

---

# Mean-Field Sampling for Cooperative Multi-Agent Reinforcement Learning

---

**Emile Anand\***

Georgia Institute of Technology  
Atlanta, GA 30308  
emile@gatech.edu

**Ishanik Karmarkar**

Stanford University  
Palo Alto, CA, 94305  
ishanik@stanford.edu

**Guannan Qu**

Carnegie Mellon University  
Pittsburgh, PA 94035  
gqu@andrew.cmu.edu

## Abstract

Designing efficient algorithms for multi-agent reinforcement learning (MARL) is fundamentally challenging because the size of the joint state and action spaces grows exponentially in the number of agents. These difficulties are exacerbated when balancing sequential global decision-making with local agent interactions. In this work, we propose a new algorithm SUBSAMPLE-MFQ (**Subsample-Mean-Field-Q-learning**) and a decentralized randomized policy for a system with  $n$  agents. For any  $k \leq n$ , our algorithm learns a policy for the system in time polynomial in  $k$ . We prove that this learned policy converges to the optimal policy on the order of  $\tilde{O}(1/\sqrt{k})$  as the number of subsampled agents  $k$  increases. In particular, this bound is independent of the number of agents  $n$ .

## 1 Introduction

Reinforcement Learning (RL) has become a popular framework to solve sequential decision making problems in unknown environments and has achieved tremendous success in a wide array of domains such as playing the game of Go [Silver et al., 2016], robotic control [Kober et al., 2013], and autonomous driving [Kiran et al., 2022, Lin et al., 2023a]. A key feature of most real-world systems is their uncertain nature, and thus, RL has emerged as a powerful tool for learning optimal policies for multi-agent systems to operate in unknown environments [Kim and Giannakis, 2017, Zhang et al., 2021, Lin et al., 2024, Anand and Qu, 2024]. While early RL works focused on the single-agent setting, multi-agent RL (MARL) has recently achieved impressive success in many applications, such as coordinating robotic swarms [Preiss et al., 2017, DeWeese and Qu, 2024], real-time bidding [Jin et al., 2018], ride-sharing [Li et al., 2019], and stochastic games [Jin et al., 2020].

Despite growing interest in MARL, extending RL to multi-agent settings poses significant computational challenges due to the *curse of dimensionality*. MARL is fundamentally difficult as agents in the real-world not only interact with the environment but also with each other [Shapley, 1953]: if each of the  $n$  agents has a state space  $S$  and action space  $A$ , the global state-action space has size  $(|S||A|)^n$ , which is exponential in  $n$ . Thus, many RL algorithms (such as temporal difference learning and tabular  $Q$ -learning) require computing and storing an  $(|S||A|)^n$ -sized  $Q$ -table [Sutton et al., 1999, Bertsekas and Tsitsiklis, 1996]. This scalability issue has been observed in a variety of MARL settings [Blondel and Tsitsiklis, 2000, Papadimitriou and Tsitsiklis, 1999, Littman, 1994].

---

\*Work partially done while a visiting student at Carnegie Mellon University and intern at Cognition AI.

An exciting line of work that addresses this intractability is mean-field MARL [Lasry and Lions, 2007, Yang et al., 2018, Gu et al., 2021, 2022, Hu et al., 2023]. The mean-field approach assumes agents are homogeneous in their state-action spaces, enabling their interactions to be approximated by a two-agent setting: here, each agent interacts with a representative “mean agent” which evolves as the empirical distribution of states of all other agents. With these assumptions, mean-field MARL learns an optimal policy with sample complexity  $O(n^{|S||A|}|S||A|)$ , which is polynomial in the number of agents. However, if  $n$  is large, this remains prohibitive even for moderate values of  $|S|$  and  $|A|$ .

Motivated by this problem, in this paper we study the cooperative setting—where agents work collaboratively to maximize a structured global reward—and ask: *Can we design a scalable MARL algorithm for learning an approximately optimal policy in a cooperative multi-agent system?*

**Contributions.** We answer this question affirmatively. Our key contributions are outlined below.

**Subsampling algorithm.** We model the problem as a Markov Decision Process (MDP) with a global agent and  $n$  local agents. We propose SUBSAMPLE-MFQ to address the challenge of MARL with a large number of local agents. SUBSAMPLE-MFQ selects  $k \leq n$  local agents to learn a deterministic policy  $\hat{\pi}_k^{\text{est}}$  by applying mean-field value iteration on the  $k$ -local-agent subsystem to learn  $\hat{Q}_k^{\text{est}}$ , which can be viewed as a smaller  $Q$  function. SUBSAMPLE-MFQ then deploys a stochastic policy  $\pi_k^{\text{est}}$ , where the global agent samples  $k$  local agents uniformly at each step and uses  $\hat{\pi}_k^{\text{est}}$  to determine its action, while each local agent samples  $k - 1$  other local agents and uses  $\hat{\pi}_k^{\text{est}}$  to determine its action.

**Sample complexity and theoretical guarantee.** As the number of local agents increases, the size of  $\hat{Q}_k$  scales polynomially with  $k$ , rather than polynomially with  $n$  as in mean-field MARL. Analogously, when the size of the local agent’s state space grows, the size of  $\hat{Q}_k$  scales exponentially with  $k$ , rather than exponentially with  $n$ , as in traditional  $Q$ -learning. The key analytic technique underlying our results is a novel MDP sampling result. This result shows that the performance gap between  $\pi_k^{\text{est}}$  and the optimal policy  $\pi^*$  is at most  $\tilde{O}(1/\sqrt{k})$ , with high probability. The choice of  $k$  reveals a fundamental trade-off between the size of the  $Q$ -table and the optimality of  $\pi_k^{\text{est}}$ . For example, if  $k$  is set to  $O(\log n)$ , SUBSAMPLE-MFQ is the first centralized MARL algorithm to achieve a *polylogarithmic* run-time in  $n$ , representing an exponential speedup over the previously best-known polytime mean-field MARL methods, while maintaining a decaying optimality gap as  $n$  gets large,

While our results are theoretical in nature, we hope SUBSAMPLE-MFQ will further exploration of sampling in Markov games, and potentially inspire new practical multi-agent algorithms.

**Related work.** MARL has a rich history, starting with early works on Markov games [Littman, 1994, Sutton et al., 1999], which are a multi-agent extension of MDPs. MARL has since been actively studied [Zhang et al., 2021] in a broad range of settings. MARL is most similar to the category of “succinctly described” MDPs [Blondel and Tsitsiklis, 2000], where the state/action space is a product space formed by the individual state/action spaces of multiple agents, and where the agents interact to maximize an objective. A recent line of research constrains the problem to sparse networked instances to enforce local interactions between agents [Qu et al., 2020a, Lin et al., 2020, Mondal et al., 2022]. In this formulation, the agents correspond to vertices on a graph who only interact with nearby agents. By exploiting Gamarnik’s correlation decay property from combinatorial optimization [Gamarnik et al., 2009], they overcome the curse of dimensionality by simplifying the problem to only search over the policy space derived from the truncated graph to learn approximately optimal solutions. However, as the underlying network structure becomes dense with many local interactions, the neighborhood of each agent gets large, and these algorithms become intractable.

**Mean-Field RL.** Under assumptions of homogeneity in the state/action spaces of the agents, the problem of densely networked multi-agent RL was studied by Yang et al. [2018], Gu et al. [2021], who approximated the solution in polynomial time with a mean-field approach where the approximation error scales in  $O(1/\sqrt{n})$ . In contrast, our work achieves *subpolynomial* runtimes by directly sampling from this mean-field distribution. Cui and Koepl [2022] introduce heterogeneity to mean-field MARL by modeling non-uniform interactions through graphons; however, these methods crucially assume the existence of graphon sequences that converge in cut-norm to the finite graph. In the cooperative setting, Subramanian et al. [2022], Cui et al. [2023] studies a mean-field setting with  $q$  types of homogeneous agents; however, their learned policy does not provably converge to the optimum.

**Other related works.** Our work is related to factored MDPs, where there is a global action affecting every agent; however, in our case, each agent has its own action [Min et al., 2023, Lauer and

Riedmiller, 2000]. Jin et al. [2020] reduces the dependence of the product action space to an additive dependence with V-learning. Our work *further* reduces the complexity of the joint state space, which has not been previously accomplished. We add to the growing literature on the Centralized Training with Decentralized Execution regime [Zhou et al., 2023], as our algorithm learns a provably approximately optimal policy using centralized information, but makes decisions using only local information during execution. Finally, one can efficiently approximate the  $Q$ -table through function approximation [Jin et al., 2021]. However, achieving theoretical bounds on the performance loss due to function approximation is intractable without strong assumptions such as linear Bellman completeness or low Bellman-Eluder dimension [Golowich and Moitra, 2024]. While our work primarily studies the finite tabular setting, we extend it to non-tabular linear MDPs in Section J.

## 2 Preliminaries

**Notation.** For  $k, n \in \mathbb{N}$  where  $k \leq n$ , let  $\binom{[n]}{k}$  denote the set of  $k$ -sized subsets of  $[n] = \{1, \dots, n\}$ . For any vector  $z \in \mathbb{R}^d$ , let  $\|z\|_1$  and  $\|z\|_\infty$  denote the standard  $\ell_1$  and  $\ell_\infty$  norms of  $z$  respectively. Let  $\|\mathbf{A}\|_1$  denote the matrix  $\ell_1$ -norm of  $\mathbf{A} \in \mathbb{R}^{n \times m}$ . Given variables  $s_1, \dots, s_n, s_\Delta := \{s_i : i \in \Delta\}$  for  $\Delta \subseteq [n]$ . We use  $\tilde{O}(\cdot)$  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  $\rho_1, \rho_2$  is given by  $\text{TV}(\rho_1, \rho_2) = \frac{1}{2} \sum_{x \in \mathcal{X}} |\rho_1(x) - \rho_2(x)|$ . Next,  $x \sim \mathcal{D}(\cdot)$  denotes that  $x$  is a random element sampled from a distribution  $\mathcal{D}$ , and we denote that  $x$  is a random sample from the uniform distribution over a finite set  $\Omega$  by  $x \sim \mathcal{U}(\Omega)$ . We include a detailed notation table in Table 1.

### 2.1 Problem formulation

We consider a system of  $n + 1$  agents, where agent  $g$  is a “global decision making agent” and the remaining  $n$  agents, denoted by  $[n]$ , are “local agents.” At time  $t$ , the agents are in state  $s(t) = (s_g(t), s_1(t), \dots, s_n(t)) \in \mathcal{S} := \mathcal{S}_g \times \mathcal{S}_l^n$ , where  $s_g(t) \in \mathcal{S}_g$  denotes the global agent’s state, and for each  $i \in [n]$ ,  $s_i(t) \in \mathcal{S}_l$  denotes the state of the  $i$ ’th local agent. The agents cooperatively select actions  $a(t) = (a_g(t), a_1(t), \dots, a_n(t)) \in \mathcal{A} := \mathcal{A}_g \times \mathcal{A}_l^n$ , where  $a_g(t) \in \mathcal{A}_g$  denotes the global agent’s action and  $a_i(t) \in \mathcal{A}_l$  denotes the  $i$ ’th local agent’s action. At time  $t + 1$ , the next state for all the agents is independently generated by stochastic transition kernels  $P_g : \mathcal{S}_g \times \mathcal{S}_g \times \mathcal{A}_g \rightarrow \Delta(\mathcal{S}_g)$  and  $P_l : \mathcal{S}_l \times \mathcal{S}_l \times \mathcal{A}_l \rightarrow \Delta(\mathcal{S}_l)$  as follows:

$$s_g(t+1) \sim P_g(\cdot | s_g(t), a_g(t)), \quad s_i(t+1) \sim P_l(\cdot | s_i(t), s_g(t), a_i(t)), \forall i \in [n]. \quad (1)$$

The system then collects a structured stage reward  $r(s(t), a(t))$  where the reward  $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$  depends on  $s(t)$  and  $a(t)$  through Equation (2), and where  $r_g$  and  $r_l$  is typically application specific.

$$r(s, a) = \underbrace{r_g(s_g, a_g)}_{\text{global component}} + \frac{1}{n} \sum_{i \in [n]} \underbrace{r_l(s_i, s_g, a_i)}_{\text{local component}} \quad (2)$$

A policy  $\pi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$  maps from states to distributions of actions such that  $a \sim \pi(\cdot | s)$ . Given  $\gamma \in (0, 1)$ , we seek to learn a policy  $\pi$  that maximizes the ( $\gamma$ -discounted) value for each  $s \in S$ :

$$V^\pi(s) = \mathbb{E}_{a(t) \sim \pi(\cdot | s)} \left[ \sum_{t=0}^{\infty} \gamma^t r(s(t), a(t)) | s(0) = s \right]. \quad (3)$$

The cardinality of the search space simplex for the optimal policy is  $|\mathcal{S}_g||\mathcal{S}_l|^n|\mathcal{A}_g||\mathcal{A}_l|^n$ , which is exponential in the number of agents, underscoring the need for efficient approximation algorithms.

To efficiently learn policies that maximize the objective, we make the following standard assumptions:

**Assumption 2.1** (Finite state/action spaces). We assume that the state and action spaces of all the agents in the MARL game are finite:  $|\mathcal{S}_l|, |\mathcal{S}_g|, |\mathcal{A}_g|, |\mathcal{A}_l| < \infty$ . Section J of the supplementary material relaxes this assumption to the non-tabular setting with infinite continuous sets.

**Assumption 2.2** (Bounded rewards). The components of the reward function are bounded. Specifically,  $\|r_g(\cdot, \cdot)\|_\infty \leq \tilde{r}_g$ , and  $\|r_l(\cdot, \cdot, \cdot)\|_\infty \leq \tilde{r}_l$ . This implies  $\|r(\cdot, \cdot)\|_\infty \leq \tilde{r}_g + \tilde{r}_l := \tilde{r}$ .

**Definition 2.3** ( $\epsilon$ -optimal policy). Given a policy simplex  $\Pi$ ,  $\pi \in \Pi$  is an  $\epsilon$ -optimal policy if  $V^\pi(s) \geq \sup_{\pi^* \in \Pi} V^{\pi^*}(s) - \epsilon$ .

**Motivating examples.** Below we give examples of two cooperative MARL settings which are naturally modeled by our setting. Our experiments in Section B reveal a monotonic improvement in the learned policies as  $k \rightarrow n$ , while providing a substantial speedup over mean-field  $Q$ -learning<sup>2</sup>.

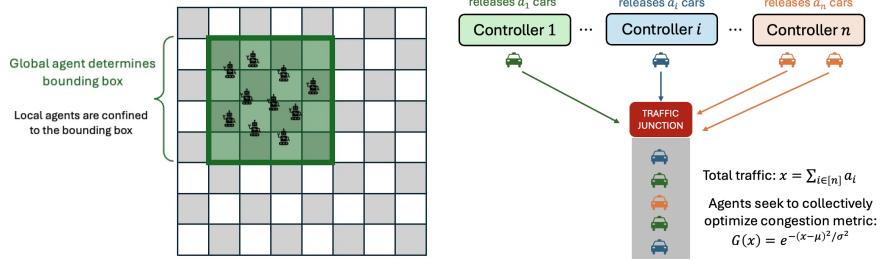


Figure 1: Bounded exploration in warehouse accidents, and traffic congestions with Gaussian squeeze.

- **Gaussian squeeze:** In this task,  $n$  homogeneous agents determine individual actions  $a_i$  to jointly maximize the objective  $r(x) = xe^{-(x-\mu)^2/\sigma^2}$ , where  $x = \sum_{i=1}^n a_i$ , and  $\mu$  and  $\sigma$  are the pre-defined mean and variance of the system. In scenarios of traffic congestion, each agent  $i \in [n]$  is a controller trying to send  $a_i$  vehicles into the main road, where controllers coordinate with each other to avoid congestion, hence avoiding either over-use or under-use, thereby contributing to the entire system. This GS problem is previously studied in Yang et al. [2018], and serves as an ablation study on the impact of subsampling for MARL.
- **Constrained exploration:** Consider an  $M \times M$  grid. Each agent’s state is a coordinate in  $[M] \times [M]$ . The state represents the *center of a  $d \times d$  box* where the global agent constrains the local agents’ movements. Initially, all agents are in the same location. At each time-step, the local agents take actions  $a_i(t) \in \mathbb{R}^2$  (e.g., up, down, left, right) to transition between states and collect rewards. The transition kernel ensures that local agents remain within the  $d \times d$  box dictated by the global agent, using knowledge of  $a_i(t)$ ,  $s_g(t)$ , and  $s_i(t)$ . In warehouse settings where shelves have collapsed, creating hazardous or inaccessible areas, we want agents to clean these areas. However, exploration in these regions may be challenging due to physical constraints or safety concerns, causing exploration in these regions to be *disincentivized* from the local agents’ perspectives. Through an appropriate design of the reward and transition functions, the global agent could guide the local agents to focus on specific  $d \times d$  grids, allowing efficient cleanup while minimizing risk.

**Capturing heterogeneity.** Following Mondal et al. [2022], Xu and Klabjan [2023], our model can capture heterogeneity in the local agents by modeling agent types as part of the state: to do this, we assign a type  $\varepsilon_i \in \mathcal{E}$  to each local agent by letting  $\mathcal{S}_l = \mathcal{E} \times \mathcal{S}'_l$ , where  $\mathcal{E}$  is a set that enumerates possible types that are treated as a fixed part of the agent’s state, and  $\mathcal{S}'_l$  is the latent state space of any local agent. The transition and reward functions can vary depending on the agent’s type. The global agent can provide unique signals to local agents of each type by letting  $s_g \in \mathcal{S}_g$  and  $a_g \in \mathcal{A}_g$  denote a state/action vector where each element corresponds to a type.

### 3 Algorithmic Approach: Subsampled Value Iteration

**Q-learning.** Our starting point is the classic Q-learning framework [Watkins and Dayan, 1992] for offline-RL, which seeks to learn the  $Q$ -function  $Q : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ . For any policy  $\pi$ ,  $Q^\pi(s, a) = \mathbb{E}^\pi[\sum_{t=0}^{\infty} \gamma^t r(s(t), a(t)) | s(0)=s, a(0)=a]$ . Initially,  $Q^0(s, a)=0$ , for all  $(s, a) \in \mathcal{S} \times \mathcal{A}$ . Then, for all  $t \in [T]$ , it is updated as  $Q^{t+1}(s, a) \leftarrow \mathcal{T}Q^t(s, a)$ , where the *Bellman operator*  $\mathcal{T}$  is

$$\mathcal{T}Q^t(s, a) = r(s, a) + \gamma \mathbb{E}_{s'_g \sim P_g(\cdot | s_g, a), s'_i \sim P_l(\cdot | s_i, s_g, a_i), \forall i \in [n]} \max_{a' \in \mathcal{A}} Q^t(s', a'). \quad (4)$$

It is well known that  $\mathcal{T}$  is  $\gamma$ -contractive, ensuring that the above updates converge to a unique  $Q^*$  such that  $\mathcal{T}Q^* = Q^*$ . The optimal policy  $\pi^* : \mathcal{S} \rightarrow \mathcal{A}$  can then be computed greedily as

<sup>2</sup>We provide supporting code for the algorithm and experiments in <https://github.com/emiletimothy/Mean-Field-Subsample-Q-Learning>

$\pi^*(s) = \arg \max_{a \in \mathcal{A}} Q^*(s, a)$ . However, the update complexity for the  $Q$ -function is  $O(|\mathcal{S}||\mathcal{A}|) = O(|\mathcal{S}_g||\mathcal{S}_l|^n|\mathcal{A}_g||\mathcal{A}_l|^n)$ , which is exponential in the number of local agents increases.

**Mean-field transformation.** To address this, mean-field MARL [Yang et al., 2018] (under homogeneity assumptions) studies the empirical distribution function  $F_{z_{[n]}} : \mathcal{Z}_l \rightarrow \mathbb{R}$ , for  $\mathcal{Z}_l := \mathcal{S}_l \times \mathcal{A}_l$ :

$$F_{z_{[n]}}(z) := \frac{1}{n} \sum_{i=1}^n \mathbf{1}\{s_i = z_s, a_i = z_a\}, \quad \forall z := (z_s, z_a) \in \mathcal{Z}_l := \mathcal{S}_l \times \mathcal{A}_l. \quad (5)$$

Let  $\mu_n(\mathcal{Z}_l) = \{\frac{b}{n} | b \in \{0, \dots, n\}\}^{|\mathcal{Z}_l|}$  be the space of  $|\mathcal{Z}_l|$ -sized vectors (or  $|\mathcal{S}_l| \times |\mathcal{A}_l|$ -sized tables), where each entry is in  $\{0, 1/n, 2/n, \dots, 1\}$ . Intuitively,  $\mu_n(\mathcal{Z}_l)$  is a discrete distribution over  $(\mathcal{S}_l, \mathcal{A}_l)$  where each probability assigned is a multiple of  $1/n$ . Then,  $F_{z_{[n]}} \in \mu_n(\mathcal{Z}_l)$  indicates the proportion of agents in each state-action pair.

As in Q-learning, in mean-field Q-learning initially  $\hat{Q}^0(s_g, s_1, a_g, a_1, F_{z_{[n]}\setminus 1}) = 0$ . At each time  $t \in [T]$ , we update  $\hat{Q}$  as  $\hat{Q}^{t+1} = \hat{\mathcal{T}}\hat{Q}^t$ , where  $\hat{\mathcal{T}}$  is the Bellman operator in *distribution space*:

$$\hat{\mathcal{T}}\hat{Q}^t(s_g, s_1, a_g, a_1, F_{z_{[n]}\setminus 1}) = r(s, a) + \gamma \mathbb{E}_{\substack{s'_g \sim P_g(\cdot | s_g, a_g) \\ s'_i \sim P_l(\cdot | s_i, s_g, a_i) \\ \forall i \in [n]}} \max_{\substack{(a'_g, a'_1, a'_{[n]\setminus 1}) \\ \in \mathcal{A}_g \times \mathcal{A}_l \times \mathcal{A}_l^{n-1}}} \hat{Q}^t(s'_g, s'_1, a'_g, a'_1, F_{z'_{[n]}\setminus 1})$$

Since the  $Q$ -function is permutation-invariant in the homogeneous local agents, one sees that for each  $t \geq 0$ ,  $Q^t(s_g, s_{[n]}, a_g, a_{[n]}) = \hat{Q}^t(s_g, s_1, a_g, a_1, F_{z_{[n]}\setminus 1})$ . In other words, mean-field Q-learning implements the same updates as standard Q-learning. However, the update complexity of  $\hat{Q}$  is only  $O(|\mathcal{S}_g||\mathcal{A}_g||\mathcal{Z}_l|n^{|\mathcal{Z}_l|})$ , which scales polynomially in  $n$  but exponentially in  $|\mathcal{Z}_l|$ .

**Remark 3.1.** Existing methods use sample complexity  $\tilde{O}(\min\{|\mathcal{S}_g||\mathcal{A}_g||\mathcal{Z}_l|^n, |\mathcal{S}_g||\mathcal{A}_g||\mathcal{Z}_l|n^{|\mathcal{Z}_l|}\})$ : one uses  $Q$ -learning if  $|\mathcal{Z}_l|^{n-1} < n^{|\mathcal{Z}_l|}$ , and mean-field value iteration otherwise. In each regime, as  $n$  scales, the update complexity becomes computationally infeasible.

To further reduce the update complexity, we propose SUBSAMPLE-MFQ to overcome the polynomial (in  $n$ ) sample complexity of mean-field  $Q$ -learning and the exponential (in  $n$ ) sample complexity of traditional  $Q$ -learning. We begin by motivating the intuition behind our algorithms.

### 3.1 Overview of approach.

**Offline Planning: Algorithm 1.** First, the global agent randomly samples a subset of local agents  $\Delta \subseteq [n]$  such that  $|\Delta| = k$ , for  $k \leq n$ . It then ignores all other local agents  $[n] \setminus \Delta$ , and performs value iteration (using  $m$  samples to update the  $Q$ -function in each iterate) to approximately learn the  $Q$ -function  $\hat{Q}_{k,m}^{\text{est}}$  and policy  $\hat{\pi}_{k,m}^{\text{est}}$  for this surrogate subsystem of  $k$  local agents. We denote the surrogate reward gained by this subsystem at each time step by  $r_\Delta : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ , where

$$r_\Delta(s, a) = r_g(s_g, a_g) + \frac{1}{|\Delta|} \sum_{i \in \Delta} r_l(s_g, s_i, a_i). \quad (6)$$

In Theorem E.3, we show that  $\|\hat{Q}_{k,m}^{\text{est}} - Q^*\|_\infty$  is Lipschitz continuous with respect to the TV-distance between the state/action pairs of the subsampled agents and the full set of  $n$  agents. Equipped with this approximation guarantee, we show how to construct an approximately optimal policy on the *full* system on  $n$  agents. In general, converting this policy on  $k$  local agents to a policy on the full  $n$ -agent system without sacrificing error guarantees can be intractable, and there is a line of works on centralized-training decentralized-execution (CTDE) [Xu and Klabjan, 2023, Zhou et al., 2023] which shows that designing performant decentralized policies can be highly non-trivial.

**Online Execution: Algorithm 2.** In order to circumvent this obstacle and convert the optimality of each agent's action in the  $k$  local-agent subsystem to an approximate optimality guarantee on the full  $n$ -agent system, we propose a randomized policy  $\pi_{k,m}^{\text{est}}$ , where the global agent samples  $\Delta \in \binom{[n]}{k}$  at each time-step to derive an action  $a_g \leftarrow \hat{\pi}_{k,m}^{\text{est}}(s_g, s_\Delta)$ , and where each local  $i$  agent samples  $k-1$  other local agents  $\Delta_i$  to derive an action  $a_i \leftarrow \hat{\pi}_{k,m}^{\text{est}}(s_g, s_i, s_{\Delta_i})$ . Finally, Theorem 4.4 shows that the policy  $\pi_{k,m}^{\text{est}}$  converges to the optimal policy  $\pi^*$  as  $k \rightarrow n$  with rate  $\tilde{O}(\frac{1}{\sqrt{k}})$ .

### 3.2 Algorithm description.

We now formally describe the algorithms. Before describing Algorithm 1 (SUBSAMPLE-MFQ: Learning) and Algorithm 2 (SUBSAMPLE-MFQ: Execution) in detail, it will be helpful to first define the *empirical distribution function*:

**Definition 3.2** (Empirical Distribution Function). For any population  $(z_1, \dots, z_n) \in \mathcal{Z}_l^n$ , where  $\mathcal{Z}_l := \mathcal{S}_l \times \mathcal{A}_l$ , define the empirical distribution function  $F_{z_\Delta} : \mathcal{Z}_l \rightarrow \mathbb{R}_+$  for all  $z := (z_s, z_a) \in \mathcal{S}_l \times \mathcal{A}_l$  and for all  $\Delta \in \binom{[n]}{k}$  by  $F_{z_\Delta}(x) := \frac{1}{k} \sum_{i \in \Delta} \mathbf{1}\{s_i = z_s, a_i = z_a\}$ .

Let  $\mu_k(\mathcal{Z}_l) := \left\{ \frac{b}{k} \mid b \in \{0, \dots, k\} \right\}^{|\mathcal{Z}_l|}$  be the space of  $|\mathcal{Z}_l|$ -sized vectors (or  $|\mathcal{S}_l| \times |\mathcal{A}_l|$ -sized tables) where each entry is in  $\{0, 1/k, 2/k, \dots, 1\}$ . Intuitively,  $\mu_k(\mathcal{Z}_l)$  is a discrete distribution over  $(\mathcal{S}_l, \mathcal{A}_l)$  where each probability assigned is a multiple of  $1/k$ . Here,  $F_{z_\Delta} \in \mu_k(\mathcal{Z}_l)$  indicates the proportion of agents (in the  $k$ -local-agent subsystem) at each state/action pair.

