions are L_P-Lipschitz)**.** *There exists $1 \leq L_P < \infty$ such that for all s, a and all $g, g' \in \mathfrak{G}(\mathcal{S})$,*

$$\mathrm{TV}(P(\cdot|s,a,g), P(\cdot|s,a,g')) \leq L_P \cdot \mathrm{TV}(g, g').$$

For $p \in \{1, \ldots, n-1\}$, define the Banach space $\mathcal{Y}_p = \{Q : \mathcal{S} \times \mathcal{A} \times \mathcal{Z}_p \to \mathbb{R}\}$ equipped with the sup-norm $\|Q\|_\infty := \sup_{s,a,z} |Q(s,a,z)|$.

Team value decomposition. Under the policy class where each agent's decision rule depends only on its local observation (s_i, g_i), the induced value function for agent i can be written as a function of s_i and g_i. Then, the team value decomposes as $V^\pi_{\text{team}}(\mathbf{s}_{1:n}, \mathbf{g}_{1:n}) = \frac{1}{n}\sum_{i=1}^n V^\pi(s_i, g_i)$, which is $\mathbb{E}[V^\pi(s,g)]$ for a uniformly random agent $i \in [n]$.

Q-function. We introduce a shared Q-function that is optimized centrally. Fix $\kappa \geq 2$ and let $\mathcal{Z}_\kappa$ denote the discrete set of empirical joint histograms on $\mathcal{S} \times \mathcal{A}$ with denominator κ. Let $\mathcal{G}_\kappa := \mathcal{P}_\kappa(\mathcal{S})$. For $z \in \mathcal{Z}_\kappa$, define its neighborhood state marginal $g_z \in \mathcal{G}_\kappa$ by $g_z(s) := \sum_{a \in \mathcal{A}} z(s,a)$ for all $s \in \mathcal{S}$. For a representative agent in state $s \in \mathcal{S}$ taking action $a \in \mathcal{A}$ and joint neighborhood histogram $z \in \mathcal{Z}_\kappa$, let

$$Q^\pi(s,a,z) = r_\ell(s,a,g_z) + \gamma \mathbb{E}_{(s',g') \sim \mathcal{J}_n(\cdot|s,a,z)}\left[V^\pi(s',g')\right],$$

where $\mathcal{J}_n$ is the induced one-step kernel, for a uniformly random agent, on (s,g) under the n-agent dynamics,

$$V^\pi(s,g) := \mathbb{E}_{a,z \sim \pi(\cdot|s,g)}[Q^\pi(s,a,z)].$$

Decentralized Execution. For $g \in \mathcal{G}_\kappa$, define the fiber $\Gamma_\kappa(g) := \{z \in \mathcal{Z}_\kappa : g_z = g\}$ and the Bellman backup given by $\mathcal{M}_\kappa Q(s,g) := \max_{a \in \mathcal{A}, z \in \Gamma_\kappa(g)} Q(s,a,z)$. The neighborhood histogram z is a latent variable generated by the environment and determined by the other agents' states and actions. It is not a directly controllable parameter for individual agents. Thus the maximization over z is not an action choice, but used as an optimization over latent variables to define a value function on (s,g) that upper-bounds achievable returns under under all completions consistent with g. The optimal (greedy) policy therefore selects an action

$$\pi^*(\cdot \mid s,g) \in \mathcal{M}_\kappa Q^*(s,g) = \arg\max_{a \in \mathcal{A}, z \in \Gamma_\kappa(g)} Q^*(s,a,z),$$

with the maximization over z serving to evaluate the action under the most favorable compatible neighborhood realization. In the fully centralized setting, optimal control would involve coordinating actions across all agents, which jointly determine the resulting neighborhood aggregate. During decentralized execution however, agents execute only the local action component and do not attempt

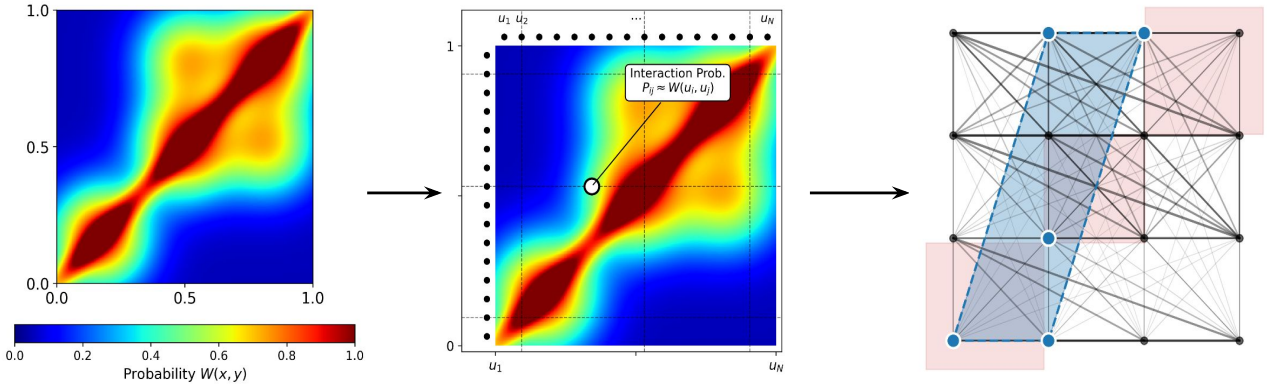


Figure 2: Schematic of Graphon Mean-Field Sampling. (Left) A continuous graphon $W(x, y)$ represents the infinite-population limit of non-uniform interactions. (Middle) The graphon with deterministic latent positions $\{\alpha_i\}_{i=1}^n \subset [0, 1]$ induces a complete weighted interaction graph on n agents with edge weights $w_{ij} = W(\alpha_i, \alpha_j)$ and $w_{ii} = 0$. These weights specify the intensity with which agent i aggregates neighbor states into its mean-field features used for learning and control. (Right) Each agent approximates its graphon-weighted neighborhood statistics by sampling a small set of agents (random subsample of $\kappa = 5$) according to the normalized weights $\bar{w}_{ij}$.

to control or realize any particular completion z.

Let $\mathcal{T}$ be the Bellman operator, where we then apply the Bellman update iteratively on each agent's Q-function. Execution is decentralized since each agent is able to locally construct an estimate $\widehat{z}_i^{(\kappa)}$. Thus, rewards and transitions depend on neighbors only through the neighborhood state marginal g_z, enabling each agent to optimize using local observations without knowing explicit identity information.

Definition 3.3 (Bellman operator $\mathcal{T}$). *For $Q \in \mathcal{Y}_{n-1}$, define*

$$\mathcal{T}Q(s, a, z) \coloneqq r_\ell(s, a, g_z) + \gamma \mathbb{E}_{(s',g') \sim \mathcal{J}_n(\cdot|s,a,z)} [\mathcal{M}_{n-1} Q(s', g')],$$

where $\mathcal{J}_n(\cdot|s, a, z)$ is the induced one-step kernel on (s', g') under the n-agent dynamics, for a uniformly random agent.

Definition 3.4 (Sampled Bellman operator $\widehat{\mathcal{T}}_\kappa$). *For $\widehat{Q}_\kappa \in \mathcal{Y}_\kappa$, define*

$$\widehat{\mathcal{T}}_\kappa \widehat{Q}_\kappa(s, a, z) \coloneqq r_\ell(s, a, g_z) + \gamma \mathbb{E}_{(s',g') \sim \mathcal{J}_\kappa(\cdot|s,a,z)} [\mathcal{M}_\kappa \widehat{Q}_\kappa(s', g')],$$

where $\mathcal{J}_\kappa(\cdot \mid s, a, z)$ is the induced one-step kernel on (s', g') generated by the $(\kappa + 1)$-agent surrogate dynamics, for a uniformly random agent.

Definition 3.5 (Empirical sampled operator $\widehat{\mathcal{T}}_{\kappa,m}$). *Given $m \in \mathbb{N}$, let $(s'_\ell, g'_\ell)_{\ell=1}^m$ be i.i.d. samples from $\mathcal{J}_\kappa(\cdot \mid s, a, z)$. Then, define*

$$\widehat{\mathcal{T}}_{\kappa,m} \widehat{Q}_{\kappa,m}(s, a, z) \coloneqq r_\ell(s, a, g_z) + \frac{\gamma}{m} \sum_{\ell=1}^{m} \mathcal{M}_\kappa \widehat{Q}_{\kappa,m}(s'_\ell, g'_\ell).$$

As a centralized-training decentralized-execution framework, GMFS proceeds as follows. First, algorithm 1 uses a generative oracle to derive an optimal Q-function on a $(\kappa + 1)$-agent surrogate model. Learning on this restricted subspace yields a policy governed by the subsampling parameter κ instead of the total population size n.

Algorithm 1 GMFS (Graphon Mean-Field Subsampling): Offline Learning

Require: Number of iterations T, subsampling parameters κ and m, and generative oracle $\mathcal{O}$.

Initialize $\widehat{Q}^{(0)}_{\kappa,m}(s, a, z) = 0, \forall (s, a, z) \in \mathcal{S} \times \mathcal{A} \times \mathcal{Z}_\kappa$
for $t = 1, \ldots, T$ **do**
  for $(s, a, z) \in \mathcal{S} \times \mathcal{A} \times \mathcal{Z}_\kappa$ **do**
    Update $\widehat{Q}^{(t+1)}_{\kappa,m}(s, a, z) = \widehat{\mathcal{T}}_{\kappa,m} \widehat{Q}^{(t)}_{\kappa,m}(s, a, z)$
Return $\widehat{Q}^{(T)}_{\kappa,m}$.

Subsequently, algorithm 2 describes how agents deploy this learned policy in a decentralized environment. During execution, each agent i independently approximates its neighborhood marginal $\widehat{g}_i^{(\kappa)}$ by sampling κ neighbors based on the graphon weights. This ensures that the computational cost per agent is independent of n, enabling scalable coordination in large-scale heterogeneous systems.

4 THEORETICAL GUARANTEES

We establish theoretical guarantees for GMFS by analyzing approximation errors due to neighborhood subsampling and finite-sample estimation. We define the properties of the sampled Bellman operators, and provide optimality bounds for the learned policy as a function of κ and m.

Bellman noise. We introduce the Bellman noise, $\epsilon_{\kappa,m}$, to account for the error in estimating the operator from finite samples. The empirical operator $\widehat{\mathcal{T}}_{\kappa,m}$ is an unbiased estimator of the sampled Bellman operator $\widehat{\mathcal{T}}_\kappa$. As shown in Lemma B.3, both $\widehat{\mathcal{T}}_\kappa$ and $\widehat{\mathcal{T}}_{\kappa,m}$ are γ-contractions with fixed-points $\widehat{Q}^*_\kappa$ and $\widehat{Q}^*_{\kappa,m}$ respectively. By the law of large numbers, $\lim_{m\to\infty} \widehat{\mathcal{T}}_{\kappa,m} = \widehat{\mathcal{T}}_\kappa$ and $\|\widehat{Q}^*_{\kappa,m} - \widehat{Q}^*_\kappa\|_\infty \to 0$ as $m \to \infty$. For finite m, we define this discrepancy as $\epsilon_{\kappa,m} \coloneqq$

Algorithm 2 GMFS (Graphon Mean-Field Subsampling): Online Execution

Require: Parameter T' for length of the game, subsampling parameter κ, graphon weights $\{\bar{w}_{ij}\}$, discount factor γ, and learned policy $\pi^T_{\kappa,m} = \mathcal{M}_\kappa \widehat{Q}^{(T)}_{\kappa,m}$.

Sample initial state $(s_1(0), \ldots, s_n(0)) \sim s_0$.
Initialize total reward $R_0 = 0$.
for $t = 0, \ldots, T' - 1$ **do**
  for $i = 1$ to n **do**
    Let $\Delta_i(t) = (J_i^1(t), \ldots, J_i^\kappa(t)) \overset{\text{(iid)}}{\sim} \bar{w}_{i,\cdot}$ on $[n] \setminus \{i\}$.
    Compute subsampled graphon-weighted mean-field features for $x \in \mathcal{S}$

$$\widehat{g}_i^{(\kappa)}(x) = \frac{1}{\kappa} \sum_{m=1}^{\kappa} \mathbb{1}\{s_{J_i^{(m)}(t)}(t) = x\}.$$

    Choose action $a_i(t) \sim \pi^T_{\kappa,m}(\cdot \mid s_i(t), \widehat{g}_i^{(\kappa)}(t))$.
  Get stage reward $\bar{R}_t \coloneqq r(\mathbf{s}_{1:n}(t), \mathbf{a}_{1:n}(t), \mathbf{g}_{1:n}(t))$.
  Let $s_i(t+1) \sim P(\cdot|s_i(t), a_i(t), g_i(t))$ for $i \in [n]$.
  $R_{t+1} = R_t + \gamma^t \cdot \bar{R}_t$.

$\|\widehat{Q}^*_{\kappa,m} - \widehat{Q}^*_\kappa\|_\infty$.

Let π^{est}_κ be the corresponding greedy GMFS policy defined in Definition D.3. First, noting that $\|V^{\pi^*} - V^{\pi^{\text{est}}_{\kappa,m}}\|_\infty \le \epsilon$ implies $\|V^{\pi^*}_{\text{team}} - V^{\pi^{\text{est}}_{\kappa,m}}_{\text{team}}\|_\infty \le \epsilon$, we show that the expected discounted cumulative reward produced by $\pi^{\text{est}}_{\kappa,m}$ is approximately optimal, with an optimality gap that decays as the sampling parameters κ and m increase.

Theorem 4.1. *For all states $s \in \mathcal{S}$ and graphon state-aggregates $g \in \mathcal{G}$, if $T \ge \frac{1}{1-\gamma} \log \frac{\|r_\ell\|_\infty \sqrt{\kappa}}{1-\gamma}$, then*

$$V^{\pi^*}(s,g) - V^{\pi^{\text{est}}_{\kappa,m}}(s,g) \le \frac{1}{20\sqrt{\kappa}(1-\gamma)} + \frac{\epsilon_{\kappa,m}}{1-\gamma} + \frac{2L_P \|r_\ell\|_\infty}{(1-\gamma)^2} \sqrt{\frac{|\mathcal{S}| \ln 2 + |\mathcal{A}| \ln \frac{20 \|r_\ell\|_\infty |\mathcal{A}| \kappa}{(1-\gamma)^2}}{2\kappa}}.$$

We generalize this result to stochastic rewards in Appendix E. To derive a final performance bound, we specify the number of samples m needed to bound $\epsilon_{\kappa,m}$.

Lemma 4.2 (Controlling the Bellman Noise). *For $\kappa \in [n]$, let the number of samples in theorem 4.1 be given by*

$$m^* = \frac{25\kappa^2\gamma^2}{(1-\gamma)^4} \|r_\ell\|^2_\infty \cdot \ln(200|\mathcal{S}|^2|\mathcal{A}|^2\kappa^{|\mathcal{S}||\mathcal{A}|}).$$

If T satisfies $T \ge \frac{2}{1-\gamma} \log \frac{\|r_\ell\|_\infty \sqrt{\kappa}}{1-\gamma}$, then we have that

$$\Pr\left[\epsilon_{\kappa,m^*} \le \frac{1}{5\sqrt{\kappa}}\right] \ge 1 - \frac{1}{100e^\kappa}.$$

See Appendix D.1 for the proof of Lemma 4.2. Combining the approximation bounds of Theorem 4.1 with the noise limits of Lemma 4.2 and defining the resulting policy as $\pi^{\text{est}}_\kappa \coloneqq \pi^{\text{est}}_{\kappa,m^*}$, we arrive at our main result in theorem 4.3:

Theorem 4.3. *Suppose $T \ge \frac{2}{1-\gamma} \log \frac{\|r_\ell\|_\infty \sqrt{\kappa}}{1-\gamma}$ and the number of samples m^* is chosen according to Lemma 4.2. Then, for all states $s \in \mathcal{S}$ and graphon state aggregates $g \in \mathcal{G}$, with probability[2] at least $1 - 1/100e^\kappa$,*

$$V^{\pi^*}(s,g) - V^{\pi^{\text{est}}_\kappa}(s,g) \le \frac{1}{4\sqrt{\kappa}(1-\gamma)} + \frac{2L_P \cdot \|r_\ell\|_\infty}{(1-\gamma)^2} \sqrt{\frac{|\mathcal{S}| \ln 2 + |\mathcal{A}| \ln \frac{20 \|r_\ell\|_\infty |\mathcal{A}| \kappa}{(1-\gamma)^2}}{2\kappa}}.$$

Sample complexity and optimality. The efficiency of GMFS is reflected in its sample complexity. For a fixed κ, algorithm 1 learns $\widehat{\pi}^{\text{est}}_k$ with an asymptotic sample complexity of $\widetilde{O}(\kappa^{|\mathcal{S}||\mathcal{A}|}|\mathcal{S}|^2|\mathcal{A}|^2)$, which is at least polynomially faster than standard Q-learning or mean-field value iteration. As shown in theorem 4.1, the optimality gap decays as $\kappa \to n$, which reveals a fundamental trade-off: increasing κ improves policy performance but increases the size of the Q-function. If we set $\kappa = O(\log n)$, the complexity becomes $\widetilde{O}((\log n)^{|\mathcal{S}||\mathcal{A}|}|\mathcal{S}|^2|\mathcal{A}|^2)$. This is an exponential speedup over the complexity of mean-field value iteration, from poly(n) to poly$(\log n)$, as well as over traditional value-iteration, where the optimality gap decays at a rate of $O(\frac{1}{\sqrt{\log n}})$.

There is evidence suggesting that the optimality gap of $\widetilde{O}(1/\sqrt{\kappa})$ is sharp. An obstacle to improving this bound is the known optimal error of $\widetilde{O}(1/\sqrt{n})$ in standard mean-field MARL. The algorithmic bottleneck in achieving a faster rate than $\widetilde{O}(1/\sqrt{\kappa})$ comes from learning the $\widehat{Q}_\kappa$ function rather than the online execution strategy. When $\kappa = n - 1$, GMFS reduces to mean-field learning with a rate of $\widetilde{O}(1/\sqrt{n})$, which matches the tight bound by Yang et al. (75). This illustrates a natural difficulty in improving the rate and provides evidence for why our bound is tight.

The poly$(\frac{1}{1-\gamma})$-dependence in our results may be loose because we do not use more complicated variance reduction techniques as in (37, 64, 65, 72) to optimize the number of samples m used to bound the Bellman error $\epsilon_{\kappa,m}$. Incorporating variance reduction would significantly increase the complexity of the algorithm and the underlying intuition. Finally, the GMFS formulation extends to off-policy Q-learning (15), which replaces the generative oracle with a stochastic approximation scheme to learn from historical data. This extension is detailed in section F, where we provide theoretical guarantees with a similar decaying optimality gap.

Generalization to infinite state spaces. In non-tabular environments with infinite state and action spaces, value-based RL methods can use function approximation to learn $\widehat{Q}_\kappa$ via deep Q-networks (66). This introduces an additional error term in the performance bound of theorem 4.1, which we analyze under a Linear MDP structure.

Assumption 4.1 (Linear MDP with infinite state spaces). *Let $\mathcal{S}$ be an infinite compact set, and assume a feature map $\phi : \mathcal{S} \times \mathcal{A} \times \mathcal{Z} \to \mathbb{R}^d$, d unknown (signed) measures $\mu = (\mu^1, \ldots, \mu^d)$ over $\mathcal{S}$, and a vector $\theta \in \mathbb{R}^d$ such that for any $(s,a,z) \in \mathcal{S} \times \mathcal{A} \times \mathcal{Z}$, we have $\mathbb{P}(\cdot|s,a,z) = \langle \phi(s,a,z), \mu(\cdot) \rangle$ and $r(s,a,g_z) = \langle \phi(s,a,z), \theta \rangle$.*

The existence of ϕ implies one can estimate the Q-function of any policy as a linear function. This assumption is used in policy

[2]The $1/100e^\kappa$ term can be replaced by an arbitrary $\delta > 0$ at the cost of attaching $\log 1/\delta$ dependencies to the error bound.

iteration methods (44), and we exploit it to obtain sample complexity bounds independently of $|\mathcal{S}|$ and $|\mathcal{A}|$. As is standard in RL, we assume bounded feature-norms (69):

Assumption 4.2 (Bounded features)**.** *We assume that* $\|\phi(s,a,z)\|_2 \leq 1$ *for all* $(s,a,z) \in \mathcal{S} \times \mathcal{A} \times \mathcal{Z}$.

Following the reduction from Ren et al. (62), we use function approximation to learn spectral features ϕ_κ for $\widehat{Q}_\kappa$. We derive a performance guarantee for the learned policy π_κ^{est}, where the optimality gap decays with κ.

Theorem 4.4. *When* π_κ^{est} *is derived from the spectral features* ϕ_κ *learned in* $\widehat{Q}_\kappa$*, and* M *is the number of samples used in the function approximation, then with probability at least* $1 - \frac{1}{50\kappa} - \frac{201}{100\sqrt{\kappa}}$*, we have*

$$V^{\pi^*}(s,g) - V^{\pi_{\kappa,m}}(s,g) \leq \widetilde{O}\left(\sqrt{\frac{d+|\mathcal{A}|}{\kappa}} + \frac{d}{\sqrt{M}} + \frac{2L_P\gamma\|r_\ell\|_\infty}{\sqrt{\kappa}}\right).$$

Although $\frac{d}{\sqrt{M}}$ grows linearly with the dimension, it is controlled by the sample budget M (i.e., chosen to scale with κ, e.g., $M \gtrsim \kappa^2 d^2$) so that it remains lower order relative to the $\widetilde{O}(1/\sqrt{\kappa})$ terms. We defer the proof of theorem 4.4 to Appendix G.

5 PROOF OUTLINE

In this section, we provide an outline of the derivation for our main results through three steps: (1) proving Lipschitz stability of the Bellman iterates, (2) bounding subsampling error via concentration inequalities, and (3) establishing a global performance guarantee using the performance difference lemma. See Appendices C and D.

Step 1: Lipschitz Continuity. We compare the Q-function evaluated under two neighborhood aggregates $z \in \mathcal{Z}$ and $\widehat{z} \in \mathcal{Z}_\kappa$, whose state marginals g_z and $g_{\widehat{z}}$ differ in total variation. Specifically, we show:

Theorem 5.1 (Lipschitz Continuity of the Bellman iterates). *Fix a subsampling parameter* $\kappa \geq 1$*. Fix* $(s,a) \in \mathcal{S} \times \mathcal{A}$*, and let* $z \in \mathcal{Z} \coloneqq \Delta(\mathcal{S} \times \mathcal{A})$*. Let* $\widehat{z} \in \mathcal{Z}_\kappa$ *be the empirical histogram of* κ *i.i.d. draws from* z*, and let* $g_z, g_{\widehat{z}}$ *be the marginals in* $\mathcal{S}$*. Then, for all* $t \in \mathbb{N}$*, we have*

$$\left|Q^t(s,a,z) - \widehat{Q}_\kappa^t(s,a,\widehat{z})\right| \leq \frac{4\|r_\ell\|_\infty}{1-\gamma} L_P \cdot \mathrm{TV}(g_z, g_{\widehat{z}}).$$

We defer the full proof of Theorem 5.1 to Appendix C.

Step 2: Concentration of the Subsampled Mean-Field. Next, we bound the discrepancy between the κ-sampled aggregate $\widehat{g}_i^{(\kappa)}$ and the true graphon-weighted mean-field g_i. We establish a concentration inequality for empirical distributions drawn from a finite population to show that with probability at least $1-\delta$:

$$\left|Q^t(s,a,z) - \widehat{Q}_\kappa^t(s,a,\widehat{z})\right| \leq \frac{4L_P\|r_\ell\|_\infty}{1-\gamma}\sqrt{\frac{|\mathcal{S}|\ln 2 + \ln\frac{2}{\delta}}{2\kappa}}.$$

This result introduces the $O(1/\sqrt{\kappa})$ rate; the proof is provided in Appendix C.1.

Step 3: Performance Difference. Finally, we combine the previous steps to bound the performance gap between the learned policy π_κ^{est} from GMFS and the optimal policy π^*. Using Lemma D.3, we obtain a uniform bound on the optimality of the estimated joint policy in terms of the discrepancy between Q^* and $\widehat{Q}_\kappa^*$. We also account for the additional error introduced by finite-sample Bellman noise $\varepsilon_{\kappa,m}$. This allows us to apply the performance difference lemma (38), yielding Theorem 4.3. The full proof is provided in Appendix D.

6 CONCLUSION

In this work, we consider the problem of learning an optimal policy in a cooperative system of n heterogeneous agents. We propose an algorithm, GMFS, which derives a policy $\widetilde{\pi}_\kappa^{\text{est}}$ where $\kappa \leq n$ is a tunable parameter for the number of agents sampled. We show that $\widetilde{\pi}_\kappa^{\text{est}}$ converges to the optimal policy π^* with a decay rate of $O(1/\sqrt{\kappa})$. To establish this result, we develop an adapted Bellman operator $\widehat{\mathcal{T}}_\kappa$ and prove its contraction property. The key technical novelty of this work lies in proving a Lipschitz continuity result for $\widehat{Q}_\kappa^*$ and leveraging the weights of a corresponding graphon. Finally, we supplement our theoretical results with motivating examples and provide additional empirical validation (see Appendix A.4) via numerical simulations of robotic control.

Limitations and Future Work. While GMFS is a scalable framework for heterogeneous MARL, there are several areas for extending its theoretical and practical scope. For instance, a natural next step is to derive matching lower bounds to demonstrate the tightness of our analysis for subsampling in graphon-weighted systems. Our analysis assumes access to a generative simulator; extending GMFS to a purely online setting with streaming interaction data and exploration-exploitation trade-offs would extend its applicability to model-free environments. It would also be interesting to study higher-order interaction structures, such as Markov random fields or hypergraph-based mean-fields, to capture context-dependent heterogeneity (73). Another direction would be to examine whether online mirror descent can be integrated with our algorithm, as in Fabian et al. (23), to get better numerical stability. Finally, it would be interesting to adapt GMFS beyond cooperative MARL to mixed cooperative-competitive environments or federated learning with non-uniform communication.

ACKNOWLEDGMENTS

This work was supported by NSF Grants CCF 2338816, CNS 2146814, CNS 2106403, CPS 2136197. SL acknowledges the support of the Kempner Institute Graduate Research Fellowship. We gratefully acknowledge insightful discussions with Yuzhou Wang, Sam van der Poel, Ishani Karmarkar, and Guannan Qu.

REFERENCES

[1] Edoardo M. Airoldi, Thiago B. Costa, and Stanley H. Chan. 2013. Stochastic blockmodel approximation of a graphon: Theory and consistent estimation. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1* (Lake Tahoe, Nevada) *(NIPS'13)*. Curran Associates Inc., Red Hook, NY, USA, 692–700.

[2] Emile Anand and Sarah Liaw. 2025. Feel-Good Thompson Sampling for Contextual Bandits: a Markov Chain Monte Carlo Showdown. arXiv:2507.15290 [cs.LG] https://arxiv.org/abs/2507.15290

[3] Emile Anand and Guannan Qu. 2024. Efficient Reinforcement Learning for Global Decision Making in the Presence of Local Agents at Scale. *arXiV* (2024). arXiv:2403.00222 [cs.LG]

[4] Emile Timothy Anand, Ishani Karmarkar, and Guannan Qu. 2025. Mean-Field Sampling for Cooperative Multi-Agent Reinforcement Learning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. San Diego, California, USA.

[5] Stefan Banach. 1922. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae* 3, 1 (1922), 133–181.

[6] Vincent D. Blondel and John N. Tsitsiklis. 2000. A Survey of Computational Complexity Results in Systems and Control. *Automatica* 36, 9 (2000), 1249–1274. https://doi.org/10.1016/S0005-1098(00)00050-9

[7] Christian Borgs, Jennifer Tour Chayes, László Miklós Lovász, Vera T. Sós, and Katalin Vesztergombi. 2007. Convergent Sequences of Dense Graphs I: Subgraph Frequencies, Metric Properties and Testing. *Advances in Mathematics* 219 (2007), 1801–1851.

[8] Peter E. Caines and Minyi Huang. 2018. Graphon Mean Field Games and the GMFG Equations. In *2018 IEEE Conference on Decision and Control (CDC)*. Miami Beach, USA, 4129–4134. https://doi.org/10.1109/CDC.2018.8619367

[9] Peter E. Caines and Minyi Huang. 2021. Graphon Mean Field Games and Their Equations. *SIAM Journal on Control and Optimization* 59, 6 (Jan. 2021), 4373–4399. https://doi.org/10.1137/20m136373x

[10] René Carmona, Mathieu Laurière, and Zongjun Tan. 2023. Model-free mean-field reinforcement learning: Mean-field MDP and mean-field Q-learning. *The Annals of Applied Probability* 33, 6B (2023), 5334 – 5381. https://doi.org/10.1214/23-AAP1949

[11] Stanley Chan and Edoardo Airoldi. 2014. A Consistent Histogram Estimator for Exchangeable Graph Models. In *Proceedings of the 31st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 32)*, Eric P. Xing and Tony Jebara (Eds.). PMLR, Bejing, China, 208–216.

[12] Olivier Chapelle and Lihong Li. 2011. An Empirical Evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger (Eds.), Vol. 24. Curran Associates, Inc., Granada, Spain.

[13] Shreyas Chaudhari, Srinivasa Pranav, Emile Anand, and José M. F. Moura. 2025. Peer-to-Peer Learning Dynamics of Wide Neural Networks. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, Hyderabad, India, 1–5. https://doi.org/10.1109/icassp49660.2025.10890126

[14] Zaiwei Chen and Siva Theja Maguluri. 2022. Sample Complexity of Policy-Based Methods under Off-Policy Sampling and Linear Function Approximation. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 151)*, Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera (Eds.). PMLR, Virtual, 11195–11214.

[15] Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. 2021. A Lyapunov theory for finite-sample guarantees of asynchronous Q-learning and TD-learning variants. *arXiv preprint arXiv:2102.01567* (2021).

[16] Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. 2021. Finite-sample analysis of off-policy TD-learning via generalized Bellman operators. *Advances in Neural Information Processing Systems* 34 (2021), 21440–21452.

[17] Zaiwei Chen, Siva Theja Maguluri, and Martin Zubeldia. 2025. Concentration of contractive stochastic approximation: Additive and multiplicative noise. *The Annals of Applied Probability* 35, 2 (2025), 1298–1352.

[18] Jingjing Cui, Yuanwei Liu, and Arumugam Nallanathan. 2020. Multi-Agent Reinforcement Learning-Based Resource Allocation for UAV Networks. *IEEE Transactions on Wireless Communications* 19, 2 (2020), 729–743. https://doi.org/10.1109/TWC.2019.2935201

[19] Kai Cui, Christian Fabian, Anam Tahir, and Heinz Koeppl. 2024. Major-Minor Mean Field Multi-Agent Reinforcement Learning. arXiv:2303.10665 [cs.LG]

[20] Kai Cui and Heinz Koeppl. 2022. Learning Graphon Mean Field Games and Approximate Nash Equilibria. In *International Conference on Learning Representations*. Virtual.

[21] Alex DeWeese and Guannan Qu. 2024. Locally Interdependent Multi-Agent MDP: Theoretical Framework for Decentralized Agents with Dynamic Dependencies. In *Forty-first International Conference on Machine Learning*. Vienna, Austria.

[22] Christian Fabian, Kai Cui, and Heinz Koeppl. 2022. Mean Field Games on Weighted and Directed Graphs via Colored Digraphons. *IEEE Control Systems Letters* 7 (2022), 877–882. https://doi.org/10.1109/LCSYS.2022.3210044

[23] Christian Fabian, Kai Cui, and Heinz Koeppl. 2023. Learning Sparse Graphon Mean Field Games. arXiv:2209.03880 [cs.MA]

[24] Christian Fabian, Kai Cui, and Heinz Koeppl. 2025. Learning Mean Field Control on Sparse Graphs. In *Proceedings of the 42nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 267)*, Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (Eds.). PMLR, Vancouver, Canada, 15637–15660.

[25] Scott Fujimoto, David Meger, and Doina Precup. 2019. Off-Policy Deep Reinforcement Learning without Exploration. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Long Beach, California, USA, 2052–2062.

[26] David Gamarnik, David A. Goldberg, and Theophane Weber. 2014. Correlation Decay in Random Decision Networks. *Math. Oper. Res.* 39, 2 (May 2014), 229–261. https://doi.org/10.1287/moor.2013.0609

[27] Shuang Gao and Peter E. Caines. 2017. The control of arbitrary size networks of linear systems via graphon limits: An initial investigation. In *2017 IEEE 56th Annual Conference on Decision and Control (CDC)*. 1052–1057. https://doi.org/10.1109/CDC.2017.8263796

[28] Noah Golowich and Ankur Moitra. 2024. The Role of Inherent Bellman Error in Offline Reinforcement Learning with Linear Function Approximation. *Reinforcement Learning Journal* 1 (2024), 302–341.

[29] Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. 2025. Mean-Field Multiagent Reinforcement Learning: A Decentralized Network Approach. *Math. Oper. Res.* 50, 1 (Feb. 2025), 506–536. https://doi.org/10.1287/moor.2022.0055

[30] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. 2021. Learning Mean-Field Games. arXiv:1901.09585 [math.OC]

[31] William L. Hamilton, Rex Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 1025–1035.

[32] Wassily Hoeffding. 1963. Probability Inequalities for Sums of Bounded Random Variables. *J. Amer. Statist. Assoc.* 58, 301 (1963), 13–30.

[33] Yuanquan Hu, Xiaoli Wei, Junji Yan, and Hengxi Zhang. 2023. Graphon mean-field control for cooperative multi-agent reinforcement learning. *Journal of the Franklin Institute* 360, 18 (2023), 14783–14805. https://doi.org/10.1016/j.jfranklin.2023.09.002

[34] Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. 2021. Bellman Eluder dimension: new rich classes of RL problems, and sample-efficient algorithms. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NIPS '21)*. Curran Associates Inc., Red Hook, NY, USA, Article 1027, 13 pages.

[35] Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. 2020. Provably efficient reinforcement learning with linear function approximation. In *Proceedings of Thirty Third Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 125)*, Jacob Abernethy and Shivani Agarwal (Eds.). PMLR, Graz, Austria, 2137–2143.

[36] Junqi Jin, Chengru Song, Han Li, Kun Gai, Jun Wang, and Weinan Zhang. 2018. Real-Time Bidding with Multi-Agent Reinforcement Learning in Display Advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) *(CIKM '18)*. Association for Computing Machinery, New York, NY, USA, 2193–2201. https://doi.org/10.1145/3269206.3272021

[37] Yujia Jin, Ishani Karmarkar, Aaron Sidford, and Jiayi Wang. 2024. Truncated Variance Reduced Value Iteration. In *Advances in Neural Information Processing Systems*. Vancouver, Canada. Also available as arXiv preprint arXiv:2405.12952.

[38] Sham M. Kakade and John Langford. 2002. Approximately Optimal Approximate Reinforcement Learning. In *International Conference on Machine Learning*. Sydney, Australia.

[39] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. 2021. Deep Reinforcement Learning for Autonomous Driving: A Survey. arXiv:2002.00444 [cs.LG]

[40] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. 2022. Deep Reinforcement Learning for Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems* 23, 6 (2022), 4909–4926. https://doi.org/10.1109/TITS.2021.3054625

[41] Robert Kleinberg. 2005. A multiple-choice secretary algorithm with applications to online auctions. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (Vancouver, British Columbia) *(SODA '05)*. Society for Industrial and Applied Mathematics, USA, 630–631.

[42] Jens Kober, J. Andrew Bagnell, and Jan Peters. 2013. Reinforcement Learning in Robotics: A Survey. *The International Journal of Robotics Research* 32, 11 (2013), 1238–1274. https://doi.org/10.1177/0278364913495721 arXiv:https://doi.org/10.1177/0278364913495721

[43] Jean-Michel Lasry and Pierre-Louis Lions. 2007. Mean Field Games. *Japanese Journal of Mathematics* 2, 1 (March 2007), 229–260. https://doi.org/10.1007/s11537-007-0657-8

[44] Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. 2020. Learning with Good Feature Representations in Bandits and in RL with a Generative Model. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, Vienna, Austria, 5662–5670.

[45] Mathieu Laurière, Sarah Perrin, Julien P'erolat, Sertan Girgin, P. Muller, Romuald Élie, Matthieu Geist, and Olivier Pietquin. 2022. Learning in Mean Field Games: A Survey. *arXiv* (2022).

[46] Minne Li, Zhiwei Qin, Yan Jiao, Yaodong Yang, Jun Wang, Chenxi Wang, Guobin Wu, and Jieping Ye. 2019. Efficient Ridesharing Order Dispatching with Mean Field Multi-Agent Reinforcement Learning. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. Association for Computing Machinery, New York, NY, USA, 983–994. https://doi.org/10.1145/3308558.3313433

[47] Yan Li, Lingxiao Wang, Jiachen Yang, Ethan Wang, Zhaoran Wang, Tuo Zhao, and Hongyuan Zha. 2021. Permutation Invariant Policy Optimization for Mean-Field Multi-Agent Reinforcement Learning: A Principled Approach. arXiv:2105.08268 [cs.LG]

[48] Yiheng Lin, James Preiss, Emile Anand, Yingying Li, Yisong Yue, and Adam Wierman. 2022. Online adaptive controller selection in time-varying systems: No-regret via contractive perturbations. *arXiv preprint arXiv:2210.12320* (2022).

[49] Yiheng Lin, James A Preiss, Emile Timothy Anand, Yingying Li, Yisong Yue, and Adam Wierman. 2023. Online Adaptive Policy Selection in Time-Varying Systems: No-Regret via Contractive Perturbations. In *Thirty-seventh Conference on Neural Information Processing Systems*. New Orleans, USA.

[50] Yiheng Lin, James A. Preiss, Fengze Xie, Emile Anand, Soon-Jo Chung, Yisong Yue, and Adam Wierman. 2024. Online Policy Optimization in Unknown Nonlinear Systems. In *Proceedings of Thirty Seventh Conference on Learning Theory (Proceedings of Machine Learning Research, Vol. 247)*, Shipra Agrawal and Aaron Roth (Eds.). PMLR, Edmonton, Canada, 3475–3522.

[51] Yiheng Lin, Guannan Qu, Longbo Huang, and Adam Wierman. 2020. Distributed Reinforcement Learning in Multi-Agent Networked Systems. *CoRR* abs/2006.06555 (2020). arXiv:2006.06555

[52] Michael L. Littman. 1994. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Machine learning proceedings (Elsevier)*. 157–163.

[53] László Lovász. 2012. *Large networks and graph limits*. Vol. 60. American Mathematical Soc.

[54] Yang Lv, Jinlong Lei, and Peng Yi. 2025. A Local Information Aggregation-Based Multiagent Reinforcement Learning for Robot Swarm Dynamic Task Allocation. *IEEE Transactions on Neural Networks and Learning Systems* 36, 6 (2025), 10437–10449. https://doi.org/10.1109/TNNLS.2025.3558282

[55] Washim Uddin Mondal, Mridul Agarwal, Vaneet Aggarwal, and Satish V. Ukkusuri. 2022. On the Approximation of Cooperative Heterogeneous Multi-Agent Reinforcement Learning (MARL) Using Mean Field Control (MFC). *Journal of Machine Learning Research* 23, 1, Article 129 (jan 2022), 46 pages.

[56] Rémi Munos. 2014. *From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning*. Vol. 7. Now Foundations and Trends. 1–129 pages. https://doi.org/10.1561/2200000038

[57] Reza Olfati-Saber, J. Alex Fax, and Richard M. Murray. 2007. Consensus and Cooperation in Networked Multi-Agent Systems. *Proc. IEEE* 95, 1 (2007), 215–233. https://doi.org/10.1109/JPROC.2006.887293

[58] Barna Pásztor, Andreas Krause, and Ilija Bogunovic. 2023. Efficient Model-Based Multi-Agent Mean-Field Reinforcement Learning. *Transactions on Machine Learning Research* (2023).

[59] James A. Preiss, Wolfgang Honig, Gaurav S. Sukhatme, and Nora Ayanian. 2017. Crazyswarm: A large nano-quadcopter swarm. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. Singapore, 3299–3304. https://doi.org/10.1109/ICRA.2017.7989376

[60] Guannan Qu, Yiheng Lin, Adam Wierman, and Na Li. 2020. Scalable Multi-Agent Reinforcement Learning for Networked Systems with Average Reward. In *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada) *(NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 175, 13 pages.

[61] Farshad Rahimi and Reza Esfanjani. 2023. Estimating tolerable communication delays for distributed optimization problems in control of heterogeneous multi-agent systems. *IET Control Theory & Applications* 18 (11 2023), n/a–n/a. https://doi.org/10.1049/cth2.12595

[62] Zhaolin Ren, Runyu Zhang, Bo Dai, and Na Li. 2025. Scalable spectral representations for multiagent reinforcement learning in network MDPs. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 258)*, Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (Eds.). PMLR, Mai Khao, Phuket, Thailand, 550–558.

[63] Carlos Riquelme, George Tucker, and Jasper Snoek. 2018. Deep Bayesian Bandits Showdown: An Empirical Comparison of Bayesian Deep Networks for Thompson Sampling. In *Proceedings of the International Conference on Learning Representations (ICLR)*. Vancouver, Canada.

[64] Aaron Sidford, Mengdi Wang, Xian Wu, Lin F. Yang, and Yinyu Ye. 2018. Near-Optimal Time and Sample Complexities for Solving Discounted Markov Decision Process with a Generative Model. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Montréal, Canada, 7739–7750.

[65] Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. 2018. Variance Reduced Value Iteration and Faster Algorithms for Solving Markov Decision Processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*. Society for Industrial and Applied Mathematics, New Orleans, LA, USA. https://doi.org/10.1137/1.9781611975031.116

[66] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* 529, 7587 (Jan. 2016), 484–489. https://doi.org/10.1038/nature16961

[67] Arvind Singh, M. Sujatha, Akshay Kadu, Mohit Bajaj, Hailu Addis, and Kota Sarada. 2025. A deep learning and IoT-driven framework for real-time adaptive resource allocation and grid optimization in smart energy systems. *Scientific Reports* 15 (06 2025). https://doi.org/10.1038/s41598-025-02649-w

[68] Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Advances in Neural Information Processing Systems*, S. Solla, T. Leen, and K. Müller (Eds.), Vol. 12. MIT Press, Denver, Colorado, USA.

[69] Volodymyr Tkachuk, Seyed Alireza Bakhtiari, Johannes Kirschner, Matej Jusup, Ilija Bogunovic, and Csaba Szepesvári. 2023. Efficient Planning in Combinatorial Action Spaces with Applications to Cooperative Multi-Agent Reinforcement Learning. *arXiV* (2023). arXiv:2302.04376 [cs.LG]

[70] Alexandre B. Tsybakov. 2008. *Introduction to Nonparametric Estimation* (1st ed.). Springer Publishing Company, Incorporated.

[71] Santosh Vempala and Andre Wibisono. 2019. Rapid convergence of the unadjusted Langevin algorithm: Isoperimetry suffices. *Advances in neural information processing systems* 32 (2019).

[72] Martin J Wainwright. 2019. Variance-reduced Q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697* (2019).

[73] Archer Wang, Emile Anand, Yilun Du, and Marin Soljačić. 2026. Unsupervised Decomposition and Recombination with Discriminator-Driven Diffusion Models. arXiv:2601.22057 [cs.CV]

[74] Min Yang, Guanjun Liu, Ziyuan Zhou, and Jiacun Wang. 2023. Partially Observable Mean Field Multi-Agent Reinforcement Learning Based on Graph Attention Network for UAV Swarms. *Drones* 7, 7 (July 2023), 476. https://doi.org/10.3390/drones7070476

[75] 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 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, Stockholm, Sweden, 5571–5580.

[76] Jing-Wen Yi, Yan wu Wang, Jiang-Wen Xiao, and Yang Chen. 2016. Consensus in second-order Markovian jump multi-agent systems via impulsive control using sampled information with heterogenous delays. *Asian Journal of Control* 18 (2016), 1940 – 1949.

[77] Fengzhuo Zhang, Vincent Y. F. Tan, Zhaoran Wang, and Zhuoran Yang. 2024. Learning regularized graphon mean-field games with unknown graphons. *J. Mach. Learn. Res.* 25, 1, Article 372 (Jan. 2024), 95 pages.

[78] Jiayi Zhang, Ziheng Liu, Yiyang Zhu, Enyu Shi, Bokai Xu, Chau Yuen, Dusit Niyato, Mérouane Debbah, Shi Jin, Bo Ai, and Xuemin Shen. 2025. Multi-Agent Reinforcement Learning in Wireless Distributed Networks for 6G. *CoRR* abs/2502.05812 (February 2025).

[79] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2021. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. In *Handbook of Reinforcement Learning and Control*. Springer, 321–384. https://doi.org/10.1007/978-3-030-60990-0_12

[80] Liancheng Zheng, Zhen Tian, Yangfan He, Shuo Liu, Huilin Chen, Fujiang Yuan, and Yanhong Peng. 2025. Enhanced Mean Field Game for Interactive Decision-Making with Varied Stylish Multi-Vehicles. arXiv:2509.00981 [cs.RO]

[81] Yihe Zhou, Shunyu Liu, Yunpeng Qing, Kaixuan Chen, Tongya Zheng, Yanhao Huang, Jie Song, and Mingli Song. 2023. Is Centralized Training with Decentralized Execution Framework Centralized Enough for MARL? arXiv:2305.17352 [cs.AI]

Outline of the Appendices.

- Section A provides our numerical simulations, experimental details, and supplemental examples;
- Section B provides mathematical background, and relevant introductory lemmas;
- Section C presents the proof of the Lipschitz continuity of the Q-function in the mean-field measure;
- Section D bounds the optimality gap between the learned stochastic policy and the optimal policy;
- Section E extends the result to stochastic rewards;
- Section F extends the result to off-policy learning;
- Section G extends the result to continuous state spaces.

Notation

Symbol	Description
$\mathbb{R}$	Space of real numbers
$\mathbb{S}^p$	The p-dimensional sphere that forms the boundary of a ball in $\mathbb{R}^{(p+1)}$
n	Number of agents in the system
$\mathbb{G} = (\mathcal{V}, \mathcal{E})$	$\lvert V\rvert$-agent interaction graph
$\mathcal{S}, \mathcal{A}$	Agent state space and agent action space
$\mathcal{G}, \mathcal{H}$	Neighboring agent state space and action space
$r_\ell(s, a, g)$	Local reward for each agent given state, action, and mean-field
$r(s, a, g)$	Team reward $\frac{1}{n}\sum_{i=1}^{n} r_\ell(\cdot)$
$\bar{w}_{ij}$	Normalized interaction weight, $\bar{w}_{ij} = \frac{w_{ij}}{\sum_{m\neq i} w_{im}}$
Δ_i	Multi-set of κ neighbors sampled for agent i
$\Gamma_\kappa(g)$	Fiber $\{z \in \mathcal{Z}_\kappa : g_z = g\}$ that completes a neighborhood aggregate g
$z \in \mathcal{Z} \coloneqq \Delta(\mathcal{S} \times \mathcal{A})$	Graphon-weighted neighborhood state-action distribution
$\widehat{z} \in \mathcal{Z}_\kappa$	Empirical histogram of κ i.i.d. draws from z
$g_z, g_{\widehat{z}}$	State marginals in $\mathcal{S}$
h_{z_i}	Action marginal $\sum_{x\in\mathcal{S}} z_i(x, u) \in \mathcal{H}$
$\widehat{z}_i^{(\kappa)}$	Empirical neighborhood state-action histogram from Δ_i
$\widehat{g}_i^{(\kappa)}$	Empirical neighborhood state marginal of $\widehat{z}_i^{(\kappa)}$
$\mathcal{J}_n$	Induced one-step kernel on (s', g') under the n-agent dynamics.
$\mathcal{J}_\kappa$	Induced one-step kernel on (s', g') generated by the $(\kappa+1)$-agent surrogate dynamics.
$Q^\pi(s, a, z), V^\pi(s, g)$	Centralized critic for agent i and value function for agent i under policy π.
Q^*, V^*	Optimal centralized Q- and V-functions
$\widehat{Q}^*_\kappa$	Optimal Q-function under κ-sampled mean-field dynamics
$\widehat{Q}^*_{\kappa,m}$	Empirical estimate of $\widehat{Q}^*_\kappa$ using m samples
$\varepsilon_{\kappa,m}$	Bellman noise: $\lVert\widehat{Q}^*_{\kappa,m} - \widehat{Q}^*_\kappa\rVert_\infty$
$\mathcal{T}$	Bellman operator for agent i (true mean-field)
$\widehat{\mathcal{T}}_\kappa$	Sampled Bellman operator (using κ-subsampled aggregates)
$\widehat{\mathcal{T}}_{\kappa,m}$	Empirical Bellman operator using m samples.
π^*	Optimal joint policy
$\pi^{\text{est}}_{\kappa,\Delta}$	Estimated joint policy under graphon-weighted subsampling
π^{est}_κ	Learned policy
$\lVert\cdot\rVert_\infty$	ℓ_∞-norm
$\gamma \in (0, 1)$	Discount factor
P	Transition kernel for an agent's next state, $s_i \sim P(\cdot \mid s_i, a_i, g_i)$

A NUMERICAL SIMULATIONS AND ADDITIONAL MOTIVATING EXAMPLES

In this section, we provide an additional conceptual formulation of `GMFS` within the context of smart grid management. We then provide an empirical evaluation of the algorithm using a cooperative robotics coordination task.

A.1 Motivating Example: Cooperative Autonomous Driving

Consider a population of n autonomous vehicles indexed by $i \in \{1, \ldots, n\}$. Each agent i is associated with a latent feature $\alpha_i \in [0, 1]$ representing its position along a road segment. Interactions are governed by a graphon $W : [0, 1]^2 \to [0, 1]$, where $W(\alpha_i, \alpha_j)$ represents the strength of the interaction between vehicles indexed by i and j. For example, we model distance-decaying influence along the road with: $W(\alpha_i, \alpha_j) = \exp(-\beta|\alpha_i - \alpha_j|)$. Each agent has state $s_i \in \mathcal{S}$, which can encode the position or velocity, and selects actions $a_i \in \mathcal{A}$, which can include steering and acceleration. Vehicle i transitions according $s_i(t+1) \sim P\big(\cdot \mid s_i(t), a_i(t), g_i(t)\big)$. At each time t, vehicle i constructs a

multiset of κ neighbors $\Delta_i(t) = (J_i^{(1)}(t), \ldots, J_i^{(\kappa)}(t))$ and forms the empirical histograms:

$$\widehat{Z}_i^{(\kappa)}(t)(s,a) = \frac{1}{\kappa}\sum_{m=1}^{\kappa} \mathbb{1}\left\{s_{J_i^{(m)}(t)}(t) = s, a_{J_i^{(m)}(t)}(t) = a\right\},$$
$$\widehat{g}_i^{(\kappa)}(t)(s) = \sum_{a\in\mathcal{A}} \widehat{Z}_i^{(\kappa)}(t)(s,a).$$

Since the distance-decay graphon induces decaying weights, the dominant contributors in this example to $g_i(t)$ are the nearby vehicles. Thus, sampling from $\bar{w}_{ij}$ targets the most informative neighbors. Increasing κ tightens concentration of $\widehat{g}_i^{(\kappa)}(t)$ while keeping the per-agent computation scalable.

A.2 Motivating Example: Cooperative Robot Coordination Task

Consider n mobile robots indexed by $i \in [n]$ operating in a warehouse. Associate each robot with a latent coordinate $\alpha_i \in [0,1]$ capturing its location along an embedding of aisles/loading zones. Let $W : [0,1]^2 \to [0,1]$ encode interaction intensity. We consider a radius graphon $W(\alpha,\beta) = \mathbb{1}\{|\alpha-\beta| \le r\}$ for some $r > 0$, so that robots primarily interact with others in nearby regions. Let $\mathcal{S}$ be a finite state space encoding each robot's local mode and task-relevant information, such as idle, moving, standing. Let $\mathcal{A}$ be a finite action space which includes move, wait, pickup. At time t, robot i observes its local state $s_i(t)$ and a neighborhood summary, selects an action $a_i(t) \in \mathcal{A}$, and transitions according to $s_i(t+1) \sim P\big(\cdot \mid s_i(t), a_i(t), g_i(t)\big)$ where $g_i(t)$ is the graphon-weighted neighborhood state feature. In decentralized execution, robot i does not enumerate all neighbors. The sampled local reward is evaluated as $r_\ell(s_i(t), a_i(t), \widehat{g}_i^{(\kappa)}(t))$, capturing congestion/collision-risk effects induced by nearby robots. The interaction locality implies that the most relevant information for robot i is concentrated among a small subset of agents with large weights $\bar{w}_{ij}$. Graphon-weighted subsampling therefore preserves the dominant neighborhood statistics while keeping per-agent computation independent of n.

A.3 Motivating Example: Energy Distribution for a Smart Grid

Modern smart grids present a natural application for GMFS due to their high topological heterogeneity and decentralized nature. Unlike traditional power systems, a substation's demand response in a smart grid is heavily influenced by immediate topological neighbors and local transmission constraints rather than the global average of the entire grid. In this conceptual setting, we consider n agents representing local substations, or energy consumers, indexed by $i \in [n]$.