**Algorithm 1** (Offline learning). Let  $m \in \mathbb{N}$  denote the sample size for the learning algorithm with sampling parameter  $k \leq n$ . As in Theorem 3.1, when  $|\mathcal{Z}_l|^k \leq |\mathcal{Z}_l|^{|\mathcal{Z}_l|}$ , the algorithm uses traditional value-iteration, and when  $|\mathcal{Z}_l|^k > |\mathcal{Z}_l|^{|\mathcal{Z}_l|}$ , it uses mean-field value iteration. We formally describe the procedure for each regime below:

**Regime with Large State/Action Space:** When  $|\mathcal{Z}_l|^k \leq |\mathcal{Z}_l|^{|\mathcal{Z}_l|}$ , we iteratively learn the optimal  $Q$ -function for a subsystem with  $k$ -local agents denoted by  $\hat{Q}_{k,m}^t : \mathcal{S}_g \times \mathcal{S}_l^k \times \mathcal{A}_g \times \mathcal{A}_l^k \rightarrow \mathbb{R}$ , which is initialized to 0. At time  $t$ , we update

$$\hat{Q}_{k,m}^{t+1}(s_g, s_\Delta, a_g, a_\Delta) = \tilde{\mathcal{T}}_{k,m} \hat{Q}_{k,m}^t(s_g, s_\Delta, a_g, a_\Delta), \quad (7)$$

where  $\tilde{\mathcal{T}}_{k,m}$  is the *empirically adapted Bellman operator* in Equation (8). Since the system is unknown,  $\tilde{\mathcal{T}}_{k,m}$  cannot directly perform the Bellman update, so it instead uses  $m$  random samples  $s_g^j \sim P_g(\cdot | s_g, a_g)$  and  $s_i^j \sim P_l(\cdot | s_i, s_g, a_i)$  for each  $j \in [m]$ ,  $i \in \Delta$  to approximate the system:

$$\tilde{\mathcal{T}}_{k,m} \hat{Q}_{k,m}^t(s_g, s_\Delta, a_g, a_\Delta) = r_\Delta(s, a) + \frac{\gamma}{m} \sum_{j \in [m]} \max_{\substack{a'_g \in \mathcal{A}_g, a'_{\Delta} \in \mathcal{A}_l^k}} \hat{Q}_{k,m}^t(s_g^j, s_\Delta^j, a'_g, a'_{\Delta}).$$

Since  $\tilde{\mathcal{T}}_{k,m}$  is  $\gamma$ -contractive, Algorithm 1 applies value iteration with  $\tilde{\mathcal{T}}$  until  $\hat{Q}_{k,m}$  converges to a fixed point satisfying  $\tilde{\mathcal{T}}_{k,m} \hat{Q}_{k,m}^{\text{est}} = \hat{Q}_{k,m}^{\text{est}}$ , yielding a deterministic policy  $\hat{\pi}_{k,m}^{\text{est}}(s_g, s_\Delta)$  where

$$\hat{\pi}_{k,m}^{\text{est}}(s_g, s_\Delta) = \arg \max_{\substack{a_g \in \mathcal{A}_g, a_\Delta \in \mathcal{A}_l^k}} \hat{Q}_{k,m}^{\text{est}}(s_g, s_\Delta, a_g, a_\Delta). \quad (8)$$

**Regime with Large Number of Agents:** When  $|\mathcal{Z}_l|^k > |\mathcal{Z}_l|^{|\mathcal{Z}_l|}$ , we learn the optimal mean-field  $Q$ -function for a  $k$  local agent system, denoted by  $\hat{Q}_{k,m}^t : \mathcal{S}_g \times \mathcal{S}_l \times \mu_{k-1}(\mathcal{Z}_l) \times \mathcal{A}_g \times \mathcal{A}_l \rightarrow \mathbb{R}$ , which is initialized to 0. At time  $t$ , we update

$$\hat{Q}_{k,m}^{t+1}(s_g, s_1, F_{z_\Delta}, a_1, a_g) = \hat{\mathcal{T}}_{k,m} \hat{Q}_{k,m}^t(s_g, s_1, F_{z_\Delta}, a_1, a_g), \quad (9)$$

where  $\hat{\mathcal{T}}_{k,m}$  is the *empirically adapted mean-field Bellman operator* in Equation (10). Similarly, as the system is unknown,  $\hat{\mathcal{T}}_{k,m}$  cannot directly perform the Bellman update and instead uses  $m$  random samples  $s_g^j \sim P_g(\cdot | s_g, a_g)$  and  $s_i^j \sim P_l(\cdot | s_i, s_g, a_i)$ ,  $\forall j \in [m]$ ,  $i \in \Delta$  to approximate the system:

$$\begin{aligned} \hat{\mathcal{T}}_{k,m} \hat{Q}_{k,m}^t(s_g, s_1, F_{z_\Delta}, a_1, a_g) &= r_\Delta(s, a) + \frac{\gamma}{m} \sum_{j \in [m]} \max_{\substack{a'_g \in \mathcal{A}_g, a'_1 \in \mathcal{A}_l, \\ F_{a'_\Delta} \in \mu_{k-1}(\mathcal{A}_l)}} \hat{Q}_{k,m}^t(s_g^j, s_1^j, F_{s_\Delta^j, a_\Delta'}, a'_1, a'_g) \end{aligned} \quad (10)$$

Since  $\hat{Q}_{k,m}^t$  depends on  $s_\Delta$  and  $a_\Delta$  through  $F_{z_\Delta}$ ,  $\hat{\mathcal{T}}_{k,m}$  is also  $\gamma$ -contractive. So, Algorithm 1 applies value iteration with  $\hat{\mathcal{T}}$  until  $\hat{Q}_{k,m}$  converges to a fixed point satisfying  $\hat{\mathcal{T}}_{k,m} \hat{Q}_{k,m}^{\text{est}} = \hat{Q}_{k,m}^{\text{est}}$ , yielding a deterministic policy  $\hat{\pi}_{k,m}^{\text{est}}(s_g, s_1, F_{s_\Delta})$ :

$$\hat{\pi}_{k,m}^{\text{est}}(s_g, s_1, F_{s_\Delta}) = \arg \max_{\substack{a_g \in \mathcal{A}_g, a_1 \in \mathcal{A}_l, F_{a_\Delta} \in \mu_{k-1}(\mathcal{A}_l)}} \hat{Q}_{k,m}^{\text{est}}(s_g, s_1, F_{s_\Delta}, a_1, a_g)$$

For  $\hat{\pi}_{k,m}^{\text{est}}(s_g, s_i, s_{\Delta \setminus i}) = a_g^*, a_i^*, a_{\Delta \setminus i}^*$ , let  $[\hat{\pi}_{k,m}^{\text{est}}(s_g, s_\Delta)]_g = a_g^*$  and  $[\hat{\pi}_{k,m}^{\text{est}}(s_g, s_i, s_{\Delta \setminus i})]_i = a_i^*$ .

---

**Algorithm 1** SUB-SAMPLE-MFQ: Learning

---

**Require:** A multi-agent system as in Section 2, number of iterations  $T$ , sampling parameters  $k \in [n]$ ,  $m \in \mathbb{N}$ , and discount factor  $\gamma \in (0, 1)$ .

- 1: Let  $\Delta = \{1, \dots, k\}$ ,  $\tilde{\Delta} = \{2, \dots, k\}$ , and  $\mu_{k-1}(\mathcal{Z}_l) = \{\frac{b}{k-1} : b \in \{0, 1, \dots, k-1\}\}^{|\mathcal{S}_l| \times |\mathcal{A}_l|}$ .
- 2: **if**  $|\mathcal{Z}_l|^k \leq |\mathcal{Z}_l|k^{|\mathcal{Z}_l|}$  **then**
- 3:   // Sub-sampled version of standard  $Q$ -learning
- 4:   Initialize  $\hat{Q}_{k,m}^0(s_g, s_\Delta, a_g, a_\Delta) = 0, \forall (s_g, s_\Delta, a_g, a_\Delta)$ .
- 5:   **for**  $t = 1$  to  $T$  **do**
- 6:     **for**  $(s_g, s_\Delta, a_\Delta, a_g) \in \mathcal{S}_g \times \mathcal{S}_\Delta^k \times \mathcal{A}_g \times \mathcal{A}_\Delta^k$  **do**
- 7:        $\hat{Q}_{k,m}^{t+1}(s_g, s_\Delta, a_g, a_\Delta) = \tilde{\mathcal{T}}_{k,m} \hat{Q}_{k,m}^t(s_g, s_\Delta, a_g, a_\Delta)$
- 8:     Let the greedy policy be  $\hat{\pi}_{k,m}^{\text{est}}(s_g, s_\Delta) := \arg \max_{a_g \in \mathcal{A}_g, a_\Delta \in \mathcal{A}_\Delta^k} \hat{Q}_{k,m}^T(s_g, s_\Delta, a_g, a_\Delta)$ .
- 9: **else**
- 10:   // Sub-sampled version of mean-field  $Q$ -learning
- 11:   Initialize  $\hat{Q}_{k,m}^0(s_g, s_1, F_{z_{\tilde{\Delta}}}, a_1, a_g) = 0, \forall (s_g, s_\Delta, a_\Delta, a_g)$ .
- 12:   **for**  $t = 1$  to  $T$  **do**
- 13:     **for**  $(s_g, s_1, F_{z_{\tilde{\Delta}}}, a_1, a_g) \in \mathcal{S}_g \times \mathcal{S}_1 \times \mu_{k-1}(\mathcal{Z}_l) \times \mathcal{A}_l \times \mathcal{A}_g$  **do**
- 14:        $\hat{Q}_{k,m}^{t+1}(s_g, s_1, F_{z_{\tilde{\Delta}}}, a_1, a_g) = \tilde{\mathcal{T}}_{k,m} \hat{Q}_{k,m}^t(s_g, s_1, F_{z_{\tilde{\Delta}}}, a_1, a_g)$
- 15:     Let the greedy policy be  $\hat{\pi}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\tilde{\Delta}}}) := \arg \max_{\substack{a_g \in \mathcal{A}_g, a_i \in \mathcal{A}_i, \\ F_{a_{\tilde{\Delta}}} \in \mu_{k-1}(\mathcal{A}_i)}} \hat{Q}_{k,m}^T(s_g, s_i, F_{z_{\tilde{\Delta}}}, a_1, a_g)$ .

---

**Remark 3.3.** Since mean-field  $Q$ -learning maintains the same updates as standard  $Q$ -learning (from Theorem C.19 which follows by noting that the  $Q$ -function is permutation-invariant in the homogeneous local agents), the deterministic policies  $\hat{\pi}_{k,m}^{\text{est}}(s_g, s_\Delta)$  and  $\hat{\pi}_{k,m}^{\text{est}}(s_g, s_1, F_{s_{\tilde{\Delta}}})$  are equivalent.

**Algorithm 2** (Online implementation). In Algorithm 2, (SUBSAMPLE-MFQ: Execution) the global agent samples local agents  $\Delta(t) \sim \mathcal{U}(\binom{[n]}{k})$  at each step to derive action  $a_g(t) = [\hat{\pi}_{k,m}^{\text{est}}(s_g, s_\Delta(t))]_g$ , and each local agent  $i$  samples other local agents  $\Delta_i(t) \sim \mathcal{U}(\binom{[n] \setminus i}{k-1})$  to derive action  $a_i(t) = [\hat{\pi}_{k,m}^{\text{est}}(s_g, s_i, s_\Delta(t))]_i$ . The system then incurs a reward  $r(s, a)$ . This procedure of first sampling agents and then applying  $\hat{\pi}_{k,m}^{\text{est}}$  is denoted by a stochastic policy  $\pi_{k,m}^{\text{est}}(a|s)$ , where  $\pi_{k,m}^{\text{est}}(a_g|s)$  is the global agent's action distribution and  $\pi_{k,m}^{\text{est}}(a_l|s)$  is the local agent's action distribution:

$$\pi_{k,m}^{\text{est}}(a_g|s) = \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \mathbf{1}(\hat{\pi}_{k,m}^{\text{est}}(s_g, s_\Delta) = a) \quad (11)$$

$$\pi_{k,m}^{\text{est}}(a_i|s) = \frac{1}{\binom{n-1}{k-1}} \sum_{\tilde{\Delta} \in \binom{[n] \setminus i}{k-1}} \mathbf{1}(\pi_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\tilde{\Delta}}}) = a_i) \quad (12)$$

The agents then transition to their next states.

## 4 Theoretical Guarantees and Analysis Approach

We now show the value of the expected discounted cumulative reward produced by  $\pi_{k,m}^{\text{est}}$  is approximately optimal, where the optimality gap decays as  $k \rightarrow n$  and  $m$  grows.

**Bellman noise.** We introduce the notion of Bellman noise, which is used in the main theorem. Note that  $\hat{\mathcal{T}}_{k,m}$  is an unbiased estimator of the adapted Bellman operator  $\hat{\mathcal{T}}_k$ ,

$$\hat{\mathcal{T}}_k \hat{Q}_k(s_g, s_\Delta, a_g, a_\Delta) = r_\Delta(s, a) + \gamma \mathbb{E}_{\substack{s'_g \sim P_g(\cdot | s_g, a_g), \\ s'_i \sim P_i(\cdot | s_i, s_g, a_i), \forall i \in \Delta}} \max_{\substack{a'_g \in \mathcal{A}_g, a'_\Delta \in \mathcal{A}_\Delta^k}} \hat{Q}_k(s'_g, s'_\Delta, a'_g, a'_\Delta). \quad (13)$$

Let  $\hat{Q}_k^0(s_g, s_\Delta, a_g, a_\Delta) = 0$ . For  $t > 0$ , let  $\hat{Q}_k^{t+1} = \hat{\mathcal{T}}_k \hat{Q}_k^t$ , where  $\hat{\mathcal{T}}_k$  is defined for  $k \leq n$  in Equation (13). Then,  $\hat{\mathcal{T}}_k$  is also a  $\gamma$ -contraction with fixed-point  $\hat{Q}_k^*$ . By the law of large numbers,  $\lim_{m \rightarrow \infty} \hat{\mathcal{T}}_{k,m} = \hat{\mathcal{T}}_k$  and  $\|\hat{Q}_{k,m}^{\text{est}} - \hat{Q}_k^*\|_\infty \rightarrow 0$  as  $m \rightarrow \infty$ . For finite  $m$ ,  $\epsilon_{k,m} := \|\hat{Q}_{k,m}^{\text{est}} - \hat{Q}_k^*\|_\infty$

---

**Algorithm 2** SUBSAMPLE-MFQ: Execution

---

**Require:** Parameter  $T'$  for the number of iterations for the decision-making sequence. Sampling parameter  $k \in [n], m \in \mathbb{N}$ . Discount factor  $\gamma$ . Policy  $\hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_\Delta})$ .

- 1: Learn  $\hat{\pi}_{k,m}^{\text{est}}$  from Algorithm 1.
- 2: Sample  $(s_g(0), s_{[n]}(0)) \sim s_0$ , where  $s_0$  is a distribution on the initial global state  $(s_g, s_{[n]})$
- 3: Initialize the total reward  $R_0 = 0$ .
- 4: Policy  $\pi_{k,m}^{\text{est}}(s)$  is defined as follows:
- 5: **for**  $t = 0$  to  $T'$  **do**
- 6:   Choose  $\Delta$  uniformly at random from  $\binom{[n]}{k}$  and let  $a_g(t) = [\hat{\pi}_{k,m}^{\text{est}}(s_g(t), s_\Delta(t))]_g$ .
- 7:   **for**  $i = 1$  to  $n$  **do**
- 8:     Choose  $\Delta_i$  uniformly at random from  $\binom{[n] \setminus i}{k-1}$  and let  $a_i(t) = [\hat{\pi}_{k,m}^{\text{est}}(s_g(t), s_i(t), s_{\Delta_i}(t))]_l$ .
- 9:     Let  $s_g(t+1) \sim P_g(\cdot | s_g(t), a_g(t))$ .
- 10:    Let  $s_i(t+1) \sim P_l(\cdot | s_i(t), s_g(t), a_i(t)), \forall i \in [n]$ .
- 11:     $R_{t+1} = R_t + \gamma^t \cdot r(s, a)$

---

is the well-studied Bellman noise. To compare the performance between  $\pi^*$  and  $\pi_k^{\text{est}}$ , we define the value function of a policy  $\pi$ :

**Definition 4.1.** For a given policy  $\pi$ , the value function  $V^\pi : \mathcal{S} \rightarrow \mathbb{R}$  for  $\mathcal{S} := \mathcal{S}_g \times \mathcal{S}_l^n$  is given by:

$$V^\pi(s) = \mathbb{E}_{a(t) \sim \pi(\cdot | s(t))} \left[ \sum_{t=0}^{\infty} \gamma^t r(s(t), a(t)) \middle| s(0) = s \right]. \quad (14)$$

Intuitively,  $V^\pi(s)$  is the expected discounted cumulative reward when starting from state  $s$  and applying actions from the policy  $\pi$  across an infinite horizon.

With the above preparations, we are primed to present our main result: a high-probability bound on the optimality gap for our learned policy  $\pi_{k,m}^{\text{est}}$  that decays with rate  $\tilde{O}(1/\sqrt{k})$ .

**Theorem 4.2.** Let  $\pi_{k,m}^{\text{est}}$  denote the learned policy deployed in SUBSAMPLE-MFQ: Execution. Then, for all  $s_0 \in \mathcal{S} := \mathcal{S}_g \times \mathcal{S}_l^n$ , we have

$$V^{\pi^*}(s_0) - V^{\pi_{k,m}^{\text{est}}}(s_0) \leq \frac{\tilde{r}}{(1-\gamma)^2} \sqrt{\frac{n-k+1}{2nk}} \sqrt{\ln \frac{40\tilde{r}|\mathcal{S}_l||\mathcal{A}_l||\mathcal{A}_g|k^{|\mathcal{A}_l|+\frac{1}{2}}}{(1-\gamma)^2}} + \frac{1}{10\sqrt{k}} + 2\epsilon_{k,m}$$

We prove Theorem 4.4 in Section G, and provide a proof sketch in Section D. We also generalize the result to stochastic rewards in Section H.

To control the Bellman noise  $\epsilon_{k,m}$ , we show that for sufficiently many samples  $m^*$ ,  $\epsilon_{k,m^*}$  also decays on the order of  $\tilde{O}(1/\sqrt{k})$  with high probability. For this, we introduce Theorem 4.3:

**Lemma 4.3** (Controlling the Bellman Noise.). For  $k \in [n]$ , let

$$m^* = 2|\mathcal{S}_g||\mathcal{A}_g||\mathcal{S}_l||\mathcal{A}_l|k^{3.5+|\mathcal{S}_l||\mathcal{A}_l|} \frac{\log(|\mathcal{S}_g||\mathcal{A}_g||\mathcal{A}_l||\mathcal{S}_l|)}{(1-\gamma)^5} \log \frac{1}{(1-\gamma)^2}$$

be the number of samples in Equation (10). If the number of iterations  $T$  satisfies  $T \geq \frac{2}{1-\gamma} \log \frac{\tilde{r}\sqrt{k}}{1-\gamma}$  then with probability at least  $1 - \frac{1}{100e^k}$ , the Bellman error satisfies  $\epsilon_{k,m^*} \leq \tilde{O}(\frac{1}{\sqrt{k}})$ .

We defer the proof of Theorem 4.3 to Section G.1. To simplify notation, let  $\pi_k^{\text{est}} := \pi_{k,m^*}^{\text{est}}$ . Then, by combining Theorem 4.3 and Theorem 4.2, we obtain our main result in Theorem 4.4:

**Theorem 4.4.** Let  $\pi_k^{\text{est}}$  denote the learned policy from SUBSAMPLE-MFQ: Execution where the number of samples  $m$  is determined in Theorem 4.3. Suppose  $T \geq \frac{2}{1-\gamma} \log \frac{\tilde{r}\sqrt{k}}{1-\gamma}$ . Then,  $\forall s_0 \in \mathcal{S} := \mathcal{S}_g \times \mathcal{S}_l^n$ , with probability at least  $1 - 1/100e^k$ , we have<sup>3</sup>,

$$V^{\pi^*}(s_0) - V^{\pi_k^{\text{est}}}(s_0) \leq \frac{\tilde{r}}{(1-\gamma)^2} \sqrt{\frac{n-k+1}{2nk} \ln \frac{40\tilde{r}|\mathcal{S}_l||\mathcal{A}_l||\mathcal{A}_g|k^{|\mathcal{A}_l|+\frac{1}{2}}}{(1-\gamma)^2}} + \frac{4}{\sqrt{k}}.$$

<sup>3</sup>The  $1/100e^k$  term can be replaced to any arbitrary  $\delta > 0$  at the cost of attaching logarithmic dependencies of  $\delta$  on the  $4/\sqrt{k}$  term in the error bound.

**Remark 4.5.** One could also derive an alternate bound that retains the  $T$  factor via a  $\gamma^T$  dependence in the final bound (which shows an exponentially decaying error with the time horizon  $T$ ). Moreover, in Theorem G.10, we show that the query complexity is on the order of  $O(mT|S_g||S_l|^k|A_g||A_l|^k)$ , and we bound  $T$  and set  $m = \text{poly}(1/(1-\gamma))$  to attain the  $1/\sqrt{k}$  rate.

**Remark 4.6.** Additionally, our  $\text{poly}(1/(1-\gamma))$ -dependence may be loose since we do not use more complicated variance reduction techniques as in Sidford et al. [2018a,b], Wainwright [2019], Jin et al. [2024] to optimize the number of samples  $m$  which is used to bound the Bellman error  $\epsilon_{k,m}$ . Moreover, incorporating variance reduction would significantly complicate the algorithm and intuition.

Our analysis hinges on two non-trivial technical steps. Firstly, for an intermediary step in Theorem E.3 we establish that TV distance, rather than the stronger KL-divergence, is the correct metric to use (as the KL-divergence between  $F_{z_\Delta}$  and  $F_{z_{[n]}}$  decays too slowly as  $k \rightarrow n$ , as we show in Theorem F.6). This requires exploiting a recent extension of the DKW inequality to sampling without replacement [Anand and Qu, 2024], and showing that the transition dynamics we study saturates the data-processing inequality for TV-distance. Secondly, we adapt the celebrated performance-difference lemma [Kakade and Langford, 2002] to our multi-agent setting, which entails a principled analysis and careful probabilistic argument (to which we refer the reader to Section G).

**Remark 4.7.** We also extend the formulation of SUBSAMPLE-MFQ to off-policy  $Q$ -learning Chen et al. [2021b], Chen and Maguluri [2022a], which replaces the generative oracle assumption with a stochastic approximation scheme that learns an approximately optimal policy using historical data. Section I provides theoretical guarantees with a similar decaying optimality gap as in Theorem 4.4.

**Remark 4.8.** The asymptotic sample complexity of Algorithm 1 for learning  $\hat{\pi}_k^{\text{est}}$  for a fixed  $k$  is  $\min\{\tilde{O}(|Z_l|^k|\mathcal{S}_g||\mathcal{A}_g|), \tilde{O}(k^{|Z_l|}|Z_l||\mathcal{S}_g||\mathcal{A}_g|)\}$ , which is at least polynomially faster the standard  $Q$ -learning or mean-field value iteration as discussed in Theorem 3.1. By Theorem 4.4, as  $k \rightarrow n$ , the optimality gap decays, revealing a fundamental trade-off in the choice of  $k$ : increasing  $k$  improves the performance of the policy, but increases the size of the  $Q$ -function. We explore this trade-off further in our experiments. If we set  $k = O(\log n)$ , this leads to a sample complexity of  $\min\{\tilde{O}(n^{\log |Z_l|}|\mathcal{S}_g||\mathcal{A}_g|), \tilde{O}((\log n)^{|Z_l|}|\mathcal{S}_g||\mathcal{A}_g|)\}$ . This is an exponential speedup on the complexity from mean-field value iteration (from  $\text{poly}(n)$  to  $\text{poly}(\log n)$ ), as well as over traditional value-iteration (from  $\exp(n)$  to  $\text{poly}(n)$ ), where the optimality gap decays to 0 with rate  $O(\frac{1}{\sqrt{\log n}})$ .

**Remark 4.9.** If  $k = O(\log n)$ , SUBSAMPLE-MFQ handles  $|\mathcal{E}| \leq O(\log n / \log \log n)$  types of local agents, since the run-time of the learning algorithm becomes  $\text{poly}(n)$ . This surpasses the previous heterogeneity capacity from Mondal et al. [2022], which only handles constant  $|\mathcal{E}| \leq O(1)$ . Increasing the agent heterogeneity using this *type formulation* does make the algorithm somewhat more expensive since it factors into the query complexity via the state space of the local agents; however, since the optimality gap of the learned policy is on the order of  $\tilde{O}(1/\sqrt{k})$  (modulo very small  $\sqrt{\log |\mathcal{S}_l||\mathcal{A}_l|}$  factors), increasing the amount of agent heterogeneity does not degrade the quality of the learned policy as the theoretical bound does not get worse. Moreover, recent methods from the graphon mean-field MARL community might be able to enable stronger heterogeneity in the system [Cui and Koeppl, 2022, Anand and Liaw, 2025, Hu et al., 2023].

In the non-tabular setting with infinite state/action spaces, one could replace the  $Q$ -learning algorithm with any arbitrary value-based RL method that learns  $\hat{Q}_k$  with function approximation [Sutton et al., 1999] such as deep  $Q$ -networks [Silver et al., 2016]. Doing so raises an additional error that factors into Theorem 4.4. We formalize this below.

**Assumption 4.10** (Linear MDP with infinite state spaces). Suppose  $\mathcal{S}_g$  and  $\mathcal{S}_l$  are infinite compact sets. Furthermore, suppose there exists a feature map  $\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$  and  $d$  unknown (signed) measures  $\mu = (\mu^1, \dots, \mu^d)$  over  $\mathcal{S}$  and a vector  $\theta \in \mathbb{R}^d$  such that for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , we have  $\mathbb{P}(\cdot|s, a) = \langle \phi(s, a), \mu(\cdot) \rangle$  and  $r(s, a) = \langle \phi(s, a), \theta \rangle$ .

The existence of  $\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$  implies one can estimate the  $Q$ -function of any policy as a linear function. This assumption is commonly used in policy iteration algorithms Lattimore et al. [2020], Wang et al. [2023], and allows one to obtain sample complexity bounds that are independent of  $|\mathcal{S}_l|$  and  $|\mathcal{A}_l|$ . Finally, as is standard in RL, we assume bounded feature-norms [Tkachuk et al., 2023]:

**Assumption 4.11** (Bounded features). We assume that  $\|\phi(s, a)\|_2 \leq 1$  for all  $(s, a) \in \mathcal{S} \times \mathcal{A}$ .

Then, through a reduction from Zhang et al. [2024], Ren et al. [2024] that uses function approximation to learn the spectral features  $\phi_k$  for  $\hat{Q}_k$ , we derive a performance guarantee for the learned policy  $\pi_k^{\text{est}}$ , where the optimality gap decays with  $k$ .

**Theorem 4.12.** *When  $\pi_k^{\text{est}}$  is derived from the spectral features  $\phi_k$  learned in  $\hat{Q}_k$ , and  $M$  is the number of samples used in the function approximation, then*

$$\Pr \left[ V^{\pi^*}(s) - V^{\pi_k^{\text{est}}}(s) \leq \tilde{O} \left( \frac{1}{\sqrt{k}} + \frac{\|\phi_k\|^5 \log 2k^2}{\sqrt{M}} + \frac{2\gamma\tilde{r} \cdot \|\phi_k\|}{(1-\gamma)\sqrt{k}} \right) \right] \geq 1 + \frac{200}{k} - \frac{201}{200\sqrt{k}}$$

We defer the proof of Theorem 4.12 to Section J.

## 5 Conclusion and Future Works

This work develops subsampling for mean field MARL in a cooperative system with a global decision-making agent and  $n$  homogeneous local agents. We propose SUBSAMPLE-MFQ which learns each agent’s best response to the mean effect from a sample of its neighbors, allowing an exponential reduction on the sample complexity of approximating a solution to the MDP. We provide a theoretical analysis on the optimality gap of the learned policy, showing that (with high probability) the learned policy converges to the optimal policy with the number of agents  $k$  sampled at the rate  $\tilde{O}(1/\sqrt{k})$ , and validate our theoretical results through numerical experiments. We show that the decay rate is maintained, on expectation, when the reward functions are stochastic and when agents learn from a single trajectory on historical data through off-policy  $Q$ -learning. Finally, we extend this result to the non-tabular setting with infinite state and action spaces under assumptions of a linear MDP model.