One could model this environment by labeling each agent with a latent position $\alpha_i \in [0,1]$, which represents geographical coordinates or a position within an infrastructural hierarchy. The interaction graphon $W : [0,1]^2 \to [0,1]$ would encode the transmission efficiency and physical connectivity between substations. As noted by Singh et al. (67), renewable energy integration requires minimizing transmission losses, which are inherently non-uniform. A graphon structure would capture this by assigning higher weights to substation pairs with high-capacity links or proximity. To ensure the resulting neighborhood sampling distribution is well-defined, we assume $\int_0^1 W(\alpha,\beta)d\beta > 0$ for all $\alpha \in [0,1]$, ensuring every node has a non-zero interaction density.

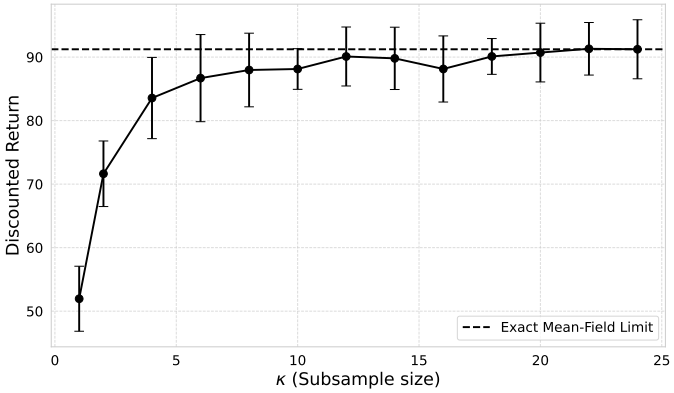
Under this formulation, each agent i maintains a state $s_i \in \mathcal{S}$ representing its current load or energy deficit and selects an action $a_i \in \mathcal{A}$ to request a specific allocation from a shared supply. At each step, agent i would reconstruct a graphon-weighted subsampled aggregate $z_i \in \mathcal{Z}_\kappa$ (Def. 3.2). The reward function $r_\ell(s_i, a_i, g_i)$ can be designed to penalize "over-allocation" costs and transmission line heating, which are non-linear functions of the local demand density. Ultimately, this example shows how GMFS could enable substations to make near-optimal allocation decisions by observing only $\kappa \ll n$ neighboring nodes, which avoids the need for a centralized controller to process the state of the entire national grid.

A.4 Evaluation on Cooperative Robot Coordination Task

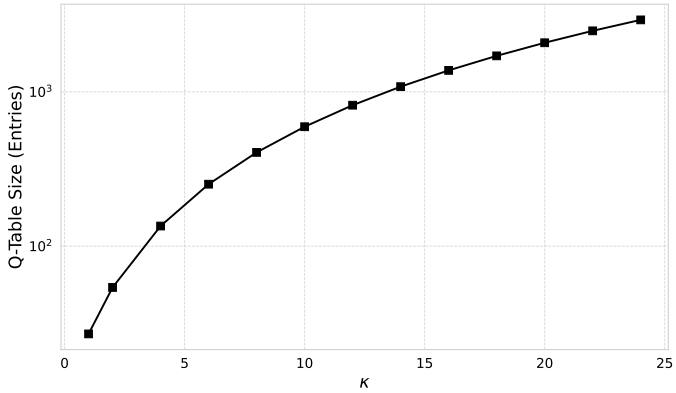
We evaluate the empirical performance of GMFS on a cooperative robot coordination task in A.2 within a spatially constrained warehouse environment. This domain is a natural fit for graphon models because robot interactions are inherently local; a robot is significantly more affected by collision risks or blocked aisles in its immediate vicinity than by the status of a robot on the opposite side of a facility.

A key difficulty in such systems is the *perception-action gap*, where agents must make decisions based on incomplete social information. In Figure 4, we visualize the time-evolution of an agent's perception. At low κ, the agent's view of its neighborhood is sparse and noisy, resulting in high-variance estimates of local congestion. As κ increases, the agent's empirical measure $\widehat{g}_i$ converges to the true geometric ball defined by the graphon. In Figure 3a, we demonstrate that GMFS allows agents to achieve near-optimal coordination even when sampling only a small fraction of the total population.

A.4.1 Experimental Setup. The robotics environment consists of $n = 25$ agents initialized on a fixed 5×5 spatial initialization grid (G) within a unit square $[0,1]^2$. The agents operate within a state space $\mathcal{S} = \{0,1,2\}$ and an action space $\mathcal{A} = \{0,1,2\}$, corresponding to *idle* (0), *transit* (1),



(a) Robotics control performance: average cumulative discounted reward V^{π} against subsample size κ (30 runs).



(b) Computational complexity scaling: Q-table size (total entries) as a function of the subsample size κ.

Figure 3: Performance-scalability tradeoff of GMFS. (Left) GMFS rapidly achieves near-optimal performance in the robotics coordination task starting from around $\kappa = 8$, approaching the full graphon mean-field baseline at $\kappa = 24$ (which corresponds to the optimal solution obtained without sampling) (22, 33). (Right) The computational cost, measured by the number of entries in the discrete neighborhood state space $\mathcal{Z}_\kappa$, grows polynomially in κ.

and *working* (2) states. Actions represent the agent's intended next state. Each evaluation run is conducted over a horizon of $H = 100$ time steps.

Compute. Experiments were implemented in Python and ran on a high-performance computing node with Intel CPUs and 1.0 TiB of DDR5 ECC RAM. We parallelized the training across subsampling parameters; the total suite took approximately 20 minutes to execute, with the exhaustive mean-field limit ($\kappa = 24$) requiring approximately 440 seconds for training.

A.4.2 Dynamics and Rewards. We define the interaction topology using a radial graphon $W(x, y)$ connecting agents based on their fixed latent positions $\alpha_i, \alpha_j \in [0, 1]^2$: $W(\alpha_i, \alpha_j) = \mathbb{1}(\|\alpha_i - \alpha_j\|_2 \leq r)$, with an interaction radius $r = 0.3$. We row-normalize these weights to induce the sampling distribution $\bar{w}_{ij} = W_{ij} / \sum_{l \neq i} W_{il}$. To handle the edge case of an isolated agent (where the row sum is zero), we define $\bar{w}_{ij}$ as the uniform distribution over all other agents $[n] \setminus \{i\}$.

The stochastic transition kernel $P(s' \mid s, a, \mu)$ simulates physical bottlenecks by making transitions to the working state ($a = 2$) congestion-dependent; the success probability is defined as $P(s' = 2 \mid a = 2, \mu) = \max(0.1, 0.9 - 0.8 \cdot \mu(2))$, with failures reverting the agent to the transit state ($s' = 1$). For all other actions ($a \in \{0, 1\}$), the agent successfully transitions to the intended state with probability 0.9, i.e., $P(s' = a \mid a, \mu) = 0.9$ and with a 0.1 probability of failure leaving the agent's state unchanged ($s' = s$).

To satisfy the Lipschitz and positive rewards assumptions in our framework, we define the reward function as:

$$r(s, a, \mu) = V(s) \cdot \max(0.4, 1.0 - L \cdot \mu(2)) - C(a). \tag{1}$$

The state values $V(s)$ are set to 10, 5, and 20 for the *idle*, *transit*, and *working* states, respectively. Action costs $C(a)$ are 0 for *idle* and *transit* and 5.0 for the *working* state. We set the congestion sensitivity to $L = 5.0$ with a minimum utility multiplier of 0.4.

A.4.3 GMFS Algorithm Configuration. We learn the optimal Q-function using offline value iteration as described. In order to maintain computational tractability for large κ, we use the state-marginal histograms, which reduces the neighborhood state space $\mathcal{Z}_\kappa$ from $\binom{\kappa+8}{8}$ (joint) to $\binom{\kappa+2}{2}$ (marginal) states. This is because the environment's rewards and transitions depend only on the neighborhood state marginal μ, so the state-marginal histogram is a sufficient statistic for the optimal Q-function in this setting. We check convergence using the Bellman error $\Delta = \|Q_{t+1} - Q_t\|_\infty$ and find that all runs reach a stable fixed point ($\Delta < 10^{-4}$) in 250 iterations. The hyperparameters used are detailed in Table 1.

B MATHEMATICAL BACKGROUND AND ADDITIONAL REMARKS

In this section, we focus on the mathematical foundations and auxiliary results required for our analysis. We first establish the contraction properties and uniform boundedness of the Bellman operators to ensure the existence and uniqueness of the optimal Q-functions. We also

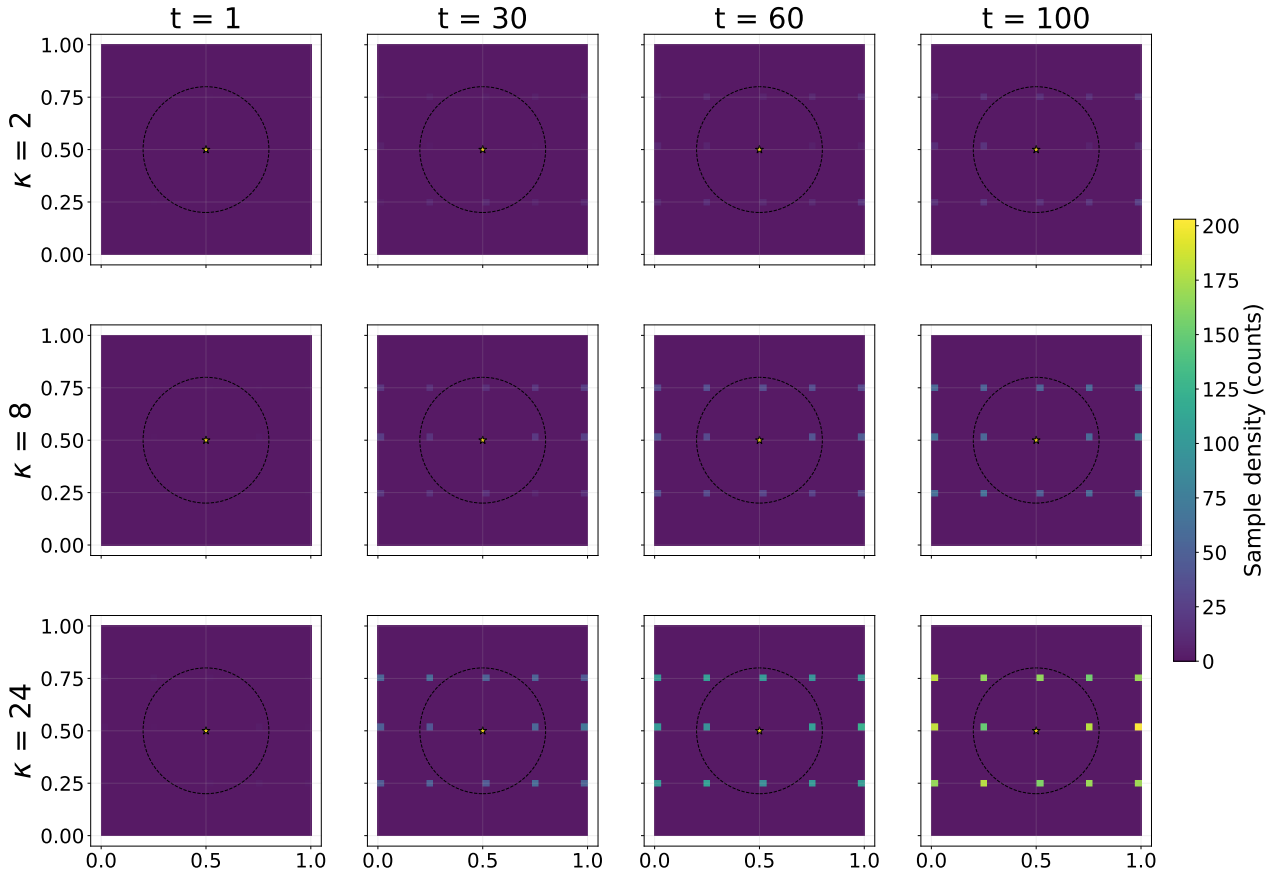


Figure 4: Perception time-evolution comparison for the focal agent (center of the 5×5 grid) under the radial graphon. Rows correspond to $\kappa \in \{2, 8, 24\}$, while columns correspond to time horizons $t \in \{1, 30, 60, 100\}$. Each panel aggregates the focal agent's sampled neighbors up to time t, with the dashed circle indicating the true interaction ball. As κ increases, the empirical neighborhood density converges faster and more uniformly to the support of the radial graphon.

Hyperparameter	Value
Discount Factor (γ)	0.95
Training Iterations (T)	250
Subsample Sizes (κ)	$\{1, 3, 6, 9, 12, 15, 18, 21, 24\}$
Monte Carlo Samples (M)	50 (per Bellman update)
Learning Rate (α)	1.0 (Exact Operator Update)
Exploration Rate (ϵ)	0.0 (Exhaustive Offline Sweep)
Independent Runs	30 seeds per κ

Table 1: GMFS Training Configuration

present an identity-aware sampling algorithm to contrast the efficiency of our mean-field reparameterization against an exhaustive labeled state representation.

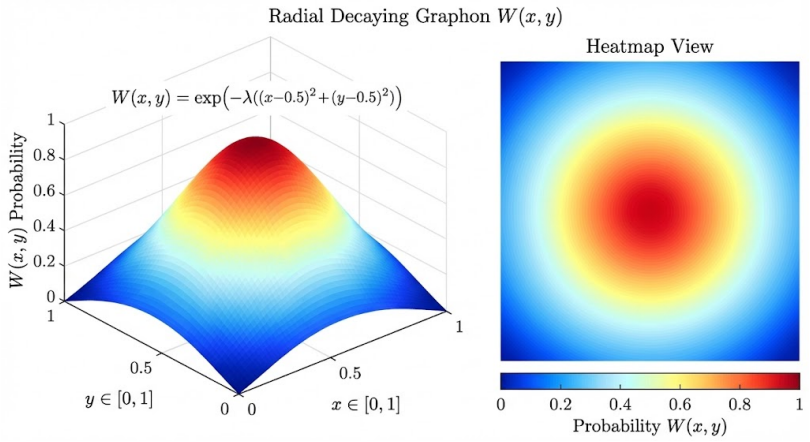


Figure 5: Visualization of a radial graphon. We note that generative AI was used to refine the aesthetics of this figure.

Definition B.1 (Lipschitz continuity)**.** *Given metric spaces $(\mathcal{X}, d_1)$ and $(\mathcal{Y}, d_2)$ and a constant $L > 0$, a map $f : \mathcal{X} \to \mathcal{Y}$ is L-Lipschitz continuous if for all $x, y \in \mathcal{X}$, $d_2(f(x), f(y)) \le L d_1(x, y)$.*

Theorem B.1. *(Banach-Caccioppoli Fixed Point Theorem (5)). Let (X, d) be a non-empty complete metric space with a γ-contraction mapping $T : X \to X$ for $\gamma \in (0, 1)$. Then, T admits a unique fixed point $x^* \in X$ such that $T(x^*) = x^*$. Moreover, for any $x_0 \in X$, $T^n(x_0) \to x^*$ as $n \to \infty$.*

Lemma B.2 (The Bellman operator and sampled Bellman operator are γ-contractions)**.** *Fix $1 \le \kappa \le n-1$. Let $\mathcal{Y}_\kappa \coloneqq \{Q : \mathcal{S} \times \mathcal{A} \times \mathcal{Z}_\kappa \to \mathbb{R}\}$ be equipped with the sup norm $\|Q\|_\infty \coloneqq \sup_{(s,a,z)} |Q(s,a,z)|$. Suppose $\gamma \in (0,1)$. Then the Bellman operator $\widehat{\mathcal{T}}_\kappa : \mathcal{Y}_\kappa \to \mathcal{Y}_\kappa$ admits a unique fixed point Q^* satisfying $\widehat{\mathcal{T}}_\kappa Q^* = Q^*$.*

Proof. We show that $\widehat{\mathcal{T}}_\kappa$ is a γ-contraction such that for all $Q_1, Q_2 \in \mathcal{Y}_\kappa$, $\|\widehat{\mathcal{T}}_\kappa Q_1 - \widehat{\mathcal{T}}_\kappa Q_2\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty$. Fix any $Q_1, Q_2 \in \mathcal{Y}_\kappa$ and any $(s,a,z) \in \mathcal{S} \times \mathcal{A} \times \mathcal{Z}_\kappa$, we have

$$
\begin{aligned}
&|\widehat{\mathcal{T}}_\kappa Q_1(s,a,z) - \widehat{\mathcal{T}}_\kappa Q_2(s,a,z)| \\
&\quad = \left| r_\ell(s,a,g_z) + \gamma \mathbb{E}_{(s',g') \sim \mathcal{J}_\kappa(\cdot|s,a,z)} [\mathcal{M}_\kappa Q_1(s',g')] - r_\ell(s,a,g_z) - \gamma \mathbb{E}_{(s',g') \sim \mathcal{J}_\kappa(\cdot|s,a,z)} [\mathcal{M}_\kappa Q_2(s',g')] \right| \\
&\quad = \left| \gamma \mathbb{E}_{(s',g')} \left[\max_{a' \in \mathcal{A}, z' \in \Gamma_\kappa(g')} Q_1(s',a',z') \right] - \gamma \mathbb{E}_{(s',g')} \left[\max_{a' \in \mathcal{A}, z' \in \Gamma_\kappa(g')} Q_2(s',a',z') \right] \right| \\
&\quad \le \gamma \mathbb{E}_{(s',g')} \left| \max_{a' \in \mathcal{A}, z' \in \Gamma_\kappa(g')} Q_1(s',a',z') - \max_{a' \in \mathcal{A}, z' \in \Gamma_\kappa(g')} Q_2(s',a',z') \right| \\
&\quad \le \gamma \|Q_1 - Q_2\|_\infty,
\end{aligned}
$$

where the the third line follows by Jensen's inequality, and the last line follows by the 1-Lipschitzness of the max operator under ℓ_∞-norm. Taking the supremum over (s,a,z) yields $\|\widehat{\mathcal{T}}_\kappa Q_1 - \widehat{\mathcal{T}}_\kappa Q_2\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty$. Since $(\mathcal{Y}_\kappa, \|\cdot\|_\infty)$ is complete, Banach's fixed point theorem implies that $\widehat{\mathcal{T}}_\kappa$ has a unique fixed point. The result follows for the original Bellman operator by taking $\kappa = n-1$, where $\widehat{\mathcal{T}}_{n-1} = \mathcal{T}$. □

Lemma B.3 (The empirical sampled Bellman operator is a γ-contraction)**.** *Fix $1 \le \kappa \le n-1$. Let $\mathcal{Y}_\kappa \coloneqq \{Q : \mathcal{S} \times \mathcal{A} \times \mathcal{Z}_\kappa \to \mathbb{R}\}$ be equipped with the sup norm $\|Q\|_\infty \coloneqq \sup_{(s,a,z)} |Q(s,a,z)|$. Suppose $\gamma \in (0,1)$. Then the sampled Bellman operator $\widehat{\mathcal{T}}_{\kappa,m} : \mathcal{Y}_\kappa \to \mathcal{Y}_\kappa$ admits a unique fixed point Q^* satisfying $\widehat{\mathcal{T}}_{\kappa,m} \widehat{Q}^*_{\kappa,m} = \widehat{Q}^*_{\kappa,m}$.*

Proof. We show that $\widehat{\mathcal{T}}_{\kappa,m}$ is a γ-contraction such that for all $Q_1, Q_2 \in \mathcal{Y}_\kappa$, $\|\widehat{\mathcal{T}}_{\kappa,m} Q_1 - \widehat{\mathcal{T}}_{\kappa,m} Q_2\|_\infty \le \gamma \|Q_1 - Q_2\|_\infty$. Fix any $Q_1, Q_2 \in \mathcal{Y}_\kappa$ and any $(s,a,z) \in \mathcal{S} \times \mathcal{A} \times \mathcal{Z}_\kappa$, we have

$$
\begin{aligned}
&|\widehat{\mathcal{T}}_{\kappa,m} Q_1(s,a,z) - \widehat{\mathcal{T}}_{\kappa,m} Q_2(s,a,z)| \\
&\quad = \left| r_\ell(s,a,g_z) + \frac{\gamma}{m} \sum_{\ell=1}^m \mathcal{M}_\kappa Q_1(s',g') - r_\ell(s,a,g_z) - \frac{\gamma}{m} \sum_{\ell=1}^m \mathcal{M}_\kappa Q_2(s',g') \right| \\
&\quad \le \frac{\gamma}{m} \sum_{\ell=1}^m \left| \max_{a' \in \mathcal{A}, z' \in \Gamma_\kappa(g')} Q_1(s',a',z') - \max_{a' \in \mathcal{A}, z' \in \Gamma_\kappa(g')} Q_2(s',a',z') \right| \\
&\quad \le \gamma \|Q_1 - Q_2\|_\infty,
\end{aligned}
$$

where the the second line follows by triangle inequality, and the last line follows by the 1-Lipschitzness of the max operator under ℓ_∞-norm. Taking the supremum over (s, a, z) yields $\|\widehat{\mathcal{T}}_{\kappa,m}Q_1 - \widehat{\mathcal{T}}_{\kappa,m}Q_2\|_\infty \le \gamma\|Q_1 - Q_2\|_\infty$. Since $(\mathcal{Y}_\kappa, \|\cdot\|_\infty)$ is complete, Banach's fixed point theorem implies that $\widehat{\mathcal{T}}_{\kappa,m}$ has a unique fixed point. Labeling the fixed point $\widehat{Q}^*_{\kappa,m}$ completes the proof. □

We next show that the Q-function is bounded throughout its iterations.

Lemma B.4. *For all $T \ge 0$, $\|Q^T\|_\infty \le \frac{\|r\|_\infty}{1-\gamma}$.*

Proof. The proof follows by induction on T. The base case follows as $Q^0 \coloneqq 0$. For the induction, note that by triangle inequality $\|Q^{T+1}\|_\infty \le \|r_\ell\|_\infty + \gamma\|Q^T\|_\infty \le \|r_\ell\|_\infty + \gamma\frac{\|r_\ell\|_\infty}{1-\gamma} = \frac{\|r_\ell\|_\infty}{1-\gamma}$, which proves the claim. □

Corollary B.5. *Observe by recursively using the γ-contractive property for T time steps, with the initializations $\widehat{Q}_\kappa = 0$ and $\widehat{Q}_{\kappa,m} = 0$, and the bounds $\|\widehat{Q}^*_\kappa\|_\infty \le \frac{\|r_\ell\|_\infty}{1-\gamma}$ and $\|\widehat{Q}^*_{\kappa,m}\|_\infty \le \frac{\|r_\ell\|_\infty}{1-\gamma}$ from Lemma B.4, that*

$$\|\widehat{Q}^*_\kappa - \widehat{Q}^T_\kappa\|_\infty \le \gamma^T \cdot \|\widehat{Q}^*_\kappa - \widehat{Q}^0_\kappa\|_\infty \le \gamma^T \frac{\|r_\ell\|_\infty}{1-\gamma}, \tag{2}$$

and

$$\|\widehat{Q}^*_{\kappa,m} - \widehat{Q}^T_{\kappa,m}\|_\infty \le \gamma^T \cdot \|\widehat{Q}^*_{\kappa,m} - \widehat{Q}^0_{\kappa,m}\|_\infty \le \gamma^T \frac{\|r_\ell\|_\infty}{1-\gamma}. \tag{3}$$

Remark B.6. *corollary B.5 characterizes the error decay between $\widehat{Q}^T_\kappa$ and $\widehat{Q}^*_\kappa$ and shows that it decays exponentially in the number of Bellman iterations by a γ^T multiplicative factor.*

B.1 Identity-Aware Sampling Algorithm and Analyses

We present an identity-aware subsampling algorithm that explicitly retains the indices of sampled neighbors within the state representation. We provide this as a simple baseline to demonstrate how subsampling and graphon-based weights can be used to construct an unbiased estimator of neighborhood-dependent rewards and transitions. We use this baseline to motivate the statistical guarantees of our main algorithm; specifically, we show that while identity-aware representations provide unbiasedness, their state-action space cardinality grows exponentially with κ. This motivates the mean-field reparameterization introduced in the main text.

Algorithm 3 Offline Learning (Identity-aware subsampling + Graphon-based Importance Sampling)

Require: Subsampling parameter $\kappa \ge 1$, iterations T, Monte Carlo samples m, proposal distribution q_i on $[n] \setminus \{i\}$, and a generative simulator for the n-agent system.
Define the labeled neighborhood space $\mathcal{Y} \coloneqq ([n] \setminus \{i\})^\kappa \times (\mathcal{S} \times \mathcal{A})^\kappa$.
Initialize $\widehat{Q}^{(0)}_\kappa(s, a, y) = 0$ for all $(s, a, y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{Y}$.
for $t = 0, 1, \ldots, T-1$ **do**
  for $(s, a, y) \in \mathcal{S} \times \mathcal{A} \times \mathcal{Y}$ **do**
    Extract sampled neighbor indices $\{J^{(1)}, \ldots, J^{(\kappa)}\}$ and their observed pairs $\{(s_j, a_j)\}_{j \in \{J^{(m)}\}}$ from y.
    Compute importance sampling weights $\rho^{(m)} = \bar{w}_{i,J^{(m)}} / q_i(J^{(m)})$ for $m = 1, \ldots, \kappa$.
    Construct the importance sampling estimate $\widehat{g}$ (or $\widehat{Z}$) using y and ρ.
    Sample $(s'_\ell, y'_\ell) \overset{\text{i.i.d.}}{\sim} \widetilde{\mathcal{J}}(\cdot \mid s, a, y)$ for $\ell = 1, \ldots, m$ {Induced kernel in the labeled space}.
    Update: $\widehat{Q}^{(t+1)}_\kappa(s, a, y) \leftarrow r_\ell(s, a, \widehat{g}) + \frac{\gamma}{m}\sum_{\ell=1}^m \max_{a' \in \mathcal{A}} \widehat{Q}^{(t)}_\kappa(s'_\ell, a', y'_\ell)$.
Return $\widehat{Q}^{(T)}_\kappa$.

Algorithm analysis. In Algorithm 3, the Monte Carlo samples can be derived from standard MCMC or uniform sampling methods (2, 12, 56, 63, 71). The main limitation of this approach is the dimensionality of the augmented state space. The number of possible neighborhood tuples in $\mathcal{Y}$ scales as $(|\mathcal{S}||\mathcal{A}|)^\kappa$. Consequently, the cardinality of the tabular identity-aware Q-function is $|\mathcal{S}||\mathcal{A}| \cdot (|\mathcal{S}||\mathcal{A}|)^\kappa = (|\mathcal{S}||\mathcal{A}|)^{\kappa+1}$.

The exponential dependence on the subsampling parameter κ motivates the mean-field reparameterization used throughout the GMFS model. Instead of learning on the labeled space $\mathcal{Y}$, we map a sampled labeled neighborhood y to an unlabeled empirical distribution $\widehat{Z}^\kappa_i$. This representation discards explicit neighbor identities but retains the aggregate information required for the reward and transition dynamics. The reparameterization thus makes learning tractable by replacing an exponentially large labeled neighborhood space with a distributional state summary whose complexity depends on $|\mathcal{S}||\mathcal{A}|$ instead.

Definition B.2 (Horvitz-Thompson estimator)**.** *Let q be a proposal distribution on a finite set $\mathcal{J}$ and let w be a target weight vector with $w(j) \geq 0$ such that $\sum_{j\in\mathcal{J}} w(j) = 1$, and $q(j) > 0$ whenever $w(j) > 0$. For any function $\phi : \mathcal{J} \to \mathbb{R}$, sample $J_1, \ldots, J_\kappa \overset{i.i.d.}{\sim} q$ and define the Horvitz-Thompson estimator*

$$\widehat{\Phi}_\kappa(\phi) \coloneqq \frac{1}{\kappa}\sum_{m=1}^{\kappa} \frac{w(J_m)}{q(J_m)}\phi(J_m). \tag{4}$$

Then $\mathbb{E}[\widehat{\Phi}_\kappa(\phi)] = \sum_{j\in\mathcal{J}} w(j)\phi(j)$.

Lemma B.7 (Unbiasedness of the neighborhood estimator)**.** *Assume $q_i(j) > 0$ for all j with $\bar{w}_{ij} > 0$. Next, define the importance sampling weights $\rho^{(m)}$ such that*

$$\rho^{(m)} = \frac{\bar{w}_{iJ^{(m)}}}{q_i(J^{(m)})}. \tag{5}$$

Then the importance weighted estimator $\widehat{Z}_i^{(\kappa)}(s,a)$ is an unbiased estimator of the true neighborhood aggregate $Z_i(s,a) = \sum_{j\neq i} \bar{w}_{ij}\mathbb{1}\{s_j = s, a_j = a\}$ under proposal distribution q_i.

Proof. By linearity of expectations, we have

$$\begin{aligned}\mathbb{E}[\widehat{Z}_i^{(\kappa)}(s,a)] &= \mathbb{E}\Big[\frac{1}{\kappa}\sum_{m=1}^{\kappa}\frac{\bar{w}_{iJ^{(m)}}}{q_i(J^{(m)})}\mathbb{1}\{s_{J_i^{(m)}} = s, a_{J_i^{(m)}} = a\}\Big]\\ &= \frac{1}{\kappa}\sum_{m=1}^{\kappa}\mathbb{E}\Big[\frac{\bar{w}_{iJ^{(m)}}}{q_i(J^{(m)})}\mathbb{1}\{s_{J_i^{(m)}} = s, a_{J_i^{(m)}} = a\}\Big],\end{aligned}$$

where

$$\mathbb{E}\Big[\frac{\bar{w}_{iJ^{(m)}}}{q_i(J^{(m)})}\mathbb{1}\{s_{J_i^{(m)}} = s, a_{J_i^{(m)}} = a\}\Big] = \mathbb{E}\Big[\frac{\bar{w}_{iJ}}{q_i(J)}\mathbb{1}\{s_J = s, a_J = a\}\Big]$$

as the $J_i^{(m)}$'s are drawn i.i.d. Therefore,

$$\begin{aligned}\mathbb{E}[\widehat{Z}_i^{(\kappa)}(s,a)] &= \mathbb{E}\Big[\frac{\bar{w}_{iJ}}{q_i(J)}\mathbb{1}\{s_J = s, a_J = a\}\Big]\\ &= \sum_{j\neq i} q_i(j)\frac{\bar{w}_{ij}}{q_i(j)}\mathbb{1}\{s_j = s, a_j = a\}\\ &= \sum_{j\neq i}\bar{w}_{ij}\mathbb{1}\{s_j = s, a_j = a\}\\ &= Z_i(s,a),\end{aligned}$$

where the second equality follows by using $J \in [n]\backslash\{i\}$. We conclude that since $\mathbb{E}\big[\widehat{Z}_i^{(\kappa)}(s,a)\big] = Z_i(s,a)$, the estimator is unbiased and thus a Horvitz-Thompson estimator.

□

Lemma B.8 (Horvitz-Thompson unbiasedness for identity-aware neighborhoods)**.** *Let q_i be a proposal distribution on $[n]\backslash\{i\}$, such that $q_i(j) > 0$ for all j with $\bar{w}_{ij} > 0$. Define the importance weights:*

$$\rho^{(m)} = \frac{\bar{w}_{iJ^{(m)}}}{q_i(J^{(m)})}$$

Then for any function $\varphi : ([n]\backslash\{i\}) \times \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, the Horvitz-Thompson estimator

$$\widehat{\Phi}_i^{(\kappa)}(\varphi) = \frac{1}{\kappa}\sum_{m=1}^{\kappa}\rho^{(m)}\varphi(J_i^{(m)}, s_{J_i^{(m)}}, a_{J_i^{(m)}}) \tag{6}$$

is unbiased for the graphon-weighted neighborhood: $\Phi_i(\varphi) = \sum_{j\neq i}\bar{w}_{ij}\varphi(j, s_j, a_j)$, i.e. $\mathbb{E}[\widehat{\Phi}_i^{(\kappa)}(\varphi)] = \Phi_i(\varphi)$

Proof. By linearity of expectation and i.i.d. sampling, we have

$$\begin{aligned}\mathbb{E}[\widehat{\Phi}_i^{(\kappa)}(\varphi)] &= \mathbb{E}\Big[\frac{\bar{w}_{ij}}{q_i(j)}\varphi(J, s_J, a_J)\Big]\\ &= \sum_{j\neq i} q_i(j)\frac{\bar{w}_{ij}}{q_i(j)}\varphi(j, s_j, a_j)\end{aligned}$$

$$= \sum_{j \neq i} \bar{w}_{ij} \varphi(j, s_j, a_j)$$
$$= \Phi_i(\varphi)$$

which proves the claim. □

C LIPSCHITZ CONTINUITY IN THE MEAN-FIELD MEASURE

When no subsampling occurs (at $\kappa = n - 1$), GMFS recovers the graphon mean-field MARL formulation corresponding to interaction graphon W. As $\kappa \to n - 1$, we show that $\widehat{Q}^*_\kappa \to Q^*$ via a Lipschitz continuity bound between Q^* and $\widehat{Q}^*_\kappa$.

Definition C.1 (Total Variation Distance)**.** *Let P and Q be discrete probability distributions over some domain Ω. Then,*

$$\mathrm{TV}(P, Q) = \frac{1}{2}\|P - Q\|_1 = \sup_{E \subseteq \Omega} \left| \Pr_P(E) - \Pr_Q(E) \right|.$$

Definition C.2 (Mixing measures)**.** *Fix $\kappa \in [n-1]$. For any fixed agent $i \in [n]$ and a sampled multiset $\Delta_i = (J_i^{(1)}, \ldots, J_\kappa^{(\kappa)})$, we define a probability measure $\mu_i^{[n]}$ on $\mathcal{S} \times \mathcal{A}$ and an empirical measure $\widehat{\mu}_{\Delta_i}$ on $\mathcal{S} \times \mathcal{A}$ by*

$$\mu_i^{[n]}(x, u) = \sum_{j \neq i} \bar{w}_{ij} \mathbb{1}\{s_j = x, a_j = u\}, \qquad \widehat{\mu}_{\Delta_i}(x, u) = \frac{1}{\kappa} \sum_{m=1}^{\kappa} \mathbb{1}\{s_{J_i^{(m)}} = x, a_{J_i^{(m)}} = u\}.$$

Theorem C.1 (Lipschitz Continuity of the Bellman iterates). *Fix a subsampling parameter $\kappa \geq 1$. Fix $(s, a) \in \mathcal{S} \times \mathcal{A}$, and let $z \in \mathcal{Z} \coloneqq \Delta(\mathcal{S} \times \mathcal{A})$. Let $\widehat{z} \in \mathcal{Z}_\kappa$ be the empirical histogram of κ i.i.d. draws from z, and let $g_z, g_{\widehat{z}}$ be the marginals in $\mathcal{S}$. Then, for all $t \in \mathbb{N}$, we have*

$$\left| Q^t(s, a, z) - \widehat{Q}^t_\kappa(s, a, \widehat{z}) \right| \leq \frac{4\|r_\ell\|_\infty}{1 - \gamma} L_P \cdot \mathrm{TV}(g_z, g_{\widehat{z}}).$$

Proof. We proceed by induction on t. At $t = 0$, we have $Q^0(s, a, z) = \widehat{Q}^0_\kappa(s, a, \widehat{z}) = 0$. At $t = 1$,

$$\begin{aligned}
|Q^1(s, a, z) - \widehat{Q}^1_\kappa(s, a, \widehat{z})| &= |\mathcal{T} Q^0(s, a, z) - \widehat{\mathcal{T}}_\kappa \widehat{Q}^0_\kappa(s, a, \widehat{z})| \\
&= \left| r_\ell(s, a, g_z) + \gamma \mathbb{E}_{\mathcal{J}_n} \max_{a', z'} Q^{(0)}(\cdot) - r_\ell(s, a, g_{\widehat{z}}) - \gamma \mathbb{E}_{\mathcal{J}_\kappa} \max_{a', \widehat{z}'} \widehat{Q}^{(0)}(\cdot) \right| \\
&= |r_\ell(s, a, g_z) - r_\ell(s, a, g_{\widehat{z}})| \\
&\leq 2\|r_\ell\|_\infty \cdot \mathrm{TV}(g_z, g_{\widehat{z}}),
\end{aligned}$$

where the last inequality follows by Assumption 3.4. This proves the base case. For $t + 1$:

$$\begin{aligned}
\left| Q^{(t+1)}(s, a, z) - \widehat{Q}^{(t+1)}_\kappa(s, a, \widehat{z}) \right| &= |r_\ell(s, a, g_z) + \gamma \mathbb{E}_{s', g' \sim J_n}[M_{n-1} Q(s', g')] - r_\ell(s, a, g_{\widehat{z}}) - \gamma \mathbb{E}_{s', \widehat{g}' \sim \mathcal{J}_\kappa}[M_\kappa \widehat{Q}_\kappa(s', \widehat{g}')]| \\
&\leq \underbrace{|r_\ell(s, a, g_z) - r_\ell(s, a, g_{\widehat{z}})|}_{\text{Term (I)}} + \gamma \underbrace{|\mathbb{E}_{s', g' \sim J_n}[M_{n-1} Q(s', g')] - \mathbb{E}_{s', \widehat{g}' \sim \mathcal{J}_\kappa}[M_\kappa \widehat{Q}_\kappa(s', \widehat{g}')]|}_{\text{Term (II)}}
\end{aligned}$$

By Assumption 3.4, Term (I) is bounded by $L_r \cdot \mathrm{TV}(\widehat{g}, g)$. We bound Term (II) by Lemma C.2,

$$|\mathbb{E}_{s', g' \sim J_n}[M_{n-1} Q(s', g')] - \mathbb{E}_{s', \widehat{g}' \sim \mathcal{J}_\kappa}[M_\kappa \widehat{Q}_\kappa(s', \widehat{g}')]| \leq \frac{4\|r_\ell\|_\infty}{1 - \gamma} L_P \cdot \mathrm{TV}(g_z, g_{\widehat{z}})$$

Then, recalling that $L_P \geq 1$ by Assumption 3.5, we have

$$\begin{aligned}
\left| Q^{(t+1)}(s, a, z) - \widehat{Q}^{(t+1)}_\kappa(s, a, \widehat{z}) \right| &\leq 2\|r_\ell\|_\infty \cdot \mathrm{TV}(g_z, g_{\widehat{z}}) + \frac{4\gamma\|r_\ell\|_\infty}{1 - \gamma} L_P \cdot \mathrm{TV}(g_z, g_{\widehat{z}}) \\
&\leq 4\|r_\ell\|_\infty \cdot L_P \cdot \mathrm{TV}(g_z, g_{\widehat{z}}) + \frac{4\gamma\|r_\ell\|_\infty}{1 - \gamma} L_P \cdot \mathrm{TV}(g_z, g_{\widehat{z}}) \\
&\leq \frac{4\|r_\ell\|_\infty}{1 - \gamma} \cdot L_P \cdot \mathrm{TV}(g_z, g_{\widehat{z}}),
\end{aligned}$$

which proves the claim. □

Definition C.3 (One-step kernel)**.** *Let $J_n(\cdot|s, a, z)$ denote the induced one-step distribution on $(s', g') \in \mathcal{S} \times \mathcal{G}$ under the n-agent dynamics, for a uniformly random agent, given $(s, a, z) \in \mathcal{S} \times \mathcal{A} \times \mathcal{Z}$.*

Definition C.4 (Surrogate one-step kernel for GMFS)**.** *For $\kappa \in \{1, \dots, n-1\}$, let $J_\kappa(\cdot|s,a,\widehat{z})$ denote the induced one-step distribution on $(s', \widehat{g}') \in \mathcal{S} \times \mathcal{G}_\kappa$ under the $(\kappa+1)$-agent surrogate dynamics, for a uniformly random agent, given $(s,a,\widehat{z}) \in \mathcal{S} \times \mathcal{A} \times \widehat{\mathcal{Z}}_\kappa$.*

For Lemma C.2, we want to show that the value term is Lipschitz in the mean-field argument, which later allows us to control the discrepancy between the original and subsampled Bellman operators.

Lemma C.2 (Lipschitz-continuity of the expected values of the Bellman iterates)**.** *Fix a subsampling parameter $\kappa \geq 1$. Fix $(s,a) \in \mathcal{S} \times \mathcal{A}$, and let $z \in \mathcal{Z} \coloneqq \Delta(\mathcal{S} \times \mathcal{A})$. Let $\widehat{z} \in \mathcal{Z}_\kappa$ be the empirical histogram of κ i.i.d. draws from z, and let $g_z, g_{\widehat{z}}$ be the marginals in $\mathcal{S}$. Then, for all $t \in \mathbb{N}$, we have*

$$|\mathbb{E}_{(s',g')\sim\mathcal{J}_n(\cdot|s,a,z)} \max_{a',z'} Q^t(s',a',z') - \mathbb{E}_{(s',\widehat{g}')\sim\mathcal{J}_\kappa(\cdot|s,a,\widehat{z})} \max_{a',\widehat{z}'} \widehat{Q}^t_\kappa(s',a',\widehat{z}')| \leq \frac{4\|r_\ell\|_\infty}{1-\gamma} L_P \cdot \mathrm{TV}(g_z, g_{\widehat{z}}),$$

where $\mathcal{J}$ is the joint stochastic kernel that generates the next state.

Proof. We proceed by induction. For the base case where $t = 0$, both Q^0 and $\widehat{Q}^0_\kappa$ are identically zero, hence the inequality holds trivially. When $t = 1$, we first note that $Q^1(s,a,z) = r_\ell(s,a,g_z)$ and we let $\widetilde{r}_*(s,g_z) = \max_a r_\ell(s,a,g_z)$. Then, we have $M_{n-1}Q^1(s',g_{z'}) = \widetilde{r}_*(s',g_{z'})$ and $M_\kappa \widehat{Q}^1_\kappa = \widetilde{r}_*(s',g_{\widehat{z}'})$. Therefore,

$$\begin{aligned}
&\left|\mathbb{E}_{(s',g')\sim\mathcal{J}_n} \max_{a',z'} [r(s',a',g_{z'})] - \mathbb{E}_{(s',\widehat{g}')\sim\mathcal{J}_\kappa} \max_{a',\widehat{z}'} [r(s',a',g_{\widehat{z}'})]\right| \\
&\qquad = \left|\mathbb{E}_{(s',g')\sim\mathcal{J}_n} \widetilde{r}_*(s',g_{z'}) - \mathbb{E}_{(s',\widehat{g}')\sim\mathcal{J}_\kappa} \widetilde{r}_*(s',g_{\widehat{z}'})\right| \\
&\qquad \leq 2\|\widetilde{r}_*\|_\infty \cdot \mathrm{TV}(J_n(\cdot|s,a,g_z), J_\kappa(\cdot|s,a,g_{\widehat{z}})) \leq 4\|r_\ell\|_\infty \cdot L_P \cdot \mathrm{TV}(g_z, g_{\widehat{z}}).
\end{aligned}$$

where the first inequality uses Lemma C.6, and the second inequality follows by triangle inequality and Assumption 3.5. Assume for $t \geq 1$:

$$\left|\mathbb{E}_{(s',g')\sim\mathcal{J}_n(\cdot|s,a,g)} [M_{n-1}Q^t(s',g')] - \mathbb{E}_{(s',\widehat{g}')\sim\mathcal{J}_\kappa(\cdot|s,a,\widehat{g})} [M_\kappa \widehat{Q}^T_\kappa(s',\widehat{g}')]\right| \leq \frac{4\|r_\ell\|_\infty}{1-\gamma} L_P \cdot \mathrm{TV}(g_z, g_{\widehat{z}'}).$$

Then, for the inductive step, using the Bellman updates for Q^{t+1} and $\widehat{Q}^{t+1}_\kappa$, we write

$$\begin{aligned}
&\left|\mathbb{E}_{(s',g')\sim\mathcal{J}_n} \max_{a',z'} Q^{t+1}(s',a',z') - \mathbb{E}_{(s',\widehat{g}')\sim\mathcal{J}_\kappa} \max_{a',\widehat{z}'} \widehat{Q}^{t+1}_\kappa(s',a',\widehat{z}')\right| \\
&\qquad = \Big|\mathbb{E}_{(s',g')\sim\mathcal{J}_n} \max_{a',z'} \left[r_\ell(s',a',g_{z'}) + \gamma \mathbb{E}_{(s'',g'')\sim\mathcal{J}_n} \max_{a'',z''} Q^t(s'',a'',z'')\right] \\
&\qquad\quad - \mathbb{E}_{(s',\widehat{g}')\sim\mathcal{J}_\kappa} \max_{a',\widehat{z}'} \left[r_\ell(s',a',g_{\widehat{z}'}) + \gamma \mathbb{E}_{(s'',\widehat{g}'')\sim\mathcal{J}_\kappa} \max_{a'',\widehat{z}''} \widehat{Q}^t_\kappa(s'',a'',\widehat{z}'')\right]\Big|
\end{aligned}$$

Then, using the fact that $\max(\cdot)$ is 1–Lipschitz, we have

$$\begin{aligned}
&\max_{a',z'} Q^{t+1}(s',a',z') - \max_{a',\widehat{z}'} \widehat{Q}^{t+1}_\kappa(s',a',\widehat{z}') \\
&\leq \max_{a',z',\widehat{z}'} \left|r_\ell(s',a',g_{z'}) - r_\ell(s',a',g_{\widehat{z}'}) + \gamma\left(\mathbb{E}_{(s'',g'')\sim\mathcal{J}_n} \max_{a'',z''} Q^t(s'',a'',z'') - \mathbb{E}_{(s''_i,\widehat{g}'')\sim\mathcal{J}_\kappa} \max_{a'',\widehat{z}''} \widehat{Q}^t_\kappa(s'',a'',\widehat{z}'')\right)\right| \\
&\leq \max_{a'} \left|r_\ell(s',a',g_{z'}) - r_\ell(s',a',g_{\widehat{z}'})\right| + \gamma \max_{a',z',\widehat{z}'} \left|\mathbb{E}_{(s'',g'')\sim\mathcal{J}_n} [\max_{a'',z''} Q^t(s'',a'',z'')] - \mathbb{E}_{(s'',\widehat{g}'')\sim\mathcal{J}_\kappa} \max_{a'',\widehat{z}''} [\widehat{Q}^t_\kappa(s'',a'',\widehat{z}'')]\right|
\end{aligned}$$

Now taking our original expectations over $\mathcal{J}_n$ and $\mathcal{J}_\kappa$, triangle inequality gives us:

$$\begin{aligned}
&\Big|\mathbb{E}_{(s',g')\sim\mathcal{J}_n} \max_{a',z'} \left[r_\ell(s',a',g_{z'}) + \gamma \mathbb{E}_{(s'',g'')\sim\mathcal{J}_n} \max_{a'',z''} Q^t(s'',a'',z'')\right] \\
&\qquad\qquad - \mathbb{E}_{(s',\widehat{g}')\sim\mathcal{J}_\kappa} \max_{a'} \left[r_\ell(s',a',g_{\widehat{z}'}) + \gamma \mathbb{E}_{(s'',\widehat{g}'')\sim\mathcal{J}_\kappa} \max \widehat{Q}^t_\kappa(s'',a'',\widehat{z}'')\right]\Big| \\
&\leq \underbrace{\left|\mathbb{E}_{\mathcal{J}_n}\left[\max_{a'} r_\ell(s',a',g_{z'})\right] - \mathbb{E}_{\mathcal{J}_\kappa}\left[\max_{a'} r_\ell(s',a',g_{\widehat{z}'})\right]\right|}_{\text{Term (I)}} \\
&\quad + \underbrace{\gamma\left|\mathbb{E}_{\mathcal{J}_n}\left[\max_{a'} \mathbb{E}_{\mathcal{J}_n}\left[\max_{a'',z''} Q^t(s'',a'',z'')\right]\right] - \mathbb{E}_{\mathcal{J}_\kappa}\left[\max_{a'} \mathbb{E}_{\mathcal{J}_\kappa}\left[\max_{a'',\widehat{z}''} \widehat{Q}^t_\kappa(s'',a'',\widehat{z}'')\right]\right]\right|}_{\text{Term (II)}}
\end{aligned}$$

Term (I) follows the same structure as in the base case ($t = 1$); hence, using Lemmas C.6 and C.4 we bound it by:

$$\begin{aligned}
\left|\mathbb{E}_{\mathcal{J}_n}\left[\max_{a'} r_\ell(s',a',g_{\widehat{z}'})\right] - \mathbb{E}_{\mathcal{J}_\kappa}\left[\max_{a'} r_\ell(s',a',g_{\widehat{z}'})\right]\right| &\leq 4\|r_\ell\|_\infty \cdot L_P \cdot \mathrm{TV}(g_{z'}, g_{\widehat{z}'}) \\
&\leq 4\|r_\ell\|_\infty \cdot L_P \cdot \mathrm{TV}(g_z, g_{\widehat{z}}),
\end{aligned}$$

where the second inequality follows by Lemma C.5.

For Term (II), the inner difference is the inductive hypothesis applied at the next step:

$$\begin{aligned}\gamma\Big|\mathbb{E}_{\mathcal{J}_n}\Big[\max_{a',z'}\mathbb{E}_{\mathcal{J}_n}\big[\max_{a'',z''}Q^t(s'',a'',z'')\big]\Big]-\mathbb{E}_{\mathcal{J}_\kappa}\Big[\max_{a',z'}\mathbb{E}_{\mathcal{J}_\kappa}\big[\max_{a'',z''}\widehat{Q}^t_\kappa(s'',a'',\widehat{z}'')\big]\Big]\Big| \\ \leq \frac{4\gamma\|r_\ell\|_\infty}{1-\gamma}L_P\cdot \mathrm{TV}(g_{z''},g_{\widehat{z}''}) \\ \leq \frac{4\gamma\|r_\ell\|_\infty}{1-\gamma}L_P\cdot \mathrm{TV}(g_z,g_{\widehat{z}}),\end{aligned}$$

where the first inequality follows from Lemma C.4 and the last inequality follows by Lemma C.5.
Then, combining Term (I) and Term (II), we get

$$\begin{aligned}|\mathbb{E}_{(s',g')\sim\mathcal{J}_n(\cdot|s,a,z)}\max_{a',z'}Q^t(s',a',z')-\mathbb{E}_{(s',\widehat{g}')\sim\mathcal{J}_\kappa(\cdot|s,a,\widehat{z})}\max_{a',\widehat{z}'}\widehat{Q}^t_\kappa(s',a',\widehat{z})| \\ \leq 4\|r_\ell\|_\infty\cdot L_P\cdot\mathrm{TV}(g_z,g_{\widehat{z}})+\frac{4\gamma\|r_\ell\|_\infty}{1-\gamma}L_P\cdot\mathrm{TV}(g_z,g_{\widehat{z}}) \\ =\frac{4\|r_\ell\|_\infty}{1-\gamma}L_P\cdot\mathrm{TV}(g_z,g_{\widehat{z}}),\end{aligned}$$

which completes the proof. □

Definition C.5 (Joint transition probability kernel)**.** *Fix an agent $i\in[n]$. Let s and a denote the state and action of agent i and $\widehat{z}$ denote its graphon-weighted state/action feature on a sample of size κ (which we denote by Δ_i). Let the state marginals of $\widehat{z}$ be $\widehat{g}$. The joint transition probability kernel for agent i and its neighborhood is then:*

$$\mathcal{J}_{\Delta_i\cup\{i\}}(s',\widehat{g}'|s,a,\widehat{z}).$$

Definition C.6 (Single Agent Transition Kernel)**.** *The single agent transition kernel for agent i is given by the marginal*

$$\mathcal{J}_1\big(s_i'|s_i,s_{\Delta_i},a_i,a_{\Delta_i}\big)=\sum_{s'_{\Delta_i}}\mathcal{J}_{\Delta_i\cup\{i\}}\big(s_i',s'_{\Delta_i}|s_i,s_{\Delta_i},a_i,a_{\Delta_i}\big).$$

Lemma C.3. *Assuming that the next state of agent i is conditionally independent of the neighbors' next states given the current states and actions, the kernel $\mathcal{J}_1$ satisfies:*

$$\mathbb{P}\big(S_i'=s_i'|s_i,s_{\Delta_i},a_i,a_{\Delta_i}\big)=\mathcal{J}_1\big(s_i'|s_i,s_{\Delta_i},a_i,a_{\Delta_i}\big)$$

Proof.

$$\begin{aligned}\mathbb{P}\big(S_i'=s_i'|s_i,s_{\Delta_i},a_i,a_{\Delta_i}\big)&=\sum_{s'_{\Delta_i}}\mathbb{P}\Big(S_i'=s_i',S_{\Delta_i}=s'_{\Delta_i}|s_i,s_{\Delta_i},a_i,a_{\Delta_i}\Big)\\&=\sum_{s'_{\Delta_i}}\mathcal{J}_{\Delta_i\cup\{i\}}\big(s_i',s'_{\Delta_i}|s_i,s_{\Delta_i},a_i,a_{\Delta_i}\big)\\&=\mathcal{J}_1\big(s_i'|s_i,s_{\Delta_i},a_i,a_{\Delta_i}\big),\end{aligned}$$

where the first equality follows by evaluating the conditional probability, and the second equality follows by definition C.6. Together, this proves the lemma. □

We now prove a contraction property in TV-distance between the stochastic kernels $\mathcal{J}_\kappa$ and $\mathcal{J}$ under a common neighborhood.