**Limitations and future work.** Our current work assumes that the global and local agents cooperate to optimize a structured reward under a specific dynamic model. While this model is more general than the federated learning setting, one direction would be to extend our algorithms and analysis to weaker network assumptions. We believe our framework, which can handle dense subgraphs, as well as expander-graph decompositions [Reingold, 2008, Anand and Umans, 2023, Anand, 2025] may be amenable for this. Secondly, our current work incorporates mild heterogeneity among agents and assumes they are cooperative; thus, another avenue would be to consider settings with competitive agents or more complex agent heterogeneity. Finally, it would be exciting to generalize this work to the online no-regret setting.

**Societal Impacts.** This work is theoretical and foundational in nature. As such, while it enables more scalable multi-agent algorithms, it is not tied to any specific applications or deployments.

**Acknowledgements.** This work was supported by NSF Grants 2154171, CAREER Award 2339112, CMU CyLab Seed Funding, and CCF 2338816. We gratefully acknowledge insightful discussions with Siva Theja Maguluri, Yiheng Lin, Jan van den Brand, Adam Wierman, Yunbum Kook, Yi Wu, Sarah Liaw, and Rishi Veerapaneni. Finally, we thank the anonymous NeurIPS reviewers for their helpful feedback.

## References

- Emile Anand. Towards the pseudorandomness of expander random walks for read-once acc0 circuits, 2025. URL <https://arxiv.org/abs/2501.07752>.
- Emile Anand and Sarah Liaw. Feel-good thompson sampling for contextual bandits: a markov chain monte carlo showdown, 2025. URL <https://arxiv.org/abs/2507.15290>.
- Emile Anand and Guannan Qu. Efficient reinforcement learning for global decision making in the presence of local agents at scale. *arXiv*, 2024. URL <https://arxiv.org/abs/2403.00222>.
- Emile Anand and Chris Umans. Pseudorandomness of the Sticky Random Walk. *Caltech Undergraduate Thesis*, 2023. URL [https://thesis.library.caltech.edu/16116/7/AnandCaltech\\_Final\\_Thesis\\_Submission2.pdf](https://thesis.library.caltech.edu/16116/7/AnandCaltech_Final_Thesis_Submission2.pdf).
- Emile Anand, Jan van den Brand, Mehrdad Ghadiri, and Daniel J. Zhang. The Bit Complexity of Dynamic Algebraic Formulas and Their Determinants. In Karl Bringmann, Martin Grohe, Gabriele Puppis, and Ola Svensson, editors, *51st International Colloquium on Automata, Languages, and Programming (ICALP 2024)*, volume 297 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 10:1–10:20, Dagstuhl, Germany, 2024. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-322-5. doi: 10.4230/LIPIcs.ICALP.2024.10. URL <https://drops.dagstuhl.de/storage/00lipics/lipics-vol297-icalp2024/LIPIcs.ICALP.2024.10/LIPIcs.ICALP.2024.10.pdf>.
- Emile Anand, Jan van den Brand, and Rose McCarty. The structural complexity of matrix-vector multiplication. *arXiv preprint arXiv:2502.21240*, 2025. URL <https://www.arxiv.org/pdf/2502.21240.pdf>.
- Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922. URL <http://eudml.org/doc/213289>.
- Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1st edition, 1996. ISBN 1886529108. URL <https://web.stanford.edu/~bvr/pubs/NDPOverview.pdf>.
- Vincent D. Blondel and John N. Tsitsiklis. A Survey of Computational Complexity Results in Systems and Control. *Automatica*, 36(9):1249–1274, 2000. ISSN 0005-1098. doi: [https://doi.org/10.1016/S0005-1098\(00\)00050-9](https://doi.org/10.1016/S0005-1098(00)00050-9). URL <https://www.sciencedirect.com/science/article/abs/pii/S0005109800000509>.
- Shreyas Chaudhari, Srinivasa Pranav, Emile Anand, and José M. F. Moura. Peer-to-peer learning dynamics of wide neural networks. *IEEE International Conference on Acoustics, Speech, and Signal Processing 2025*, 2024. URL <https://arxiv.org/abs/2409.15267>.
- Zaiwei Chen and Siva Theja Maguluri. Sample complexity of policy-based methods under off-policy sampling and linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 11195–11214. PMLR, 2022a. URL <https://proceedings.mlr.press/v151/chen22i/chen22i.pdf>.
- Zaiwei Chen and Siva Theja Maguluri. Sample complexity of policy-based methods under off-policy sampling and linear function approximation. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 11195–11214. PMLR, 28–30 Mar 2022b. URL <https://proceedings.mlr.press/v151/chen22i/chen22i.pdf>.
- Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of off-policy td-learning via generalized bellman operators. *Advances in Neural Information Processing Systems*, 34:21440–21452, 2021a. URL <https://proceedings.neurips.cc/paper/2021/file/b2eeb7362ef83deff5c7813a67e14f0a-Paper.pdf>.

- Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. A lyapunov theory for finite-sample guarantees of asynchronous q-learning and td-learning variants. *arXiv preprint arXiv:2102.01567*, 2021b. URL <https://pubsonline.informs.org/doi/10.1287/opre.2022.0249>.
- Zaiwei Chen, Siva Theja Maguluri, and Martin Zubeldia. Concentration of contractive stochastic approximation: Additive and multiplicative noise. *The Annals of Applied Probability*, 35(2):1298–1352, 2025.
- Kai Cui and Heinz Koepll. Learning Graphon Mean Field Games and Approximate Nash Equilibria. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=0sgnt1pKDOz>.
- Kai Cui, Christian Fabian, and Heinz Koepll. Multi-Agent Reinforcement Learning via Mean Field Control: Common Noise, Major Agents and Approximation Properties, 2023. URL <https://arxiv.org/abs/2303.10665>.
- Alex DeWeese and Guannan Qu. Locally interdependent multi-agent MDP: Theoretical framework for decentralized agents with dynamic dependencies. In *Forty-first International Conference on Machine Learning*, 2024. URL <https://arxiv.org/pdf/2406.06823.pdf>.
- Thomas G. Dietterich, George Trimponias, and Zhitang Chen. Discovering and removing exogenous state variables and rewards for reinforcement learning, 2018. URL <https://arxiv.org/abs/1806.01584>.
- A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *The Annals of Mathematical Statistics*, 27(3):642 – 669, 1956. doi: 10.1214/aoms/1177728174. URL <https://doi.org/10.1214/aoms/1177728174>.
- Dylan J Foster, Alexander Rakhlin, Ayush Sekhari, and Karthik Sridharan. On the Complexity of Adversarial Decision Making. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL <https://openreview.net/forum?id=pgBpQYss2ba>.
- Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019. URL <https://arxiv.org/pdf/1812.02900.pdf>.
- David Gamarnik, David Goldberg, and Theophane Weber. Correlation Decay in Random Decision Networks, 2009. URL <https://www.jstor.org/stable/24540896>.
- Subhashis Ghosal and Aad van der Vaart. *Fundamentals of Nonparametric Bayesian Inference: Space of Probability Densities*, page 516–527. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2017. URL <https://www.cambridge.org/core/books/fundamentals-of-nonparametric-bayesian-inference/C96325101025D308C9F31F4470DEA2E8>.
- Noah Golowich and Ankur Moitra. The role of inherent bellman error in offline reinforcement learning with linear function approximation, 2024. URL <https://arxiv.org/abs/2406.11686>.
- Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-Field Controls with Q-Learning for Cooperative MARL: Convergence and Complexity Analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1168–1196, 2021. doi: 10.1137/20M1360700. URL <https://doi.org/10.1137/20M1360700>.
- Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-Field Multi-Agent Reinforcement Learning: A Decentralized Network Approach, 2022. URL <https://arxiv.org/pdf/2108.02731.pdf>.
- Yuanquan Hu, Xiaoli Wei, Junji Yan, and Hengxi Zhang. Graphon Mean-Field Control for Cooperative Multi-Agent Reinforcement Learning. *Journal of the Franklin Institute*, 360(18):14783–14805, 2023. ISSN 0016-0032. URL <https://doi.org/10.1016/j.jfranklin.2023.09.002>.

- Chi Jin, Zhuoran Yang, Zhaoran Wang, and Michael I Jordan. Provably efficient reinforcement learning with linear function approximation. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2137–2143. PMLR, 09–12 Jul 2020. URL <https://proceedings.mlr.press/v125/jin20a.html>.
- Chi Jin, Qinghua Liu, and Sobhan Miryoosefi. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms, 2021. URL <https://arxiv.org/abs/2102.00815>.
- Junqi Jin, Chengru Song, Han Li, Kun Gai, Jun Wang, and Weinan Zhang. Real-time bidding with multi-agent reinforcement learning in display advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM ’18, page 2193–2201, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360142. doi: 10.1145/3269206.3272021. URL <https://dl.acm.org/doi/10.1145/3269206.3272021>.
- Yujia Jin, Ishani Karmarkar, Aaron Sidford, and Jiayi Wang. Truncated variance reduced value iteration. In *Advances in Neural Information Processing Systems*, 2024. URL [https://proceedings.neurips.cc/paper\\_files/paper/2024/file/d529b943af3dba734f8a7d49efcb6d09-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2024/file/d529b943af3dba734f8a7d49efcb6d09-Paper-Conference.pdf). Also available as arXiv preprint arXiv:2405.12952.
- Sham Kakade and John Langford. Approximately Optimal Approximate Reinforcement Learning. In Claude Sammut and Achim Hoffman, editors, *Proceedings of the Nineteenth International Conference on Machine Learning (ICML 2002)*, pages 267–274, San Francisco, CA, USA, 2002. Morgan Kauffman. ISBN 1-55860-873-7. URL <http://ttic.uchicago.edu/~sham/papers/r1/aoarl.pdf>.
- Seung-Jun Kim and Geogios B. Giannakis. An Online Convex Optimization Approach to Real-time Energy Pricing for Demand Response. *IEEE Transactions on Smart Grid*, 8(6):2784–2793, 2017. doi: 10.1109/TSG.2016.2539948. URL <https://ieeexplore.ieee.org/document/7438918>.
- B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep Reinforcement Learning for Autonomous Driving: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2022. doi: 10.1109/TITS.2021.3054625. URL <https://ieeexplore.ieee.org/document/9351818>.
- Robert Kleinberg. A multiple-choice secretary algorithm with applications to online auctions. In *Proceedings of the Sixteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA ’05, page 630–631, USA, 2005. Society for Industrial and Applied Mathematics. ISBN 0898715857. URL <https://dl.acm.org/doi/10.5555/1070432.1070519>.
- Jens Kober, J. Andrew Bagnell, and Jan Peters. Reinforcement Learning in Robotics: A Survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013. doi: 10.1177/0278364913495721. URL <https://doi.org/10.1177/0278364913495721>.
- Kasper Green Larsen, Omar Montasser, and Nikita Zhivotovskiy. Derandomizing multi-distribution learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL <https://openreview.net/forum?id=twYE75Mnkt>.
- Jean-Michel Lasry and Pierre-Louis Lions. Mean Field Games. *Japanese Journal of Mathematics*, 2(1):229–260, March 2007. ISSN 1861-3624. doi: 10.1007/s11537-007-0657-8. URL <https://link.springer.com/article/10.1007/s11537-007-0657-8>.
- Tor Lattimore, Csaba Szepesvari, and Gellert Weisz. Learning with good feature representations in bandits and in RL with a generative model. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5662–5670. PMLR, 13–18 Jul 2020. URL <https://proceedings.mlr.press/v119/lattimore20a.html>.
- Martin Lauer and Martin A. Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML ’00, page 535–542, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc. ISBN 1558607072. URL <https://dl.acm.org/doi/10.5555/645529.658113>.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample Complexity of Asynchronous Q-Learning: Sharper Analysis and Variance Reduction. *IEEE Transactions on Information Theory*, 68(1):448–473, 2022. doi: 10.1109/TIT.2021.3120096. URL <https://arxiv.org/abs/2006.03041>.

Minne Li, Zhiwei Qin, Yan Jiao, Yaodong Yang, Jun Wang, Chenxi Wang, Guobin Wu, and Jieping Ye. Efficient ridesharing order dispatching with mean field multi-agent reinforcement learning. In *The World Wide Web Conference, WWW ’19*, page 983–994, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366748. doi: 10.1145/3308558.3313433. URL <https://doi.org/10.1145/3308558.3313433>.

Yiheng Lin, Guannan Qu, Longbo Huang, and Adam Wierman. Distributed Reinforcement Learning in Multi-Agent Networked Systems. *CoRR*, abs/2006.06555, 2020. URL <https://arxiv.org/abs/2006.06555>.

Yiheng Lin, James A Preiss, Emile Timothy Anand, Yingying Li, Yisong Yue, and Adam Wierman. Online Adaptive Policy Selection in Time-Varying Systems: No-Regret via Contractive Perturbations. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023a. URL <https://openreview.net/forum?id=hDajsofjRM>.

Yiheng Lin, James A Preiss, Emile Timothy Anand, Yingying Li, Yisong Yue, and Adam Wierman. Learning-augmented Control via Online Adaptive Policy Selection: No Regret via contractive Perturbations. In *ACM SIGMETRICS, Workshop on Learning-augmented Algorithms: Theory and Applications 2023*, 2023b. URL <https://learning-augmented-algorithms.github.io/papers/sigmetrics23-lata-posters-paper5.pdf>.

Yiheng Lin, James A. Preiss, Fengze Xie, Emile Anand, Soon-Jo Chung, Yisong Yue, and Adam Wierman. Online policy optimization in unknown nonlinear systems. In Shipra Agrawal and Aaron Roth, editors, *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pages 3475–3522. PMLR, 30 Jun–03 Jul 2024. URL <https://proceedings.mlr.press/v247/lin24a.html>.

Michael L. Littman. Markov Games as a Framework for Multi-Agent Reinforcement Learning. In *Machine learning proceedings*, Elsevier, page 157–163, 1994. URL <https://www.sciencedirect.com/science/article/abs/pii/B9781558603356500271>.

P. Massart. The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *The Annals of Probability*, 18(3):1269 – 1283, 1990. doi: 10.1214/aop/1176990746. URL <https://projecteuclid.org/journals/annals-of-probability/volume-18/issue-3/The-Tight-Constant-in-the-Dvoretzky-Kiefer-Wolfowitz-Inequality/10.1214/aop/1176990746.full>.

Yifei Min, Jiafan He, Tianhao Wang, and Quanquan Gu. Cooperative Multi-Agent Reinforcement Learning: Asynchronous Communication and Linear Function Approximation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 24785–24811. PMLR, 23–29 Jul 2023. URL <https://proceedings.mlr.press/v202/min23a.html>.

Washim Uddin Mondal, Mridul Agarwal, Vaneet Aggarwal, and Satish V. Ukkusuri. On the Approximation of Cooperative Heterogeneous Multi-Agent Reinforcement Learning (MARL) Using Mean Field Control (MFC). *Journal of Machine Learning Research*, 23(1), jan 2022. ISSN 1532-4435. URL <https://jmlr.org/papers/volume23/21-1312/21-1312.pdf>.

Michael Naaman. On the Tight Constant in the Multivariate Dvoretzky–Kiefer–Wolfowitz Inequality. *Statistics & Probability Letters*, 173:109088, 2021. ISSN 0167-7152. doi: <https://doi.org/10.1016/j.spl.2021.109088>. URL <https://www.sciencedirect.com/science/article/pii/S016771522100050X>.

Christos H. Papadimitriou and John N. Tsitsiklis. The Complexity of Optimal Queuing Network Control. *Mathematics of Operations Research*, 24(2):293–305, 1999. ISSN 0364765X, 15265471. URL <https://www.jstor.org/stable/3690486>.

James A. Preiss, Wolfgang Honig, Gaurav S. Sukhatme, and Nora Ayanian. Crazyswarm: A large nano-quadcopter swarm. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3299–3304, 2017. doi: 10.1109/ICRA.2017.7989376. URL <https://ieeexplore.ieee.org/document/7989376/>.

Guannan Qu, Yiheng Lin, Adam Wierman, and Na Li. Scalable Multi-Agent Reinforcement Learning for Networked Systems with Average Reward. In *Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS’20*, Red Hook, NY, USA, 2020a. Curran Associates Inc. ISBN 9781713829546. URL <https://proceedings.neurips.cc/paper/2020/file/168efc366c449fab9c2843e9b54e2a18-Paper.pdf>.

Guannan Qu, Adam Wierman, and Na Li. Scalable Reinforcement Learning of Localized Policies for Multi-Agent Networked Systems. In Alexandre M. Bayen, Ali Jadbabaie, George Pappas, Pablo A. Parrilo, Benjamin Recht, Claire Tomlin, and Melanie Zeilinger, editors, *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 256–266. PMLR, 10–11 Jun 2020b. URL <https://proceedings.mlr.press/v120/qu20a.html>.

Omer Reingold. Undirected Connectivity in Log-Space. *J. ACM*, 55(4), sep 2008. ISSN 0004-5411. doi: 10.1145/1391289.1391291. URL <https://doi.org/10.1145/1391289.1391291>.

Zhaolin Ren, Runyu Zhang, Bo Dai, and Na Li. Scalable spectral representations for network multiagent control, 2024. URL <https://arxiv.org/abs/2410.17221>.

Ross D. Shachter. Bayes-ball: The rational pastime (for determining irrelevance and requisite information in belief networks and influence diagrams), 2013.

Lloyd S. Shapley. A value for n-person games. In *Contributions to the Theory of Games*, 1953. URL <https://www.rand.org/content/dam/rand/pubs/papers/2021/P295.pdf>.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin F. Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving discounted markov decision process with a generative model. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 7739–7750, 2018a.

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. 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)*, New Orleans, LA, USA, 2018b. Society for Industrial and Applied Mathematics. doi: 10.1137/1.9781611975031.116.

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. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature*, 529(7587):484–489, January 2016. ISSN 1476-4687. doi: 10.1038/nature16961. URL <https://www.nature.com/articles/nature16961>.

Sriram Ganapathi Subramanian, Matthew E. Taylor, Mark Crowley, and Pascal Poupart. Decentralized mean field games, 2022. URL <https://arxiv.org/abs/2112.09099>.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL [https://proceedings.neurips.cc/paper\\_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf).

Volodymyr Tkachuk, Seyed Alireza Bakhtiari, Johannes Kirschner, Matej Jusup, Ilija Bogunovic, and Csaba Szepesvári. Efficient planning in combinatorial action spaces with applications to cooperative multi-agent reinforcement learning. *arXiv*, 2023. URL <https://proceedings.mlr.press/v206/tkachuk23a/tkachuk23a.pdf>.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519. URL <https://link.springer.com/book/10.1007/b13794>.

- Martin J Wainwright. Variance-reduced q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.
- Jiangxing Wang, Deheng Ye, and Zongqing Lu. More centralized training, still decentralized execution: Multi-agent conditional policy factorization, 2023. URL <https://arxiv.org/abs/2209.12681>.
- Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3):279–292, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992698. URL <https://link.springer.com/article/10.1007/BF00992698>.
- Mengfan Xu and Diego Klabjan. Decentralized randomly distributed multi-agent multi-armed bandit with heterogeneous rewards. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL <https://openreview.net/forum?id=DqfdhM64LI>.
- Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean Field Multi-Agent Reinforcement Learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5571–5580. PMLR, 10–15 Jul 2018. URL <https://proceedings.mlr.press/v80/yang18d.html>.
- Hongming Zhang, Tongzheng Ren, Chenjun Xiao, Dale Schuurmans, and Bo Dai. Provable representation with efficient planning for partially observable reinforcement learning. In *Proceedings of the 41st International Conference on Machine Learning*, ICML’24. JMLR.org, 2024. URL <https://dl.acm.org/doi/10.5555/3692070.3694540>.
- Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms, 2021. URL <https://arxiv.org/abs/1911.10635>.
- Yihe Zhou, Shunyu Liu, Yunpeng Qing, Kaixuan Chen, Tongya Zheng, Yanhao Huang, Jie Song, and Mingli Song. Is centralized training with decentralized execution framework centralized enough for marl?, 2023. URL <https://arxiv.org/abs/2305.17352>.

## A Mathematical Background and Additional Remarks

### Outline of the Appendices.

- Section B presents numerical simulations on the performance of SUBSAMPLE-MFQ on the Gaussian squeeze and constrained exploration tasks.
- Section C presents notation and basic lemmas involving the learned  $\hat{Q}_k$ -function, as well as a stable and practical implementation of Algorithm 1.
- Section D presents a proof sketch of our main result in Theorem 4.4.
- Section E presents the proof of the Lipschitz continuity between  $\hat{Q}_k$  and  $Q^*$ .
- Section F presents our bound on the TV-distance in Theorem D.2.
- Section G proves the bound on the optimality gap between the learned policy  $\tilde{\pi}_{k,m}^{\text{est}}$  and the optimal policy  $\pi^*$ .
- Section H presents an extension of the result to stochastic rewards.
- Section I presents an extension of the result to off-policy learning that follows by using historical data.
- Section J presents an extension of the result to continuous state/action spaces under a linear MDP assumption.
- Section K presents some practical derandomized variants with randomness-sharing.

Table 1: Important notations in this paper.

Notation	Meaning
$\ \cdot\ _1$	$\ell_1$ (Manhattan) norm;
$\ \cdot\ _\infty$	$\ell_\infty$ norm;
$\mathbb{Z}_+$	The set of strictly positive integers;
$\mathbb{R}^d$	The set of $d$ -dimensional reals;
$[m]$	The set $\{1, \dots, m\}$ , where $m \in \mathbb{Z}_+$ ;
$\binom{[m]}{k}$	The set of $k$ -sized subsets of $\{1, \dots, m\}$ ;
$a_g$	$a_g \in \mathcal{A}_g$ is the action of the global agent;
$s_g$	$s_g \in \mathcal{S}_g$ is the state of the global agent;
$a_1, \dots, a_n$	$a_1, \dots, a_n \in \mathcal{A}_l^n$ are the actions of the local agents $1, \dots, n$ ;
$s_1, \dots, s_n$	$s_1, \dots, s_n \in \mathcal{S}_l^n$ are the states of the local agents $1, \dots, n$ ;
$a$	$a = (a_g, a_1, \dots, a_n) \in \mathcal{A}_g \times \mathcal{A}_l^n$ is the tuple of actions of all agents;
$s$	$s = (s_g, s_1, \dots, s_n) \in \mathcal{S}_g \times \mathcal{S}_l^n$ is the tuple of states of all agents;
$z_i$	$z_i = (s_i, a_i) \in \mathcal{Z}_l$ , for $i \in [n]$ ;
$\mu_k(\mathcal{Z}_l)$	$\mu_k(\mathcal{Z}_l) = \{0, 1/k, 2/k, \dots, 1\}^{ \mathcal{Z}_l }$ ;
$\mu(\mathcal{Z}_l)$	$\mu(\mathcal{Z}_l) := \mu_n(\mathcal{Z}_l) \{0, 1/n, 2/n, \dots, 1\}^{ \mathcal{Z}_l }$ ;
$s_\Delta$	For $\Delta \subseteq [n]$ , and a collection of variables $\{s_1, \dots, s_n\}$ , $s_\Delta := \{s_i : i \in \Delta\}$ ;
$\sigma(z_\Delta, z'_\Delta)$	Product sigma-algebra generated by sequences $z_\Delta$ and $z'_\Delta$ ;
$\pi^*$	$\pi^*$ is the optimal deterministic policy function such that $a = \pi^*(s)$ ;
$\hat{\pi}_k^*$	$\hat{\pi}_k^*$ is the optimal deterministic policy function on a constrained system of $ \Delta  = k$ local agents;
$\tilde{\pi}_k^{\text{est}}$	$\tilde{\pi}_k^{\text{est}}$ is the stochastic policy map learned with parameter $k$ such that $a \sim \tilde{\pi}_k^{\text{est}}(s)$ ;
$P_g(\cdot   s_g, a_g)$	$P_g(\cdot   s_g, a_g)$ is the stochastic transition kernel for the state of the global agent;
$P_l(\cdot   a_i, s_i, s_g)$	$P_l(\cdot   a_i, s_i, s_g)$ is the stochastic transition kernel for the state of local agent $i \in [n]$ ;
$r_g(s_g, a_g)$	$r_g$ is the global agent's component of the reward;
$r_l(s_i, s_g, a_i)$	$r_l$ is the component of the reward for local agent $i \in [n]$ ;
$r(s, a)$	$r(s, a) = r_g(s_g, a_g) + \frac{1}{n} \sum_{i \in [n]} r_l(s_i, s_g, a_i)$ is the reward of the system;
$r_\Delta(s, a)$	$r_\Delta(s, a) = r_g(s_g, a_g) + \frac{1}{ \Delta } \sum_{i \in \Delta} r_l(s_i, s_g, a_i)$ is the constrained system's reward with $ \Delta  = k$ local agents;
$\mathcal{T}$	$\mathcal{T}$ is the centralized Bellman operator;
$\hat{\mathcal{T}}_k$	$\hat{\mathcal{T}}_k$ is the Bellman operator on a constrained system of $ \Delta  = k$ local agents;
$\Pi^\Theta(y)$	$\ell_1$ projection of $y$ onto set $\Theta$ ;

**Definition A.1** (Lipschitz continuity). Given metric spaces  $(\mathcal{X}, d_{\mathcal{X}})$  and  $(\mathcal{Y}, d_{\mathcal{Y}})$  and a constant  $L > 0$ , a map  $f : \mathcal{X} \rightarrow \mathcal{Y}$  is  $L$ -Lipschitz continuous if for all  $x, y \in \mathcal{X}$ ,  $d_{\mathcal{Y}}(f(x), f(y)) \leq L \cdot d_{\mathcal{X}}(x, y)$ .

**Theorem A.2** (Banach-Caccioppoli fixed point theorem [Banach, 1922]). Consider the metric space  $(\mathcal{X}, d_{\mathcal{X}})$ , and  $T : \mathcal{X} \rightarrow \mathcal{X}$  such that  $T$  is a  $\gamma$ -Lipschitz continuous mapping for  $\gamma \in (0, 1)$ . Then, by the Banach-Caccioppoli fixed-point theorem, there exists a unique fixed point  $x^* \in \mathcal{X}$  for which  $T(x^*) = x^*$ . Additionally,  $x^* = \lim_{s \rightarrow \infty} T^s(x_0)$  for any  $x_0 \in \mathcal{X}$ .

## B Numerical Experiments

This section provides numerical simulations for the examples outlined in Section 2. All experiments were run on a 2-core CPU server with 12GB RAM. We chose a parameter complexity for each simulation that was sufficient to emphasize characteristics of the theory, such as the complexity improvement and the decaying optimality gap.<sup>4</sup>

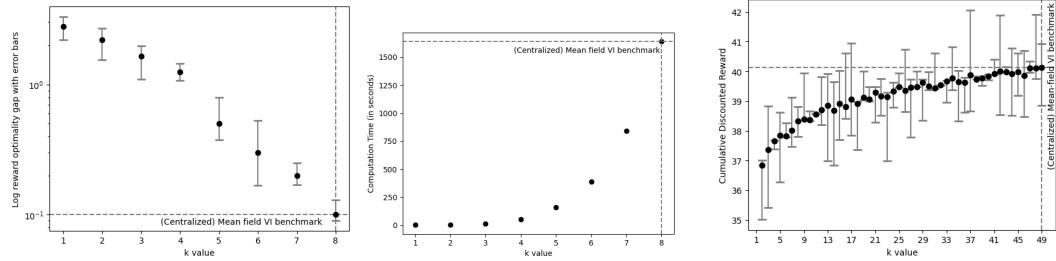


Figure 2: a) Reward optimality gap (log scale) with  $\pi_{k,m}^{\text{est}}$  running 300 iterations. b) Computation time (in minutes) against sampling parameter  $k$ , for  $k \leq n = 8$ , to learn policy  $\hat{\pi}_{k,m}^{\text{est}}$ . c) Discounted cumulative rewards for  $k \leq n = 50$ .