Lemma C.4 (Coupling stability of the surrogate kernel)**.** *Fix $(s,a)\in\mathcal{S}\times\mathcal{A}$. Let $z\in\mathcal{P}(\mathcal{S}\times\mathcal{A})$ and let $\widehat{z}\in\mathcal{Z}_\kappa$ be any empirical histogram; write $g:=g_z$ and $\widehat{g}:=g_{\widehat{z}}$. Then, under the dynamics of $\mathcal{J}_n$ and $\mathcal{J}_\kappa$, there exists a coupling of (S_0',g') and $(\widehat{S}_0',\widehat{g}')$ such that $\Pr[S_0'\neq\widehat{S}_0']\leq\mathrm{TV}(P(\cdot\mid s,a,g),P(\cdot\mid s,a,\widehat{g}))\leq L_P\cdot\mathrm{TV}(g,\widehat{g})$ and*

$$\mathbb{E}\big[\mathrm{TV}(g',\widehat{g}')\big]\leq L_P\cdot\mathrm{TV}(g,\widehat{g}). \tag{7}$$

Proof. For the focal agent, take a maximal coupling of $P(\cdot\mid s,a,g)$ and $P(\cdot\mid s,a,\widehat{g})$, so that

$$\Pr[S_0'\neq\widehat{S}_0']=\mathrm{TV}(P(\cdot\mid s,a,g),P(\cdot\mid s,a,\widehat{g})).$$

Then, applying the Lipschitz continuity of the transition kernels from Assumption 3.5 completes the proof of the first inequality. For the second inequality, we consider the neighborhood marginal. Couple the two constructions by using the same i.i.d. draws $(X_m,U_m)\sim z$ and,

conditional on each (X_m, U_m), couple $X'_m \sim P(\cdot \mid X_m, U_m, g)$ with $\widehat{X}'_m \sim P(\cdot \mid X_m, U_m, \widehat{g})$ via an optimal coupling. Then, from Assumption 3.5, we have that

$$\begin{aligned}\Pr[X'_m \neq \widehat{X}'_m \mid X_m, U_m] &= \mathrm{TV}(P(\cdot \mid X_m, U_m, g), P(\cdot \mid X_m, U_m, \widehat{g})) \\ &\leq L_P \cdot \mathrm{TV}(g, \widehat{g}),.\end{aligned}$$

Next, let $D \coloneqq \sum_{m=1}^{\kappa} \mathbb{1}\{X'_m \neq \widehat{X}'_m\}$ be the number of mismatches. A single mismatch can change the empirical histogram by at most $2/\kappa$ in ℓ_1, hence at most $1/\kappa$ in TV. Therefore, deterministically, we have

$$\|g' - \widehat{g}'\|_1 \leq \frac{2D}{\kappa} \implies \frac{1}{2}\|g' - \widehat{g}'\|_1 \leq \frac{D}{\kappa}.$$

Finally, taking expectations and using linearity, we have

$$\mathbb{E}[\mathrm{TV}(g', \widehat{g}')] \leq \frac{1}{\kappa} \sum_{m=1}^{\kappa} \Pr[X'_m \neq \widehat{X}'_m] \leq L_P \cdot \mathrm{TV}(g, \widehat{g}),$$

which completes the proof. □

Lemma C.5 (Total variation contraction under mixture kernels)**.** *Let $(\mathcal{Z}, \mathcal{A})$ be the index measurable space and $(\mathcal{X}, \mathcal{B})$ be the output measurable space. Let $J_1(\cdot|z)$ be a Markov kernel from $\mathcal{Z}$ to $\mathcal{X}$. For probability measures μ, ν on $\mathcal{Z}$, define the mixtures $J_\mu(B) = \int_{\mathcal{Z}} J_1(B|z)\mu(\mathrm{d}z)$ and $J_\nu(B) \coloneqq \int_{\mathcal{Z}} J_1(B|z)\nu(\mathrm{d}z)$ for all $B \in \mathcal{B}$. Then,*

$$\mathrm{TV}(\mathcal{J}_\mu, \mathcal{J}_\nu) \leq \mathrm{TV}(\mu, \nu). \tag{8}$$

Proof. Note that the lemma is a special case of the data processing inequality for f-divergences, where $f(t) = \frac{1}{2}|t-1|$. To prove it, we use that for probability measures P, Q on a measurable space,

$$\mathrm{TV}(P, Q) = \sup_B |P(B) - Q(B)| = \frac{1}{2} \sup_{\|f\|_\infty \leq 1} |f\mathrm{d}(P - Q)|.$$

So, fix any measurable $B \in \mathcal{B}$ and let $\delta \coloneqq \mu - \nu$ be a signed measure on $\mathcal{Z}$, where $\delta(\mathcal{Z}) = 0$. Then,

$$\begin{aligned}J_\mu(B) - J_\nu(B) &= \int_{\mathcal{Z}} J_1(B|z)\delta(\mathrm{d}z) \\ &= \frac{1}{2}\int (2J_1(B|z) - 1)\mathrm{d}\delta,\end{aligned}$$

where the last equality follows from $\int 1 d\delta = \delta(\mathcal{Z}) = 0$. Then, define $\phi_B(z) \coloneqq 2J_1(B|z) - 1$. Since $J_1(B|z) \in [0, 1]$, we have $\phi_B(z) \in [-1, 1]$ and hence $\|\phi_B\|_\infty \leq 1$.
Therefore, we write

$$\begin{aligned}|J_\mu(B) - J_\nu(B)| &= \frac{1}{2}|\phi_B \mathrm{d}(\mu - \nu)| \\ &\leq \frac{1}{2} \sup_{\|f\|_\infty \leq 1} |f\mathrm{d}(\mu - \nu)| \\ &= \mathrm{TV}(\mu, \nu).\end{aligned}$$

Since this holds for every $B \in \mathcal{B}$, taking the supremum gives

$$\mathrm{TV}(J_\mu, J_\nu) = \sup_B |J_\mu(B) - J_\nu(B)| \leq \mathrm{TV}(\mu, \nu),$$

which proves the claim. □

Lemma C.6 (Reward expectation mismatch under subsampling)**.** *Let $\widetilde{r} : \mathcal{S} \times \mathcal{G} \to \mathbb{R}$ be a bounded local immediate reward with $\|r_\ell\|_\infty < \infty$. Let $\mu_{[n]}$ and $\widehat{\mu}_\Delta$ are distributions on the index space from Lemma C.5. Next, define the induced transition kernels*

$$\mathcal{J}_n(\cdot) = \int \mathcal{J}_1(\cdot \mid z)\mu_{[n]}(\mathrm{d}z) \tag{9}$$

and

$$\mathcal{J}_\kappa(\cdot) = \int \mathcal{J}_1(\cdot \mid z)\widehat{\mu}_\Delta(\mathrm{d}z), \tag{10}$$

where $\mathcal{J}_1(\cdot \mid z)$ is the single-agent transition kernel. The reward expectation mismatch is then given by

$$|\mathbb{E}_{\mathcal{J}_n}[\widetilde{r}(x, g')] - \mathbb{E}_{\mathcal{J}_\kappa}[\widetilde{r}(x, \widehat{g}']| \leq 2\|r_\ell\|_\infty \cdot \mathrm{TV}(\mu_{[n]}, \widehat{\mu}_\Delta).$$

Proof. By expanding the expectations and using $|\widetilde{r}| \le \|r_\ell\|_\infty$, we have

$$\begin{aligned}|\mathbb{E}_{\mathcal{J}_n}[\widetilde{r}(x,g')] - \mathbb{E}_{\mathcal{J}_\kappa}[\widetilde{r}(x,\widetilde{g}')]| &\le \sum_{(x,g')\in\mathcal{S}\times\mathcal{G}} \left|\mathcal{J}_n(x,g'|\cdot) - \mathcal{J}_\kappa(x,g'|\cdot)\right||\widetilde{r}_\ell(x,g')| \\ &\le \|r_\ell\|_\infty \cdot \sum_{(x,g')\sim\mathcal{S}\times\mathcal{G}} |\mathcal{J}_n(x,g'|\cdot) - \mathcal{J}_\kappa(x,g'|\cdot)| \\ &= 2\|r_\ell\|_\infty \cdot \mathrm{TV}(\mathcal{J}_n, \mathcal{J}_\kappa)\end{aligned}$$

Now we bound the total variation distance between the two kernels. Since $\mathcal{J}_n$ and $\mathcal{J}_\kappa$ are mixture kernels induced by $\mu_{[n]}$ and $\widehat{\mu}_\Delta$, lemma C.5 gives $\mathrm{TV}(\mathcal{J}_n, \mathcal{J}_\kappa) \le \mathrm{TV}(\mu_{[n]}, \widehat{\mu}_\Delta)$. Hence, the bound on our expectation mismatch is

$$|\mathbb{E}_{J_n}[\widetilde{r}(x,g')] - \mathbb{E}_{J_\kappa}[\widetilde{r}(x,\widetilde{g}']| \le 2\|r_\ell\|_\infty \cdot \mathrm{TV}(\mu_{[n]}, \widehat{\mu}_\Delta),$$

which completes our proof. □

C.1 Concentration on the Subsampled Mean-Field Features

Fix $\kappa \in [n-1]$. We now wish to bound the TV-distance between the true graphon-weighted neighborhood feature g and its subsampled estimate $\widehat{g}$. Since g is a probability mass over a finite discrete space $\mathcal{S}$ and $\widehat{g}$ is the empirical distribution formed by κ i.i.d. samples drawn from g_i, we can argue a concentration bound using a similar argument as in Anand and Qu (3).

We begin by stating the probability model.

Lemma C.7 (i.i.d. neighbor-state sampling)**.** *Fix an agent i and condition on the global joint state $s(t) = (s_1(t), \dots, s_n(t))$. Define the ground-truth graphon-weighted neighborhood state distribution for $x \in \mathcal{S}$ by*

$$g_i(x) \coloneqq \sum_{j\neq i} \bar{w}_{ij} \mathbb{1}\{s_j(t) = x\},$$

where $\bar{w}_{ij}$ are the normalized interaction weights from Definition 2.1. Then, sample neighbor indices $J_1^{(1)}, \dots, J_i^{(\kappa)} \sim \bar{w}_i$ and define the sampled neighbor states $X_m \coloneqq s_{J_i^{(m)}} \in \mathcal{S}$. Then, conditional on s, the random variables $X_1, \dots, X_\kappa$ are i.i.d. with law g_i. In other words, $\Pr[X_m = x|s] = g_i(x)$, for all $x \in \mathcal{S}$ and $m = 1, \dots, \kappa$.

Proof. The proof follows by noting that

$$\begin{aligned}\Pr[X_m = x|s] &= \sum_{j\neq i} \Pr[J_i^{(m)} = j]\mathbb{1}\{s_j = x\} \\ &= \sum_{j\neq i} \bar{w}_{ij}\mathbb{1}\{s_j = x\} \\ &= g_i(x),\end{aligned}$$

where independence follows from the i.i.d sampling of $J_i^{(m)}$. □

Theorem C.8 (TV concentration for empirical distributions on a finite alphabet). *Let $\mathcal{S}$ be a finite set, and let $X_1, \dots, X_\kappa$ be i.i.d. samples from a distribution $p \in \Delta(\mathcal{S})$. Let $\widehat{p}$ be the empirical distribution given by $\widehat{p}(x) \coloneqq \frac{1}{\kappa}\sum_{m=1}^{\kappa} \mathbb{1}\{X_m = x\}$. Then, for any $\epsilon > 0$, we have that with probability at least $1-\delta$,*

$$\mathrm{TV}(\widehat{p}, p) \le \sqrt{\frac{|\mathcal{S}|\ln 2 + \ln(2/\delta)}{2\kappa}}$$

Proof. Recall that

$$\mathrm{TV}(\widehat{p}, p) = \sup_{E\subseteq\mathcal{S}} |\widehat{p}(E) - p(E)|,$$

where $\widehat{p}(E) \coloneqq \sum_{x\in E}\widehat{p}(x)$. Fix any subset $E \subseteq S$, and define $Y_m \coloneqq \mathbb{1}\{X_m \in E\} \in \{0,1\}$. Then, $Y_1, \dots, Y_\kappa$ are i.i.d. Bernoulli with mean $p(E)$ and $\widehat{p}(E) = \frac{1}{\kappa}\sum_{m=1}^{\kappa} Y_m$. Then, by Hoeffding's inequality (32), we have

$$\Pr[\mathrm{TV}(\widehat{p}, p) \ge \epsilon] \le 2^{|S|+1}\exp(-2\kappa\epsilon^2).$$

Now, taking a union bound over all subsets $E \subseteq \mathcal{S}$, we have

$$\Pr\left[\sup_{E\subseteq\mathcal{S}} |\widehat{p}(E) - p(E)| \ge \epsilon\right] \le \left|\sum_{E\subseteq\mathcal{S}} 1\right| \cdot 2\exp(-2\kappa\epsilon^2)$$

$$= 2^{|\mathcal{S}|+1} \exp(-2\kappa\epsilon^2).$$

Finally, reparameterizing by setting $\delta = 2^{|\mathcal{S}|+1} \exp(-2\kappa\epsilon^2)$ yields the claim. □

Corollary C.9. *Fix a subsampling parameter $\kappa \geq 1$ and error parameter $\delta \in (0,1)$. Fix $(s,a) \in \mathcal{S} \times \mathcal{A}$, and let $z \in \mathcal{Z} \coloneqq \Delta(\mathcal{S} \times \mathcal{A})$. Let $\widehat{z} \in \mathcal{Z}_\kappa$ be the empirical histogram of κ i.i.d. draws from z, and let $g_z, g_{\widehat{z}}$ be the marginals in $\mathcal{S}$. Then, for all $t \in \mathbb{N}$, we have that with probability at least $1-\delta$,*

$$\left| Q^t(s,a,z) - \widehat{Q}^t_\kappa(s,a,\widehat{z}) \right| \leq \frac{4L_P \|r_\ell\|_\infty}{1-\gamma} \cdot \sqrt{\frac{|\mathcal{S}| \ln 2 + \ln(2/\delta)}{2\kappa}}.$$

D PERFORMANCE GAP BETWEEN OPTIMAL AND SUBSAMPLED POLICIES

In this section, we relate the discrepancy between the optimal Q-function Q^* and its subsampled fixed point $\widehat{Q}^*_\kappa$ to the performance gap between the optimal policy π^* and the estimated GMFS policy π^{est}_κ. We consider a state space $\mathcal{S} \times \mathcal{G}_\kappa$ and an action space $\mathcal{A} \times \mathcal{H}_\kappa$. Specifically, for each (s,g), feasible actions are completions of the fiber $z \in \Gamma_\kappa(g)$. A policy $\pi(\cdot \mid s,g)$ is thus a distribution over the product space $(a,z) \in \mathcal{A} \times \Gamma_\kappa(g)$.

Definition D.1 (Value Function)**.** *Given a policy π, the value function is given by*

$$V^\pi(s,g) \coloneqq \mathbb{E}_{(a,z)\sim\pi(\cdot|s,g)}[Q^\pi(s,a,z)].$$

We define π^* as the optimal policy induced by the optimal Q-function Q^*.

Definition D.2 (Optimal policy π^*)**.** *For each $(s,g) \in \mathcal{S} \times \mathcal{G}_\kappa$, the optimal greedy policy is:*

$$\pi^*(\cdot \mid s,g) \in \arg\max_{(a,z)\in\mathcal{A}\times\Gamma_\kappa(g)} Q^*(s,a,z),$$

where the associated Bellman backup is $\mathcal{M}_\kappa Q(s,g) \coloneqq \max_{a\in\mathcal{A}, z\in\Gamma_\kappa(g)} Q(s,a,z)$.

Definition D.3 (Estimated GMFS policy π^{est}_κ)**.** *Let $\widehat{Q}_\kappa$ be the Q-function obtained after T Bellman updates in Algorithm 1. For any $(s,\widehat{g}) \in \mathcal{S} \times \mathcal{G}_\kappa$, the estimated greedy joint action for the agent and its fiber completion is:*

$$(a,\widehat{z}) \coloneqq \arg\max_{a\in\mathcal{A}, \widehat{z}\in\Gamma_\kappa(g)} \widehat{Q}_\kappa(s,a,\widehat{z}).$$

We next introduce the celebrated performance difference lemma from Kakade and Langford (38), which is a key tool for bounding the optimality gap of our learned policy.

Lemma D.1 (Performance Difference Lemma, (38))**.** *Given two policies π and π', for any initial state s_0, we have:*

$$V^\pi(s_0) - V^{\pi'}(s_0) = \frac{1}{1-\gamma} \mathbb{E}_{s\sim d^\pi_{s_0}} \left[\mathbb{E}_{a\sim\pi(\cdot|s)}[A^{\pi'}(s,a)] \right],$$

where $A^{\pi'}(s,a) = Q^{\pi'}(s,a) - V^{\pi'}(s)$ is the advantage function, and $d^\pi_s(s') = (1-\gamma)\sum_{h=0}^\infty \gamma^h \Pr^\pi_h[s',s]$ where $\Pr^\pi_h[s',s]$ is the probability of π reaching state s' at time step h when starting from state s.

We are now ready to formulate the proof for Theorem 4.1.

Theorem D.2. *Fix $\delta \in (0,1)$. For all states $s \in \mathcal{S}$ and graphon state aggregates $g \in \mathcal{G}$, if $T \geq \frac{1}{1-\gamma} \log \frac{\|r_\ell\|_\infty \sqrt{\kappa}}{1-\gamma}$, then*

$$V^{\pi^*}(s,g) - V^{\pi^{\text{est}}_\kappa}(s,g) \leq \frac{2L_P\|r_\ell\|_\infty}{(1-\gamma)^2} \sqrt{\frac{|\mathcal{S}| \ln 2 + \ln(2/\delta)}{2\kappa}} + \frac{\epsilon_{\kappa,m}}{1-\gamma} + \frac{2\|r_\ell\|_\infty}{(1-\gamma)^2} |\mathcal{A}| \cdot \kappa^{|\mathcal{A}|} \delta.$$

Proof. Applying the Performance Difference Lemma, let $A^{\pi'}(s,a,z) = Q^{\pi'}(s,a,z) - V^{\pi'}(s,g)$ be the advantage function of policy π' at (s,a,z). Let $d^\pi_{(s,g)}$ denote the discounted occupancy measure over the space $\mathcal{S} \times \mathcal{A} \times \mathcal{Z}$ induced by policy π starting from (s,g). Then,

$$\begin{aligned}
V^{\pi^*}(s_0,g_0) - V^{\pi^{\text{est}}_\kappa}(s_0,g_0) &= \frac{1}{1-\gamma} \mathbb{E}_{(s,a,z)\sim d^{\pi^{\text{est}}_\kappa}_{(s_0,g_0)}} \left[\mathbb{E}_{a\sim\pi^*(\cdot|s,g_z)} A^{\pi^*}(s,a,z) \right] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{(s,a,z)\sim d^{\pi^{\text{est}}_\kappa}_{(s_0,g_0)}} \left[\mathbb{E}_{a\sim\pi^*(\cdot|s,g_z)} Q^{\pi^*}(s,a,z) - \mathbb{E}_{a\sim\pi^{\text{est}}_\kappa(\cdot|s,g)} Q^{\pi^*}(s,a,z) \right] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{(s,a,z)\sim d^{\pi^{\text{est}}_\kappa}} \left[Q^{\pi^*}(s,g,\pi^*(\cdot|s,g)) - \mathbb{E}_{a\sim\pi^{\text{est}}_\kappa(\cdot|s,g)} Q^{\pi^*}(s,a,z) \right]. \qquad (11)
\end{aligned}$$

From Definition 3.1, $\bar{w}_{i,j}$ is the normalized sampling distribution over agent i's neighbors. For each agent i, $\Delta_i = (J_i^{(1)}, ...J_i^{(k-1)}), J_i \sim \bar{w}_{i,j}$. Using the law of total expectation:

$$\begin{aligned}
&\mathbb{E}_{a\sim\pi_\kappa^{\text{est}}(\cdot|s,g)} Q^{\pi^*}(s,a,z) \\
&= \mathbb{E}_\Delta \mathbb{E}_{a\sim\prod_{i=1}^n \widehat{\pi}_\kappa^{\text{est}}(\cdot|s,g)} \left[Q^{\pi^*}(s,a,z)\right] = \sum_{\Delta_1}\cdots\sum_{\Delta_n}\Big(\prod_{i=1}^n\prod_{r=1}^{\kappa-1}\bar{w}_{i,J_i^{(r)}}\Big)\sum_{a\in\mathcal{A}^n} Q^{\pi^*}(s,a,z)\prod_{i=1}^n \widehat{\pi}_\kappa^{\text{est}}(a_i|s,\widehat{g})
\end{aligned}$$

Then

$$\begin{aligned}
&Q^*(s,g,\pi^*(\cdot\mid s,g)) - \mathbb{E}_{a\sim\pi_\kappa^{\text{est}}(\cdot|s,g)} Q^*(s,a,z) \\
&\qquad = \sum_{\Delta_1}\cdots\sum_{\Delta_n}\Big(\prod_{i=1}^n\prod_{r=1}^{\kappa-1}\bar{w}_{i,J_i^{(r)}}\Big)\Big(Q^*(s,g,\pi^*(\cdot\mid s,g)) - \sum_{a\in\mathcal{A}^n} Q^*(s,a,z)\prod_{i=1}^n \widehat{\pi}_\kappa^{\text{est}}(a_i\mid s,\widehat{g})\Big)
\end{aligned}$$

Plugging back into Eq. (11):

$$\begin{aligned}
&V^{\pi^*}(s,g) - V^{\pi_\kappa^{\text{est}}}(s,g) \\
&\quad = \frac{1}{1-\gamma}\mathbb{E}_{(s',a',z')\sim d_{(s,g)}^{\pi_\kappa^{\text{est}}}}\Big[\sum_{\Delta_1}\cdots\sum_{\Delta_n}\Big(\prod_{i=1}^n\prod_{r=1}^{\kappa-1}\bar{w}_{i,J_i^{(r)}}\Big)\Big(Q^*(s',g',\pi^*(\cdot\mid s',g')) \\
&\qquad\qquad - \sum_{a\in\mathcal{A}^n} Q^*(s',a',z')\prod_{i=1}^n \widehat{\pi}_\kappa^{\text{est}}(a_i\mid s',\widehat{g}')\Big)\Big] \\
&\quad = \frac{1}{1-\gamma}\mathbb{E}_{(s',a',z')\sim d_{(s,g)}^{\pi_\kappa^{\text{est}}}}\Big[\sum_{\Delta_1}\cdots\sum_{\Delta_n}\Big(\prod_{i=1}^n\prod_{r=1}^{\kappa-1}\bar{w}_{i,J_i^{(r)}}\Big)\Big(Q^*(s',g',\pi^*(\cdot\mid s',g')) - Q^*(s',g',\pi_{\kappa,\Delta}^{\text{est}}(\cdot|s',\widehat{g}'))\Big)\Big]
\end{aligned}$$

Now we apply Lemma D.3.

$$\begin{aligned}
&V^{\pi^*}(s,g) - V^{\pi_\kappa^{\text{est}}}(s,g) \\
&\quad \le \frac{1}{1-\gamma}\mathbb{E}_{(s',a',z')\sim d_{(s,g)}^{\pi_\kappa^{\text{est}}}}\Bigg[\sum_{\Delta_1}\cdots\sum_{\Delta_n}\Big(\prod_{i=1}^n\prod_{r=1}^{\kappa-1}\bar{w}_{i,J_i^{(r)}}\Big)\cdot \\
&\qquad\qquad \Bigg(\frac{1}{n}\sum_{i=1}^n\Big|Q^*(s',g',\pi^*(\cdot|s',g')) - \widehat{Q}_\kappa^{\text{est}}\big(s',\widehat{g}',\pi^*(\cdot|s',\widehat{g}')\big)\big|_{\{i\}\cup\Delta_i}\Big)\Big| \\
&\qquad\qquad + \frac{1}{n}\sum_{i=1}^n\Big|\widehat{Q}_\kappa^{\text{est}}\big(s',\widehat{g}',\pi_{\kappa,\Delta}^{\text{est}})(\cdot|s,g)\big|_{\{i\}\cup\Delta_i}\big) - Q^*(s',g',\pi_{\kappa,\Delta}^{\text{est}})(\cdot|s,g)\Big|\Bigg)\Bigg] \\
&\quad \le \frac{1}{1-\gamma}\mathbb{E}_{(s',a',z')\sim d_{(s,g)}^{\pi_\kappa^{\text{est}}}}\Big[\mathbb{E}_\Delta\Big[\frac{1}{n}\sum_{i=1}^n\Big|Q^*(s',g',\pi^*(\cdot|s,g)) - \widehat{Q}_\kappa^{\text{est}}\big(s',\widehat{g}',\pi^*(\cdot|s,g)\big|_{\{i\}\cup\Delta_i}\big)\Big|\Big] \\
&\qquad\qquad + \mathbb{E}_\Delta\Big[\frac{1}{n}\sum_{i=1}^n\Big|\widehat{Q}_\kappa^{\text{est}}\big(s',\widehat{g}',\pi_{\kappa,\Delta}^{\text{est}}(\cdot|s,\widehat{g})\big|_{\{i\}\cup\Delta_i}\big) - Q^*(s',g',\pi_{\kappa,\Delta}^{\text{est}}(\cdot|s,\widehat{g}))\Big|\Big]\Big]
\end{aligned}$$

Now we apply Lemma D.4, where we set $\mathcal{D} = (s',g') \sim d_{(s,g)}^{\pi_\kappa^{\text{est}}}$. Applying Lemma D.4 to the first term gives

$$\begin{aligned}
&\frac{1}{1-\gamma}\mathbb{E}_{(s',a',z')\sim d_{(s,g)}^{\pi_\kappa^{\text{est}}}}\Big[\mathbb{E}_\Delta\Big[\frac{1}{n}\sum_{i=1}^n\Big|Q^*(s',g',\pi^*) - \widehat{Q}_\kappa^{\text{est}}\big(s',\widehat{g}',\pi^*(\cdot|s,\widehat{g})\big|_{\{i\}\cup\Delta_i}\big)\Big|\Big] \\
&\qquad\qquad \le \frac{L_P\|r_\ell\|_\infty}{(1-\gamma)^2}\sqrt{\frac{|\mathcal{S}|\ln 2 + \ln(2/\delta)}{2\kappa}} + \frac{\epsilon_{\kappa,m}}{1-\gamma} + \frac{\|r_\ell\|_\infty}{(1-\gamma)^2}|\mathcal{A}|\cdot\kappa^{|\mathcal{A}|}\delta_t
\end{aligned}$$