### B.1 Constrained Exploration

Let  $\mathcal{S}_g = \mathcal{S}_l = [6]^2$  where for  $s_a \in \{\mathcal{S}_g, \mathcal{S}_l\}$ ,  $s_a^{(i)}$  denotes the projection of  $s_a$  onto its  $i$ 'th coordinate. Further, let  $\mathcal{A}_l = \mathcal{A}_g = \{\text{"up"}, \text{"down"}, \text{"right"}, \text{"left"}\}$  given formally by  $\{(1), (-1), (0), (0)\}$ . Let  $\Pi_D(x)$  denote the  $\ell_1$ -projection of  $x$  onto set  $D$ . Then, let  $s_g^{t+1}(s_g^t, a_g^t) = \Pi_{\mathcal{S}_g}(s_g^t + a_g^t)$ ,  $s_i^{t+1}(s_i^t, s_g^t, a_i^t) = \Pi_{\mathcal{S}_l}(s_i^t + |s_i^t - s_g^t| + a_i^t)$ . We let the global agent's reward be  $r_g(s_g, a_g) = 2 \sum_{i=1}^3 \mathbb{1}\{a_g = (1), s_g^{(1)} \leq i\} + 2 \sum_{i=3}^6 \mathbb{1}\{a_g = (-1), s_g^{(1)} > i\}$ , and the local agent's rewards be  $r_l(s_i, s_g, a_i) = \mathbb{1}\{a_i^t \neq 0\} + 6 \cdot \mathbb{1}\{s_i^{(1)} = s_g^{(1)}\} + 2 \cdot \|s_i^t - s_g^t\|_1$ .

Intuitively, the reward function is designed to force the global agent to oscillate vertically in the grid, while forcing the local agents to oscillate horizontally around the global agent, thereby simulating a constrained exploration. This model has been studied previously in Lin et al. [2023b].

For this task, we ran a simulation with  $n = 8$  agents, with  $m = 20$  samples in the empirically adapted Bellman operator. We provide simulation results in Figure 2a. We observe monotonic improvements in the cumulative discounted rewards as  $k \rightarrow n$ . Since  $k = n$  recovers value-iteration and mean-field MARL algorithms, the reward at  $k = n$  is the baseline we compare our algorithm to. When  $k < n$ , we observe that the reward accrued by SUBSAMPLE-MFQ is only marginally less than the reward gained by value-iteration.

### B.2 Gaussian Squeeze (GS)

In this task,  $n$  homogeneous agents determine their individual action  $a_i$  to jointly maximize the objective  $r(x) = xe^{-(x-\mu)^2/\sigma^2}$ , where  $x = \sum_{i=1}^n a_i$ ,  $a_i \in \{0, \dots, 9\}$ , and  $\mu$  and  $\sigma$  are the pre-defined mean and variance of the system. We consider a homogeneous system devoid of a global

<sup>4</sup>We provide supporting code for the algorithm and experiments in <https://github.com/emiletimothy/Mean-Field-Subsample-Q-Learning>

agent to match the mean-field setting in Yang et al. [2018], where the richness of our model setting can still express GS. We use the global agent to model the state of the system, given by the number of vehicles in each controller’s lane.

We set  $\mathcal{S}_l = [4]$ ,  $\mathcal{A}_l = \{0, \dots, 4\}$ , and  $\mathcal{S}_g = \mathcal{A}_l$  such that  $s_g = \lceil \frac{1}{n} \sum_{i=1}^n s_i \rceil$ , and  $\mathcal{A}_g = \{0\}$ . The transition functions are given by  $s_i(t+1) = s_i(t) - 1 \cdot \mathbb{1}\{s_i(t) > s_g(t)\} + \text{Ber}(p)$  and  $s_g(t+1) = \lceil \frac{1}{n} \sum_{i=1}^n s_i(t+1) \rceil$ , where  $\text{Ber}(p)$  is a Bernoulli random variable with parameter  $p > 0$ . Finally, the global agent’s reward function is given by  $r_g(s_g, a_g) = -s_g$  and the local agent’s reward function is given by  $r_l(s_i, s_g, a_i) = s_i \cdot e^{-(s_i - s_g)^2/4} - \max\{s_i, a_i\}$ .

For this task, we ran a small-scale simulation with  $n = 8$  agents, and a large-scale simulation with  $n = 50$  agents, and used  $m = 20$  samples in the empirical Bellman operator. We provide simulation results in Figure 2b and Figure 2c, where Figure 2b demonstrates the exponential improvement in computational complexity of SUBSAMPLE-MFQ, and Figure 2c demonstrates a monotonic improvement in the cumulative rewards, consistent with Theorem 4.4. Here, both metrics outperform the mean-field value iteration benchmark.

## C Notation and Basic Lemmas

For convenience, we restate below the various Bellman operators under consideration.

**Definition C.1** (Bellman Operator  $\mathcal{T}$ ).

$$\mathcal{T}Q^t(s, a) := r(s, a) + \gamma \mathbb{E}_{\substack{s'_g \sim P_g(\cdot | s_g, a_g), \\ s'_i \sim P_l(\cdot | s_i, s_g, a_i), \forall i \in [n]}} \max_{a' \in \mathcal{A}} Q^t(s', a') \quad (15)$$

**Definition C.2** (Adapted Bellman Operator  $\hat{\mathcal{T}}_k$ ). The adapted Bellman operator updates a smaller  $Q$  function (which we denote by  $\hat{Q}_k$ ), for a surrogate system with the global agent and  $k \in [n]$  local agents denoted by  $\Delta$ , using mean-field value iteration and  $j \in \Delta$  such that:

$$\hat{\mathcal{T}}_k \hat{Q}_k^t(s_g, s_j, F_{z_{\Delta \setminus j}}, a_j, a_g) := r_\Delta(s, a) + \gamma \mathbb{E}_{\substack{s'_g \sim P_g(\cdot | s_g, a_g), \\ s'_i \sim P_l(\cdot | s_i, s_g, a_i), \forall i \in \Delta}} \max_{a' \in \mathcal{A}} \hat{Q}_k^t(s'_g, s'_j, F_{z'_{\Delta \setminus j}}, a'_j, a'_g) \quad (16)$$

**Definition C.3** (Empirical Adapted Bellman Operator  $\hat{\mathcal{T}}_{k,m}$ ). The empirical adapted Bellman operator  $\hat{\mathcal{T}}_{k,m}$  empirically estimates the adapted Bellman operator update using mean-field value iteration by drawing  $m$  random samples of  $s_g \sim P_g(\cdot | s_g, a_g)$  and  $s_i \sim P_l(\cdot | s_i, s_g, a_i)$  for  $i \in \Delta$ , where for  $\ell \in [m]$ , the  $\ell$ ’th random sample is given by  $s_g^\ell$  and  $s_\Delta^\ell$ , and  $j \in \Delta$ :

$$\hat{\mathcal{T}}_{k,m} \hat{Q}_{k,m}^t(s_g, s_j, F_{z_{\Delta \setminus j}}, a_j, a_g) := r_\Delta(s, a) + \frac{\gamma}{m} \sum_{\ell \in [m]} \max_{a' \in \mathcal{A}} \hat{Q}_{k,m}^t(s_g^\ell, s_j^\ell, F_{z_\Delta^\ell}, a_j^\ell, a_g^\ell) \quad (17)$$

**Lemma C.4.** For any  $\Delta \subseteq [n]$  such that  $|\Delta| = k$ , suppose  $0 \leq r_\Delta(s, a) \leq \tilde{r}$ . Then, for all  $t \in \mathbb{N}$ ,  $\hat{Q}_k^t \leq \frac{\tilde{r}}{1-\gamma}$ .

*Proof.* The proof follows by induction on  $t$ . The base case follows from  $\hat{Q}_k^0 := 0$ . For the induction, note that by the triangle inequality  $\|\hat{Q}_k^{t+1}\|_\infty \leq \|r_\Delta\|_\infty + \gamma \|\hat{Q}_k^t\|_\infty \leq \tilde{r} + \gamma \frac{\tilde{r}}{1-\gamma} = \frac{\tilde{r}}{1-\gamma}$ .  $\square$

**Remark C.5.** By the law of large numbers,  $\lim_{m \rightarrow \infty} \hat{\mathcal{T}}_{k,m} = \hat{\mathcal{T}}_k$ , where the error decays in  $O(1/\sqrt{m})$  by the Chernoff bound. Also,  $\hat{\mathcal{T}}_n := \mathcal{T}$ . Further, Theorem C.4 is independent of the choice of  $k$ . Therefore, for  $k = n$ , this implies an identical bound on  $Q^t$ . An identical argument implies the same bound on  $\hat{Q}_{k,m}^t$ .

$\mathcal{T}$  satisfies a  $\gamma$ -contractive property under the infinity norm [Watkins and Dayan, 1992]. We similarly show that  $\hat{\mathcal{T}}_k$  and  $\hat{\mathcal{T}}_{k,m}$  satisfy a  $\gamma$ -contractive property under infinity norm in Theorems C.6 and C.7.

**Lemma C.6.**  $\hat{\mathcal{T}}_k$  satisfies the  $\gamma$ -contractive property under infinity norm:

$$\|\hat{\mathcal{T}}_k \hat{Q}'_k - \hat{\mathcal{T}}_k \hat{Q}_k\|_\infty \leq \gamma \|\hat{Q}'_k - \hat{Q}_k\|_\infty \quad (18)$$

*Proof.* Suppose we apply  $\hat{\mathcal{T}}_k$  to  $\hat{Q}_k(s_g, F_{z_\Delta}, a_g)$  and  $\hat{Q}'_k(s_g, F_{z_\Delta}, a_g)$  for  $|\Delta| = k$ . Then:

$$\begin{aligned} & \|\hat{\mathcal{T}}_k \hat{Q}'_k - \hat{\mathcal{T}}_k \hat{Q}_k\|_\infty \\ &= \gamma \max_{\substack{s_g \in \mathcal{S}_g, \\ a_g \in \mathcal{A}_g, \\ F_{z_\Delta} \in \mu_k(\mathcal{Z}_l)}} \left| \mathbb{E}_{\substack{s'_g \sim P_g(\cdot | s_g, a_g), \\ s'_i \sim P_l(\cdot | s_i, s_g, a_i), \\ \forall i \in \Delta}} \max_{a' \in \mathcal{A}} \hat{Q}'_k(s'_g, F_{z'_\Delta}, a'_g) - \mathbb{E}_{\substack{s'_g \sim P_g(\cdot | s_g, a_g), \\ s'_i \sim P_l(\cdot | s_i, s_g, a_i), \\ \forall i \in \Delta'}} \max_{a' \in \mathcal{A}} \hat{Q}_k(s'_g, F_{z'_\Delta}, a'_g) \right| \\ &\leq \gamma \max_{\substack{s'_g \in \mathcal{S}_g, \\ F_{z_\Delta} \in \mu_k(\mathcal{Z}_l), \\ a' \in \mathcal{A}}} |\hat{Q}'_k(s'_g, F_{z'_\Delta}, a'_g) - \hat{Q}_k(s'_g, F_{z'_\Delta}, a'_g)| = \gamma \|\hat{Q}'_k - \hat{Q}_k\|_\infty \end{aligned}$$

The equality cancels common  $r_\Delta(s, a)$  terms in each operator. The second line uses Jensen's inequality, maximizes over actions, and bounds expected values with the maximizers of the random variables.  $\square$

**Lemma C.7.**  $\hat{\mathcal{T}}_{k,m}$  satisfies the  $\gamma$ -contractive property under infinity norm.

*Proof.* Similarly to Theorem C.6, suppose we apply  $\hat{\mathcal{T}}_{k,m}$  to  $\hat{Q}_{k,m}(s_g, F_{z_\Delta}, a_g)$  and  $\hat{Q}'_{k,m}(s_g, F_{z_\Delta}, a_g)$ . Then:

$$\begin{aligned} \|\hat{\mathcal{T}}_{k,m} \hat{Q}_k - \hat{\mathcal{T}}_{k,m} \hat{Q}'_k\|_\infty &= \frac{\gamma}{m} \left\| \sum_{\ell \in [m]} \left( \max_{a' \in \mathcal{A}} \hat{Q}_k(s_g^\ell, F_{z_\Delta^\ell}, a'_g) - \max_{a' \in \mathcal{A}} \hat{Q}'_k(s_g^\ell, F_{z_\Delta^\ell}, a'_g) \right) \right\|_\infty \\ &\leq \gamma \max_{a'_g \in \mathcal{A}_g, s'_g \in \mathcal{S}_g, z_\Delta \in \mathcal{Z}_l^k} |\hat{Q}_k(s'_g, F_{z'_\Delta}, a'_g) - \hat{Q}'_k(s'_g, F_{z'_\Delta}, a'_g)| \\ &= \gamma \|\hat{Q}_k - \hat{Q}'_k\|_\infty \end{aligned}$$

The first inequality uses the triangle inequality and the general property  $|\max_{a \in A} f(a) - \max_{b \in B} f(b)| \leq \max_{c \in A} |f(a) - f(b)|$ . The last line recovers the definition of infinity norm.  $\square$

**Remark C.8.** The  $\gamma$ -contractivity of  $\hat{\mathcal{T}}_k$  and  $\hat{\mathcal{T}}_{k,m}$  attracts the trajectory between two  $\hat{Q}_k$  and  $\hat{Q}_{k,m}$  functions on the same state-action tuple by  $\gamma$  at each step. Repeatedly applying the Bellman operators produces a unique fixed-point from the Banach fixed-point theorem which we introduce in Theorems C.9 and C.10.

**Definition C.9** ( $\hat{Q}_k^*$ -function). Suppose  $\hat{Q}_k^0 := 0$  and let  $\hat{Q}_k^{t+1}(s_g, F_{z_\Delta}, a_g) = \hat{\mathcal{T}}_k \hat{Q}_k^t(s_g, F_{z_\Delta}, a_g)$  for  $t \in \mathbb{N}$ . Denote the fixed-point of  $\hat{\mathcal{T}}_k$  by  $\hat{Q}_k^*$  such that  $\hat{\mathcal{T}}_k \hat{Q}_k^*(s_g, F_{z_\Delta}, a_g) = \hat{Q}_k^*(s_g, F_{z_\Delta}, a_g)$ .

**Definition C.10** ( $\hat{Q}_{k,m}^*$ -function). Let  $\hat{Q}_{k,m}^0 := 0$  and  $\hat{Q}_{k,m}^{t+1}(s_g, F_{z_\Delta}, a_g) = \hat{\mathcal{T}}_{k,m} \hat{Q}_{k,m}^t(s_g, F_{z_\Delta}, a_g)$  for  $t \in \mathbb{N}$ . Then, the Banach-Cacciopoli fixed-point of the adapted Bellman operator  $\hat{\mathcal{T}}_{k,m}$  is  $\hat{Q}_{k,m}^*$  such that  $\hat{\mathcal{T}}_{k,m} \hat{Q}_{k,m}^*(s_g, F_{z_\Delta}, a_g) = \hat{Q}_{k,m}^*(s_g, F_{z_\Delta}, a_g)$ .

**Corollary C.11.** Observe that by recursively using the  $\gamma$ -contractive property for  $T$  time steps:

$$\|\hat{Q}_k^* - \hat{Q}_k^T\|_\infty \leq \gamma^T \cdot \|\hat{Q}_k^* - \hat{Q}_k^0\|_\infty \quad (19)$$

$$\|\hat{Q}_{k,m}^* - \hat{Q}_{k,m}^T\|_\infty \leq \gamma^T \cdot \|\hat{Q}_{k,m}^* - \hat{Q}_{k,m}^0\|_\infty \quad (20)$$

Further, noting that  $\hat{Q}_k^0 = \hat{Q}_{k,m}^0 := 0$ ,  $\|\hat{Q}_k^*\|_\infty \leq \frac{\tilde{r}}{1-\gamma}$ , and  $\|\hat{Q}_{k,m}^*\|_\infty \leq \frac{\tilde{r}}{1-\gamma}$  from Theorem C.4:

$$\|\hat{Q}_k^* - \hat{Q}_k^T\|_\infty \leq \gamma^T \frac{\tilde{r}}{1-\gamma}, \quad (21)$$

$$\|\hat{Q}_{k,m}^* - \hat{Q}_{k,m}^T\|_\infty \leq \gamma^T \frac{\tilde{r}}{1-\gamma} \quad (22)$$

**Remark C.12.** Theorem C.11 characterizes the error decay between  $\hat{Q}_k^T$  and  $\hat{Q}_k^*$  and shows that it decays exponentially in the number of Bellman iterations by a  $\gamma^T$  multiplicative factor.

Furthermore, we characterize the maximal policies greedy policies obtained from  $Q^*$ ,  $\hat{Q}_k^*$ , and  $\hat{Q}_{k,m}^*$ .

**Definition C.13** (Optimal policy  $\pi^*$ ). The greedy policy derived from  $Q^*$  is

$$\pi^*(s) := \arg \max_{a \in \mathcal{A}} Q^*(s, a). \quad (23)$$

**Definition C.14** (Optimal subsampled policy  $\hat{\pi}_k^*$ ). The greedy policy from  $\hat{Q}_k^*$  is

$$\hat{\pi}_k^*(s_g, s_i, F_{s_{\Delta \setminus i}}) := \arg \max_{(a_g, a_i, F_{a_{\Delta \setminus i}}) \in \mathcal{A}_g \times \mathcal{A}_i \times \mu_{k-1}(\mathcal{A}_l)} \hat{Q}_k^*(s_g, s_i, F_{z_{\Delta \setminus i}}, a_i, a_g). \quad (24)$$

**Definition C.15** (Optimal empirically subsampled policy  $\hat{\pi}_{k,m}^{\text{est}}$ ). The greedy policy from  $\hat{Q}_{k,m}^{\text{est}}$  is given by

$$\hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_{\Delta}}) := \arg \max_{(a_g, a_i, F_{a_{\Delta \setminus i}}) \in \mathcal{A}_g \times \mathcal{A}_i \times \mu_{k-1}(\mathcal{A}_l)} \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{z_{\Delta \setminus i}}, a_i, a_g). \quad (25)$$

Figure 4 details the analytic flow on how we use the empirical adapted Bellman operator to perform value iteration on  $\hat{Q}_{k,m}$  to get  $\hat{Q}_{k,m}^{\text{est}}$  which approximates  $Q^*$ .

Algorithm 3 gives a stable implementation of Algorithm 1 with learning rates  $\{\eta_t\}_{t \in [T]}$ . Algorithm 3 is provably numerical stable under fixed-point arithmetic [Anand et al., 2024, 2025].  $\hat{Q}_{k,m}^t$  in Algorithm 3 is  $\gamma$ -contractive as in Theorem C.6, given an appropriately conditioned sequence of learning rates  $\eta_t$ :

---

**Algorithm 3** Stable (Practical) Implementation of Algorithm 1: SUBSAMPLE-Q: Learning

---

**Require:** A multi-agent system as described in Section 2. Parameter  $T$  for the number of iterations in the initial value iteration step. Hyperparameter  $k \in [n]$ . Discount parameter  $\gamma \in (0, 1)$ . Oracle  $\mathcal{O}$  to sample  $s'_g \sim P_g(\cdot | s_g, a_g)$  and  $s_i \sim P_l(\cdot | s_i, s_g, a_i)$  for all  $i \in [n]$ . Learning rate sequence  $\{\eta_t\}_{t \in [T]}$  where  $\eta_t \in (0, 1]$ .

- 1: Let  $\Delta = [k]$ .
  - 2: **for**  $(s_g, s_1, F_{z_{\Delta \setminus 1}}) \in \mathcal{S}_g \times \mathcal{S}_l \times \mu_{k-1}(\mathcal{Z}_l)$  **do**
  - 3:     **for**  $(a_g, a_1) \in \mathcal{A}_g \times \mathcal{A}_l$  **do**
  - 4:         Set  $\hat{Q}_{k,m}^0(s_g, s_1, F_{z_{\Delta \setminus 1}}, a_1, a_g) = 0$
  - 5:     **for**  $t = 1$  to  $T$  **do**
  - 6:         **for**  $(s_g, s_1, F_{z_{\Delta \setminus 1}}) \in \mathcal{S}_g \times \mathcal{S}_l \times \mu_{k-1}(\mathcal{Z}_l)$  **do**
  - 7:             **for**  $(a_g, a_1) \in \mathcal{A}_g \times \mathcal{A}_l$  **do**
  - 8:
- $$\begin{aligned} \hat{Q}_{k,m}^{t+1}(s_g, s_1, F_{z_{\Delta \setminus 1}}, a_g, a_1) &= (1 - \eta_t) \hat{Q}_{k,m}^t(s_g, s_1, F_{z_{\Delta \setminus 1}}, a_g, a_1) \\ &\quad + \eta_t \hat{\mathcal{T}}_{k,m} \hat{Q}_{k,m}^t(s_g, s_1, F_{z_{\Delta \setminus 1}}, a_g, a_1) \end{aligned}$$

- 9: Let the approximate policy be

$$\hat{\pi}_{k,m}^T(s_g, s_1, F_{s_{\Delta \setminus 1}}) = \arg \max_{(a_g, a_1, a_{\Delta \setminus 1}) \in \mathcal{A}_g \times \mathcal{A}_l \times \mu_{k-1}(\mathcal{A}_l)} \hat{Q}_{k,m}^T(s_g, s_1, F_{z_{\Delta \setminus 1}}, a_g, a_1).$$


---

**Theorem C.16.** As  $T \rightarrow \infty$ , if  $\sum_{t=1}^T \eta_t = \infty$ , and  $\sum_{t=1}^T \eta_t^2 < \infty$ , then  $Q$ -learning converges to the optimal  $Q$  function asymptotically with probability 1.

Furthermore, finite-time guarantees with the learning rate and sample complexity have been shown in Chen and Maguluri [2022b], which when adapted to our  $\hat{Q}_{k,m}$  framework in Algorithm 3 yields:

**Theorem C.17** (Chen and Maguluri [2022b]). For all  $t = \{1, \dots, T\}$  and for any  $\epsilon > 0$ , if the learning rate sequence  $\eta_t$  satisfies  $\eta_t = (1 - \gamma)^4 \epsilon^2$  and  $T = \frac{|\mathcal{S}_l| |\mathcal{A}_l| k^{|S_l|+|\mathcal{A}_l|} |\mathcal{S}_g| |\mathcal{A}_g|}{(1-\gamma)^5} \epsilon^2$ , then

$$\|\hat{Q}_{k,m}^T - \hat{Q}_{k,m}^{\text{est}}\| \leq \epsilon.$$

**Definition C.18** (Star Graph  $S_n$ ). For  $n \in \mathbb{N}$ , the star graph  $S_n$  is the complete bipartite graph  $K_{1,n}$ .

$S_n$  captures the graph density notion by saturating the set of neighbors of the central node. Such settings find applications beyond RL as well [Chaudhari et al., 2024, Anand and Qu, 2024, Li et al., 2019]. The cardinality of the search space simplex for the optimal policy is exponential in  $n$ , so it cannot be naively modeled by an MDP: we need to exploit the symmetry of the local agents. This intuition allows our subsampling algorithm to run in polylogarithmic time (in  $n$ ). Some works leverage an exponential decaying property that truncates the search space for policies over immediate neighborhoods of agents; however, this still relies on the assumption that the graph neighborhood for the agent is sparse [Qu et al., 2020a,b]; however,  $S_n$  is not locally sparse; hence, previous methods do not apply to this problem instance.

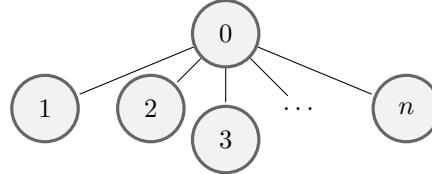


Figure 3: Star graph  $S_n$

Finally, we argue why mean-field value iteration faithfully performs the same update as value iteration.

**Lemma C.19** (Equivalence of Mean-Field Value Iteration and Value Iteration). Since the local agents  $1, \dots, n$  are all homogeneous in their state/action spaces, the  $\hat{Q}_k$ -function only depends on them through their empirical distribution  $F_{z_\Delta}$ , proving the lemma. Therefore, for the remainder of the paper, we will use  $\hat{Q}_k(s_g, F_{z_\Delta}, a_g) := \hat{Q}_k(s_g, s_i, F_{z_{\Delta \setminus i}}, a_g, a_i) := \hat{Q}_k(s_g, s_\Delta, a_g, a_\Delta)$  interchangeably, unless making a remark about the computational complexity of learning each function.

## D Proof Sketch

This section details an outline for the proof of Theorem 4.4, as well as some key ideas. At a high level, our SUBSAMPLE-MFQ framework recovers exact mean-field  $Q$  learning and traditional value iteration when  $k = n$  and as  $m \rightarrow \infty$ . Further, as  $k \rightarrow n$ ,  $\hat{Q}_k^*$  should intuitively get closer to  $Q^*$  from which the optimal policy is derived. Thus, the proof is divided into three major steps: firstly, we prove a Lipschitz continuity bound between  $\hat{Q}_k^*$  and  $\hat{Q}_n^*$  in terms of the total variation (TV) distance between  $F_{z_\Delta}$  and  $F_{z_{[n]}}$ . Next, we bound the TV distance between  $F_{z_\Delta}$  and  $F_{z_{[n]}}$ . Finally, we bound the value differences between  $\pi_k^{\text{est}}$  and  $\pi^*$  by bounding  $Q^*(s, \pi^*(s)) - Q^*(s, \pi_k^{\text{est}}(s))$  and then using the performance difference lemma from Kakade and Langford [2002].

**Step 1: Lipschitz Continuity Bound.** To compare  $\hat{Q}_k^*(s_g, F_{s_\Delta}, a_g)$  with  $Q^*(s, a_g)$ , we prove a Lipschitz continuity bound between  $\hat{Q}_k^*(s_g, F_{s_\Delta}, a_g)$  and  $\hat{Q}_{k'}^*(s_g, F_{s_{\Delta'}}, a_g)$  with respect to the TV distance measure between  $s_\Delta \in \binom{[n]}{k}$  and  $s_{\Delta'} \in \binom{[n]}{k'}$ :

**Theorem D.1** (Lipschitz continuity in  $\hat{Q}_k^*$ ). *For all  $(s, a) \in \mathcal{S} \times \mathcal{A}$ ,  $\Delta \in \binom{[n]}{k}$  and  $\Delta' \in \binom{[n]}{k'}$ ,*

$$|\hat{Q}_k^*(s_g, F_{z_\Delta}, a_g) - \hat{Q}_{k'}^*(s_g, F_{z_{\Delta'}}, a_g)| \leq \frac{2}{1-\gamma} \|r_l(\cdot, \cdot)\|_\infty \cdot \text{TV}(F_{z_\Delta}, F_{z_{\Delta'}})$$

We defer the proof of Theorem D.1 to Appendix E. See Figure 4 for a comparison between the  $\hat{Q}_k^*$  learning and estimation process, and the exact  $Q$ -learning framework.

**Step 2: Bounding Total Variation (TV) Distance.** We bound the TV distance between  $F_{z_\Delta}$  and  $F_{z_{[n]}}$ , where  $\Delta \in \mathcal{U}(\binom{[n]}{k})$ . This task is equivalent to bounding the discrepancy between the empirical distribution and the distribution of the underlying finite population. When each  $i \in \Delta$  is uniformly sampled *without* replacement, we use Theorem F.3 from Anand and Qu [2024] which generalizes the Dvoretzky-Kiefer-Wolfowitz (DKW) concentration inequality for empirical distribution functions.

Using this, we show:

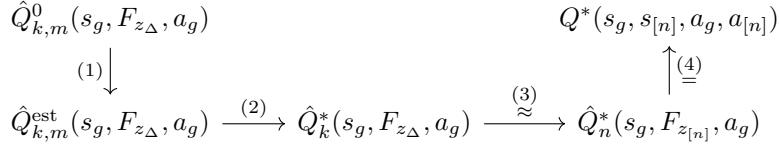


Figure 4: Flow of the algorithm and relevant analyses in learning  $Q^*$ . Here, (1) follows by performing Algorithm 1 (SUBSAMPLE-MFQ: Learning) on  $\hat{Q}_{k,m}^0$ . (2) follows from Theorem 4.3. (3) follows from the Lipschitz continuity and total variation distance bounds in Theorems D.1 and D.2. Finally, (4) follows from noting that  $\hat{Q}_n^* = Q^*$ .