Applying Lemma D.4 to the second term, where π^* is replaced by $\pi_{\kappa,\Delta}^{\text{est}}$ is valid since Lemma D.4 is uniform over the local join actions union-bounded by Lemma D.5, gives us the bound

$$\begin{aligned}
&\frac{1}{1-\gamma}\mathbb{E}_{(s',a',z')\sim d_{(s,g)}^{\pi_\kappa^{\text{est}}}}\mathbb{E}_\Delta\Big[\frac{1}{n}\sum_{i=1}^n\Big|\widehat{Q}_\kappa^{\text{est}}\big(s',\widehat{g}',\pi_{\kappa,\Delta}^{\text{est}}(\cdot|s,\widehat{g})\big|_{\{i\}\cup\Delta_i}\big) - Q^*(s',g',\pi_{\kappa,\Delta}^{\text{est}})(\cdot|s,g)\Big|\Big]\Big] \\
&\qquad\qquad \le \frac{L_P\|r_\ell\|_\infty}{(1-\gamma)^2}\sqrt{\frac{|\mathcal{S}|\ln 2 + \ln(2/\delta)}{2\kappa}} + \frac{\epsilon_{\kappa,m}}{1-\gamma} + \frac{\|r_\ell\|_\infty}{(1-\gamma)^2}|\mathcal{A}|\cdot\kappa^{|\mathcal{A}|}\delta_t,
\end{aligned}$$

which completes the proof. □

Lemma D.3 (Uniform bound on Q^* under graphon-weighted subsampling)**.** *Fix any state $(s, g) \in \mathcal{S} \times \mathcal{G}$. For each agent $i \in [n]$, and with subsampling given by Definition 3.1, we let F_{Δ_i} denote the corresponding sampled feature. The estimated joint action selection:*

$$\pi^{\text{est}}_{\kappa,\Delta}(\cdot|s, g) = \prod_{i=1}^{n} \widehat{\pi}^{\text{est}}_{\kappa}(a_i|s, a, F_{\Delta_i})$$

Then,

$$\begin{aligned} Q^*(s, g, \pi^*(\cdot|s, \widehat{g})) - Q^*(s, g, \pi^{\text{est}}_{\kappa,\Delta}(\cdot|s, g)) \le \frac{1}{n}\sum_{i=1}^{n} \Big| Q^*(s, g, \pi^*(\cdot|s, g)) - \widehat{Q}^{\text{est}}_{\kappa}\Big(s, \widehat{g}, \pi^*(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\Big)\Big| \\ + \Big|\widehat{Q}^{\text{est}}_{\kappa}\Big(s, \widehat{g}, \pi^{\text{est}}_{\kappa,\Delta}(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\Big) - Q^*(s, g, \pi^{\text{est}}_{\kappa,\Delta}(\cdot|s, g))\Big| \end{aligned}$$

Proof.

$$\begin{aligned} Q^*(s, g, \pi^*(\cdot|s, g)) - Q^*(s, g, \pi^{\text{est}}_{\kappa,\Delta}(\cdot|s, g)) &= \frac{1}{n}\sum_{i=1}^{n} \widehat{Q}^{\text{est}}_{\kappa}\Big(s, \widehat{g}, \pi^{\text{est}}_{\kappa,\Delta}(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\Big) - \frac{1}{n}\sum_{i=1}^{n} \widehat{Q}^{\text{est}}_{\kappa}\Big(s, \widehat{g}, \pi^{\text{est}}_{\kappa,\Delta}(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\Big) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} \widehat{Q}^{\text{est}}_{\kappa}\Big(s, \widehat{g}, \pi^*(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\Big) - \frac{1}{n}\sum_{i=1}^{n} \widehat{Q}^{\text{est}}_{\kappa}\Big(s, \widehat{g}, \pi^*(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\Big) \\ &\le \Big|Q^*(s, g, \pi^*(\cdot|s, g)) - \frac{1}{n}\sum_{i=1}^{n} \widehat{Q}^{\text{est}}_{\kappa}\Big(s, \widehat{g}, \pi^*(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\Big)\Big| \\ &\quad + \Big|\frac{1}{n}\sum_{i=1}^{n} \widehat{Q}^{\text{est}}_{\kappa}\Big(s, \widehat{g}, \pi^{\text{est}}_{\kappa,\Delta}(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\Big) - Q^*(s, g, \pi^{\text{est}}_{\kappa,\Delta}(\cdot|s, g))\Big| \\ &\le \frac{1}{n}\sum_{i=1}^{n} \Big|Q^*(s, g, \pi^*(\cdot|s, g)) - \widehat{Q}^{\text{est}}_{\kappa}\Big(s, \widehat{g}, \pi^*(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\Big)\Big| \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} \Big|\widehat{Q}^{\text{est}}_{\kappa}\Big(s, \widehat{g}, \pi^{\text{est}}_{\kappa,\Delta}(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\Big) - Q^*(s, g, \pi^{\text{est}}_{\kappa,\Delta}(\cdot|s, g))\Big|, \end{aligned}$$

which completes the proof. □

Lemma D.4. *Let r_ℓ denote the local reward function used by the (graphon-weighted) sampled Bellman operator. For any arbitrary distribution $\mathcal{D}$ of states $(s, g) \in \mathcal{S} \times \mathcal{G}$ and for any Δ_i generated by Definition 3.1 and $\delta \in (0, 1]$,*

$$\mathbb{E}_{(s,g)\sim\mathcal{D}}\left[\frac{1}{n}\sum_{i=1}^{n}\Big|Q^*(s, g, \pi^*(\cdot \mid s, g)) - \widehat{Q}^{\text{est}}_{\kappa}\big(s, \widehat{g}, \pi^*(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\big)\Big|\right] \le \frac{L_P\|r_\ell\|_\infty}{1-\gamma}\Phi_{\kappa,\delta} + \epsilon_{\kappa,m} + \frac{\|r_\ell\|_\infty}{1-\gamma}|\mathcal{A}| \cdot \kappa^{|\mathcal{A}|}\delta_t$$

where

$$\Phi_{\kappa,\delta} \coloneqq \sqrt{\frac{|\mathcal{S}|\ln 2 + \ln(2/\delta)}{2\kappa}}.$$

Proof. By linearity of expectation,

$$\begin{aligned} &\mathbb{E}_{(s,g)\sim\mathcal{D}}\Big[\frac{1}{n}\sum_{i=1}^{n}\Big|Q^*(s, g, \pi^*(\cdot \mid s, g)) - \widehat{Q}^{\text{est}}_{\kappa}\big(s, \widehat{g}, \pi^*(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\big)\Big|\Big] \\ &\qquad = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{(s,g)\sim\mathcal{D}}\Big[\Big|Q^*(s, g, \pi^*(\cdot \mid s, g)) - \widehat{Q}^{\text{est}}_{\kappa}\big(s, \widehat{g}, \pi^*(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\big)\Big|\Big] \end{aligned}$$

Define the indicator function $\mathcal{I} : [n] \times \mathcal{S} \times \mathbb{N} \times (0, 1] \to \{0, 1\}$ by

$$\mathcal{I}_i(s, g, \Delta_i, \delta) \coloneqq \mathbb{1}\left\{\Big|Q^*(s, g, \pi^*(\cdot \mid s, g)) - \widehat{Q}^{\text{est}}_{\kappa}\big(s, \widehat{g}, \pi^*(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\big)\Big| \le \frac{L_P\|r_\ell\|_\infty}{1-\gamma}\Phi_{\kappa,\delta} + \epsilon_{\kappa,m}\right\}$$

The expected difference

$$\begin{aligned} &\mathbb{E}_{(s,g)\sim\mathcal{D}}\Big[\Big|Q^*(s, g, \pi^*(\cdot \mid s, g)) - \widehat{Q}^{\text{est}}_{\kappa}\big(s, \widehat{g}, \pi^*(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\big)\Big|\Big] \\ &\quad = \mathbb{E}_{(s,g)\sim\mathcal{D}}\Big[\Big|Q^*(s, g, \pi^*(\cdot \mid s, g)) - \widehat{Q}^{\text{est}}_{\kappa}\big(s, \widehat{g}, \pi^*(\cdot|s, \widehat{g})\big|_{\{i\}\cup\Delta_i}\big)\Big|\Big|\mathcal{I}_i\Big] \end{aligned}$$

$$+ \mathbb{E}_{(s,g)\sim\mathcal{D}}\left[\left|Q^*(s,g,\pi^*(\cdot \mid s,g)) - \widehat{Q}^{\text{est}}_\kappa\big(s,\widehat{g},\pi^*(\cdot|s,\widehat{g})\big|_{\{i\}\cup\Delta_i}\big)\right|(1-\mathcal{I}_i)\right]$$

$$\leq \frac{L_P\|r_\ell\|_\infty}{1-\gamma}\Phi_{\kappa,\delta} + \epsilon_{\kappa,m}$$
$$+ \frac{\|r_\ell\|_\infty}{1-\gamma}\Pr\left[\left|Q^*(s,g,\pi^*(\cdot \mid s,g)) - \widehat{Q}^{\text{est}}_\kappa\big(s,\widehat{g},\pi^*(\cdot|s,\widehat{g})\big|_{\{i\}\cup\Delta_i}\big)\right| > \frac{L_P\|r_\ell\|_\infty}{1-\gamma}\Phi_{\kappa,\delta} + \epsilon_{\kappa,m}\right]$$

where we used the general property for a random variable X and constant c that $\mathbb{E}[X] = \mathbb{E}[X\mathbb{1}\{X \leq c\}] + \mathbb{E}[(1 - \mathbb{1}\{X \leq c\})X]$. Now we apply the union bound from Lemma D.5 with $T = 1$ and parameter δ which implies that uniformly over $i \in [n]$ and over the local join actions indexed in the lemma, there exists for each fixed i

$$\Pr\left(\left|Q^*(s,g,\pi^*) - \widehat{Q}^*_\kappa(s,\widehat{g},\pi^*(\cdot|s,\widehat{g}))|_{\{i\}\cup\Delta_i)}\right| > \frac{L_P}{1-\gamma}\|r_\ell\|_\infty\Phi_{\kappa,\delta}\right) \leq |\mathcal{A}|\cdot\kappa^{|\mathcal{A}|}\delta_t$$

Thus we have

$$\mathbb{E}_{(s,g)\sim\mathcal{D}}\left[\left|Q^*(s,g,\pi^*(\cdot \mid s,g)) - \widehat{Q}^{\text{est}}_\kappa\big(s,\widehat{g},\pi^*(\cdot|s,\widehat{g})\big)\big|_{\{i\}\cup\Delta_i}\big)\right|\right] \leq \frac{L_P\|r_\ell\|_\infty}{1-\gamma}\Phi_{\kappa,\delta} + \epsilon_{\kappa,m} + \frac{\|r_\ell\|_\infty}{1-\gamma}|\mathcal{A}|\cdot\kappa^{|\mathcal{A}|}\delta_t$$

which completes the proof after averaging over $i \in [n]$. □

Lemma D.5 (Union bound under graphon-weighted subsampling)**.** *Fix $(s,g) \in \mathcal{S}\times\mathcal{G}$. Let $\delta_t, \ldots \delta_T \in (0,1)$ be given. Then define for each $t \in [T]$:*

$$\Phi_{\kappa,t} \coloneqq \sqrt{\frac{|\mathcal{S}|\ln 2 + \ln(2/\delta_t)}{2\kappa}}$$

For each $t \in [T]$ and each local joint action $a_{\{i\}\cup\Delta_i} \in \mathcal{A}^{\{i\}\cup\Delta_i}$, define the deviation event

$$B_t^{a_{\{i\}\cup\Delta_i},\Delta_i} \coloneqq \left\{|Q^*(s,g,\pi^*(\cdot|s,g)) - \widehat{Q}^*_\kappa(s,\widehat{g},\pi^{\text{est}}_\kappa(\cdot|s,\widehat{g})| > \frac{L_P}{1-\gamma}\cdot\Phi_{\kappa,t}\|r_\ell(\cdot,\cdot)\|_\infty + \epsilon_{\kappa,m}\right\}.$$

Define the bad event at time t as the union over all indices, including the support of the distribution used to generate Δ_i:

$$B_t = \bigcup_{a_{\{i\}\cup\Delta_i}\in\mathcal{A}^{\{i\}\cup\Delta_i}}\ \bigcup_{\Delta_i\in\text{Supp}(\tilde{w}_{i,\cdot}^{\otimes(k-1)})} B_t^{a_{\{i\}\cup\Delta_i},\Delta_i}.$$

Next, let $B = \bigcup_{t=1}^T B_t$. Then the probability that no bad event B_t occurs is

$$\Pr(B^c) \ \geq\ 1 - |\mathcal{A}|\cdot\kappa^{|\mathcal{A}|}\sum_{t=1}^T \delta_t.$$

Proof.

$$\begin{aligned}
&\left|Q^*(s,g,\pi^*) - \widehat{Q}^{\text{est}}_\kappa(s,g,i,F_{\Delta_i},a_{\{i\}\cup\Delta_i})\right| \\
&\qquad\leq \left|Q^*(s,g,\pi^*) - \widehat{Q}^*_\kappa(s,\widehat{g},\pi^*(\cdot|s,\widehat{g}))\right| + \left|\widehat{Q}^*_\kappa(s,g,\pi^*) - \widehat{Q}^{\text{est}}_\kappa(s,\widehat{g},\pi^*(\cdot|s,\widehat{g}))\right| \\
&\qquad\leq \left|Q^*(s,g,\pi^*) - \widehat{Q}^*_\kappa(s,\widehat{g},\pi^*(\cdot|s,\widehat{g})\right| + \left\|\widehat{Q}^*_\kappa(s,g,\pi^*) - \widehat{Q}^{\text{est}}_\kappa(s,\widehat{g},\pi^*(\cdot|s,\widehat{g}))\right\|_\infty \\
&\qquad\leq \left|Q^*(s,g,\pi^*) - \widehat{Q}^*_\kappa(s,\widehat{g},\pi^*(\cdot|s,\widehat{g}))\right| + \epsilon_{\kappa,m}
\end{aligned}$$

By Lemma C.8, with probability at least $1-\delta_t$,

$$\begin{aligned}
\left|Q^*(s,g,\pi^*) - \widehat{Q}^*_\kappa(s,\widehat{g},\pi^*(\cdot|s,\widehat{g}))\right| &\leq \frac{L_P}{1-\gamma}\cdot\Phi_{\kappa,t}\|r_\ell(\cdot,\cdot)\|_\infty \\
&\leq \frac{L_P\|r_\ell\|_\infty}{1-\gamma}\sqrt{\frac{|\mathcal{S}|\ln 2 + \ln(2/\delta_t)}{2\kappa}}
\end{aligned}$$

Therefore, $B_t^{a_{\{i\}\cup\Delta_i},\Delta_i}$ occurs with probability at most δ_t. Now let us define the empirical action distribution induced by the sampled neighborhood actions. For this, let $\widehat{g}_a \in \mu_\kappa(\mathcal{A})$ where $\mu_\kappa(\mathcal{A}) \coloneqq \left\{\nu \in \mathcal{P}(\mathcal{A}) : \nu(a) \in \left\{0, \frac{1}{\kappa-1}, \ldots, 1\right\}\right\}$. Since the local estimator depends on $\{a_j\}_{j\in\Delta_i}$ only through the empirical measure $\widehat{g}_a$), union bounding across all events parameterized by (i,Δ_i) is covered by union bounding across the finite set of possible empirical distributions $\widehat{g}_a \in \mu_\kappa(\mathcal{A})$. For fixed t, now union bound across the index sets in B_t:

$$\Pr[B_t] = \Pr\left[\bigcup_{a_{\{i\}\cup\Delta_i}\in\mathcal{A}^{\{i\}\cup\Delta_i}}\ \bigcup_{\Delta_i\in\text{Supp}(\tilde{w}_{i,\cdot}^{\otimes(k-1)})} B_t^{a_{\{i\}\cup\Delta_i},\Delta_i}\right]$$

$$\leq \sum_{a_{\{i\}\cup\Delta_i} \in \mathcal{A}^{\{i\}\cup\Delta_i}} \sum_{\widehat{g}_a \in \mu(\mathcal{A})} \delta_t$$
$$\leq |\mathcal{A}| \cdot |\mu_\kappa(\mathcal{A})| \cdot \delta_t$$

Each $\widehat{g}_a$ corresponds to a count vector $(c_a)_{a\in\mathcal{A}} \in \mathbb{N}^{|\mathcal{A}|}$ with $\sum_{a\in\mathcal{A}} c_a = \kappa - 1$ (where c_a is how many sampled neighbors took action a), and hence we have

$$|\mu_\kappa(\mathcal{A})| = \binom{(\kappa-1)+|\mathcal{A}|-1}{|\mathcal{A}|-1} \leq \kappa^{|\mathcal{A}|-1} \leq \kappa^{|\mathcal{A}|}$$

giving us $\Pr(B_t) \leq |\mathcal{A}| \cdot \kappa^{|\mathcal{A}|}\delta_t$. Finally, applying the union bound over T gives us

$$\Pr[B] = \Pr\left[\bigcup_{t=1}^{T} B_t\right] \leq \sum_{t=1}^{T} \Pr[B_t] \leq |\mathcal{A}| \cdot \kappa^{|\mathcal{A}|} \sum_{t=1}^{T} \delta_t.$$

Therefore, we have $\Pr(\bar{B}) \geq 1 - |\mathcal{A}| \cdot \kappa^{|\mathcal{A}|} \sum_{t=1}^{T} \delta_t$, which completes the proof. □

Corollary D.6 (Optimizing Parameters)**.**

$$V^{\pi^*}(s,g) - V^{\pi_\kappa^{\text{est}}}(s,g) \leq \frac{2L_P \cdot \|r_\ell\|_\infty}{(1-\gamma)^2}\sqrt{\frac{|\mathcal{S}|\ln 2 + \ln(2/\delta)}{2\kappa}} + \frac{2\|r_\ell\|_\infty}{(1-\gamma)^2}|\mathcal{A}| \cdot \kappa^{|\mathcal{A}|}\delta + \frac{\epsilon_{\kappa,m}}{1-\gamma}.$$

Setting $\delta = \frac{(1-\gamma)^2}{20\|r_\ell\|_\infty |\mathcal{A}|\kappa^{|\mathcal{A}|+1/2}}$ *recovers a decaying optimality gap on the order*

$$V^{\pi^*}(s,g) - V^{\pi_\kappa^{\text{est}}}(s,g) \leq \frac{2L_P \cdot \|r_\ell\|_\infty}{(1-\gamma)^2}\sqrt{\frac{|\mathcal{S}|\ln 2 + |\mathcal{A}|\ln\frac{20\|r_\ell\|_\infty|\mathcal{A}|\kappa}{(1-\gamma)^2}}{2\kappa}} + \frac{1}{10\sqrt{\kappa}} + \frac{\epsilon_{\kappa,m}}{1-\gamma}.$$

Finally, using the probabilistic bound from Lemma D.9 that with probability at least $1 - \frac{1}{100e^\kappa}$, $\epsilon_{\kappa,m} \leq \frac{2}{\sqrt{\kappa}}$, *we get*

$$V^{\pi^*}(s,g) - V^{\pi_\kappa^{\text{est}}}(s,g) \leq \frac{2L_P \cdot \|r_\ell\|_\infty}{(1-\gamma)^2}\sqrt{\frac{|\mathcal{S}|\ln 2 + |\mathcal{A}|\ln\frac{20\|r_\ell\|_\infty|\mathcal{A}|\kappa}{(1-\gamma)^2}}{2\kappa}} + \frac{21}{10\sqrt{\kappa}},$$

which completes the proof of theorem 4.3.

D.1 Bounding the Bellman Error

To bound the Bellman error $\epsilon_{k,m}$, we first recall Hoeffding's inequality (32).

Lemma D.7 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables such that $a_i \leq X_i \leq b_i$ almost surely. Let $S_n = \sum_{i=1}^n X_i$. Then, for all $t > 0$, we have that*

$$\Pr[|S_n - \mathbb{E}[S_n]| \geq t] \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Lemma D.8. *Fix $\kappa \geq 1$ and let $\widehat{\mathcal{T}}_\kappa$ and $\widehat{\mathcal{T}}_{\kappa,m}$ be as in Definitions 3.4 and 3.5. Let $\widehat{Q}^*_\kappa$ denote the unique fixed point of $\widehat{\mathcal{T}}_\kappa$ and $\widehat{Q}^*_{\kappa,m}$ the unique fixed point of $\widehat{\mathcal{T}}_{\kappa,m}$. Let $N_\kappa := |\mathcal{S}|^2|\mathcal{A}|^2\kappa^{|\mathcal{S}||\mathcal{A}|}$. Then for any $\rho \in (0,1)$, with probability at least $1-\rho$ over the sampling used to form $\widehat{\mathcal{T}}_{\kappa,m}$, we have*

$$\epsilon_{\kappa,m} := \|\widehat{Q}^*_{\kappa,m} - \widehat{Q}^*_\kappa\|_\infty \leq \frac{\gamma\|r_\ell\|_\infty}{(1-\gamma)^2}\sqrt{\frac{2\ln(2N_\kappa/\rho)}{m}}.$$

Proof. We first control the Bellman operator's deviation at the fixed point $\widehat{Q}^*_\kappa$. For any $(s,a,z) \in \mathcal{S}\times\mathcal{A}\times\mathcal{Z}_\kappa$, define

$$Y := M_\kappa \widehat{Q}^*_\kappa(S', g'),$$

where $(S', g') \sim J_\kappa(\cdot \mid s, a, z)$. By Lemma B.4, we have $\|\widehat{Q}^*_\kappa\|_\infty \leq \frac{\|r_\ell\|_\infty}{1-\gamma}$, hence $|Y| \leq \frac{\|r_\ell\|_\infty}{1-\gamma}$ almost surely.
The empirical operator uses i.i.d. samples $Y_1, \ldots, Y_m$ of Y and forms their average. By Hoeffding's inequality,

$$\Pr\left[\left|\frac{1}{m}\sum_{\ell=1}^{m} Y_\ell - \mathbb{E}[Y]\right| \geq \eta\right] \leq 2\exp\Big(-\frac{m(1-\gamma)^2\eta^2}{2\|r_\ell\|_\infty^2}\Big).$$

Taking a union bound over all N_κ tuples (s,a,z) gives

$$\Pr\Big[\|\widehat{\mathcal{T}}_{\kappa,m}\widehat{Q}^*_\kappa - \widehat{\mathcal{T}}_\kappa\widehat{Q}^*_\kappa\|_\infty \geq \gamma\eta\Big] \leq 2N_\kappa\exp\Big(-\frac{m(1-\gamma)^2\eta^2}{2\|r_\ell\|^2}\Big).$$

Setting the right-hand side to ρ and solving for η, we have

$$\eta = \frac{\|r_\ell\|_\infty}{1-\gamma}\sqrt{\frac{2\ln(2N_\kappa/\rho)}{m}}.$$

Thus, with probability at least $1-\rho$, we have

$$\|\widehat{\mathcal{T}}_{\kappa,m}\widehat{Q}^*_\kappa - \widehat{\mathcal{T}}_\kappa\widehat{Q}^*_\kappa\|_\infty \le \gamma\frac{\|r_\ell\|_\infty}{1-\gamma}\sqrt{\frac{2\ln(2N_\kappa/\rho)}{m}}.$$

Finally, using the contraction bound since $\widehat{Q}^*_{\kappa,m} = \widehat{T}_{\kappa,m}\widehat{Q}^*_{\kappa,m}$, we have

$$\begin{aligned}\|\widehat{Q}^*_{\kappa,m} - \widehat{Q}^*_\kappa\|_\infty &\le \frac{1}{1-\gamma}\|\widehat{\mathcal{T}}_{\kappa,m}\widehat{Q}^*_\kappa - \widehat{\mathcal{T}}_\kappa\widehat{Q}^*_\kappa\|_\infty \\ &\le \gamma\frac{\|r_\ell\|_\infty}{(1-\gamma)^2}\sqrt{\frac{2\ln(2N_\kappa/\rho)}{m}},\end{aligned}$$

which yields the stated bound.

□

Lemma D.9. *If* $T = \frac{2}{1-\gamma}\log\frac{\|r_\ell\|_\infty\sqrt{\kappa}}{1-\gamma}$, GMFS: Learning *runs in time* $\widetilde{O}(T|\mathcal{A}|^3|\mathcal{S}|^3\kappa^{2+2|\mathcal{S}||\mathcal{A}|}\|r_\ell\|_\infty)$*, while accruing a Bellman noise* $\epsilon_{\kappa,m} \le \frac{1}{5\sqrt{\kappa}}$ *with probability at least* $1-\frac{1}{100e^\kappa}$, .

Proof. We first prove that $\|\widehat{Q}^T_\kappa - \widehat{Q}^*_\kappa\|_\infty \le \frac{1}{\sqrt{\kappa}}$.
For this, it suffices to show $\gamma^T\frac{\|r_\ell\|_\infty}{1-\gamma} \le \frac{1}{\sqrt{\kappa}} \implies \gamma^T \le \frac{1-\gamma}{\|r_\ell\|_\infty\sqrt{\kappa}}$. Then, using $\gamma = 1-(1-\gamma) \le e^{-(1-\gamma)}$, it again suffices to show $e^{-(1-\gamma)T} \le \frac{1-\gamma}{\|r_\ell\|_\infty\sqrt{\kappa}}$. Taking logarithms, we have

$$\begin{aligned}\exp(-T(1-\gamma)) &\le \frac{1-\gamma}{\|r_\ell\|_\infty\sqrt{\kappa}} \\ -T(1-\gamma) &\le \log\frac{1-\gamma}{\|r_\ell\|_\infty\sqrt{\kappa}} \\ T &\ge \frac{1}{1-\gamma}\log\frac{\|r_\ell\|_\infty\sqrt{\kappa}}{1-\gamma}\end{aligned}$$

Since $T = \frac{2}{1-\gamma}\log\frac{\|r_\ell\|_\infty\sqrt{\kappa}}{1-\gamma} > \frac{1}{1-\gamma}\log\frac{\|r_\ell\|_\infty\sqrt{\kappa}}{1-\gamma}$, the condition holds and $\|\widehat{Q}^T_\kappa - \widehat{Q}^*_\kappa\|_\infty \le \frac{1}{\sqrt{\kappa}}$.
Then, rearranging Lemma D.8 and incorporating the convergence error of the $\widehat{Q}_\kappa$-function, one has that with probability at least $1-\rho$,

$$\epsilon_{\kappa,m} \le \frac{1}{\sqrt{\kappa}} + \gamma\frac{\|r_\ell\|_\infty}{(1-\gamma)^2}\sqrt{\frac{2\ln(2N_\kappa/\rho)}{m}}. \tag{12}$$

If we desire $\epsilon_{\kappa,m} \le \frac{c}{\sqrt{\kappa}}$, it suffices to choose

$$m^* \ge \frac{2\gamma^2}{(1-\gamma)^4}\|r_\ell\|^2_\infty \cdot \frac{\kappa}{c^2} \cdot \ln\left(\frac{2|\mathcal{S}||\mathcal{A}||\mathcal{Z}_\kappa|}{\rho}\right).$$

Letting $\rho = \frac{1}{100e^\kappa}$, $c = \frac{1}{5}$, and using $|\mathcal{Z}_\kappa| \le |\mathcal{S}||\mathcal{A}|\kappa^{|\mathcal{S}||\mathcal{A}|}$, we have that

$$m^* \ge \frac{25\kappa^2\gamma^2}{(1-\gamma)^4}\|r_\ell\|^2_\infty \cdot \ln\left(200|\mathcal{S}|^2|\mathcal{A}|^2\kappa^{|\mathcal{S}||\mathcal{A}|}\right) \tag{13}$$

attains a Bellman error of $\epsilon_{\kappa,m} \le \frac{1}{5\sqrt{\kappa}}$ with probability at least $1-\frac{1}{100e^\kappa}$.

Finally, the runtime of our learning algorithm is

$$O(mT|\mathcal{S}|^2|\mathcal{A}|^2\kappa^{|\mathcal{S}||\mathcal{A}|}) = \widetilde{O}(|\mathcal{A}|^3|\mathcal{S}|^3\kappa^{2+2|\mathcal{S}||\mathcal{A}|}\|r_\ell\|_\infty),$$

which is still polynomial in κ, proving the claim. □

E EXTENSION TO STOCHASTIC REWARDS

As in Anand et al. (4), this section extends the GMFS framework to environments where rewards are stochastic. While the primary analysis in this work assumes deterministic local rewards, many real-world multi-agent systems, such as sensor noise, result in rewards drawn from a probability distribution.

Suppose we are given a family of distributions $\{\mathcal{L}_{s_i,a_i,g_i}\}_{(s_i,a_i,g_i)\in\mathcal{S}\times\mathcal{A}\times\mathfrak{G}(\mathcal{S}),\forall i\in[n]}$. For joint states, actions, and neighborhood aggregates $(\mathbf{s},\mathbf{a},\mathbf{g})\in\mathcal{S}^n\times\mathcal{A}^n\times\mathfrak{G}(\mathcal{S})^n$, let $R(\mathbf{s},\mathbf{a},\mathbf{g})$ denote a stochastic team reward of the form:

$$R(\mathbf{s},\mathbf{a},\mathbf{g}) = \frac{1}{n}\sum_{i\in[n]} r_\ell(s_i,a_i,g_i), \tag{14}$$

where each local reward is an independent random variable $r_\ell(s_i,a_i,g_i)\sim\mathcal{L}_{s_i,a_i,g_i}$. We assume that these distributions are uniformly bounded.

Assumption E.1 (Bounded Stochastic Rewards)**.** *Define the union of the supports of all reward distributions as:*

$$\bar{\mathcal{L}} = \bigcup_{(s,a,g)\in\mathcal{S}\times\mathcal{A}\times\mathfrak{G}(\mathcal{S})} \mathrm{supp}(\mathcal{L}_{s,a,g}),$$

where $\mathrm{supp}(\mathcal{D})$ *denotes the support of distribution* $\mathcal{D}$*. Let* $\widehat{\mathcal{L}} = \sup(\bar{\mathcal{L}})$ *and* $\widetilde{\mathcal{L}} = \inf(\bar{\mathcal{L}})$*. We assume that* $-\infty < \widetilde{\mathcal{L}} \le \widehat{\mathcal{L}} < \infty$*, and that these bounds are known a priori.*

To handle this stochasticity, we introduce a randomized version of our empirical operator.

Definition E.1 (Randomized Empirical Bellman Operator)**.** *Let* $\widehat{\mathcal{T}}^{\mathrm{rand}}_{\kappa,m}$ *be the randomized empirical adapted Bellman operator such that:*

$$\widehat{\mathcal{T}}^{\mathrm{rand}}_{\kappa,m}\widehat{Q}^t_{\kappa,m}(s,a,z) = \widetilde{r}_\ell(s,a,g_z) + \frac{\gamma}{m}\sum_{\ell\in[m]} \mathcal{M}_\kappa\widehat{Q}_{\kappa,m}(s'_\ell,g'_\ell), \tag{15}$$

where $\widetilde{r}_\ell(s,a,g_z)$ *is a single sample drawn from* $\mathcal{L}_{s,a,g_z}$*.*

GMFS with Stochastic Rewards. Our proposed extension of GMFS averages Ξ independent samples of the randomized operator $\widehat{\mathcal{T}}^{\mathrm{rand}}_{\kappa,m}$ to update the Q-function. One can show that $\widehat{\mathcal{T}}^{\mathrm{rand}}_{\kappa,m}$ remains a contraction operator with modulus γ. Then, by the Banach Fixed Point Theorem, the operator $\widehat{\mathcal{T}}^{\mathrm{rand}}_{\kappa,m}$ admits a unique fixed point $\widehat{Q}^{\mathrm{rand}}_{\kappa,m}$ toward which the iterates converge.

Algorithm 4 GMFS (Graphon Mean-Field Subsampling): Offline Learning with Stochastic Rewards

Require: Number of iterations T, subsampling parameters κ and m, discount parameter γ, averaging parameter Ξ, and generative oracle O.

Initialize $\widehat{Q}^{(0)}_{\kappa,m}(s,a,z) = 0$ for all $(s,a,z)\in\mathcal{S}\times\mathcal{A}\times\mathcal{Z}_\kappa$.
for $t = 0,\dots,T-1$ **do**
  for $(s,a,z)\in\mathcal{S}\times\mathcal{A}\times\mathcal{Z}_\kappa$ **do**
    $\rho = 0$
    for $\xi = 1,\dots,\Xi$ **do**
      Sample a realization of the randomized operator: $\rho = \rho + \widehat{\mathcal{T}}^{\mathrm{rand}}_{\kappa,m}\widehat{Q}^t_{\kappa,m}(s,a,z)$
    Update $\widehat{Q}^{(t+1)}_{\kappa,m}(s,a,z) = \rho/\Xi$
Return $\widehat{Q}^{(T)}_{\kappa,m}$.

To bound the error introduced by the stochasticity of the rewards, we recall a standard concentration result.

Theorem E.1 (Hoeffding's Theorem (70)). *Let* $X_1,\dots,X_n$ *be independent random variables such that* $a_i \le X_i \le b_i$ *almost surely. Let* $S_n = \sum_{i=1}^n X_i$*. Then, for all* $\epsilon > 0$*:*

$$\Pr[|S_n - \mathbb{E}[S_n]| \ge \epsilon] \le 2\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n (b_i-a_i)^2}\right). \tag{16}$$

Lemma E.2 (Uniform concentration of averaged stochastic rewards)**.** *Under Assumption E.1, define* $\Delta_L \coloneqq \widehat{L} - \widetilde{L}$*. For any* $\delta\in(0,1)$ *and any averaging parameter* $\Xi\in\mathbb{N}$*, let*

$$\widetilde{r}_\ell(s,a,g) \coloneqq \frac{1}{\Xi}\sum_{\xi=1}^{\Xi} r_\ell^{(\xi)}(s,a,g), \tag{17}$$

where $r_\ell^{(\xi)}(s,a,g) \overset{\text{i.i.d.}}{\sim} L_{s,a,g}$*. Next, let* $\bar{r}_\ell(s,a,g) \coloneqq \mathbb{E}[r_\ell(s,a,g)]$*. Then with probability at least* $1-\delta$*, we have*

$$\sup_{(s,a,g)\in\mathcal{S}\times\mathcal{A}\times\mathcal{G}_\kappa} \left|\widetilde{r}_\ell(s,a,g) - \bar{r}_\ell(s,a,g)\right| \le \Delta_L\sqrt{\frac{\ln\left(2|\mathcal{S}||\mathcal{A}||\mathcal{G}_\kappa|/\delta\right)}{2\Xi}}. \tag{18}$$

Moreover, if $\sup_{s,a,g}|\widetilde{r}_\ell - \bar{r}_\ell| \le \varepsilon_r$*, then* $\|\widehat{Q}^{*,\mathrm{avg}}_\kappa - \widehat{Q}^*_\kappa\|_\infty \le \frac{\varepsilon_r}{1-\gamma}$*, and the performance bound of theorem 4.3 degrades by at most* $\frac{\varepsilon_r}{(1-\gamma)^2}$*.*

Proof. Fix any (s, a, g). Then, the random variables $r_\ell^{(\xi)}(s, a, g)$ are i.i.d. and bounded in the range $[\widetilde{L}, \widehat{L}]$. Then, by Hoeffding's inequality,

$$\Pr\left[\left|\widetilde{r}_\ell(s,a,g) - \bar{r}_\ell(s,a,g)\right| \geq \varepsilon\right] \leq 2\exp\left(-\frac{2\Xi\varepsilon^2}{\Delta_L^2}\right).$$

Union-bounding over the finite set $S \times A \times G_\kappa$ gives

$$\Pr\left[\sup_{s,a,g}\left|\widetilde{r}_\ell(s,a,g) - \bar{r}_\ell(s,a,g)\right| \geq \varepsilon\right] \leq 2|S||A||G_\kappa|\exp\Big(-\frac{2\Xi\varepsilon^2}{\Delta_L^2}\Big).$$

Reparameterizing the RHS to δ and solving for ε yields the first claim. For the second claim, it suffices to note that replacing the reward function in the Bellman operator changes the operator by at most ε_r in ℓ_∞. Since the Bellman operator is a γ-contraction, our fixed-point perturbation yields that $\|\widehat{Q}_\kappa^{*,\text{avg}} - \widehat{Q}_\kappa^*\|_\infty \leq \frac{\varepsilon_r}{1-\gamma}$. The same bound transfers to value function by the performance difference lemma, while picking up another $\frac{1}{1-\gamma}$ factor. □

Remark E.3. *From Lemma E.2, we have that in order to keep the reward-averaging contribution to be at most $\frac{1}{\sqrt{\kappa}}$ with probability at least $1-\delta$, it suffices to choose*

$$\Xi \geq \frac{\Delta_L^2\kappa}{2} \cdot \ln\left(\frac{2|\mathcal{S}||\mathcal{A}||\mathcal{G}_\kappa|}{\delta}\right). \tag{19}$$

Through this averaging argument, we observe that as the subsampling parameter κ increases, the optimality gap decays to zero while the probability of success approaches one. The argument can be strengthened by estimating $\widehat{\mathcal{L}}$ and $\widetilde{\mathcal{L}}$ using order statistics to bound estimation errors (41)). This extension would be an essential step in transitioning this framework to a fully online learning setting via a stochastic approximation scheme. Furthermore, one could incorporate variance-based analysis (37), leveraging the skewness of the reward distribution and allowing for the assignment of optimism or pessimism scores to the resulting estimates.

F EXTENSION TO OFF-POLICY LEARNING

A limitation of the planning approach in Algorithm 1 is that it computes $\widehat{Q}_\kappa^*$ by assuming access to a generative oracle for the transition functions P_g, P_l and the reward function $r(\cdot,\cdot,\cdot)$. In certain realistic RL applications, a generative oracle like such is unavailable. Instead, it is more desirable to perform off-policy learning, where the agent learns from *historical data* (25). In this setting, the agents learn the target policy $\widehat{\pi}_\kappa^*$ using a dataset generated by a different behavior policy π_b (the strategy used to explore the environment). There is a significant body of work on the theoretical guarantees in off-policy learning (14–17).

In fact, these previous results are amenable to transforming guarantees about offline Q-learning to off-policy Q-learning, typically at the cost of $\log|\mathcal{S}||\mathcal{A}|$ factors in the sample complexity or runtime. Therefore, this section demonstrates that our previous results satisfy the necessary conditions to extend the GMFS framework to the off-policy Q-learning for the subsampled $\widehat{Q}_\kappa$-function. We show that, in expectation, the learned policy π_κ maintains a decaying optimality gap of $\widetilde{O}(1/\sqrt{\kappa})$, where the randomness is over the heuristic behavior policy π_b.
The off-policy $\widehat{Q}_\kappa$-learning algorithm is an iterative procedure to estimate the optimal $\widehat{Q}_\kappa$-function as follows: first, a sample trajectory $\{(s_t, a_t, z_t)\}_{t\geq 0}$ is collected using a suitable behavior policy π_b. After initializing $\widehat{Q}_\kappa^0 : \mathcal{S}\times\mathcal{A}\times\mathcal{Z}_\kappa \to \mathbb{R}$, the iterate $\widehat{Q}_\kappa^t(s,a,z)$ is updated for each $t \geq 0$ according to:

$$\widehat{Q}_\kappa^{t+1}(s,a,z) = (1-\alpha_t)\widehat{Q}_\kappa^t(s,a,z) + \alpha_t\Big(r_\ell(s,a,g_z) + \gamma\mathcal{M}_\kappa\widehat{Q}_\kappa^t(s',g')\Big), \tag{20}$$

where $\alpha_t \in (0,1)$ is the learning rate. Note that the update in eq. (20) is sample-based; it does not require an expectation over the transition dynamics and can be computed directly from historical data. To ensure convergence, we make the following standard ergodicity assumption:

Assumption F.1 (Ergodicity of Behavior Policy)**.** *The behavior policy π_b satisfies $\pi_b(a|s,z) > 0$ for all $(s,a,z) \in \mathcal{S}\times\mathcal{A}\times\mathcal{Z}_\kappa$. Additionally, the Markov chain $\mathcal{M} = \{(s_t, z_t)\}_{t\geq 0}$ induced by π_b is irreducible and aperiodic with stationary distribution μ and mixing time: $t_\delta(\mathcal{M}) = \min\{t \geq 0 : \max_{(s,z)\in\mathcal{S}\times\mathcal{Z}_\kappa}\|P^t((s,z),\cdot) - \mu(\cdot)\|_{\text{TV}} \leq \delta\}$. There are many heuristics for constructing such behavior policies are well-established in the literature (25).*

Theorem F.1. *Let π_κ be the policy learned through off-policy $\widehat{Q}_\kappa$-learning. Under Assumption F.1, with probability at least $1 - \frac{1}{100e^\kappa}$, we have:*

$$\begin{aligned}\mathbb{E}[V^{\pi^*}(s_0) - V^{\pi_\kappa}(s_0)] &\leq \frac{L_P\|r_\ell\|_\infty}{(1-\gamma)^2}\sqrt{\frac{|\mathcal{S}|\ln 2 + |\mathcal{A}|\ln\frac{20\|r_\ell\|_\infty|\mathcal{A}|\kappa}{1-\gamma}}{2\kappa}}\sqrt{\ln\frac{40\|r_\ell\|_\infty|\mathcal{S}||\mathcal{A}|\kappa^{|\mathcal{A}||\mathcal{S}|+\frac{1}{2}}}{(1-\gamma)^2}} + \frac{21}{10\sqrt{\kappa}} \\ &= \widetilde{O}(1/\sqrt{\kappa}),\end{aligned}$$

where the expectation is taken over the stochasticity of the behavior policy π_b.

Proposition F.1 (Analytical Properties of the Subsampled Operator)**.** *The following properties hold for the subsampled Q-function and the associated Markov chain:*

(1) For any $s \in \mathcal{S}, a \in \mathcal{A}$ and neighborhood marginals $g, g' \in \mathcal{G}_\kappa$, $\|\widehat{Q}_\kappa(s,a,z) - \widehat{Q}_\kappa(s,a,z')\| \le \frac{L_P}{1-\gamma}\|r_\ell\|_\infty \cdot \mathrm{TV}(g,g')$ (Theorem C.1).

(2) $\|\widehat{Q}_\kappa\|_\infty \le \frac{\|r_\ell\|_\infty}{1-\gamma}$ (Lemma B.4).

(3) $\|\widehat{\mathcal{T}}_\kappa Q - \widehat{\mathcal{T}}_\kappa Q'\|_\infty \le \gamma\|Q - Q'\|_\infty$ (Lemma B.3).

(4) The Markov chain $\mathcal{M}$ induced by π_b satisfies the rapid mixing property defined in Assumption F.1.

By treating the single-trajectory update of the $\widehat{Q}_\kappa$-function as a noisy approximation of the expected update from the ideal Bellman operator, Chen et al. (15) uses Markovian stochastic approximation to bound the mean-squared error $\mathbb{E}[\|\widehat{Q}_\kappa^T - \widehat{Q}_\kappa^*\|_\infty^2]$. We restate their result adapted to our subsampled regime:

Theorem F.2 (Theorem 3.1 in Chen et al. (15) adapted to GMFS). *Suppose the learning rate is constant, $\alpha_t = \alpha$ for all $t \ge 0$, and is chosen such that $\alpha t_\alpha(\mathcal{M}) \le c_{Q,0}\frac{(1-\gamma)^2}{\log|\mathcal{S}||\mathcal{G}_\kappa|}$, where $c_{Q,0}$ is a numerical constant. Then, under the properties in Proposition F.1, for all $t \ge t_\alpha(\mathcal{M})$, we have:*

$$\mathbb{E}[\|\widehat{Q}_\kappa^t - \widehat{Q}_\kappa^*\|_\infty^2] \le c_{Q,1}\left(1 - \frac{(1-\gamma)\alpha}{2}\right)^{t - t_\alpha(\mathcal{M})} + c_{Q,2}\frac{\log \kappa^{|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^2}\alpha t_\alpha(\mathcal{M}),$$

where $c_{Q,1} = 3\left(\frac{\|r_\ell\|_\infty}{1-\gamma} + 1\right)^2$ and $c_{Q,2} = 912e\left(\frac{3\|r_\ell\|_\infty}{1-\gamma} + 1\right)^2$. The expectation is taken over the stochasticity of the behavior policy π_b.

Corollary F.3 (Corollary 3.2 in Chen et al. (15) adapted to our setting)**.** *To ensure that $\mathbb{E}[\|\widehat{Q}_\kappa^t - \widehat{Q}_\kappa^*\|_\infty] \le \frac{1}{100\sqrt{\kappa}}$, the required number of iterations t satisfies:*

$$t > \widetilde{O}\left(\frac{\kappa \log^2(100\sqrt{\kappa})|\mathcal{S}||\mathcal{A}|\kappa^{|\mathcal{A}||\mathcal{S}|}}{(1-\gamma)^5}\right).$$

With this sample complexity, we recover an expected value analog of theorem 4.3 via the triangle inequality.

Corollary F.4. *For $\delta \in (0,1)$, with probability at least $1-\delta$, we have:*

$$\mathbb{E}[\widehat{Q}_\kappa^*(s,a,\widehat{z}) - Q_n^*(s,a,z)] \le \frac{L_P\|r_\ell\|_\infty}{1-\gamma}\sqrt{\frac{|\mathcal{S}|\ln 2 + \ln(2/\delta)}{2\kappa}},$$

where the expectation is taken over the stochasticity of the behavior policy π_b.

In turn, following the derivation in the proof of theorem 4.3, it is straightforward to verify that this yields a bound on the expected performance difference for off-policy learning.

Corollary F.5. *With probability at least $1 - \frac{1}{100e^\kappa}$, the expected performance gap satisfies:*

$$\begin{aligned}\mathbb{E}[V^{\pi^*}(s,g) - V^{\pi_\kappa}(s,g)] &\le \frac{L_P\|r_\ell\|_\infty}{(1-\gamma)^2}\sqrt{\frac{|\mathcal{S}|\ln 2 + |\mathcal{A}|\ln\frac{20\|r_\ell\|_\infty|\mathcal{A}|\kappa}{1-\gamma}}{2\kappa}}\sqrt{\ln\frac{40\|r_\ell\|_\infty|\mathcal{S}||\mathcal{A}|\kappa^{|\mathcal{A}||\mathcal{S}|+\frac{1}{2}}}{(1-\gamma)^2}} + \frac{21}{10\sqrt{\kappa}} \\ &= \widetilde{O}(1/\sqrt{\kappa}),\end{aligned}$$

where the expectation is taken over the stochasticity of the behavior policy π_b.

G EXTENSION TO CONTINUOUS STATE SPACES

Multi-agent settings in which agents operate in continuous state space have numerous applications in optimization, control, and synchronization (48–50, 59). Therefore, this section is devoted to extending the tabular analysis of section 4 to non-tabular environments with a compact (and possibly continuous) state space. The main technical differences from the finite setting are that the state space $\mathcal{S}$ is uncountable, hence one must work in function spaces, and the κ-sampled mean-field state space $\mathcal{G}_\kappa$ is infinite when $\mathcal{S}$ is continuous, which requires all the union bounds over $|\mathcal{G}_\kappa|$ (for instance, in Lemma D.5) to be replaced with covering-number arguments. For this section, we keep the mean-field as part of the state. Concretely, the representative agent's state is $x = (s,g)$ where $s \in \mathcal{S}$ is the agent's local state and g is a mean-field (neighborhood) distribution over $\mathcal{S}$, which can exist over $\mathcal{G}$ or $\mathcal{G}_\kappa$.

Definition G.1 (Augmented mean-field state space)**.** Let $(\mathcal{S}, d_{\mathcal{S}})$ be a compact metric space with Borel σ-algebra $\mathcal{B}(\mathcal{S})$. Let $\mathcal{A}$ be a finite action set (the extension to compact $\mathcal{A}$ is analogous via an additional action-space covering). Let $\mathcal{M} \coloneqq \mathcal{P}(\mathcal{S})$ be the set of Borel probability measures on $\mathcal{S}$. For $\kappa \in \mathbb{N}$, define the κ-empirical mean-field class

$$\mathcal{M}_\kappa \coloneqq \left\{ \frac{1}{\kappa} \sum_{m=1}^{\kappa} \delta_{x_m} : x_1, \ldots, x_\kappa \in \mathcal{S} \right\} \subset \mathcal{M},$$

which is uncountable whenever $\mathcal{S}$ is infinite. We define the augmented (mean-field) state spaces $\mathcal{X} \coloneqq \mathcal{S} \times \mathcal{M}$ and $\mathcal{X}_\kappa \coloneqq \mathcal{S} \times \mathcal{M}_\kappa$, equipped with the product Borel σ-algebras. Then, a Markov policy is a measurable map $\pi : \mathcal{X} \to \Delta(\mathcal{A})$.

Definition G.2 (κ-surrogate MDP)**.** Fix a discount $\gamma \in (0, 1)$. We consider an augmented MDP on $\mathcal{X}$ with reward $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and transition kernel $\bar{P}$ from $\mathcal{X} \times \mathcal{A}$ to $\mathcal{X}$. In GMFS, $\bar{P}$ is the one-step kernel of (s, g) under the mean-field dynamics. Likewise, define the κ-surrogate MDP on $\mathcal{X}_\kappa$ with reward r_κ and kernel $\bar{P}_\kappa$, where $\bar{P}_\kappa$ is the distributional one-step kernel induced by the $(\kappa + 1)$-agent surrogate construction that outputs (s', g') where s' is the focal agent's next state and g' is the empirical mean-field of κ sampled neighbors.

Definition G.3 (Bellman operator on $\mathcal{X}$)**.** For any bounded measurable $Q : \mathcal{X} \times \mathcal{A} \times \mathcal{Z} \to \mathbb{R}$, define the optimal Bellman operator on $\mathcal{X}$ by:

$$\mathcal{T}V(x) \coloneqq \max_{a \in \mathcal{A}} \left\{ r_\ell(x, a) + \gamma \int_{\mathcal{X}} V(x') \bar{P}(dx' \mid x, a) \right\}. \tag{21}$$

Similarly define $\mathcal{T}_\kappa$ on bounded $V : \mathcal{X}_\kappa \to \mathbb{R}$ by replacing $(r, \bar{P})$ with $(r_\kappa, \bar{P}_\kappa)$.

Lemma G.1. *$\mathcal{T}$ is a γ-contraction in the sup-norm, and therefore has a fixed point V^*.*

Proof. Consider $V_1, V_2 : \mathcal{X} \to \mathbb{R}$. Then, we have that

$$\begin{aligned}
|\mathcal{T}V_1 - \mathcal{T}V_2| &= \left| \max_{a \in \mathcal{A}} \left\{ r_\ell(x, a) + \gamma \int_{\mathcal{X}} V_1(x') \bar{P}(\mathrm{d}x' | x, a) \right\} - \max_{a \in \mathcal{A}} \left\{ r_\ell(x, a) + \gamma \int_{\mathcal{X}} V_2(x') \bar{P}(\mathrm{d}x' | x, a) \right\} \right| \\
&\leq \gamma \cdot \max_{a \in \mathcal{A}} \left| \int_{\mathcal{X}} V_1(x') P(\mathrm{d}x' | x, a) - \int_{\mathcal{X}} V_2(x') P(\mathrm{d}x' | x, a) \right| \\
&\leq \gamma \cdot \|V_1 - V_2\|_\infty,
\end{aligned}$$

where the first inequality follows by the 1-Lipschitzness of the max operator. □

We now impose a linear MDP structure on the augmented state spaces $\mathcal{X}$ and $\mathcal{X}_\kappa$).

Definition G.4 (Linear mean-field MDP on $\mathcal{X}$)**.** *The augmented MDP $(\mathcal{X}, \mathcal{A}, \bar{P}, r, \gamma)$ is linear of dimension d if there exists a measurable feature map $\varphi : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$, a vector $\theta \in \mathbb{R}^d$, and a collection of d signed measures $\mu = (\mu_1, \ldots, \mu_d)$ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ such that for all $(x, a) \in \mathcal{X} \times \mathcal{A}$ and all measurable $B \subseteq \mathcal{X}$, we have $r(x, a) = \langle \varphi(x, a), \theta \rangle$ and $\bar{P}(B \mid x, a) = \langle \varphi(x, a), \mu(B) \rangle$, where $\mu(B) \coloneqq (\mu_1(B), \ldots, \mu_d(B)) \in \mathbb{R}^d$.*

Assumption G.1 (Bounded features and parameters)**.** *Recalling Assumption 4.2, we assume WLOG that $\|\varphi(x, a)\|_2 \leq 1$ for all (x, a), $\|\theta\|_2 \leq \Theta$, and $\|\mu(\mathcal{X})\|_2 \leq \mathsf{M}$ for some finite constants Θ and M.*

Assumption G.2 (Positivity/normalization)**.** *To enforce regularity on the probability kernel φ, we assume that for all measurable B, $\langle \varphi(x, a), \mu(\mathcal{X}) \rangle = 1$ and $\langle \varphi(x, a), \mu(B) \rangle \geq 0$.*

The above assumptions imply that Q^π is linear in V^π.

Lemma G.2 (Linearity of Q^π in the augmented linear MDP)**.** *Fix $\gamma \in (0, 1)$, and let $(\mathcal{X}, \mathcal{A}, P, r)$ be the augmented discounted MDP where $\mathcal{A}$ is finite. Then for any stationary policy π, there exists $w^\pi \in \mathbb{R}^d$ such that for all $(x, a) \in \mathcal{X} \times \mathcal{A}$,*

$$Q^\pi(x, a) = \langle \varphi(x, a), w^\pi \rangle. \tag{22}$$

Moreover, we define the $d \times d$ matrix A_π where $(A_\pi)_{ij} \coloneqq \int_{\mathcal{X}} \mu_i(\mathrm{d}x) \bar{\varphi}_{\pi,j}(x)$. If $\|\theta\|_2 \leq \Theta$, $\|A_\pi\|_2 \leq 1$ and $\|\varphi(x, a)\|_2 \leq 1$ for all (x, a), then $\|w^\pi\|_2 \leq \frac{\Theta}{1-\gamma}$.