**Theorem D.2.** *Given a finite population  $\mathcal{Z} = (z_1, \dots, z_n)$  for  $\mathcal{Z} \in \mathcal{Z}_l^n$ , let  $\Delta \subseteq [n]$  be a uniformly random sample from  $\mathcal{Z}$  of size  $k$  chosen without replacement. Fix  $\epsilon > 0$ . Then, for all  $x \in \mathcal{Z}_l$ :*

$$\Pr \left[ \sup_{x \in \mathcal{Z}_l} \left| \frac{1}{|\Delta|} \sum_{i \in \Delta} \mathbb{1}\{z_i = x\} - \frac{1}{n} \sum_{i \in [n]} \mathbb{1}\{z_i = x\} \right| \leq \epsilon \right] \geq 1 - 2|\mathcal{Z}_l|e^{-\frac{2kn\epsilon^2}{n-k+1}}.$$

Then, by Theorem D.2 and the definition of total variation distance from Section 2, we have that for  $\delta \in (0, 1]$ , with probability at least  $1 - \delta$ ,

$$\text{TV}(F_{s_\Delta}, F_{s_{[n]}}) \leq \sqrt{\frac{n-k+1}{8nk} \ln \frac{2|\mathcal{Z}_l|}{\delta}} \quad (26)$$

We then apply this result to our MARL setting by studying the rate of decay of the objective function between the learned policy  $\pi_k^{\text{est}}$  and the optimal policy  $\pi^*$  (Theorem 4.4).

**Step 3: Performance Difference Lemma to Complete the Proof.** As a consequence of the prior two steps and Lemma 4.3,  $Q^*(s, a')$  and  $\hat{Q}_{k,m}^{\text{est}}(s_g, F_{z_\Delta}, a'_g)$  become similar as  $k \rightarrow n$ . We further prove that the value generated by the policies  $\pi^*$  and  $\pi_k^{\text{est}}$  must also be very close (where the residue shrinks as  $k \rightarrow n$ ). We then use the well-known performance difference lemma [Kakade and Langford, 2002] which we restate in Appendix G.1. A crucial theorem needed to use the performance difference lemma is a bound on  $Q^*(s', \pi^*(s')) - Q^*(s', \hat{\pi}_k^{\text{est}}(s'_g, F_{s'_\Delta}))$ .

Therefore, we formulate and prove Theorem D.3 which yields a probabilistic bound on this difference, where the randomness is over the choice of  $\Delta \in \binom{[n]}{k}$ :

**Theorem D.3.** *For a fixed  $s' \in \mathcal{S} := \mathcal{S}_g \times \mathcal{S}_l^n$  and for  $\delta \in (0, 1]$ , with probability at least  $1 - 2|\mathcal{A}_g|k^{|\mathcal{A}_l|}\delta$ :*

$$Q^*(s', \pi^*(s')) - Q^*(s', \hat{\pi}_{k,m}^{\text{est}}(s'_g, F_{s'_\Delta})) \leq \frac{2\|r_l(\cdot, \cdot)\|_\infty}{1-\gamma} \sqrt{\frac{n-k+1}{2nk} \ln \left( \frac{2|\mathcal{Z}_l|}{\delta} \right)} + 2\epsilon_{k,m}.$$

We defer the proof of Theorem D.3 and finding optimal value of  $\delta$  to Theorem G.8 in the Appendix. Using Theorem D.3 and the performance difference lemma leads to Theorem 4.4.

## E Proof of Lipschitz-continuity Bound

This section proves a Lipschitz-continuity bound between  $\hat{Q}_k^*$  and  $Q^*$  and includes a framework to compare  $\frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^*(s_g, s_\Delta, a_g)$  and  $Q^*(s, a_g)$  in Theorem E.12.

Let  $z_i := (s_i, a_i) \in \mathcal{Z}_l := \mathcal{S}_l \times \mathcal{A}_l$  and  $z_\Delta = \{z_i : i \in \Delta\} \in \mathcal{Z}_l^k$ . For  $z_i = (s_i, a_i)$ , let  $z_i(s) = s_i$ , and  $z_i(a) = a_i$ . With abuse of notation, note that  $\hat{Q}_k^T(s_g, a_g, s_\Delta, a_\Delta)$  is equal to  $\hat{Q}_k^T(s_g, a_g, z_\Delta)$ .

The following definitions will be relevant to the proof of Theorem E.3.

**Definition E.1** (Empirical Distribution Function). For all  $z \in \mathcal{Z}_l^{|\Delta|}$  and  $(s', a') \in \mathcal{S}_l \times \mathcal{A}_l$ , where  $\Delta \subseteq [n]$ ,

$$F_{z_\Delta}(s', a') = \frac{1}{|\Delta|} \sum_{i \in \Delta} \mathbb{1}\{z_i(s) = s', z_i(a) = a'\}$$

**Definition E.2** (Total Variation Distance). Let  $P$  and  $Q$  be discrete probability distribution over some domain  $\Omega$ . Then,

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

**Theorem E.3** ( $\hat{Q}_k^T$  is  $\frac{2}{1-\gamma} \|r_l(\cdot, \cdot)\|_\infty$ -Lipschitz continuous with respect to  $F_{z_\Delta}$  in total variation distance). Suppose  $\Delta, \Delta' \subseteq [n]$  such that  $|\Delta| = k$  and  $|\Delta'| = k'$ . Then:

$$\left| \hat{Q}_k^T(s_g, a_g, F_{z_\Delta}) - \hat{Q}_{k'}^T(s_g, a_g, F_{z_{\Delta'}}) \right| \leq \left( \sum_{t=0}^{T-1} 2\gamma^t \right) \|r_l(\cdot, \cdot, \cdot)\|_\infty \cdot \text{TV}(F_{z_\Delta}, F_{z_{\Delta'}})$$

*Proof.* We prove this inductively. First, note that  $\hat{Q}_k^0(\cdot, \cdot, \cdot) = \hat{Q}_{k'}^0(\cdot, \cdot, \cdot) = 0$  from the initialization step, which proves the lemma for  $T = 0$ , since  $\text{TV}(\cdot, \cdot) \geq 0$ . At  $T = 1$ :

$$\begin{aligned} |\hat{Q}_k^1(s_g, a_g, F_{z_\Delta}) - \hat{Q}_{k'}^1(s_g, a_g, F_{z_{\Delta'}})| &= \left| \hat{\mathcal{T}}_k \hat{Q}_k^0(s_g, a_g, F_{z_\Delta}) - \hat{\mathcal{T}}_{k'} \hat{Q}_{k'}^0(s_g, a_g, F_{z_{\Delta'}}) \right| \\ &= \left| r(s_g, F_{s_\Delta}, a_g) + \gamma \mathbb{E}_{s'_g, s'_{\Delta'}} \max_{a'_g \in \mathcal{A}_g} \hat{Q}_k^0(s'_g, F_{s'_{\Delta'}}, a'_g) \right. \\ &\quad \left. - r(s_g, F_{s_{\Delta'}}, a_g) - \gamma \mathbb{E}_{s'_g, s'_{\Delta'}} \max_{a'_g \in \mathcal{A}_g} \hat{Q}_{k'}^0(s'_g, F_{s'_{\Delta'}}, a'_g) \right| \\ &= |r(s_g, a_g, F_{z_\Delta}) - r(s_g, a_g, F_{z_{\Delta'}})| \\ &= \left| \frac{1}{k} \sum_{i \in \Delta} r_l(s_g, z_i) - \frac{1}{k'} \sum_{i \in \Delta'} r_l(s_g, z_i) \right| \\ &= \left| \mathbb{E}_{z_l \sim F_{z_\Delta}} r_l(s_g, z_l) - \mathbb{E}_{z'_l \sim F_{z_{\Delta'}}} r_l(s_g, z'_l) \right| \end{aligned}$$

In the first and second equalities, we use the time evolution property of  $\hat{Q}_k^1$  and  $\hat{Q}_{k'}$  by applying the adapted Bellman operators  $\hat{\mathcal{T}}_k$  and  $\hat{\mathcal{T}}_{k'}$  to  $\hat{Q}_k^0$  and  $\hat{Q}_{k'}$ , respectively, and expanding. In the third and fourth equalities, we note that  $\hat{Q}_k^0(\cdot, \cdot, \cdot) = \hat{Q}_{k'}^0(\cdot, \cdot, \cdot) = 0$ , and subtract the common ‘global component’ of the reward function.

Then, noting the general property that for any function  $f : \mathcal{X} \rightarrow \mathcal{Y}$  for  $|\mathcal{X}| < \infty$  we can write  $f(x) = \sum_{y \in \mathcal{X}} f(y) \mathbb{1}\{y = x\}$ , we have:

$$\begin{aligned} |\hat{Q}_k^1(s_g, a_g, F_{z_\Delta}) - \hat{Q}_{k'}^1(s_g, a_g, F_{z_{\Delta'}})| &= \left| \mathbb{E}_{z_l \sim F_{z_\Delta}} \left[ \sum_{z \in \mathcal{Z}} r_l(s_g, z) \mathbb{1}\{z_l = z\} \right] - \mathbb{E}_{z'_l \sim F_{z_{\Delta'}}} \left[ \sum_{z \in \mathcal{Z}} r_l(s_g, z) \mathbb{1}\{z'_l = z\} \right] \right| \\ &= \left| \sum_{z \in \mathcal{Z}} r_l(s_g, z) \cdot (\mathbb{E}_{z_l \sim F_{z_\Delta}} \mathbb{1}\{z_l = z\} - \mathbb{E}_{z'_l \sim F_{z_{\Delta'}}} \mathbb{1}\{z'_l = z\}) \right| \\ &= \left| \sum_{z \in \mathcal{Z}} r_l(s_g, z) \cdot (F_{z_\Delta}(z) - F_{z_{\Delta'}}(z)) \right| \\ &\leq \left| \max_{z \in \mathcal{Z}} r_l(s_g, z) \right| \cdot \sum_{z \in \mathcal{Z}} |F_{z_\Delta}(z) - F_{z_{\Delta'}}(z)| \\ &\leq 2 \|r_l(\cdot, \cdot)\|_\infty \cdot \text{TV}(F_{z_\Delta}, F_{z_{\Delta'}}) \end{aligned}$$

The second equality follows from the linearity of expectations, and the third equality follows by noting that for any random variable  $X \sim \mathcal{X}$ ,  $\mathbb{E}_X \mathbb{1}[X = x] = \Pr[X = x]$ . The first inequality follows from an application of the triangle inequality and the Cauchy-Schwarz inequality, and the second inequality uses the definition of TV distance. Thus, at  $T = 1$ ,  $\hat{Q}$  is  $(2 \|r_l(\cdot, \cdot)\|_\infty)$ -Lipschitz continuous in TV distance, proving the base case.

Assume that for  $T \leq t' \in \mathbb{N}$ :

$$\left| \hat{Q}_k^T(s_g, a_g, F_{z_\Delta}) - \hat{Q}_{k'}^T(s_g, a_g, F_{z_{\Delta'}}) \right| \leq \left( \sum_{t=0}^{T-1} 2\gamma^t \right) \|r_l(\cdot, \cdot)\|_\infty \cdot \text{TV}(F_{z_\Delta}, F_{z_{\Delta'}})$$

Then, inductively:

$$\begin{aligned} & |\hat{Q}_k^{T+1}(s_g, a_g, F_{z_\Delta}) - \hat{Q}_{k'}^{T+1}(s_g, a_g, F_{z_{\Delta'}})| \\ & \leq \left| \frac{1}{|\Delta|} \sum_{i \in \Delta} r_l(s_g, z_i) - \frac{1}{|\Delta'|} \sum_{i \in \Delta'} r_l(s_g, z_i) \right| \\ & \quad + \gamma \left| \mathbb{E}_{s'_g, s'_{\Delta}} \max_{\substack{a'_g \in \mathcal{A}_g, \\ a'_{\Delta} \in \mathcal{A}_l^k}} \hat{Q}_k^T(s'_g, a'_g, F_{z'_{\Delta}}) - \mathbb{E}_{s'_g, s'_{\Delta'}} \max_{\substack{a'_g \in \mathcal{A}_g, \\ a'_{\Delta'} \in \mathcal{A}_l^{k'}}} \hat{Q}_{k'}^T(s'_g, a'_g, F_{z'_{\Delta'}}) \right| \\ & \leq 2\|r_l(\cdot, \cdot)\|_\infty \cdot \text{TV}(F_{z_\Delta}, F_{z_{\Delta'}}) \\ & \quad + \gamma \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k} \max_{\substack{a'_g \in \mathcal{A}_g, \\ a'_{\Delta} \in \mathcal{A}_l^k}} \hat{Q}_k^T(s'_g, a'_g, F_{z'_{\Delta}}) - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}} \max_{\substack{a'_g \in \mathcal{A}_g, \\ a'_{\Delta'} \in \mathcal{A}_l^{k'}}} \hat{Q}_{k'}^T(s'_g, a'_g, F_{z'_{\Delta'}}) \right| \\ & \leq 2\|r_l(\cdot, \cdot)\|_\infty \cdot \text{TV}(F_{z_\Delta}, F_{z_{\Delta'}}) + \gamma \left( \sum_{\tau=0}^{T-1} 2\gamma^\tau \right) \|r_l(\cdot, \cdot)\|_\infty \cdot \text{TV}(F_{z_\Delta}, F_{z_{\Delta'}}) \\ & = \left( \sum_{\tau=0}^T 2\gamma^\tau \right) \|r_l(\cdot, \cdot)\|_\infty \cdot \text{TV}(F_{z_\Delta}, F_{z_{\Delta'}}) \end{aligned}$$

In the first inequality, we rewrite the expectations over the states as the expectation over the joint transition probabilities. The second inequality then follows from Theorem E.14. To apply it to Theorem E.14, we conflate the joint expectation over  $(s_g, s_{\Delta \cup \Delta'})$  and reduce it back to the original form of its expectation. Finally, the third inequality follows from Theorem E.5.

By the inductive hypothesis, the claim is proven.  $\square$

**Definition E.4.** [Joint Stochastic Kernels] The joint stochastic kernel on  $(s_g, s_\Delta)$  for  $\Delta \subseteq [n]$  where  $|\Delta| = k$  is defined as  $\mathcal{J}_k : \mathcal{S}_g \times \mathcal{S}_l^k \times \mathcal{S}_g \times \mathcal{A}_g \times \mathcal{S}_l^k \times \mathcal{A}_l^k \rightarrow [0, 1]$ , where

$$\mathcal{J}_k(s'_g, s'_{\Delta} | s_g, a_g, s_\Delta, a_\Delta) := \Pr[(s'_g, s'_{\Delta}) | s_g, a_g, s_\Delta, a_\Delta] \quad (27)$$

**Lemma E.5.** For all  $T \in \mathbb{N}$ , for any  $a_g \in \mathcal{A}_g, s_g \in \mathcal{S}_g, s_\Delta \in \mathcal{S}_l^k, a_\Delta \in \mathcal{A}_l^k, a'_\Delta \in \mathcal{A}_l^k$ , and for all joint stochastic kernels  $\mathcal{J}_k$  as defined in Theorem E.4:

$$\begin{aligned} & \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_\Delta, a_\Delta)} \max_{a'_g, a'_\Delta} \hat{Q}_k^T(s'_g, a'_g, F_{z'_{\Delta}}) \right. \\ & \left. - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \max_{a'_g, a'_{\Delta'}} \hat{Q}_{k'}^T(s'_g, a'_g, F_{z'_{\Delta'}}) \right| \leq \left( \sum_{\tau=0}^{T-1} 2\gamma^\tau \right) \|r_l(\cdot, \cdot)\|_\infty \text{TV}(F_{z_\Delta}, F_{z_{\Delta'}}) \end{aligned}$$

*Proof.* We prove this inductively. At  $T = 0$ , the statement is true since  $\hat{Q}_k^0(\cdot, \cdot, \cdot) = \hat{Q}_{k'}^0(\cdot, \cdot, \cdot) = 0$  and  $\text{TV}(\cdot, \cdot) \geq 0$ .

At  $T = 1$ ,

$$\begin{aligned} & \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_\Delta, a_\Delta)} \max_{a'_g, a'_\Delta} \hat{Q}_k^1(s'_g, a'_g, s'_{\Delta}, a'_\Delta) \right. \\ & \quad \left. - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \max_{a'_g, a'_{\Delta'}} \hat{Q}_{k'}^1(s'_g, a'_g, s'_{\Delta'}, a'_{\Delta'}) \right| \\ & = \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_\Delta, a_\Delta)} \max_{a'_g, a'_\Delta} \left[ r_g(s'_g, a'_g) + \frac{\sum_{i \in \Delta} r_l(s'_i, a'_i, s'_g)}{k} \right] \right. \\ & \quad \left. - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \max_{a'_g, a'_{\Delta'}} \left[ r_g(s'_g, a'_g) + \frac{\sum_{i \in \Delta'} r_l(s'_i, a'_i, s'_g)}{k'} \right] \right| \end{aligned}$$

$$\begin{aligned}
&= \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \max_{a'_{\Delta}} \frac{\sum_{i \in \Delta} r_l(s'_i, a'_i, s'_g)}{k} \right. \\
&\quad \left. - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \max_{a'_{\Delta'}} \frac{\sum_{i \in \Delta'} r_l(s'_i, a'_i, s'_g)}{k'} \right| \\
&= \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \frac{1}{k} \sum_{i \in \Delta} \tilde{r}_l(s'_i, s'_g) - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \frac{1}{k'} \sum_{i \in \Delta'} \tilde{r}_l(s'_i, s'_g) \right|
\end{aligned}$$

In the last equality, we note that

$$\max_{a'_{\Delta}} \sum_{i \in \Delta} r_l(s'_i, a'_i, s'_g) = \sum_{i \in \Delta} \max_{a'_{\Delta}} r_l(s'_i, a'_i, s'_g) = \sum_{i \in \Delta} \max_{a'_i} r_l(s'_i, a'_i, s'_g) = \sum_{i \in \Delta} \tilde{r}_l(s'_i, s'_g),$$

where  $\tilde{r}_l(s'_i, s'_g) := \max_{a'_i} r_l(s'_i, a'_i, s'_g)$ .

Then, we have:

$$\begin{aligned}
&\left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \frac{1}{k} \sum_{i \in \Delta} \tilde{r}_l(s'_i, s'_g) - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \frac{1}{k'} \sum_{i \in \Delta'} \tilde{r}_l(s'_i, s'_g) \right| \\
&= \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \frac{1}{k} \sum_{x \in \mathcal{S}_l} \tilde{r}_l(x, s'_g) F_{s'_{\Delta}}(x) \right. \\
&\quad \left. - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \frac{1}{k'} \sum_{x \in \mathcal{S}_l} \tilde{r}_l(x, s'_g) F_{s'_{\Delta'}}(x) \right| \\
&= \left| \mathbb{E}_{s'_g \sim \sum_{s'_{\Delta \cup \Delta'} \in \mathcal{S}_l} |\Delta \cup \Delta'| \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot, s'_{\Delta \cup \Delta'} | s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'})} \right. \\
&\quad \left. \sum_{x \in \mathcal{S}_l} \tilde{r}_l(x, s'_g) \mathbb{E}_{s'_{\Delta \cup \Delta'} \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot, s'_{\Delta \cup \Delta'} | s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'})} [F_{s'_{\Delta}}(x) - F_{s'_{\Delta'}}(x)] \right| \\
&\leq \|\tilde{r}_l(\cdot, \cdot)\|_{\infty} \cdot \mathbb{E}_{s'_g \sim \sum_{s'_{\Delta \cup \Delta'} \in \mathcal{S}_l} |\Delta \cup \Delta'| \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot, s'_{\Delta \cup \Delta'} | s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'})} \\
&\quad \sum_{x \in \mathcal{S}_l} |\mathbb{E}_{s'_{\Delta \cup \Delta'} | s'_g} F_{s'_{\Delta}}(x) - \mathbb{E}_{s'_{\Delta \cup \Delta'} | s'_g} F_{s'_{\Delta'}}(x)| \\
&\leq \|r_l(\cdot, \cdot)\|_{\infty} \cdot \mathbb{E}_{s'_g \sim \sum_{s'_{\Delta \cup \Delta'} \in \mathcal{S}_l} |\Delta \cup \Delta'| \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot, s'_{\Delta \cup \Delta'} | s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'})} \\
&\quad \sum_{x \in \mathcal{S}_l} |\mathbb{E}_{s'_{\Delta \cup \Delta'} | s'_g} F_{s'_{\Delta}}(x) - \mathbb{E}_{s'_{\Delta \cup \Delta'} | s'_g} F_{s'_{\Delta'}}(x)| \\
&\leq 2\|r_l(\cdot, \cdot)\|_{\infty} \cdot \mathbb{E}_{s'_g \sim \sum_{s'_{\Delta \cup \Delta'} \in \mathcal{S}_l} |\Delta \cup \Delta'| \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot, s'_{\Delta \cup \Delta'} | s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'})} \\
&\quad \text{TV}(\mathbb{E}_{s'_{\Delta \cup \Delta'} | s'_g} F_{s'_{\Delta}}, \mathbb{E}_{s'_{\Delta \cup \Delta'} | s'_g} F_{s'_{\Delta'}}) \\
&\leq 2\|r_l(\cdot, \cdot)\|_{\infty} \cdot \text{TV}(F_{z_{\Delta}}, F_{z'_{\Delta}})
\end{aligned}$$

The first equality follows from noting that  $f(x) = \sum_{x' \in \mathcal{X}} f(x') \mathbb{1}\{x = x'\}$  and from Fubini-Tonelli's inequality which allows us to swap the order of summations as the summand is finite. The second equality uses the law of total expectation. The first inequality uses Jensen's inequality and the triangle inequality. The second inequality uses  $\|\tilde{r}_l(\cdot, \cdot)\|_{\infty} \leq \|r_l(\cdot, \cdot)\|_{\infty}$  which holds as  $\|r_l\|_{\infty}$  is the infinite-norm of the local reward functions and is therefore atleast as large any other element in the image of  $r_l$ . The third inequality follows from the definition of total variation distance, and the final inequality follows from Theorem E.7. This proves the base case.

Then, assume that for  $T \leq t' \in \mathbb{N}$ , for all joint stochastic kernels  $\mathcal{J}_k$  and  $\mathcal{J}_{k'}$ , and for all  $a'_g \in \mathcal{A}_g, a'_{\Delta} \in \mathcal{A}_l^k$ :

$$\begin{aligned}
&\left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \max_{a'_g, a'_{\Delta}} \hat{Q}_k^T(s'_g, a'_g, s'_{\Delta}, a'_{\Delta}) - \right. \\
&\quad \left. \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \max_{a'_g, a'_{\Delta'}} \hat{Q}_{k'}^T(s'_g, a'_g, s'_{\Delta'}, a'_{\Delta'}) \right| \leq 2 \left( \sum_{t=0}^{T-1} \gamma^t \right) \|r_l(\cdot, \cdot)\|_{\infty} \text{TV}(F_{z_{\Delta}}, F_{z_{\Delta'}})
\end{aligned}$$

For the remainder of the proof, we adopt the shorthand  $\mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}}$  to denote  $\mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})}$ , and  $\mathbb{E}_{(s''_g, s''_{\Delta}) \sim \mathcal{J}}$  to denote  $\mathbb{E}_{(s''_g, s''_{\Delta}) \sim \mathcal{J}(\cdot, \cdot | s'_g, a'_g, s'_{\Delta}, a'_{\Delta})}$ . Then, inductively, we have:

$$\left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \max_{a'_g, a'_{\Delta}} \hat{Q}_k^{T+1}(s'_g, a'_g, s'_{\Delta}, a'_{\Delta}) - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \max_{a'_g, a'_{\Delta'}} \hat{Q}_{k'}^{T+1}(s'_g, a'_g, s'_{\Delta'}, a'_{\Delta'}) \right| = (\text{I}) - (\text{II}),$$

where

$$(\text{I}) = \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \max_{a'_g, a'_{\Delta}} \left[ r_{\Delta}(s'_g, a'_g, s'_{\Delta}, a'_{\Delta}) + \gamma \mathbb{E}_{(s''_g, s''_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s'_g, a'_g, s'_{\Delta}, a'_{\Delta})} \max_{a''_g, a''_{\Delta}} \hat{Q}_k^T(s''_g, s''_{\Delta}, a''_g, a''_{\Delta}) \right] \right|$$

and

$$(\text{II}) = \left| \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \max_{a'_g, a'_{\Delta'}} \left[ r_{\Delta'}(s'_g, a'_g, s'_{\Delta'}, a'_{\Delta'}) + \gamma \mathbb{E}_{(s''_g, s''_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s'_g, a'_g, s'_{\Delta'}, a'_{\Delta'})} \max_{a''_g, a''_{\Delta'}} \hat{Q}_{k'}^T(s''_g, s''_{\Delta'}, a''_g, a''_{\Delta'}) \right] \right|$$

Let

$$\begin{aligned} \tilde{a}_g, \tilde{a}_{\Delta} &= \arg \max_{a'_g \in \mathcal{A}_g, a'_{\Delta} \in \mathcal{A}_l^k} \left[ r_{\Delta}(s'_g, a'_g, s'_{\Delta}, a'_{\Delta}) + \gamma \mathbb{E}_{(s''_g, s''_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s'_g, a'_g, s'_{\Delta}, a'_{\Delta})} \max_{a''_g, a''_{\Delta}} \hat{Q}_k^T(s''_g, s''_{\Delta}, a''_g, a''_{\Delta}) \right], \end{aligned}$$

and

$$\begin{aligned} \hat{a}_g, \hat{a}_{\Delta'} &= \arg \max_{a'_g \in \mathcal{A}_g, a'_{\Delta'} \in \mathcal{A}_l^k} \left[ r_{\Delta'}(s'_g, a'_g, s'_{\Delta'}, a'_{\Delta'}) + \gamma \mathbb{E}_{(s''_g, s''_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s'_g, a'_g, s'_{\Delta'}, a'_{\Delta'})} \max_{a''_g, a''_{\Delta'}} \hat{Q}_{k'}^T(s''_g, s''_{\Delta'}, a''_g, a''_{\Delta'}) \right] \end{aligned}$$

Then, define  $\tilde{a}_{\Delta \cup \Delta'} \in \mathcal{A}_l^{|\Delta \cup \Delta'|}$  by

$$\tilde{a}_i = \begin{cases} \tilde{a}_i, & \text{if } i \in \Delta, \\ \hat{a}_i, & \text{if } i \in \Delta' \setminus \Delta \end{cases}$$

Similarly, define  $\hat{a}_{\Delta \cup \Delta'} \in \mathcal{A}_l^{|\Delta \cup \Delta'|}$  by

$$\hat{a}_i = \begin{cases} \hat{a}_i, & i \in \Delta' \\ \tilde{a}_i, & i \in \Delta \setminus \Delta' \end{cases}$$

Suppose  $(\text{I}) \geq (\text{II})$ . Then,

$$\begin{aligned} &\left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \max_{a'_g, a'_{\Delta}} \hat{Q}_k^{T+1}(s'_g, a'_g, s'_{\Delta}, a'_{\Delta}) - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \max_{a'_g, a'_{\Delta'}} \hat{Q}_{k'}^{T+1}(s'_g, a'_g, s'_{\Delta'}, a'_{\Delta'}) \right| \end{aligned}$$

$$\begin{aligned}
&= \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \hat{Q}_k^{T+1}(s'_g, \tilde{a}_g, s'_{\Delta}, \tilde{a}_{\Delta}) \right. \\
&\quad \left. - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \hat{Q}_{k'}^{T+1}(s'_g, \hat{a}_g, s'_{\Delta'}, \hat{a}_{\Delta'}) \right| \\
&\leq \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \hat{Q}_k^{T+1}(s'_g, \tilde{a}_g, s'_{\Delta}, \tilde{a}_{\Delta}) \right. \\
&\quad \left. - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \hat{Q}_{k'}^{T+1}(s'_g, \tilde{a}_g, s'_{\Delta'}, \tilde{a}_{\Delta'}) \right| \\
&= \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \left[ r_{\Delta}(s'_g, s'_{\Delta}, \tilde{a}_g, \tilde{a}_{\Delta}) \right. \right. \\
&\quad \left. + \gamma \mathbb{E}_{(s''_g, s''_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s'_g, \tilde{a}_g, s'_{\Delta}, \tilde{a}_{\Delta})} \max_{a''_g, a''_{\Delta}} \hat{Q}_k^T(s''_g, a''_g, s''_{\Delta}, a''_{\Delta}) \right] \\
&\quad \left. - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \left[ r_{\Delta'}(s'_g, s'_{\Delta'}, \tilde{a}_g, \tilde{a}_{\Delta'}) \right. \right. \\
&\quad \left. + \gamma \mathbb{E}_{(s''_g, s''_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s'_g, \tilde{a}_g, s'_{\Delta'}, \tilde{a}_{\Delta'})} \max_{a''_g, a''_{\Delta'}} \hat{Q}_{k'}^T(s''_g, a''_g, s''_{\Delta'}, a''_{\Delta'}) \right] \right| \\
&\leq \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} r_{\Delta}(s'_g, s'_{\Delta}, \tilde{a}_g, \tilde{a}_{\Delta}) \right. \\
&\quad \left. - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} r_{\Delta'}(s'_g, s'_{\Delta'}, \tilde{a}_g, \tilde{a}_{\Delta'}) \right| \\
&\quad + \gamma \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \mathbb{E}_{(s''_g, s''_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s'_g, \tilde{a}_g, s'_{\Delta}, \tilde{a}_{\Delta})} \max_{a''_g, a''_{\Delta}} \hat{Q}_k^T(s''_g, a''_g, s''_{\Delta}, a''_{\Delta}) \right. \\
&\quad \left. - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \mathbb{E}_{(s''_g, s''_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s'_g, \tilde{a}_g, s'_{\Delta'}, \tilde{a}_{\Delta'})} \max_{a''_g, a''_{\Delta'}} \hat{Q}_{k'}^T(s''_g, a''_g, s''_{\Delta'}, a''_{\Delta'}) \right| \\
&= \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \frac{1}{k} \sum_{i \in \Delta} r_l(s'_i, s'_g, \tilde{a}_i) \right. \\
&\quad \left. - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \frac{1}{k'} \sum_{i \in \Delta'} r_l(s'_i, s'_g, \tilde{a}_i) \right| \\
&\quad + \gamma \left| \mathbb{E}_{(s''_g, s''_{\Delta}) \sim \tilde{\mathcal{J}}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \max_{a''_g, a''_{\Delta}} \hat{Q}_k^T(s''_g, a''_g, s''_{\Delta}, a''_{\Delta}) \right. \\
&\quad \left. - \mathbb{E}_{(s''_g, s''_{\Delta'}) \sim \tilde{\mathcal{J}}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \max_{a''_g, a''_{\Delta'}} \hat{Q}_{k'}^T(s''_g, a''_g, s''_{\Delta'}, a''_{\Delta'}) \right| \\
&\leq 2 \|r_l(\cdot, \cdot)\|_{\infty} \cdot \text{TV}(F_{z_{\Delta}}, F_{z_{\Delta'}}) + \gamma \left( \sum_{t=0}^T 2\gamma^t \right) \|r_l(\cdot, \cdot)\|_{\infty} \cdot \text{TV}(F_{z_{\Delta}}, F_{z_{\Delta'}}) \\
&= \left( \sum_{t=0}^{T+1} 2\gamma^t \right) \|r_l(\cdot, \cdot)\|_{\infty} \cdot \text{TV}(F_{z_{\Delta}}, F_{z_{\Delta'}})
\end{aligned}$$

The first equality rewrites the equations with their respective maximizing actions. The first inequality upper-bounds this difference by allowing all terms to share the common action  $\tilde{a}$ . Using the Bellman equation, the second equality expands  $\hat{Q}_k^{T+1}$  and  $\hat{Q}_{k'}^{T+1}$ . The second inequality follows from the triangle inequality. The third equality follows by subtracting the common  $r_g(s'_g, \tilde{a}_g)$  terms from the reward and noting that the two expectation terms  $\mathbb{E}_{s'_g, s'_{\Delta}} \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta}) \mathbb{E}_{s''_g, s''_{\Delta}} \sim \mathcal{J}_k(\cdot, \cdot | s_g, \tilde{a}_g, s_{\Delta}, \tilde{a}_{\Delta})$  can be combined into a single expectation  $\mathbb{E}_{s''_g, s''_{\Delta}} \sim \tilde{\mathcal{J}}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})$ .

In the special case where  $a_{\Delta} = \tilde{a}_{\Delta}$ , we derive a closed form expression for  $\tilde{\mathcal{J}}_k$  in Theorem E.11. To justify the third inequality, the second term follows from the induction hypothesis and the first term

follows from the following derivation:

$$\begin{aligned}
& \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \frac{\sum_{i \in \Delta} r_l(s'_i, s'_g, \tilde{a}_i)}{k} - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \frac{\sum_{i \in \Delta'} r_l(s'_i, s'_g, \tilde{a}_i)}{k'} \right| \\
&= \left| \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \frac{\sum_{i \in \Delta} r_l^{\tilde{a}}(s'_i, s'_g)}{k} - \mathbb{E}_{(s'_g, s'_{\Delta'}) \sim \mathcal{J}_{k'}(\cdot, \cdot | s_g, a_g, s_{\Delta'}, a_{\Delta'})} \frac{\sum_{i \in \Delta'} r_l^{\tilde{a}}(s'_i, s'_g)}{k'} \right| \\
&\leq \|r_l^{\tilde{a}}\|_{\infty} \mathbb{E}_{s'_g \sim \sum_{s'_{\Delta \cup \Delta'} \in \mathcal{S}_l^{|\Delta \cup \Delta'|}} \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot, \cdot | s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'}) \text{TV}(\mathbb{E}_{s'_{\Delta \cup \Delta'} | s'_g} F_{s_{\Delta}}, \mathbb{E}_{s'_{\Delta \cup \Delta'} | s'_g} F_{s_{\Delta'}}) \\
&\leq 2\|r_l(\cdot, \cdot)\|_{\infty} \cdot \text{TV}(F_{z_{\Delta}}, F_{z_{\Delta'}})
\end{aligned}$$

The above derivation follows from the same argument as in the base case where  $r_l^{\tilde{a}}(s'_i, s'_g) := r_l(s'_i, s'_g, \tilde{a}_i)$  for any  $i \in \Delta \cup \Delta'$ . Similarly, if (I) < (II), an analogous argument that replaces  $\tilde{a}_{\Delta}$  with  $\hat{a}_{\Delta}$  yields the same result.

Therefore, by the induction hypothesis, the claim is proven.  $\square$

**Remark E.6.** Given a joint transition probability function  $\mathcal{J}_{|\Delta \cup \Delta'|}$  as defined in Theorem E.4, we can recover the transition function for a single agent  $i \in \Delta \cup \Delta'$  given by  $\mathcal{J}_1$  using the law of total probability and the conditional independence between  $s_i$  and  $s_g \cup s_{[n] \setminus i}$  in Equation (28). This characterization is crucial in Theorem E.7 and Theorem E.5:

$$\mathcal{J}_1(\cdot | s'_g, s_g, a_g, s_i, a_i) = \sum_{s'_{\Delta \cup \Delta' \setminus i} \sim \mathcal{S}_l^{|\Delta \cup \Delta'|-1}} \mathcal{J}_{|\Delta \cup \Delta'|}(s'_{\Delta \cup \Delta' \setminus i} \circ s'_i | s'_g, s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'}) \quad (28)$$

Here, we use a conditional independence property proven in Theorem E.10.

**Lemma E.7.** The TV-distance between the next-step expected empirical distribution functions is bounded by the TV-distance between the existing empirical distribution functions.

$$\begin{aligned}
& \text{TV} \left( \mathbb{E}_{s'_{\Delta \cup \Delta'} | s'_g \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot | s'_g, s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'})} F_{s'_{\Delta}}, \mathbb{E}_{s'_{\Delta \cup \Delta'} | s'_g \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot | s'_g, s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'})} F_{s'_{\Delta'}} \right) \\
&\leq \text{TV}(F_{z_{\Delta}}, F_{z_{\Delta'}})
\end{aligned}$$

*Proof.* From the definition of total variation distance, we have:

$$\begin{aligned}
& \text{TV} \left( \mathbb{E}_{s'_{\Delta \cup \Delta'} | s'_g \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot | s'_g, s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'})} F_{s'_{\Delta}}, \mathbb{E}_{s'_{\Delta \cup \Delta'} | s'_g \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot | s'_g, s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'})} F_{s'_{\Delta'}} \right) \\
&= \frac{1}{2} \sum_{x \in \mathcal{S}_l} \left| \mathbb{E}_{s'_{\Delta \cup \Delta'} | s'_g \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot | s'_g, s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'})} F_{s'_{\Delta}}(x) - \mathbb{E}_{s'_{\Delta \cup \Delta'} | s'_g \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot | s'_g, s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'})} F_{s'_{\Delta'}}(x) \right| \\
&= \frac{1}{2} \sum_{x \in \mathcal{S}_l} \left| \frac{1}{k} \sum_{i \in \Delta} \mathcal{J}_1(x | s'_g, s_g, a_g, s_i, a_i) - \frac{1}{k'} \sum_{i \in \Delta'} \mathcal{J}_1(x | s'_g, s_g, a_g, s_i, a_i) \right| \\
&= \frac{1}{2} \sum_{x \in \mathcal{S}_l} \left| \frac{1}{k} \sum_{a_l \in \mathcal{A}_l} \sum_{s_l \in \mathcal{S}_l} \sum_{i \in \Delta} \mathcal{J}_1(x | s'_g, s_g, a_g, s_i, a_i) \mathbb{1}_{\{s_i = s_l\}} \right. \\
&\quad \left. - \frac{1}{k'} \sum_{a_l \in \mathcal{A}_l} \sum_{s_l \in \mathcal{S}_l} \sum_{i \in \Delta'} \mathcal{J}_1(x | s'_g, s_g, a_g, s_i, a_i) \mathbb{1}_{\{s_i = s_l\}} \right| \\
&= \frac{1}{2} \sum_{x \in \mathcal{S}_l} \left| \frac{1}{k} \sum_{a_l \in \mathcal{A}_l} \sum_{s_l \in \mathcal{S}_l} \sum_{i \in \Delta} \mathcal{J}_1(x | s'_g, s_g, a_g, \cdot, \cdot) \vec{\mathbb{1}}_{\{s_i = s_l\}} \right. \\
&\quad \left. - \frac{1}{k'} \sum_{a_l \in \mathcal{A}_l} \sum_{s_l \in \mathcal{S}_l} \sum_{i \in \Delta'} \mathcal{J}_1(x | s'_g, s_g, a_g, \cdot, \cdot) \vec{\mathbb{1}}_{\{s_i = s_l\}} \right|
\end{aligned}$$

The second equality uses Theorem E.9. The third equality uses the property that each local agent can only have one state/action pair, and the fourth equality vectorizes the indicator variables. Then,

$$\begin{aligned}
& \text{TV} \left( \mathbb{E}_{s'_{\Delta \cup \Delta'}} | s'_g \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot | s'_g, s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'}) F_{s'_{\Delta}} , \mathbb{E}_{s'_{\Delta \cup \Delta'}} | s'_g \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot | s'_g, s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'}) F_{s'_{\Delta'}} \right) \\
& \leq \frac{1}{2} \sum_{x \in \mathcal{S}_l} \sum_{a_l \in \mathcal{A}_l} \sum_{s_l \in \mathcal{S}_l} \left| \frac{1}{k} \sum_{i \in \Delta} \mathcal{J}_1(x | s'_g, s_g, a_g, \cdot, \cdot) \vec{\mathbb{1}}_{\{s_i = s_l\}} - \frac{1}{k'} \sum_{i \in \Delta'} \mathcal{J}_1(x | s'_g, s_g, a_g, \cdot, \cdot) \vec{\mathbb{1}}_{\{s_i = s_l\}} \right| \\
& = \frac{1}{2} \sum_{a_l \in \mathcal{A}_l} \sum_{s_l \in \mathcal{S}_l} \left\| \frac{1}{k} \sum_{i \in \Delta} \mathcal{J}_1(\cdot | s'_g, s_g, a_g, \cdot, \cdot) \vec{\mathbb{1}}_{\{s_i = s_l\}} - \frac{1}{k'} \sum_{i \in \Delta'} \mathcal{J}_1(\cdot | s'_g, s_g, a_g, \cdot, \cdot) \vec{\mathbb{1}}_{\{s_i = s_l\}} \right\|_1 \\
& \leq \frac{1}{2} \sum_{a_l \in \mathcal{A}_l} \sum_{s_l \in \mathcal{S}_l} \left\| \frac{1}{k} \sum_{i \in \Delta} \vec{\mathbb{1}}_{\{s_i = s_l\}} - \frac{1}{k'} \sum_{i \in \Delta'} \vec{\mathbb{1}}_{\{s_i = s_l\}} \right\|_1 \cdot \|\mathcal{J}_1(\cdot | s'_g, s_g, a_g, \cdot, \cdot)\|_1 \\
& = \frac{1}{2} \sum_{a_l \in \mathcal{A}_l} \sum_{s_l \in \mathcal{S}_l} \left\| \frac{1}{k} \sum_{i \in \Delta} \mathbb{1}_{\{s_i = s_l\}} - \frac{1}{k'} \sum_{i \in \Delta'} \mathbb{1}_{\{s_i = s_l\}} \right\|_1 \\
& = \frac{1}{2} \sum_{(s_l, a_l) \in \mathcal{S}_l \times \mathcal{A}_l} \left| \frac{1}{k} \sum_{i \in \Delta} \mathbb{1}_{\{s_i = s_l\}} - \frac{1}{k'} \sum_{i \in \Delta'} \mathbb{1}_{\{s_i = s_l\}} \right| \\
& = \frac{1}{2} \sum_{z_l \in \mathcal{S}_l \times \mathcal{A}_l} |F_{z_\Delta}(z_l) - F_{z_{\Delta'}}(z_l)| \\
& := \text{TV}(F_{z_\Delta}, F_{z_{\Delta'}})
\end{aligned}$$

The first inequality uses the triangle inequality. The second inequality and fifth equality follow from Hölder's inequality and the sum of the probabilities from the stochastic transition function  $\mathcal{J}_1$  being equal to 1. The sixth equality uses Fubini-Tonelli's theorem which applies as the total variation distance measure is bounded from above by 1. The final equality recovers the total variation distance for the variable  $z = (s, a) \in \mathcal{S}_l \times \mathcal{A}_l$  across agents  $\Delta$  and  $\Delta'$ , which proves the claim.  $\square$

Next, recall the data processing inequality.

**Lemma E.8** (Data Processing Inequality.). *Let  $A$  and  $B$  be random variables over some domain  $S$ . Let  $f$  be some function (not necessarily deterministically) mapping from  $S$  to any codomain  $T$ . Then, every  $f$ -divergence  $\chi$  satisfies*

$$\chi(f(A) \| f(B)) \leq \chi(A \| B)$$

**Remark:** We show an analog of the data processing inequality. Under this lens, Theorem E.7 saturates the data-processing relation for the TV distance in our multi-agent setting.

**Lemma E.9.**

$$\mathbb{E}_{s'_{\Delta \cup \Delta'}} | s'_g \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot | s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'}) F_{s'_{\Delta}}(x) = \frac{1}{k} \sum_{i \in \Delta} \mathcal{J}_1(x | s'_g, s_g, a_g, s_i, a_i)$$

*Proof.* By expanding on the definition of the empirical distribution function  $F_{s_{\Delta'}}(x) = \frac{1}{|\Delta|} \sum_{i \in \Delta} \mathbb{1}\{s_i = x\}$ , we have:

$$\begin{aligned}
& \mathbb{E}_{s_{\Delta \cup \Delta'}} | s'_g \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot | s'_g, s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'}) F_{s'_{\Delta}}(x) \\
& = \frac{1}{k} \sum_{i \in \Delta} \mathbb{E}_{s_{\Delta \cup \Delta'}} | s'_g \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot | s'_g, s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'}) \mathbb{1}\{s'_i = x\} \\
& = \frac{1}{k} \sum_{i \in \Delta} \mathbb{E}_{s_{\Delta \cup \Delta'}} | s'_g \sim \mathcal{J}_1(\cdot | s'_g, s_g, a_g, s_i, a_i) \mathbb{1}\{s'_i = x\} \\
& = \frac{1}{k} \sum_{i \in \Delta} \mathcal{J}_1(x | s'_g, s_g, a_g, s_i, a_i)
\end{aligned}$$

The second equality follows from the conditional independence of  $s'_i$  from  $s_{\Delta \cup \Delta' \setminus i}$  from Theorem E.10, and the final equality uses the fact that the expectation of an indicator random variable is the probability distribution function of the random variable.  $\square$

**Lemma E.10.** *The distribution  $s'_{\Delta \cup \Delta' \setminus i} | s'_g, s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'}$  is conditionally independent to the distribution  $s'_i | s'_g, s_g, a_g, s_i, a_i$  for any  $i \in \Delta \cup \Delta'$ .*

*Proof.* We direct the interested reader to the Bayes-Ball theorem in Shachter [2013] for proving conditional independence. For ease of exposition, we restate the two rules in Shachter [2013] that introduce the notion of  $d$ -separations which implies conditional independence. Suppose we have a causal graph  $G = (V, E)$  where the vertex set  $V = [p]$  for  $p \in \mathbb{N}$  is a set of variables and the edge set  $E \subseteq 2^V$  denotes dependence through connectivity. Then, the two rules that establish  $d$ -separations are as follows:

1. For  $x, y \in V$ , we say that  $x, y$  are  $d$ -connected if there exists a path  $(x, \dots, y)$  that can be traced without traversing a pair of arrows that point at the same vertex.
2. We say that  $x, y \in V$  are  $d$ -connected, conditioned on a set of variables  $Z \subseteq V$ , if there is a path  $(x, \dots, y)$  that does not contain an event  $z \in Z$  that can be traced without traversing a pair of arrows that point at the same vertex.

If  $x, y \in V$  is not  $d$ -connected through any such path, then  $x, y$  is  $d$ -separated, which implies conditional independence.

Let  $Z = \{s'_g, s_g, a_g, s_i, a_i\}$  be the set of variables we condition on. Then, the below figure demonstrates the causal graph for the events of interest.

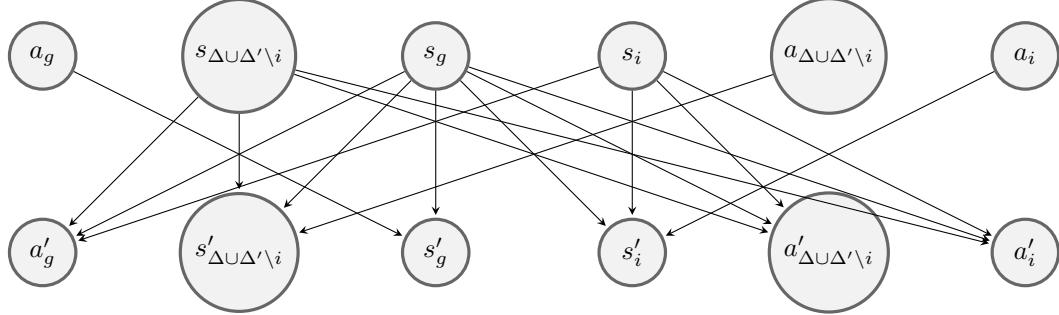


Figure 5: Causal graph to demonstrate the dependencies between variables.

1. Observe that all paths (through undirected edges) stemming from  $s'_{\Delta \cup \Delta' \setminus i} \rightarrow s_{\Delta \cup \Delta' \setminus i}$  pass through  $s_i \in Z$  which is blocked.
2. All other paths from  $s'_{\Delta \cup \Delta' \setminus i}$  to  $s'_i$  pass through  $s_g \cup s'_g \in Z$ .

Therefore,  $s'_{\Delta \cup \Delta' \setminus i}$  and  $s'_i$  are  $d$ -separated by  $Z$ . Hence, by Shachter [2013],  $s'_{\Delta \cup \Delta' \setminus i}$  and  $s'_i$  are conditionally independent.  $\square$

**Lemma E.11.** *For any joint transition probability function  $\mathcal{J}_k$  on  $s_g \in \mathcal{S}_g, s_\Delta \in \mathcal{S}_l^k, a_\Delta \in \mathcal{A}_l^k$  where  $|\Delta| = k$ , given by  $\mathcal{J}_k : \mathcal{S}_g \times \mathcal{S}_l^{|\Delta|} \times \mathcal{S}_g \times \mathcal{A}_g \times \mathcal{S}_l^{|\Delta|} \times \mathcal{A}_l^{|\Delta|} \rightarrow [0, 1]$ , we have:*

$$\begin{aligned} & \mathbb{E}_{(s'_g, s'_\Delta) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_\Delta, a_\Delta)} \left[ \mathbb{E}_{(s''_g, s''_\Delta) \sim \mathcal{J}_k(\cdot, \cdot | s'_g, a_g, s'_\Delta, a_\Delta)} \max_{a''_g \in \mathcal{A}_g, a''_\Delta \in \mathcal{A}_l^k} \hat{Q}_k^T(s''_g, s''_\Delta, a''_g, a''_\Delta) \right] \\ &= \mathbb{E}_{(s''_g, s''_\Delta) \sim \mathcal{J}_k^2(\cdot, \cdot | s_g, a_g, s_\Delta, a_\Delta)} \max_{a''_g \in \mathcal{A}_g} \hat{Q}_k^T(s''_g, s''_\Delta, a''_g, a''_\Delta) \end{aligned}$$

*Proof.* By expanding the expectations:

$$\begin{aligned}
& \mathbb{E}_{(s'_g, s'_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \left[ \mathbb{E}_{(s''_g, s''_{\Delta}) \sim \mathcal{J}_k(\cdot, \cdot | s'_g, a_g, s'_{\Delta}, a_{\Delta})} \max_{a''_g \in \mathcal{A}_g, a''_{\Delta} \in \mathcal{A}_l^k} \hat{Q}_k^T(s''_g, s''_{\Delta}, a''_g, a''_{\Delta}) \right] \\
&= \sum_{(s'_g, s'_{\Delta}) \in \mathcal{S}_g \times \mathcal{S}_l^{|\Delta|}} \sum_{(s''_g, s''_{\Delta}) \in \mathcal{S}_g \times \mathcal{S}_l^{|\Delta|}} \mathcal{J}_k[s'_g, s'_{\Delta}, s_g, a_g, s_{\Delta}, a_{\Delta}] \mathcal{J}_k[s''_g, s''_{\Delta}, s'_g, a_g, s'_{\Delta}, a_{\Delta}] \\
&\quad \max_{a''_g \in \mathcal{A}_g, a''_{\Delta} \in \mathcal{A}_l^k} \hat{Q}_k^T(s''_g, s''_{\Delta}, a''_g, a''_{\Delta}) \\
&= \sum_{(s''_g, s''_{\Delta}) \in \mathcal{S}_g \times \mathcal{S}_l^{|\Delta|}} \mathcal{J}_k^2[s''_g, s''_{\Delta}, s_g, a_g, s_{\Delta}, a_{\Delta}] \max_{a''_g \in \mathcal{A}_g, a''_{\Delta} \in \mathcal{A}_l^k} \hat{Q}_k^T(s''_g, s''_{\Delta}, a''_g, a''_{\Delta}) \\
&= \mathbb{E}_{(s''_g, s''_{\Delta}) \sim \mathcal{J}_k^2(\cdot, \cdot | s_g, a_g, s_{\Delta}, a_{\Delta})} \max_{a''_g \in \mathcal{A}_g, a''_{\Delta} \in \mathcal{A}_l^k} \hat{Q}_k^T(s''_g, s''_{\Delta}, a''_g, a''_{\Delta})
\end{aligned}$$

In the second equality, the right-stochasticity of  $\mathcal{J}_k$  implies the right-stochasticity of  $\mathcal{J}_k^2$ .

Further, observe that  $\mathcal{J}_k[s'_g, s'_{\Delta}, s_g, a_g, s_{\Delta}, a_{\Delta}] \mathcal{J}_k[s''_g, s''_{\Delta}, s'_g, a_g, s'_{\Delta}, a_{\Delta}]$  denotes the probability of the transitions  $(s_g, s_{\Delta}) \rightarrow (s'_g, s'_{\Delta}) \rightarrow (s''_g, s''_{\Delta})$  with actions  $a_g, a_{\Delta}$  at each step, where the joint state evolution is governed by  $\mathcal{J}_k$ . Thus,

$$\sum_{(s'_g, s'_{\Delta}) \in \mathcal{S}_g \times \mathcal{S}_l^{|\Delta|}} \mathcal{J}_k[s'_g, s'_{\Delta}, s_g, a_g, s_{\Delta}, a_{\Delta}] \mathcal{J}_k[s''_g, s''_{\Delta}, s'_g, a_g, s'_{\Delta}, a_{\Delta}] = \mathcal{J}_k^2[s''_g, s''_{\Delta}, s_g, a_g, s_{\Delta}, a_{\Delta}]$$

since  $\sum_{(s'_g, s'_{\Delta}) \in \mathcal{S}_g \times \mathcal{S}_l^{|\Delta|}} \mathcal{J}_k[s'_g, s'_{\Delta}, s_g, a_g, s_{\Delta}, a_{\Delta}] \mathcal{J}_k[s''_g, s''_{\Delta}, s'_g, a_g, s'_{\Delta}, a_{\Delta}]$  is the stochastic probability function corresponding to the two-step evolution of the joint states from  $(s_g, s_{\Delta})$  to  $(s''_g, s''_{\Delta})$  under actions  $a_g, a_{\Delta}$ . This can be thought of as an analogously to the fact that a 1 on the  $(i, j)$ 'th entry on the square of a 0/1 adjacency matrix of a graph represents the fact that there is a path of distance 2 between vertices  $i$  and  $j$ .

Finally, the third equality recovers the definition of the expectation, with respect to the joint probability function  $\mathcal{J}_k^2$ .  $\square$

We next show  $\lim_{T \rightarrow \infty} \mathbb{E}_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^T(s_g, s_{\Delta}, a_g, a_{\Delta}) = \mathbb{E}_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^*(s_g, s_{\Delta}, a_g, a_{\Delta}) = Q^*(s, a)$ .

**Lemma E.12.** *The  $Q^*$  function is the average value of each of the  $\binom{n}{k}$  sub-sampled  $\hat{Q}_k^*$ -functions.*

$$Q^*(s, a) - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^T(s_g, s_{\Delta}, a_g, a_{\Delta}) \leq \gamma^T \cdot \frac{\tilde{r}}{1 - \gamma}$$

*Proof.* We bound the differences between  $\hat{Q}_k^T$  at each Bellman iteration of our approximation to  $Q^*$ .

Note that:

$$\begin{aligned}
& Q^*(s, a) - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^T(s_g, s_{\Delta}, a_g, a_{\Delta}) \\
&= \mathcal{T}Q^*(s, a) - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \hat{\mathcal{T}}_k \hat{Q}_k^{T-1}(s_g, s_{\Delta}, a_g, a_{\Delta}) \\
&= r_{[n]}(s_g, s_{[n]}, a_g) + \gamma \mathbb{E}_{\substack{s'_g \sim P_g(\cdot | s_g, a_g), \\ s'_i \sim P_l(\cdot | s_i, a_i, s_g), \forall i \in [n]}} \max_{a'_g \in \mathcal{A}_g, a'_{[n]} \in \mathcal{A}_l^n} Q^*(s', a') \\
&\quad - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \left[ r_{\Delta}(s_g, s_{\Delta}, a_g, a_{\Delta}) + \gamma \mathbb{E}_{\substack{s'_g \sim P_g(\cdot | s_g, a_g), \\ s'_i \sim P_l(\cdot | s_i, a_i, s_g), \forall i \in \Delta}} \max_{a'_g \in \mathcal{A}_g, a'_{\Delta} \in \mathcal{A}_l^k} Q_k^T(s'_g, s'_{\Delta}, a'_g, a'_{\Delta}) \right]
\end{aligned}$$

Next, observe that  $r_{[n]}(s_g, s_{[n]}, a_g, a_{[n]}) = \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} r_{[\Delta]}(s_g, s_{\Delta}, a_g, a_{\Delta})$ .

To prove this, we write:

$$\begin{aligned}
\frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} r_{[\Delta]}(s_g, s_\Delta, a_g, a_\Delta) &= \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} (r_g(s_g, a_g) + \frac{1}{k} \sum_{i \in \Delta} r_l(s_i, a_i, s_g)) \\
&= r_g(s_g, a_g) + \frac{\binom{n-1}{k-1}}{k \binom{n}{k}} \sum_{i \in [n]} r_l(s_i, a_i, s_g) \\
&= r_g(s_g, a_g) + \frac{1}{n} \sum_{i \in [n]} r_l(s_i, a_i, s_g) := r_{[n]}(s_g, s_{[n]}, a_g, a_{[n]})
\end{aligned}$$

In the second equality, we reparameterized the sum to count the number of times each  $r_l(s_i, s_g)$  was added for each  $i \in \Delta$ . To do this, we count the number of  $(k-1)$  other agents that could form a  $k$ -tuple with agent  $i$ , and there are  $(n-1)$  candidates from which we choose the  $(k-1)$  agents. In the last equality, we expanded and simplified the binomial coefficients.