Proof. We fix a stationary policy π and define the policy evaluation Bellman operator on bounded functions $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$:

$$\mathcal{T}^\pi Q(x, a) \coloneqq r(x, a) + \gamma \int_{\mathcal{X}} \sum_{a' \in \mathcal{A}} \pi(a' \mid x') Q(x', a') P(dx' \mid x, a), \tag{23}$$

where $\mathcal{T}$ is a contractive operator with module γ.

We first show that the linear class is invariant under $\mathcal{T}^\pi$. For any $w \in \mathbb{R}^d$, define the linear function $Q_w(x,a) \coloneqq \langle \varphi(x,a), w\rangle$ and let $\bar{\varphi}_\pi(x) \coloneqq \sum_{a\in\mathcal{A}} \pi(a \mid x)\varphi(x,a) \in \mathbb{R}^d$ denote the policy-averaged feature map. We now compute $\mathcal{T}^\pi Q_w$. Using the linearity of the reward, we have $r(x,a) = \langle \varphi(x,a), \theta\rangle$. For the transition term, define the bounded measurable function

$$V_w(x') \coloneqq \sum_{a'\in\mathcal{A}} \pi(a' \mid x') Q_w(x',a') = \bar{\varphi}_\pi(x')^\top w. \tag{24}$$

Then, for any bounded measurable $f : \mathcal{X} \to \mathbb{R}$ we have

$$\int_{\mathcal{X}} f(x')P(dx' \mid x,a) = \int_{\mathcal{X}} f(x')\langle \varphi(x,a), \mu(dx')\rangle = \left\langle \varphi(x,a), \int_{\mathcal{X}} f(x')\mu(dx')\right\rangle.$$

Apply this with $f = V_w$, we obtain

$$\begin{aligned}
\int_{\mathcal{X}} V_w(x')P(dx' \mid x,a) &= \left\langle \varphi(x,a), \int_{\mathcal{X}} V_w(x')\mu(dx')\right\rangle \\
&= \left\langle \varphi(x,a), \int_{\mathcal{X}} \left(\bar{\varphi}_\pi(x')^\top w\right)\mu(dx')\right\rangle \\
&= \left\langle \varphi(x,a), \left(\int_{\mathcal{X}} \mu(dx')\bar{\varphi}_\pi(x')^\top\right) w\right\rangle \\
&= \langle \varphi(x,a), A_\pi w\rangle.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\mathcal{T}^\pi Q_w(x,a) &= \langle \varphi(x,a), \theta\rangle + \gamma\langle \varphi(x,a), A_\pi w\rangle \\
&= \langle \varphi(x,a), \theta + \gamma A_\pi w\rangle \\
&= Q_{\theta+\gamma A_\pi w}(x,a).
\end{aligned}$$

Thus the linear class $\{Q_w : w \in \mathbb{R}^d\}$ is invariant under $\mathcal{T}^\pi$. Next, to prove the linear representation, initialize $Q^{(0)} \equiv 0 = Q_{w^{(0)}}$ with $w^{(0)} \coloneqq 0$ and define the iterates $Q^{(t+1)} \coloneqq \mathcal{T}^\pi Q^{(t)}$ for $t \geq 0$. Then, inductively, there exists $w^{(t)} \in \mathbb{R}^d$ such that $Q^{(t)} = Q_{w^{(t)}}$ and $w^{(t+1)} = \theta + \gamma A_\pi w^{(t)}$. Since $\mathcal{T}^\pi$ is a γ-contraction, we have that $Q^{(t)} \to Q^\pi$ in the infinity norm. Moreover, because the image $\{Q_w\}$ is a finite-dimensional linear subspace of bounded functions, it is closed in $\|\cdot\|_\infty$ and hence the limit Q^π must also belong to this subspace, i.e., there exists at least one w^π such that $Q^\pi = Q_{w^\pi}$, thereby proving the linear representation. Finally, assuming $\|\theta\|_2 \leq \Theta$ and $\|A_\pi\|_2 \leq 1$, we see that $w^{(t)} = \sum_{h=0}^{t-1} \gamma^h A_\pi^h \theta$. Hence, using submultiplicativity and $\|A_\pi\|_2 \leq 1$, we have

$$\begin{aligned}
\|w^{(t)}\|_2 &\leq \sum_{h=0}^{t-1} \gamma^h \|A_\pi\|_2^h \|\theta\|_2 \\
&\leq \|\theta\|_2 \sum_{h=0}^{t-1} \gamma^h \\
&\leq \frac{\|\theta\|_2}{1-\gamma} \leq \frac{\Theta}{1-\gamma},
\end{aligned}$$

proving the lemma. □

We now address the function-space policy evaluation and define the projected Bellman equation.

Definition G.5 (Projected Bellman equation)**.** Fix a policy π, and let ν be any reference distribution on $\mathcal{X} \in L_2(\nu)$, where $L_2(\nu)$ is the Hilbert space equipped with inner product

$$\langle f, h\rangle_{L_2(\nu)} = \int f(x)h(x)\nu(dx). \tag{25}$$

Then, letting $\bar{\varphi}_\pi(x) \coloneqq \mathbb{E}_{a\sim\pi(\cdot|x)}[\varphi(x,a)]$, we define the feature operator $\Phi : L_2(\nu) \leftarrow \mathbb{R}^d$ by $(\Phi w)(x) \coloneqq \langle \bar{\varphi}_\pi(x), w\rangle$. Let Π_ν denote the orthogonal projection in $L_2(\nu)$ onto range(Φ). Then, the projected Bellman equation is given by

$$\Phi w \coloneqq \Pi_\nu \mathcal{T}^\pi(\Phi w), \tag{26}$$

where

$$\mathcal{T}^\pi V(x) \coloneqq \mathbb{E}_{a\sim\pi(\cdot|x)}\left[r(x,a) + \gamma \int V(x')\bar{P}(dx' \mid x,a)\right]. \tag{27}$$

Remark G.3. *Taking first-order optimality conditions yields the normal equations*

$$\mathbb{E}_{x\sim\nu}\Big[\bar{\varphi}_\pi(x)\bar{\varphi}_\pi(x)^\top\Big]w = \mathbb{E}_{x\sim\nu}\Big[\bar{\varphi}_\pi(x)\big(r_\pi(x)+\gamma\mathbb{E}[V_w(x')\mid x]\big)\Big],$$

with $r_\pi(x) \coloneqq \mathbb{E}_{a\sim\pi}[r(x,a)]$ *and* $V_w(x) \coloneqq \langle\bar{\varphi}_\pi(x), w\rangle$.

We now show how to handle $\mathcal{M}_\kappa$ being uncountable by using covering numbers instead of $|\mathcal{G}_\kappa|$. Specifically, in the tabular setting, we could union-bound over $(s,g)\in\mathcal{S}\times\mathcal{G}_\kappa$ because $\mathcal{G}_\kappa$ was finite. However, in this setting where $\mathcal{S}$ is continuous, $\mathcal{M}_\kappa$ is uncountable. Therefore, we instead use a covering-number (metric-entropy) bound. For any subset $\mathcal{V}$ of a metric space (V, d), let $N(\varepsilon,\mathcal{V},d)$ denote the minimum size of an ε-net. We observe that, under a linear MDP, all concentration arguments needed for policy evaluation/control can be reduced to controlling random quantities indexed by the feature vectors $\varphi(x,a)$, not by (x,a) directly.

Remark G.4. Consider the feature image

$$\mathcal{V}_\kappa \coloneqq \{\varphi(x,a) : x\in\mathcal{X}_\kappa, a\in\mathcal{A}\} \subseteq \mathbb{R}^d. \tag{28}$$

By Assumption G.1, we have that $\mathcal{V}_\kappa$ is contained in the unit Euclidean ball $\mathbb{B}_2^d$. Specifically, for every $\varepsilon\in(0,1]$, we have

$$N(\varepsilon,\mathcal{V}_\kappa,\|\cdot\|_2) \le N(\varepsilon,\mathbb{B}_2^d,\|\cdot\|_2) \le \left(\frac{3}{\varepsilon}\right)^d. \tag{29}$$

Therefore, we can now replace any argument that previously union-bounded over $|\mathcal{S}||\mathcal{G}_\kappa|$ by a union bound over an ε-net of $\mathcal{V}_\kappa$ with complexity $d\log(3/\varepsilon)$.

We are now ready to state our result which decomposes the subsampling error and estimation error which arises from learning/using features in a non-tabular space. First, in the linear-MDP setting, it suffices to control the induced error in *feature expectations*. We assume that the only way the mean-field enters the dynamics is through the d-dimensional feature vector.

Assumption G.3. *There exists a bounded map* $\psi:\mathcal{S}\times\mathcal{A}\times\mathcal{G}\to\mathbb{R}^d$ *with* $\|\psi(s,a,\cdot)\|_2\le 1$ *such that for all* $x=(s,g)\in\mathcal{X}$ *and* $a\in\mathcal{A}$,

$$\varphi(x,a) = \int_{\mathcal{S}} \psi(s,a,u)g(du). \tag{30}$$

As a result, if $\widehat{g} = \frac{1}{\kappa}\sum_{m=1}^{\kappa}\delta_{U_m}$ *with* $U_m \overset{i.i.d.}{\sim} g$, *then* $\varphi((s,\widehat{g}),a) = \frac{1}{\kappa}\sum_{m=1}^{\kappa}\psi(s,a,U_m)$.

Remark G.5. *Under Assumption G.3, the vector Hoeffding bound yields* $\|\varphi((s,\widehat{g}),a)-\varphi((s,g),a)\|_2 = O\big(\sqrt{d/\kappa}\big)$.

Next, let $\widehat{\varphi}$ be a learned feature map (e.g. from a spectral factorization procedure). We quantify its error by

$$\varepsilon_{\text{rep}} \coloneqq \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \|\widehat{\varphi}(x,a)-\varphi(x,a)\|_2. \tag{31}$$

Let $\widehat{Q}_\kappa$ be the value function returned by any consistent linear evaluation/planning procedure using M samples from a generative model on the κ-surrogate MDP, producing a greedy policy $\widehat{\pi}_\kappa$.

Theorem G.6 (Continuous-state extension with covering numbers). *Let* $\widehat{\pi}_\kappa$ *be a greedy policy computed on the* κ*-surrogate using learned features* $\widehat{\varphi}$ *with representation error* $\varepsilon_{\text{rep}} \coloneqq \sup_{x,a}\|\widehat{\varphi}_\kappa(x,a)-\varphi_\kappa(x,a)\|_2$ *and* M *samples. Then for any* $\delta\in(0,1)$, *we have that with probability at least* $1-3\delta$,

$$\sup_{x\in\mathcal{X}}\Big[V^*(x)-V^{\widehat{\pi}_\kappa}(x)\Big] \le \widetilde{O}\left(\frac{1}{(1-\gamma)^3}\sqrt{\frac{d+|\mathcal{A}|+\log(1/\delta)}{\kappa}} + \frac{1}{(1-\gamma)^2}\sqrt{\frac{d\log\left(\frac{N(\varepsilon)}{\delta}\right)}{M}} + \frac{\varepsilon}{(1-\gamma)^2} + \frac{\varepsilon_{\text{rep}}}{(1-\gamma)^2}\right),$$

where $\mathcal{V}_\kappa=\{\varphi(x,a): x\in\mathcal{X}_\kappa, a\in\mathcal{A}\}$ *and* $N(\cdot)$ *is the covering number. Moreover, since* $\mathcal{V}_\kappa\subseteq\mathbb{B}_2^d$, *we can upper bound the entropy term via* (29) *as* $\log N(\varepsilon,\mathcal{V}_\kappa,\|\cdot\|_2) \le d\log(3/\varepsilon)$, *so the statistical term is at most* $\widetilde{O}(d/\sqrt{M})$.

Proof. Let $\mathcal{T}$ and $\widetilde{\mathcal{T}}$ be γ-contractions on $(\mathcal{B}(\mathcal{X}\times\mathcal{A}),\|\cdot\|_\infty)$ with unique fixed points Q^* and $\widetilde{Q}^*$. Then,

$$\begin{aligned}\|Q^*-\widetilde{Q}^*\|_\infty &= \|\mathcal{T}Q^*-\widetilde{\mathcal{T}}\widetilde{Q}^*\|_\infty\\ &\le \|\mathcal{T}Q^*-\widetilde{\mathcal{T}}Q^*\|_\infty + \|\widetilde{\mathcal{T}}Q^*-\widetilde{\mathcal{T}}\widetilde{Q}^*\|_\infty\\ &\le \|(\mathcal{T}-\widetilde{\mathcal{T}})Q^*\|_\infty + \gamma\|Q^*-\widetilde{Q}^*\|_\infty.\end{aligned}$$

Let $\widehat{Q}$ be any bounded function and let $\widehat{\pi}$ be greedy with respect to $\widehat{Q}$, i.e. $\widehat{\pi}(\cdot\mid x)\in\arg\max_{a\in\mathcal{A}}\widehat{Q}(x,a)$. Then, for the *true* MDP,

$$\sup_{x\in\mathcal{X}}\big(V^*(x)-V^{\widehat{\pi}}(x)\big) \le \frac{2}{1-\gamma}\|\widehat{Q}-Q^*\|_\infty. \tag{32}$$

We will apply (32) with $\widehat{Q} = \widehat{Q}^*_\kappa$. Then, by the triangle inequality, $\|\widehat{Q}^*_\kappa - Q^*\|_\infty \le \|Q^*_\kappa - Q^*\|_\infty + \|\widehat{Q}^*_\kappa - Q^*_\kappa\|_\infty$. We bound each term separately. We first bound the subsampling bias $\|Q^*_\kappa - Q^*\|_\infty$ by using $\|Q^*_\kappa - Q^*\|_\infty \le \frac{1}{1-\gamma}\|(\mathcal{T}_\kappa - \mathcal{T})Q^*\|_\infty$. Fix (x, a) and let $f^*(x') \coloneqq \max_{a'} Q^*(x', a')$. Then,

$$\begin{aligned}(\mathcal{T}_\kappa - \mathcal{T})Q^*(x,a) &= \langle \varphi_\kappa(x,a) - \varphi(x,a), \theta\rangle + \gamma\left\langle \varphi_\kappa(x,a) - \varphi(x,a), \int_{\mathcal{X}} f^*(x')\mu(\mathrm{d}x')\right\rangle \\ &= \left\langle \varphi_\kappa(x,a) - \varphi(x,a),\ \theta + \gamma\int_{\mathcal{X}} f^*(x')\mu(\mathrm{d}x')\right\rangle \\ &\le \|\varphi_\kappa(x,a) - \varphi(x,a)\|_2 \cdot \left\|\theta + \gamma\int f^*\mathrm{d}\mu\right\|_2,\end{aligned}$$

where the last inequality uses Cauchy-Schwarz. We then bound the second factor using our boundedness assumption while using $\|f^*\|_\infty \le \|Q^*\|_\infty \le \frac{1}{1-\gamma}$. Specifically, for each coordinate, we have

$$\left|\int f^*\mathrm{d}\mu_i\right| \le \|\mu_i\|_{\mathrm{TV}}\|f^*\|_\infty \implies \left\|\int f^* d\mu\right\|_2 \le \|\mu\|_{\mathrm{TV},2}\|f^*\|_\infty \le \frac{1}{1-\gamma}, \tag{33}$$

Therefore, we have

$$\begin{aligned}\left\|\theta + \gamma\int f^* d\mu\right\|_2 &\le \|\theta\|_2 + \gamma\left\|\int f^* d\mu\right\|_2 \\ &\le 1 + \frac{\gamma}{1-\gamma} = \frac{1}{1-\gamma}.\end{aligned}$$

Note that by Lemma G.8, there exists a construction of φ_κ from κ i.i.d. samples such that for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\Delta_\kappa \coloneqq \sup_{(x,a)\in\mathcal{X}\times\mathcal{A}} \|\varphi_\kappa(x,a) - \varphi(x,a)\|_2 \ \le\ C_1\sqrt{\frac{d + |\mathcal{A}| + \log(1/\delta)}{\kappa}}, \tag{34}$$

for a universal constant C_1 . So, taking the supremum over (x, a) yields $\|(\mathcal{T}_\kappa - \mathcal{T})Q^*\|_\infty \le \frac{\Delta_\kappa}{1-\gamma}$. Hence, we have that

$$\|Q^*_\kappa - Q^*\|_\infty \le \frac{\Delta_\kappa}{(1-\gamma)^2} \le \frac{C_1}{(1-\gamma)^2}\sqrt{\frac{d + |\mathcal{A}| + \log(1/\delta)}{\kappa}}.$$

Next, note that by Lemma G.7, we can compute an approximate optimal Q-function for M_κ in the linear class using M samples such that with probability at least $1-\delta$,

$$\|\widehat{Q}^*_{\kappa,M} - Q^*_\kappa\|_\infty \le \frac{C_2}{1-\gamma}\cdot\sqrt{\frac{d\log\left(\frac{N(\varepsilon)}{\delta}\right)}{M}} + \frac{C_2}{1-\gamma}\cdot\varepsilon. \tag{35}$$

Finally, the representation error contributes at most $\frac{C_3}{1-\gamma}\varepsilon_{\mathrm{rep}}$ (additively) to $\|\widehat{Q}^*_\kappa - Q^*_\kappa\|_\infty$. Therefore, together, we have that with probability at least $1-3\delta$,

$$\|\widehat{Q}^*_\kappa - Q^*\|_\infty \le \frac{C_1}{(1-\gamma)^2}\sqrt{\frac{d + |\mathcal{A}| + \log(1/\delta)}{\kappa}} + \frac{C_2}{1-\gamma}\cdot\sqrt{\frac{d\log\left(\frac{N(\varepsilon)}{\delta}\right)}{M}} + \frac{C_2}{1-\gamma}\cdot\varepsilon + \frac{C_3}{1-\gamma}\varepsilon_{\mathrm{rep}}.$$

Then, using $\sup_x(V^*(x) - V^{\widehat{\pi}_\kappa}(x)) \le \frac{2}{1-\gamma}\|\widehat{Q}^*_\kappa - Q^*\|_\infty$ proves the theorem. □

Finally, we show how to produce a uniform concentration over $\mathcal{V}_\kappa$ via covering numbers.

Lemma G.7. *Let $\mathcal{V} \subseteq \mathbb{R}^d$ be any set with $\sup_{v\in\mathcal{V}}\|v\|_2 \le 1$. Let $Z_1, \dots, Z_M$ be i.i.d. random vectors in $[-1,1]^d$ with mean $\mathbb{E}[Z_1] = 0$. Define $\overline{Z} \coloneqq \frac{1}{M}\sum_{m=1}^M Z_m$. Fix $\varepsilon \in (0,1]$ and let $N(\varepsilon) \coloneqq N(\varepsilon, \mathcal{V}, \|\cdot\|_2)$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have that*

$$\sup_{v\in\mathcal{V}} |\langle v, \overline{Z}\rangle| \ \le\ \sqrt{\frac{2d\log\left(\frac{2N(\varepsilon)}{\delta}\right)}{M}} + 2\varepsilon\sqrt{d}.$$

Proof. Let $\mathcal{N}$ be an ε-net of $\mathcal{V}$ in $\|\cdot\|_2$ of size $|\mathcal{N}| = N(\varepsilon)$. For any fixed $u \in \mathcal{N}$, define the scalar random variables $Y_m(u) \coloneqq \langle u, Z_m\rangle$. Since $Z_m \in [-1,1]^d$ and $\|u\|_2 \le 1$, we have

$$\|u\|_1 \le \sqrt{d}\|u\|_2 \le \sqrt{d} \implies |Y_m(u)| \le \|u\|_1 \le \sqrt{d}$$

almost surely. By Hoeffding's inequality, we have

$$\Pr\left[|\langle u, \overline{Z}\rangle| \ge t\right] = \Pr\left(\left|\frac{1}{M}\sum_{m=1}^M Y_m(u)\right| \ge t\right)$$

$$\leq 2\exp\Big(-\frac{2Mt^2}{(2\sqrt{d})^2}\Big) = 2\exp\Big(-\frac{Mt^2}{2d}\Big).$$

Union-bounding over all $u \in \mathcal{N}$ gives $\Pr\big[\sup_{u\in\mathcal{N}} |\langle u, \overline{Z}\rangle| \geq t\big] \leq 2N(\varepsilon)\exp(-\frac{Mt^2}{2d})$. Then, setting the right-hand side to δ yields that with probability at least $1-\delta$,

$$\sup_{u\in\mathcal{N}} |\langle u, \overline{Z}\rangle| \leq \sqrt{\frac{2d\log(\frac{2N(\varepsilon)}{\delta})}{M}}.$$

We now extend from the net to all $v \in \mathcal{V}$. For any $v \in \mathcal{V}$, choose $u \in \mathcal{N}$ with $\|v-u\|_2 \leq \varepsilon$. Then

$$\begin{aligned}|\langle v, \overline{Z}\rangle| &\leq |\langle u, \overline{Z}\rangle| + |\langle v-u, \overline{Z}\rangle| \\ &\leq \sup_{u\in\mathcal{N}} |\langle u, \overline{Z}\rangle| + \|v-u\|_2\|\overline{Z}\|_2.\end{aligned}$$

Since each coordinate of $\overline{Z}$ lies in $[-1,1]$, we have that $\|\overline{Z}\|_2 \leq \sqrt{d}$; hence, $|\langle v-u, \overline{Z}\rangle| \leq \varepsilon\sqrt{d}$. Applying the same argument to $-\overline{Z}$ yields the factor $2\varepsilon\sqrt{d}$ in the symmetric bound, which completes the proof. □

Lemma G.8 (Construction of φ_κ and a $\kappa^{-1/2}$ perturbation bound)**.** *Let $p \coloneqq d + |\mathcal{A}|$, and consider the measurable mean-field embedding map $\psi : \Xi \to \mathbb{R}^p$ such that $\|\psi(\xi)\|_2 \leq 1$ for all $\xi \in \Xi$. Then, consider the population mean-field feature as $u(m) \coloneqq \mathbb{E}_{\xi\sim m}\big[\psi(\xi)\big] \in \mathbb{R}^p$, such that $\varphi((s,m),a) = \Phi(s,a,u(m))$. Next, given κ i.i.d. samples $\xi_1,\ldots,\xi_\kappa \sim m$, define the empirical embedding $\widehat{u}_\kappa(m) \coloneqq \frac{1}{\kappa}\sum_{j=1}^{\kappa}\psi(\xi_j)$ and set $\varphi_\kappa((s,m),a) \coloneqq \Phi(s,a,\widehat{u}_\kappa(m))$. Then for any (x,a) and $\delta \in (0,1)$, with probability at least $1-\delta$, we have that for a universal constant C_1,*

$$\|\varphi_\kappa(x,a) - \varphi(x,a)\|_2 \leq C_1\sqrt{\frac{p+\log(1/\delta)}{\kappa}} = C_1\sqrt{\frac{d+|\mathcal{A}|+\log(1/\delta)}{\kappa}}. \tag{36}$$

Proof. Fix any $(x,a) = ((s,m),a)$. Then, from Lemma C.5, we have

$$\|\varphi_\kappa(x,a) - \varphi(x,a)\|_2 = \|\Phi(s,a,\widehat{u}_\kappa(m)) - \Phi(s,a,u(m))\|_2 \leq \|\widehat{u}_\kappa(m) - u(m)\|_2.$$

Let $\Delta \coloneqq \widehat{u}_\kappa(m) - u(m) = \frac{1}{\kappa}\sum_{j=1}^{\kappa}\big(\psi(\xi_j) - \mathbb{E}[\psi(\xi_j)]\big)$. For any fixed unit vector $v \in \mathbb{S}^{p-1}$, the scalar random variables $Y_j(v) \coloneqq \langle v, \psi(\xi_j)\rangle$ satisfy $|Y_j(v)| \leq \|v\|_2\|\psi(\xi_j)\|_2 \leq 1$ almost surely. Hence, by Hoeffding's inequality,

$$\Pr\big[\big|\langle v, \Delta\rangle\big| \geq t\big] = \Pr\left[\left|\frac{1}{\kappa}\sum_{j=1}^{\kappa}(Y_j(v) - \mathbb{E}[Y_j(v)])\right| \geq t\right] \leq 2\exp\left(-\frac{\kappa t^2}{2}\right).$$

Let $\mathcal{N}$ be a $(1/2)$-net of $\mathbb{S}^{p-1}$ in $\|\cdot\|_2$, such that $|\mathcal{N}| \leq 5^p$ (62). Union-bounding, we have

$$\Pr\left[\sup_{v\in\mathcal{N}} |\langle v, \Delta\rangle| \geq t\right] \leq 2|\mathcal{N}|\exp\left(-\frac{\kappa t^2}{2}\right) \leq 2\cdot 5^p\exp\left(-\frac{\kappa t^2}{2}\right).$$

Reparameterizing the RHS to δ and solving for t, we get that $t = \sqrt{\frac{2(p\log 5 + \log(2/\delta))}{\kappa}}$. Now, we extend from the net to the whole sphere. So, for any $v \in \mathbb{S}^{p-1}$ pick $\widetilde{v} \in \mathcal{N}$ with $\|v-\widetilde{v}\|_2 \leq 1/2$. Then, we have

$$\begin{aligned}|\langle v, \Delta\rangle| &\leq |\langle \widetilde{v}, \Delta\rangle| + |\langle v-\widetilde{v}, \Delta\rangle| \\ &\leq \sup_{u\in\mathcal{N}} |\langle u, \Delta\rangle| + \|v-\widetilde{v}\|_2\|\Delta\|_2 \\ &\leq \sup_{u\in\mathcal{N}} |\langle u, \Delta\rangle| + \tfrac{1}{2}\|\Delta\|_2.\end{aligned}$$

Taking supremum over v gives $\|\Delta\|_2 \leq 2\sup_{u\in\mathcal{N}} |\langle u, \Delta\rangle|$. Therefore, with probability at least $1-\delta$,

$$\|\Delta\|_2 \leq 2t = 2\sqrt{\frac{2(\,p\log 5 + \log(2/\delta)\,)}{\kappa}} \leq C_1\sqrt{\frac{p+\log(1/\delta)}{\kappa}}$$

proving the claim. □

Consequently, in the continuous state-action setting, as $\kappa \to n$ and $M \to \infty$, the value function of the estimated policy converges to the optimal value function, i.e., $V^{\pi_{\kappa,m}}(s,g) \to V^{\pi^*}(s,g)$. Intuitively, as the subsampling parameter κ approaches the population size n, the optimality gap diminishes following the concentration arguments in Theorem 4.3. As the number of samples M increases, the linear function approximation error vanishes, which allows for the exact recovery of the spectral representations.