So, we have that:

$$\begin{aligned}
&\sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left[ Q^*(s, a) - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^T(s_g, s_{[n]}, a_g, a_\Delta) \right] \\
&= \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left[ \mathcal{T}Q^*(s, a) - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \hat{\mathcal{T}}_k \hat{Q}_k^{T-1}(s_g, s_\Delta, a_g, a_\Delta) \right] \\
&= \gamma \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \left[ \mathbb{E}_{\substack{s'_g \sim P(\cdot|s_g, a_g), \\ s'_i \sim P_l(\cdot|s_i, a_i, s_g), \forall i \in [n]}} \max_{a' \in \mathcal{A}} Q^*(s', a') \right. \\
&\quad \left. - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \mathbb{E}_{\substack{s'_g \sim P_g(\cdot|s_g, a_g), \\ s'_i \sim P_l(\cdot|s_i, a_i, s_g), \forall i \in \Delta}} \max_{\substack{a'_g \in \mathcal{A}_g, \\ a'_\Delta \in \mathcal{A}_l^k}} \hat{Q}_k^{T-1}(s'_g, s'_\Delta, a'_g, a'_\Delta) \right] \\
&= \gamma \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathbb{E}_{\substack{s'_g \sim P_g(\cdot|s_g, a_g), \\ s'_i \sim P_l(\cdot|s_i, a_i, s_g), \forall i \in [n]}} \left[ \max_{a' \in \mathcal{A}} Q^*(s', a') \right. \\
&\quad \left. - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \max_{\substack{a'_g \in \mathcal{A}_g, a'_\Delta \in \mathcal{A}_l^k}} \hat{Q}_k^{T-1}(s'_g, s'_\Delta, a'_g, a'_\Delta) \right] \\
&\leq \gamma \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathbb{E}_{\substack{s'_g \sim P_g(\cdot|s_g, a_g), \\ s'_i \sim P_l(\cdot|s_i, a_i, s_g), \forall i \in [n]}} \max_{\substack{a'_g \in \mathcal{A}_g, a'_{[n]} \in \mathcal{A}_l^n}} \left[ Q^*(s', a') - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^{T-1}(s'_g, s'_\Delta, a'_g, a'_\Delta) \right] \\
&\leq \gamma \sup_{(s',a') \in \mathcal{S} \times \mathcal{A}} \left[ Q^*(s', a') - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^{T-1}(s'_g, s'_\Delta, a'_g, a'_\Delta) \right]
\end{aligned}$$

We justify the first inequality by noting the general property that for positive vectors  $v, v'$  for which  $v \succeq v'$  which follows from the triangle inequality:

$$\begin{aligned}
\left\| v - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} v' \right\|_\infty &\geq \left\| \|v\|_\infty - \left\| \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} v' \right\|_\infty \right\| \\
&= \|v\|_\infty - \left\| \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} v' \right\|_\infty \\
&\geq \|v\|_\infty - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \|v'\|_\infty
\end{aligned}$$

Thus, applying this bound recursively, we get:

$$\begin{aligned}
Q^*(s, a) & - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^T(s_g, s_\Delta, a_g, a_\Delta) \\
& \leq \gamma^T \sup_{(s', a') \in \mathcal{S} \times \mathcal{A}} \left[ Q^*(s', a') - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^0(s'_g, s'_\Delta, a'_g, a'_\Delta) \right] \\
& = \gamma^T \sup_{(s', a') \in \mathcal{S} \times \mathcal{A}} Q^*(s', a') \\
& = \gamma^T \cdot \frac{\tilde{r}}{1 - \gamma}
\end{aligned}$$

The first inequality follows from the  $\gamma$ -contraction property of the update procedure, and the ensuing equality follows from our bound on the maximum possible value of  $Q$  from Theorem C.4 and noting that  $\hat{Q}_k^0 := 0$ .

Therefore, as  $T \rightarrow \infty$ ,

$$Q^*(s, a_g) - \frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^T(s_g, s_\Delta, a_g) \rightarrow 0,$$

which proves the lemma.  $\square$

**Corollary E.13.** Since  $\frac{1}{\binom{n}{k}} \sum_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^T(s_g, s_\Delta, a_g, a_\Delta) := \mathbb{E}_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^T(s_g, s_\Delta, a_g, a_\Delta)$ , we therefore get:

$$Q^*(s, a) - \lim_{T \rightarrow \infty} \mathbb{E}_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^T(s_g, s_\Delta, a_g, a_\Delta) \leq \lim_{T \rightarrow \infty} \gamma^T \cdot \frac{\tilde{r}}{1 - \gamma}$$

Consequently,

$$Q^*(s, a) - \mathbb{E}_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^*(s_g, s_\Delta, a_g, a_\Delta) \leq 0$$

Further, we have that

$$Q^*(s, a) - \mathbb{E}_{\Delta \in \binom{[n]}{k}} \hat{Q}_k^*(s_g, s_\Delta, a_g, a_\Delta) = 0,$$

since  $\hat{Q}_k^*(s_g, s_\Delta, a_g, a_\Delta) \leq Q^*(s, a)$ .

**Lemma E.14.** The absolute difference between the expected maximums between  $\hat{Q}_k$  and  $\hat{Q}_{k'}$  is atmost the maximum of the absolute difference between  $\hat{Q}_k$  and  $\hat{Q}_{k'}$ , where the expectations are taken over any joint distributions of states  $\mathcal{J}$ , and the maximums are taken over the actions.

$$\begin{aligned}
& \left| \mathbb{E}_{(s'_g, s'_{\Delta \cup \Delta'}) \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot, \cdot | s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'})} \left[ \max_{a'_g \in \mathcal{A}_g, a'_\Delta \in \mathcal{A}_l^k} \hat{Q}_k^T(s'_g, s'_\Delta, a'_g, a'_\Delta) \right. \right. \\
& \quad \left. \left. - \max_{a'_g \in \mathcal{A}_g, a'_{\Delta'} \in \mathcal{A}_l^{k'}} \hat{Q}_{k'}^T(s'_g, s'_{\Delta'}, a'_g, a'_{\Delta'}) \right] \right| \\
& \leq \max_{a'_g \in \mathcal{A}_g, a'_{\Delta \cup \Delta'} \in \mathcal{A}_l^{|\Delta \cup \Delta'|}} \left| \mathbb{E}_{(s'_g, s'_{\Delta \cup \Delta'}) \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot, \cdot | s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'})} \left[ \hat{Q}_k^T(s'_g, s'_\Delta, a'_g, a'_\Delta) \right. \right. \\
& \quad \left. \left. - \hat{Q}_{k'}^T(s'_g, s'_{\Delta'}, a'_g, a'_{\Delta'}) \right] \right|
\end{aligned}$$

*Proof.* Denote:

$$a_g^*, a_\Delta^* := \arg \max_{\substack{a'_g \in \mathcal{A}_g, \\ a'_\Delta \in \mathcal{A}_l^k}} \hat{Q}_k^T(s'_g, F_{s'_\Delta}, a'_g, a'_\Delta)$$

$$\tilde{a}_g^*, \tilde{a}_{\Delta'}^* := \arg \max_{\substack{a'_g \in \mathcal{A}_g, \\ a'_{\Delta'} \in \mathcal{A}_l^{k'}}} \hat{Q}_{k'}^T(s'_g, F_{s'_{\Delta'}}, a'_g, a'_{\Delta'})$$

We extend  $a_{\Delta}^*$  to  $a_{\Delta'}^*$  by letting  $a_{\Delta' \setminus \Delta}^*$  be the corresponding  $\Delta \setminus \Delta'$  variables from  $\tilde{a}_{\Delta}^*$  in  $\mathcal{A}_t^{|\Delta' \setminus \Delta|}$ . For the remainder of this proof, we adopt the shorthand  $\mathbb{E}_{s'_g, s'_{\Delta \cup \Delta'}}$  to refer to  $\mathbb{E}_{(s'_g, s'_{\Delta \cup \Delta'}) \sim \mathcal{J}_{|\Delta \cup \Delta'|}(\cdot, \cdot | s_g, a_g, s_{\Delta \cup \Delta'}, a_{\Delta \cup \Delta'})}$ . Then, if  $\mathbb{E}_{s'_g, s'_{\Delta \cup \Delta'}} \max_{a'_g \in \mathcal{A}_g, a'_{\Delta} \in \mathcal{A}_t^k} \hat{Q}_k^T(s'_g, F_{z'_\Delta}, a'_g) - \mathbb{E}_{s'_g, s'_{\Delta \cup \Delta'}} \max_{a'_g \in \mathcal{A}_g, a'_{\Delta'} \in \mathcal{A}_t^k} \hat{Q}_{k'}^T(s'_g, F_{z'_{\Delta'}}, a'_g) > 0$ , we have:

$$\begin{aligned} & \left| \mathbb{E}_{s'_g, s'_{\Delta \cup \Delta'}} \max_{a'_g \in \mathcal{A}_g, a'_{\Delta} \in \mathcal{A}_t^k} \hat{Q}_k^T(s'_g, F_{z'_\Delta}, a'_g) - \mathbb{E}_{s'_g, s'_{\Delta \cup \Delta'}} \max_{a'_g \in \mathcal{A}_g, a'_{\Delta'} \in \mathcal{A}_t^k} \hat{Q}_{k'}^T(s'_g, F_{z'_{\Delta'}}, a'_g) \right| \\ &= \mathbb{E}_{s'_g, s'_{\Delta \cup \Delta'}} \hat{Q}_k^T(s'_g, s'_{\Delta}, a_g^*, a_{\Delta}^*) - \mathbb{E}_{s'_g, s'_{\Delta \cup \Delta'}} \hat{Q}_{k'}^T(s'_g, s'_{\Delta'}, \tilde{a}_g^*, \tilde{a}_{\Delta'}^*) \\ &\leq \mathbb{E}_{s'_g, s'_{\Delta \cup \Delta'}} \hat{Q}_k^T(s'_g, s'_{\Delta}, a_g^*, a_{\Delta}^*) - \mathbb{E}_{s'_g, s'_{\Delta \cup \Delta'}} \hat{Q}_{k'}^T(s'_g, s'_{\Delta'}, a_g^*, a_{\Delta'}^*) \\ &\leq \max_{\substack{a'_g \in \mathcal{A}_g, \\ a_{\Delta \cup \Delta'} \in \mathcal{A}_t^{|\Delta \cup \Delta'|}}} \left| \mathbb{E}_{s'_g, s'_{\Delta \cup \Delta'}} \hat{Q}_k^T(s'_g, s'_{\Delta}, a'_g, a'_{\Delta}) - \mathbb{E}_{s'_g, s'_{\Delta \cup \Delta'}} \hat{Q}_{k'}^T(s'_g, s'_{\Delta'}, a'_g, a'_{\Delta'}) \right| \end{aligned}$$

We observe that if the opposite inequality holds (i.e.,  $\mathbb{E}_{s'_g, s'_{\Delta \cup \Delta'}} \max_{a'_g \in \mathcal{A}_g, a'_{\Delta} \in \mathcal{A}_t^k} \hat{Q}_k^T(s'_g, s'_{\Delta}, a'_g, a'_{\Delta}) - \mathbb{E}_{s'_g, s'_{\Delta \cup \Delta'}} \max_{a'_g \in \mathcal{A}_g, a'_{\Delta'} \in \mathcal{A}_t^k} \hat{Q}_{k'}^T(s'_g, s'_{\Delta'}, a'_g, a'_{\Delta'}) < 0$ ), an analogous argument by replacing  $a_g^*$  with  $\tilde{a}_g^*$  and  $a_{\Delta}^*$  with  $\tilde{a}_{\Delta}^*$  yields an identical bound.  $\square$

**Lemma E.15.** Suppose  $z, z' \geq 1$ . Consider functions  $\Gamma : \Theta_1 \times \Theta_2 \times \dots \times \Theta_z \times \Theta^* \rightarrow \mathbb{R}$  and  $\Gamma' : \Theta'_1 \times \Theta'_2 \times \dots \times \Theta'_{z'} \times \Theta^* \rightarrow \mathbb{R}$ , where  $\Theta_1, \dots, \Theta_z$  and  $\Theta'_1, \dots, \Theta'_{z'}$  are finite sets. Consider a probability distribution function  $\mu_{\Theta_i}$  for  $i \in [z]$  and  $\mu'_{\Theta_i}$  for  $i \in [z']$ . Then:

$$\begin{aligned} & \left| \mathbb{E}_{\theta_1 \sim \mu_{\Theta_1}} \max_{\substack{\theta^* \in \Theta^* \\ \theta_z \sim \mu_{\Theta_z}}} \Gamma(\theta_1, \dots, \theta_z, \theta^*) - \mathbb{E}_{\theta_1 \sim \mu'_{\Theta_1}} \max_{\substack{\theta^* \in \Theta^* \\ \theta_{z'} \sim \mu'_{\Theta_{z'}}}} \Gamma'(\theta_1, \dots, \theta_{z'}, \theta^*) \right| \\ &\leq \max_{\theta^* \in \Theta^*} \left| \mathbb{E}_{\theta_1 \sim \mu_{\Theta_1}} \Gamma(\theta_1, \dots, \theta_z, \theta^*) - \mathbb{E}_{\theta_1 \sim \mu'_{\Theta_1}} \Gamma'(\theta_1, \dots, \theta_{z'}, \theta^*) \right| \end{aligned}$$

*Proof.* Let  $\hat{\theta}^* := \arg \max_{\theta^* \in \Theta^*} \Gamma(\theta_1, \dots, \theta_z, \theta^*)$  and  $\tilde{\theta}^* := \arg \max_{\theta^* \in \Theta^*} \Gamma'(\theta_1, \dots, \theta_{z'}, \theta^*)$ .

If  $\mathbb{E}_{\theta_1 \sim \mu_{\Theta_1}, \dots, \theta_z \sim \mu_{\Theta_z}} \max_{\theta^* \in \Theta^*} \Gamma(\theta_1, \dots, \theta_z, \theta^*) - \mathbb{E}_{\theta_1 \sim \mu'_{\Theta_1}, \dots, \theta_{z'} \sim \mu'_{\Theta_{z'}}} \max_{\theta^* \in \Theta^*} \Gamma'(\theta_1, \dots, \theta_{z'}, \theta^*) > 0$ , then:

$$\begin{aligned} & \left| \mathbb{E}_{\theta_1 \sim \mu_{\Theta_1}} \max_{\substack{\theta^* \in \Theta^* \\ \theta_z \sim \mu_{\Theta_z}}} \Gamma(\theta_1, \dots, \theta_z, \theta^*) - \mathbb{E}_{\theta_1 \sim \mu'_{\Theta_1}} \max_{\substack{\theta^* \in \Theta^* \\ \theta_{z'} \sim \mu'_{\Theta_{z'}}}} \Gamma'(\theta_1, \dots, \theta_{z'}, \theta^*) \right| \\ &= \mathbb{E}_{\theta_1 \sim \mu_{\Theta_1}} \Gamma(\theta_1, \dots, \theta_z, \hat{\theta}^*) - \mathbb{E}_{\theta_1 \sim \mu'_{\Theta_1}} \Gamma'(\theta_1, \dots, \theta_{z'}, \tilde{\theta}^*) \\ &\leq \mathbb{E}_{\theta_1 \sim \mu_{\Theta_1}} \Gamma(\theta_1, \dots, \theta_z, \hat{\theta}^*) - \mathbb{E}_{\theta_1 \sim \mu'_{\Theta_1}} \Gamma'(\theta_1, \dots, \theta_{z'}, \hat{\theta}^*) \\ &\leq \max_{\theta^* \in \Theta^*} \left| \mathbb{E}_{\theta_1 \sim \mu_{\Theta_1}} \Gamma(\theta_1, \dots, \theta_z, \theta^*) - \mathbb{E}_{\theta_1 \sim \mu'_{\Theta_1}} \Gamma'(\theta_1, \dots, \theta_{z'}, \theta^*) \right| \end{aligned}$$

Here, we replace each  $\theta^*$  with the maximizers of their corresponding terms, and upper bound them by the maximizer of the larger term. Next, we replace  $\hat{\theta}^*$  in both expressions with the maximizer choice  $\theta^*$  from  $\Theta^*$ , and further bound the expression by its absolute value.

If  $\mathbb{E}_{\theta_1 \sim \mu_{\Theta_1}, \dots, \theta_z \sim \mu_{\Theta_z}} \max_{\theta^* \in \Theta^*} \Gamma(\theta_1, \dots, \theta_z, \theta^*) - \mathbb{E}_{\theta_1 \sim \mu'_{\Theta_1}, \dots, \theta_{z'} \sim \mu'_{\Theta_{z'}}} \max_{\theta^* \in \Theta^*} \Gamma'(\theta_1, \dots, \theta_{z'}, \theta^*)$  is negative, then an analogous argument that replaces  $\hat{\theta}^*$  with  $\tilde{\theta}^*$  yields the same result.  $\square$

## F Bounding Total Variation Distance

As  $|\Delta| \rightarrow n$ , we prove that the total variation (TV) distance between the empirical distribution of  $z_\Delta$  and  $z_{[n]}$  goes to 0. Here, recall that  $z_i \in \mathcal{Z} = \mathcal{S}_l \times \mathcal{A}_l$ , and  $z_\Delta = \{z_i : i \in \Delta\}$  for  $\Delta \in \binom{[n]}{k}$ . Before bounding the total variation distance between  $F_{z_\Delta}$  and  $F_{z_{[n]}}$ , we first introduce Lemma C.5 of Anand and Qu [2024] which can be viewed as a generalization of the Dvoretzky-Kiefer-Wolfowitz concentration inequality, for sampling without replacement. We first make an important remark.

**Remark F.1.** First, observe that if  $\Delta$  is an independent random variable uniformly supported on  $\binom{[n]}{k}$ , then  $s_\Delta$  and  $a_\Delta$  are also independent random variables uniformly supported on the global state  $\binom{s_{[n]}}{k}$  and the global action  $\binom{a_{[n]}}{k}$ . To see this, let  $\psi_1 : [n] \rightarrow \mathcal{S}_l$  where  $\psi_1(i) = s_i$  and  $\xi_1 : [n] \rightarrow \mathcal{A}_l$  where  $\xi_1(i) = a_i$ . This naturally extends to  $\psi_k : [n]^k \rightarrow \mathcal{S}_l^k$ , where  $\psi_k(i_1, \dots, i_k) = (s_{i_1}, \dots, s_{i_k})$  and  $\xi_k : [n]^k \rightarrow \mathcal{A}_l^k$ , where  $\xi_k(i_1, \dots, i_k) = (a_{i_1}, \dots, a_{i_k})$  for all  $k \in [n]$ . Then, the independence of  $\Delta$  implies the independence of the generated  $\sigma$ -algebra. Further,  $\psi_k$  and  $\xi_k$  (which are a Lebesgue measurable function of a  $\sigma$ -algebra) are sub-algebras, implying that  $s_\Delta$  and  $a_\Delta$  must also be independent random variables.

For reference, we present the multidimensional Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Dvoretzky et al. [1956], Massart [1990], Naaman [2021]) which bounds the difference between an empirical distribution function for a set  $B_\Delta$  and  $B_{[n]}$  when each element of  $\Delta$  for  $|\Delta| = k$  is sampled uniformly at random from  $[n]$  with replacement.

**Theorem F.2** (Multi-dimensional Dvoretzky-Kiefer-Wolfowitz (DFW) inequality [Dvoretzky et al., 1956, Naaman, 2021]). *Suppose  $B \subset \mathbb{R}^d$  and  $\epsilon > 0$ . If  $\Delta \subseteq [n]$  is sampled uniformly with replacement, then*

$$\Pr \left[ \sup_{x \in B} \left| \frac{1}{|\Delta|} \sum_{i \in \Delta} \mathbb{1}\{B_i = x\} - \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{B_i = x\} \right| < \epsilon \right] \geq 1 - d(n+1)e^{-2|\Delta|\epsilon^2}.$$

Lemma C.5 of Anand and Qu [2024] generalizes the DKW inequality for sampling without replacement:

**Lemma F.3** (Sampling without replacement analogue of the DKW inequality, Lemma C.5 in Anand and Qu [2024]). *Consider a finite population  $\mathcal{X} = (x_1, \dots, x_n) \in \mathcal{B}_l^n$  where  $\mathcal{B}_l$  is a finite set. Let  $\Delta \subseteq [n]$  be a random sample of size  $k$  chosen uniformly and without replacement. Then, for all  $x \in \mathcal{B}_l$ :*

$$\Pr \left[ \sup_{x \in \mathcal{B}_l} \left| \frac{1}{|\Delta|} \sum_{i \in \Delta} \mathbb{1}\{x_i = x\} - \frac{1}{n} \sum_{i \in [n]} \mathbb{1}\{x_i = x\} \right| < \epsilon \right] \geq 1 - 2|\mathcal{B}_l|e^{-\frac{2|\Delta|n\epsilon^2}{n-|\Delta|+1}}$$

**Lemma F.4.** *With probability atleast  $1 - 2|\mathcal{S}_l||\mathcal{A}_l|e^{-\frac{8kn\epsilon^2}{n-k+1}}$ ,*

$$\text{TV}(F_{z_\Delta}, F_{z_{[n]}}) \leq \epsilon$$

*Proof.* Recall that  $z := (s_l, a_l) \in \mathcal{S}_l \times \mathcal{A}_l$ . From Theorem F.3, substituting  $\mathcal{B}_l = \mathcal{Z}_l$  yields:

$$\begin{aligned} \Pr \left[ \sup_{z_l \in \mathcal{Z}_l} |F_{z_\Delta}(z_l) - F_{z_{[n]}}(z_l)| \leq 2\epsilon \right] &\geq 1 - 2|\mathcal{Z}_l|e^{-\frac{8kn\epsilon^2}{n-k+1}} \\ &= 1 - 2|\mathcal{S}_l||\mathcal{A}_l|e^{-\frac{8kn\epsilon^2}{n-k+1}}, \end{aligned}$$

which yields the proof.  $\square$

We now present an alternate bound for the total variation distance, where the distance *actually goes* to 0 as  $|\Delta| \rightarrow n$ . For this, we use the fact that the total variation distance between two product distributions is subadditive.

**Lemma F.5** (Lemma B.8.1 of Ghosal and van der Vaart [2017]. Subadditivity of TV distance for Product Distributions). *Let  $P$  and  $Q$  be product distributions over some domain  $S$ . Let  $\alpha_1, \dots, \alpha_d$  be the marginal distributions of  $P$  and  $\beta_1, \dots, \beta_q$  be the marginal distributions of  $Q$ . Then,*

$$\|P - Q\|_1 \leq \sum_{i=1}^d \|\alpha_i - \beta_i\|_1$$

**Lemma F.6** (KL-divergence decays too slowly).

$$\text{TV}(F_{z_\Delta}, F_{z_{[n]}}) \leq \sqrt{1 - \frac{|\Delta|}{n}}$$

*Proof.* By the symmetry of the total variation distance metric, we have that

$$\text{TV}(F_{z_{[n]}}, F_{z_\Delta}) = \text{TV}(F_{z_\Delta}, F_{z_{[n]}}).$$

From the Bretagnolle-Huber inequality Tsybakov [2008] we have that

$$\text{TV}(f, g) = \sqrt{1 - e^{-D_{\text{KL}}(f\|g)}}.$$

Here,  $D_{\text{KL}}(f\|g)$  is the Kullback-Leibler (KL) divergence metric between probability distributions  $f$  and  $g$  over the sample space, which we denote by  $\mathcal{X}$  and is given by

$$D_{\text{KL}}(f\|g) := \sum_{x \in \mathcal{X}} f(x) \ln \frac{f(x)}{g(x)} \quad (29)$$

Thus, from Equation (29):

$$\begin{aligned} D_{\text{KL}}(F_{z_\Delta}\|F_{z_{[n]}}) &= \sum_{z \in \mathcal{Z}_l} \left( \frac{1}{|\Delta|} \sum_{i \in \Delta} \mathbb{1}\{z_i = z\} \right) \ln \frac{n \sum_{i \in \Delta} \mathbb{1}\{z_i = z\}}{|\Delta| \sum_{i \in [n]} \mathbb{1}\{z_i = z\}} \\ &= \frac{1}{|\Delta|} \sum_{z \in \mathcal{Z}_l} \left( \sum_{i \in \Delta} \mathbb{1}\{z_i = z\} \right) \ln \frac{n}{|\Delta|} \\ &\quad + \frac{1}{|\Delta|} \sum_{z \in \mathcal{Z}_l} \left( \sum_{i \in \Delta} \mathbb{1}\{z_i = z\} \right) \ln \frac{\sum_{i \in \Delta} \mathbb{1}\{z_i = z\}}{\sum_{i \in [n]} \mathbb{1}\{z_i = z\}} \\ &= \ln \frac{n}{|\Delta|} + \frac{1}{|\Delta|} \sum_{z \in \mathcal{Z}_l} \left( \sum_{i \in \Delta} \mathbb{1}\{z_i = z\} \right) \ln \frac{\sum_{i \in \Delta} \mathbb{1}\{z_i = z\}}{\sum_{i \in [n]} \mathbb{1}\{z_i = z\}} \\ &\leq \ln \left( \frac{n}{|\Delta|} \right) \end{aligned}$$

In the third line, we note that  $\sum_{z \in \mathcal{Z}_l} \sum_{i \in \Delta} \mathbb{1}\{z_i = z\} = |\Delta|$  since each local agent contained in  $\Delta$  must have some state/action pair contained in  $\mathcal{Z}_l$ . In the last line, we note that  $\sum_{i \in \Delta} \mathbb{1}\{z_i = z\} \leq \sum_{i \in [n]} \mathbb{1}\{z_i = z\}$ , For all  $z \in \mathcal{Z}_l$ , and thus the summation of logarithmic terms in the third line is negative.

Finally, using this bound in the Bretagnolle-Huber inequality yields the lemma.  $\square$

**Corollary F.7.** From Theorem F.6, setting  $\Delta = [n]$  also recovers  $\text{TV}(F_{z_\Delta}, F_{z_{[n]}}) = 0$ .

**Theorem F.8.** With probability atleast  $1 - \delta$  for  $\delta \in (0, 1)^2$ :

$$\left| \hat{Q}_k^*(s_g, F_{s_\Delta}, a_g, F_{a_\Delta}) - \hat{Q}_n^*(s_g, F_{s_{[n]}}, a_g, F_{a_{[n]}}) \right| \leq \frac{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}}{1 - \gamma} \sqrt{\frac{n - k + 1}{8kn}} \cdot \|r_l(\cdot, \cdot)\|_\infty$$

*Proof.* From combining the total variation distance bound in Theorem F.4 and the Lipschitz continuity bound in Theorem E.3 with  $\sum_{t=0}^T \gamma^T \leq \frac{1}{1-\gamma}$  for  $\gamma \in (0, 1)$ , we have:

$$\Pr \left[ \left| \hat{Q}_k^*(s_g, F_{s_\Delta}, a_g, F_{a_\Delta}) - \hat{Q}_n^*(s_g, F_{s_{[n]}}, a_g, F_{a_{[n]}}) \right| \leq \frac{2\epsilon}{1 - \gamma} \cdot \|r_l(\cdot, \cdot)\|_\infty \right] \geq 1 - 2|\mathcal{S}_l||\mathcal{A}_l|e^{-\frac{8kn\epsilon^2}{n-k+1}}$$

Then, reparameterizing  $1 - 2|\mathcal{S}_l||\mathcal{A}_l|e^{-\frac{8kn\epsilon^2}{n-k+1}}$  into  $1 - \delta$  to get  $\epsilon = \sqrt{\frac{n-k+1}{8kn} \ln \left( \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta} \right)}$  gives that with probability at least  $1 - \delta$ ,

$$\left| \hat{Q}_k^*(s_g, F_{s_\Delta}, a_g, F_{a_\Delta}) - \hat{Q}_n^*(s_g, F_{s_{[n]}}, a_g, F_{a_{[n]}}) \right| \leq \frac{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}}{1 - \gamma} \sqrt{\frac{n - k + 1}{8kn}} \cdot \|r_l(\cdot, \cdot)\|_\infty,$$

proving the claim.  $\square$

## G Using the Performance Difference Lemma

In general, convergence analysis requires the guarantee that a stationary optimal policy exists. Fortunately, when working with the empirical distribution function, the existence of a stationary optimal policy is guaranteed when the state/action spaces are finite or countably infinite. However, lifting the knowledge of states onto the continuous empirical distribution function space, and designing a policy on the lifted space is still analytically challenging. To circumvent this, Gu et al. [2021] creates lifted  $\epsilon$ -nets and does kernel regression to obtain convergence guarantees. Moreover, our result has a similar flavor to MDPs with dynamic exogenous inputs from learning theory, [Dietterich et al., 2018, Foster et al., 2022, Anand and Qu, 2024], wherein our subsampling algorithm treats each sampled state as an endogenous state.

Here, our analytic approach bears a stark difference, wherein we analyze the *sampled* structure of the mean-field empirical distribution function, rather than studying the structure of the lifted space. For this, we leverage the classical Performance Difference Lemma, which we restate below for completeness.

**Lemma G.1** (Performance Difference Lemma, Kakade and Langford [2002]). *Given policies  $\pi_1, \pi_2$ , with corresponding value functions  $V^{\pi_1}, V^{\pi_2}$ :*

$$V^{\pi_1}(s) - V^{\pi_2}(s) = \frac{1}{1-\gamma} \mathbb{E}_{\substack{s' \sim d_s^{\pi_1} \\ a' \sim \pi_1(\cdot|s')}} [A^{\pi_2}(s', a')]$$

Here,  $A^{\pi_2}(s', a') := Q^{\pi_2}(s', a') - V^{\pi_2}(s')$  and  $d_s^{\pi_1}(s') = (1-\gamma) \sum_{h=0}^{\infty} \gamma^h \Pr_h^{\pi_1}[s', s]$  where  $\Pr_h^{\pi_1}[s', s]$  is the probability of  $\pi_1$  reaching state  $s'$  at time step  $h$  starting from state  $s$ .

We denote our learned policy  $\pi_{k,m}^{\text{est}}$  where:

$$\pi_{k,m}^{\text{est}}(s_g, s_{[n]}) = (\pi_{k,m}^{g,\text{est}}(s_g, s_{[n]}), \pi_{k,m}^{1,\text{est}}(s_g, s_{[n]}), \dots, \pi_{k,m}^{n,\text{est}}(s_g, s_{[n]})) \in \mathcal{P}(\mathcal{A}_g) \times \mathcal{P}(\mathcal{A}_l)^n,$$

where  $\pi_{k,m}^{g,\text{est}}(s_g, s_{[n]}) = \hat{\pi}_{k,m}^{g,\text{est}}(s_g, s_u, F_{s_{\Delta \setminus u}})$  is the global agent's action and  $\pi_{k,m}^{i,\text{est}}(s_g, s_{[n]}) := \hat{\pi}_{k,m}^{i,\text{est}}(s_g, s_i, F_{s_{\Delta_i}})$  is the action of the  $i$ 'th local agent. Here,  $\Delta_i$  is a random variable supported on  $\binom{[n] \setminus i}{k-1}$ ,  $u$  is a random variable uniformly distributed on  $[n]$ , and  $\Delta$  is a random variable uniformly distributed on  $[n] \setminus u$ . Then, denote the optimal policy  $\pi^*$  given by

$$\pi^*(s) = (\pi_g^*(s_g, s_{[n]}), \pi_1^*(s_g, s_{[n]}), \dots, \pi_n^*(s_g, s_{[n]})) \in \mathcal{P}(\mathcal{A}_g) \times \mathcal{P}(\mathcal{A}_l)^n,$$

where  $\pi_g^*(s_g, s_{[n]})$  is the global agent's action, and  $\pi_i^*(s_g, s_{[n]}) := \pi^*(s_g, s_i, F_{s_{\Delta_i}})$  is the action of the  $i$ 'th local agent. Next, in order to compare the difference in the performance of  $\pi^*(s)$  and  $\pi_{k,m}^{\text{est}}(s_g, s_{[n]})$ , we define the value function of a policy  $\pi$  to be the infinite-horizon  $\gamma$ -discounted rewards, (denoted by  $V^\pi$ ) as follows:

**Definition G.2.** The value function  $V^\pi : \mathcal{S} \rightarrow \mathbb{R}$  of a given policy  $\pi$ , for  $\mathcal{S} := \mathcal{S}_g \times \mathcal{S}_l^n$  is:

$$V^\pi(s) = \mathbb{E}_{a(t) \sim \pi(\cdot|s(t))} \left[ \sum_{t=0}^{\infty} \gamma^t r(s(t), a(t)) | s(0) = s \right]. \quad (30)$$

**Theorem G.3.** For the optimal policy  $\pi^*$  and the learned policy  $\pi_{k,m}^{\text{est}}$ , for any state  $s_0 \in \mathcal{S}$ , we have:

$$V^{\pi^*}(s_0) - V^{\pi_{k,m}^{\text{est}}}(s_0) \leq \frac{\tilde{r}}{(1-\gamma)^2} \sqrt{\frac{n-k+1}{2nk}} \sqrt{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}} + 2\epsilon_{k,m} + \frac{2\tilde{r}}{(1-\gamma)^2} |\mathcal{A}_g| k^{|\mathcal{A}_l|} \delta$$

*Proof.* Applying the performance difference lemma to the policies gives us:

$$\begin{aligned} V^{\pi^*}(s_0) - V^{\pi_{k,m}^{\text{est}}}(s_0) &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi_{k,m}^{\text{est}}}} \mathbb{E}_{a \sim \pi_{k,m}^{\text{est}}(\cdot|s)} [V^{\pi^*}(s) - Q^{\pi^*}(s, a)] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi_{k,m}^{\text{est}}}} \left[ \mathbb{E}_{a' \sim \pi^*(\cdot|s)} Q^{\pi^*}(s, a') - \mathbb{E}_{a \sim \pi_{k,m}^{\text{est}}(\cdot|s)} Q^{\pi^*}(s, a) \right] \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi_{k,m}^{\text{est}}}} \left[ Q^{\pi^*}(s, \pi^*(\cdot|s)) - \mathbb{E}_{a \sim \pi_{k,m}^{\text{est}}(\cdot|s)} Q^{\pi^*}(s, a) \right] \end{aligned}$$

Next, by the law of total expectation,

$$\begin{aligned} & \mathbb{E}_{a \sim \pi_{k,m}^{\text{est}}(\cdot|s)} [Q^*(s, a)] \\ &= \sum_{\Delta \in \binom{[n]}{k}} \sum_{\substack{\Delta^1, \dots, \Delta^n: \\ \Delta^i \in \binom{[n] \setminus i}{k-1}}} \frac{1}{\binom{n}{k} \binom{n-1}{k-1}^n} Q^*(s, \hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_\Delta})_g, \hat{\pi}_{k,m}^{\text{est}}(s_g, s_1, F_{s_{\Delta^1}}), \dots, \hat{\pi}_{k,m}^{\text{est}}(s_g, s_n, F_{s_{\Delta^n}})) \end{aligned}$$

Therefore,

$$\begin{aligned} & Q^{\pi^*}(s, \pi^*(\cdot|s)) - \mathbb{E}_{a \sim \pi_{k,m}^{\text{est}}(\cdot|s)} Q^*(s, a) \\ &= \sum_{\Delta \in \binom{[n]}{k}} \sum_{\Delta^1 \in \binom{[n] \setminus 1}{k-1}} \sum_{\Delta^2 \in \binom{[n] \setminus 2}{k-1}} \dots \sum_{\Delta^n \in \binom{[n] \setminus n}{k-1}} \frac{1}{\binom{n}{k} \binom{n-1}{k-1}^n} \left( Q^*(s, \pi^*(\cdot|s)) \right. \\ &\quad \left. - Q^*(s, \hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_\Delta})_g, \hat{\pi}_{k,m}^{\text{est}}(s_g, s_1, F_{s_{\Delta^1}}), \dots, \hat{\pi}_{k,m}^{\text{est}}(s_g, s_n, F_{s_{\Delta^n}})) \right) \end{aligned}$$

Therefore, by grouping the equations above, we have:

$$\begin{aligned} & V^{\pi^*}(s_0) - V^{\tilde{\pi}_{k,m}^{\text{est}}}(s_0) \\ &\leq \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi_{k,m}^{\text{est}}}} \mathbb{E}_{a \sim \tilde{\pi}_{k,m}^{\text{est}}(\cdot|s)} \sum_{\Delta \in \binom{[n]}{k}} \sum_{\substack{\Delta^i \in \binom{[n] \setminus i}{k-1}, \\ \forall i \in [n]}} \frac{1}{n \binom{n}{k} \binom{n-1}{k-1}^n} \left( \sum_{i \in [n]} \left| Q^*(s, \pi^*(s)) \right. \right. \\ &\quad \left. \left. - \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \pi^*(s)_g, \{\pi^*(s)_j\}_{j \in \{i, \Delta^i\}}) \right| \right. \\ &\quad \left. + \frac{1}{\binom{n}{k} \binom{n-1}{k-1}^n} \sum_{i \in [n]} \left| \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_\Delta})_g, \{\hat{\pi}_{k,m}^{\text{est}}(s_g, s_j, \Delta^j)\}_{j \in \{i, \Delta^i\}}) \right. \right. \\ &\quad \left. \left. - Q^*(s, \hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_\Delta})_g, \{\hat{\pi}_{k,m}^{\text{est}}(s_g, s_j, F_{s_{\Delta^j}})\}_{j \in [n]}) \right| \right) \end{aligned}$$

Theorem G.4 shows a uniform bound on

$$Q^*(s, \pi^*(\cdot|s)) - Q^*(s, \hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_\Delta})_g, \hat{\pi}_{k,m}^{\text{est}}(s_g, s_1, F_{s_{\Delta^1}}), \dots, \hat{\pi}_{k,m}^{\text{est}}(s_g, s_n, F_{s_{\Delta^n}}))$$

(independent of  $\Delta^i, \forall i \in [n]$ ), allowing the sums and the counts in the denominator will cancel out. Observe that  $\hat{\pi}_{k,m}^{\text{est}} : \mathcal{S} \rightarrow \mathcal{A}$  and  $\pi^* : \mathcal{S} \rightarrow \mathcal{A}$  are deterministic functions. Therefore, denote  $a = \pi^*(s)$ . Then, from Theorem G.6,

$$\begin{aligned} & \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi_{k,m}^{\text{est}}}} \mathbb{E}_{a \sim \pi_{k,m}^{\text{est}}(\cdot|s)} \sum_{\Delta \in \binom{[n]}{k}} \sum_{\Delta^1 \in \binom{[n] \setminus 1}{k-1}} \dots \sum_{\Delta^n \in \binom{[n] \setminus n}{k-1}} \frac{1}{n \binom{n}{k} \binom{n-1}{k-1}^n} \sum_{i \in [n]} \left| Q^*(s, \pi^*(s)) \right. \\ &\quad \left. - \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \pi^*(s)_g, \{\pi^*(s)_j\}_{j \in \{i, \Delta^i\}}) \right| \\ &= \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi_{k,m}^{\text{est}}}} \sum_{\Delta \in \binom{[n]}{k}} \sum_{\substack{\Delta^1, \dots, \Delta^n: \\ \Delta^i \in \binom{[n] \setminus i}{k-1}}} \frac{1}{\binom{n}{k} \binom{n-1}{k-1}^n} \sum_{i \in [n]} \left| \hat{Q}_n^*(s_g, F_{s_{[n]}})_g, \hat{\pi}_n^*(s_g, F_{s_{[n]}})_g, \hat{\pi}_n^*(s_g, F_{s_{[n]}})_{1:n} \right. \\ &\quad \left. - \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \pi^*(s)_g, \{\pi^*(s)_j\}_{j \in \{i, \Delta^i\}}) \right| \\ &\leq \frac{\tilde{r}}{(1-\gamma)^2} \sqrt{\frac{n-k+1}{8nk}} \sqrt{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}} + \epsilon_{k,m} + \frac{\tilde{r}}{(1-\gamma)^2} |\mathcal{A}_g| k^{|\mathcal{A}_l|} \delta \end{aligned}$$

Similarly, from Theorem G.7, we have that

$$\begin{aligned} & \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_{s_0}^{\pi_{k,m}^{\text{est}}}} \sum_{\Delta \in \binom{[n]}{k}} \sum_{\Delta^1 \in \binom{[n] \setminus 1}{k-1}} \cdots \sum_{\Delta^n \in \binom{[n] \setminus n}{k-1}} \frac{\frac{1}{n \binom{n}{k}}}{\binom{n-1}{k-1}^n} \left| \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \right. \\ & \left. \hat{\pi}_{k,m}^{\text{est}}(s_g, s_{\Delta})_g, \{\hat{\pi}_{k,m}^{\text{est}}(s_g, s_j, F_{s_{\Delta^j}})\}_{j \in [n]}) - Q^*(s, \hat{\pi}_{k,m}^{\text{est}}(s_g, s_{\Delta})_g, \{\hat{\pi}_{k,m}^{\text{est}}(s_g, s_j, F_{s_{\Delta^j}})\}_{j \in [n]}) \right| \\ & \leq \frac{\tilde{r}}{(1-\gamma)^2} \sqrt{\frac{n-k+1}{8nk}} \sqrt{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}} + \epsilon_{k,m} + \frac{\tilde{r}}{(1-\gamma)^2} |\mathcal{A}_g| k^{|\mathcal{A}_l|} \delta \end{aligned}$$

Hence, combining the above inequalities, we get:

$$V^{\pi^*}(s_0) - V^{\pi_{k,m}^{\text{est}}}(s_0) \leq \frac{2\tilde{r}}{(1-\gamma)^2} \sqrt{\frac{n-k+1}{8nk}} \sqrt{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}} + 2\epsilon_{k,m} + \frac{2\tilde{r}}{(1-\gamma)^2} |\mathcal{A}_g| k^{|\mathcal{A}_l|} \delta$$

which yields the claim. We defer parameter optimization to Theorem G.8.  $\square$

**Lemma G.4** (Uniform Bound on  $Q^*$  with different actions). *For all  $s \in \mathcal{S}, \Delta \in \binom{[n]}{k}$  and  $\Delta^i \in \binom{[n] \setminus i}{k-1}$  for  $i \in [n]$ , we have:*

$$\begin{aligned} & Q^*(s, \pi^*(\cdot|s)) - Q^*(s, \hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_{\Delta}})_g, \{\hat{\pi}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}})\}_{i \in [n]}) \\ & \leq \frac{1}{n} \sum_{i \in [n]} \left| Q^*(s, \pi^*(s)) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \pi^*(s)_g, \{\pi^*(s)_j\}_{j \in \{i, \Delta^i\}}) \right| \\ & \quad + \left| \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_{\Delta}})_g, \{\hat{\pi}_{k,m}^{\text{est}}(s_g, s_j, F_{s_{\Delta^j}})\}_{j \in \{i, \Delta^i\}}) \right. \\ & \quad \left. - Q^*(s, \hat{\pi}_{k,m}^{\text{est}}(s_g, s_{\Delta})_g, \{\hat{\pi}_{k,m}^{\text{est}}(s_g, s_j, F_{s_{\Delta^j}})\}_{j \in [n]}) \right| \end{aligned}$$

*Proof.* Observe that

$$\begin{aligned} & Q^*(s, \pi^*(\cdot|s)) - Q^*(s, \hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_{\Delta}})_g, \{\hat{\pi}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}})\}_{i \in [n]}) \\ & \leq \frac{1}{n} \sum_{i \in [n]} \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_{\Delta}})_g, \hat{\pi}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}), \{\hat{\pi}_{k,m}^{\text{est}}(s_g, s_j, F_{s_{\Delta^j}})\}_{j \in \Delta^i}) \\ & \quad - \frac{1}{n} \sum_{i \in [n]} \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_{\Delta}})_g, \hat{\pi}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}), \{\hat{\pi}_{k,m}^{\text{est}}(s_g, s_j, F_{s_{\Delta^j}})\}_{j \in \Delta^i}) \\ & \quad + \frac{1}{n} \sum_{i \in [n]} \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \pi^*(s)_g, \pi^*(s)_i, \{\pi^*(s)_j\}_{j \in \Delta^k}) \\ & \quad - \frac{1}{n} \sum_{i \in [n]} \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \pi^*(s)_g, \pi^*(s)_i, \{\pi^*(s)_j\}_{j \in \Delta^k}) \\ & \leq \left| Q^*(s, \pi^*(s)) - \frac{1}{n} \sum_{i \in [n]} \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \pi^*(s)_g, \pi^*(s)_i, \{\pi^*(s)_j\}_{j \in \Delta^i}) \right| \\ & \quad + \left| \frac{1}{n} \sum_{i \in [n]} \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_{\Delta}})_g, \{\hat{\pi}_{k,m}^{\text{est}}(s_g, s_j, F_{s_{\Delta^j}})\}_{j \in \{i, \Delta^i\}}) \right. \\ & \quad \left. - Q^*(s, \hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_{\Delta}})_g, \{\hat{\pi}_{k,m}^{\text{est}}(s_g, s_j, F_{s_{\Delta^j}})\}_{j \in [n]}) \right| \\ & \leq \frac{1}{n} \sum_{i \in [n]} \left| Q^*(s, \pi^*(s)) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \pi^*(s)_g, \{\pi^*(s)_j\}_{j \in \{i, \Delta^i\}}) \right| \end{aligned}$$

$$+ \frac{1}{n} \sum_{i \in [n]} \left| \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \hat{\pi}_{k,m}^{\text{est}}(s_g, F_{s_{\Delta}})_g, \{\hat{\pi}_{k,m}^{\text{est}}(s_g, s_j, \Delta^j)\}_{j \in \{i, \Delta^i\}}) \right. \\ \left. - Q^*(s, \hat{\pi}_{k,m}^{\text{est}}(s_g, s_{\Delta}), \{\hat{\pi}_{k,m}^{\text{est}}(s_g, s_j, F_{s_{\Delta^j}})\}_{j \in [n]}) \right|,$$

which proves the claim.  $\square$

**Lemma G.5.** Fix  $s \in \mathcal{S} := \mathcal{S}_g \times \mathcal{S}_l^n$ . For each  $j \in [n]$ , suppose we are given  $T$ -length sequences of random variables  $\{\Delta_1^j, \dots, \Delta_T^j\}_{j \in [n]}$ , distributed uniformly over the support  $\binom{[n] \setminus j}{k-1}$ . Further, suppose we are given a fixed sequence  $\delta_1, \dots, \delta_T$ , where each  $\delta_t \in (0, 1]$  for  $t \in [T]$ . Let

$$\Phi_{k,t} = \frac{1}{1-\gamma} \sqrt{\frac{n-k+1}{8kn} \ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta_t}}.$$

Then, for each action  $a = (a_g, a_{[n]}) = \pi^*(s)$ , for  $t \in [T]$  and  $j \in [n]$ , define deviation events  $B_t^{a_g, a_j, \Delta_t^j}$  such that:

$$B_t^{a_g, a_j, \Delta_t^j} := \left\{ \left| Q^*(s_g, s_j, F_{z_{[n] \setminus j}}, a_j, a_g) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_j, F_{z_{\Delta_t^j}}, a_j, a_g) \right| > \Phi_{k,t} \cdot \|r_l(\cdot, \cdot)\|_{\infty} + \epsilon_{k,m} \right\} \quad (31)$$

For  $i \in [T]$ , we define bad-events  $B_t$  (which is a union over each deviation event):

$$B_t = \bigcup_{j \in [n]} \bigcup_{a_g \in \mathcal{A}_g} \bigcup_{a_{j, \Delta_t^j} \in \mathcal{A}_l^k} B_t^{a_g, a_j, \Delta_t^j}$$

Next, denote  $B = \bigcup_{i=1}^T B_i$ . Then, the probability that no bad event  $B_t$  occurs is:

$$\Pr[\bar{B}] := 1 - \Pr[B] \geq 1 - |\mathcal{A}_g| k^{|\mathcal{A}_l|} \sum_{i=1}^T \delta_i$$

*Proof.*

$$\begin{aligned} & |Q^*(s_g, s_j, F_{z_{[n] \setminus j}}, a_g, a_j) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_j, F_{z_{\Delta_t^j}}, a_g, a_j)| \\ &= \left| Q^*(s_g, s_j, F_{z_{[n] \setminus j}}, a_g, a_j) - \hat{Q}_k^*(s_g, s_j, F_{z_{\Delta_t^j}}, a_g, a_j) \right. \\ &\quad \left. + \hat{Q}_k^*(s_g, s_j, F_{z_{\Delta_t^j}}, a_g, a_j) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_j, F_{z_{\Delta_t^j}}, a_g, a_j) \right| \\ &\leq \left| Q^*(s_g, s_j, F_{z_{[n] \setminus j}}, a_g, a_j) - \hat{Q}_k^*(s_g, s_j, F_{z_{\Delta_t^j}}, a_g, a_j) \right| \\ &\quad + \left| \hat{Q}_k^*(s_g, s_j, F_{z_{\Delta_t^j}}, a_g, a_j) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_j, F_{z_{\Delta_t^j}}, a_g, a_j) \right| \\ &\leq \left| Q^*(s_g, s_j, F_{z_{[n] \setminus j}}, a_g, a_j) - \hat{Q}_k^*(s_g, s_j, F_{z_{\Delta_t^j}}, a_g, a_j) \right| + \epsilon_{k,m} \end{aligned}$$

The first inequality above follows from the triangle inequality, and the second inequality uses

$$\begin{aligned} & \left| \hat{Q}_k^*(s_g, s_j, F_{z_{\Delta_t^j}}, a_g, a_j) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_j, F_{z_{\Delta_t^j}}, a_g, a_j) \right| \\ &\leq \left\| \hat{Q}_k^*(s_g, s_j, F_{z_{\Delta_t^j}}, a_g, a_j) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_j, F_{z_{\Delta_t^j}}, a_g, a_j) \right\|_{\infty} \\ &\leq \epsilon_{k,m}, \end{aligned}$$

where the  $\epsilon_{k,m}$  follows from Theorem 4.3. From Theorem F.8, we have that with probability at least  $1 - \delta_t$ ,

$$\begin{aligned} & \left| Q^*(s_g, s_j, F_{z_{[n] \setminus j}}, a_g, a_j) - \hat{Q}_k^*(s_g, s_j, F_{z_{\Delta_t^j}}, a_g, a_j) \right| \leq \Phi_{k,t} \cdot \|r_l(\cdot, \cdot)\|_{\infty} \\ &= \frac{1}{1-\gamma} \sqrt{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta_t}} \sqrt{\frac{n-k+1}{8kn}} \cdot \|r_l(\cdot, \cdot)\|_{\infty} \end{aligned}$$

So, event  $B_t^{a_g, a_{j, \Delta_t^j}}$  occurs with probability atmost  $\delta_t$ .

Here, observe that if we union bound across all the events parameterized by the empirical distributions of  $a_g, a_{j, \Delta_t^j} \in \mathcal{A}_g \times \mathcal{A}_l^k$  given by  $F_{a_{\Delta_t^j}}$ , this also forms a covering of the choice of variables  $F_{s_{\Delta_t^j}}$  by agents  $j \in [m]$ , and therefore across all choices of  $\Delta_t^1, \dots, \Delta_t^n$  (subject to the permutation invariance of local agents) for a fixed  $t$ .

Thus, from the union bound, we get:

$$\Pr[B_t] \leq \sum_{a_g \in \mathcal{A}_g} \sum_{F_{a_{\Delta_t}} \in \mu_k(\mathcal{A}_l)} \Pr[B_t^{a_g, a_{1, \Delta_t^1}}] \leq |\mathcal{A}_g| k^{|\mathcal{A}_l|} \Pr[B_t^{a_g, a_{1, \Delta_t^1}}]$$

Applying the union bound again proves the lemma:

$$\Pr[\bar{B}] \geq 1 - \sum_{t=1}^T \Pr[B_t] \geq 1 - |\mathcal{A}_g| k^{|\mathcal{A}_l|} \sum_{t=1}^T \delta_t,$$

which proves the claim.  $\square$

**Lemma G.6.** For any arbitrary distribution  $\mathcal{D}$  of states  $\mathcal{S} := \mathcal{S}_g \times \mathcal{S}_l^n$ , for any  $\Delta^i \in \binom{[n] \setminus i}{k-1}$  for  $i \in [n]$  and for  $\delta \in (0, 1]$  we have:

$$\begin{aligned} & \mathbb{E}_{s \sim \mathcal{D}} \left[ \sum_{\Delta \in \binom{[n]}{k}} \sum_{\Delta^1 \in \binom{[n] \setminus 1}{k-1}} \cdots \sum_{\Delta^n \in \binom{[n] \setminus n}{k-1}} \frac{1}{\binom{n}{k}} \frac{1}{\binom{n-1}{k-1}^n} \sum_{i \in [n]} \frac{1}{n} \left| Q^*(s, \pi^*(s)) \right. \right. \\ & \quad \left. \left. - \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta_i}}, \pi^*(s)_g, \{\pi^*(s)_j\}_{j \in \{i, \Delta^i\}}) \right| \right] \\ & \leq \frac{\tilde{r}}{1-\gamma} \sqrt{\frac{n-k+1}{8nk}} \sqrt{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}} + \epsilon_{k,m} + \frac{\tilde{r}}{1-\gamma} |\mathcal{A}_g| k^{|\mathcal{A}_l|} \delta \end{aligned}$$

*Proof.* By the linearity of expectations, observe that:

$$\begin{aligned} & \mathbb{E}_{s \sim \mathcal{D}} \left[ \sum_{\Delta \in \binom{[n]}{k}} \sum_{\Delta^1 \in \binom{[n] \setminus 1}{k-1}} \cdots \sum_{\Delta^n \in \binom{[n] \setminus n}{k-1}} \frac{1}{\binom{n}{k}} \frac{1}{\binom{n-1}{k-1}^n} \sum_{i \in [n]} \frac{1}{n} \left| Q^*(s, \pi^*(s)) \right. \right. \\ & \quad \left. \left. - \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta_i}}, \pi^*(s)_g, \{\pi^*(s)_j\}_{j \in \{i, \Delta^i\}}) \right| \right] \\ & = \sum_{\Delta \in \binom{[n]}{k}} \sum_{\Delta^1 \in \binom{[n] \setminus 1}{k-1}} \cdots \sum_{\Delta^n \in \binom{[n] \setminus n}{k-1}} \frac{1}{\binom{n}{k}} \frac{1}{\binom{n-1}{k-1}^n} \sum_{i \in [n]} \frac{1}{n} \mathbb{E}_{s \sim \mathcal{D}} \left| Q^*(s, \pi^*(s)) \right. \\ & \quad \left. - \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta_i}}, \pi^*(s)_g, \{\pi^*(s)_j\}_{j \in \{i, \Delta^i\}}) \right| \end{aligned}$$

Let

$$\Phi_{k,\delta} = \sqrt{\frac{n-k+1}{8nk}} \sqrt{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}}.$$

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

$$\mathcal{I}(i, s, k, \delta) :=$$

$$\mathbb{1} \left\{ \left| Q^*(s, \pi^*(s)) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta_i}}, \pi^*(s)_g, \{\pi^*(s)_j\}_{j \in \{i, \Delta^i\}}) \right| \leq \frac{\|r_l(\cdot, \cdot)\|_\infty}{1-\gamma} \Phi_{k,\delta} + \epsilon_{k,m} \right\}$$

The expected difference between  $Q^*(s', \pi^*(s'))$  and  $\hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta_i}}, \pi^*(s)_g, \{\pi^*(s)_j\}_{j \in \{i, \Delta^i\}})$  is bounded as follows:

$$\begin{aligned} & \mathbb{E}_{s \sim \mathcal{D}} \left| Q^*(s', \pi^*(s')) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta_i}}, \pi^*(s)_g, \{\pi^*(s)_j\}_{j \in \{i, \Delta^i\}}) \right| \\ & = \mathbb{E}_{s \sim \mathcal{D}} \left[ \mathcal{I}(i, s, k, \delta) \left| Q^*(s', \pi^*(s')) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta_i}}, \pi^*(s)_g, \{\pi^*(s)_j\}_{j \in \{i, \Delta^i\}}) \right| \right] \\ & \quad + \mathbb{E}_{s \sim \mathcal{D}} \left[ (1 - \mathcal{I}(i, s, k, \delta)) \left| Q^*(s', \pi^*(s')) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta_i}}, \pi^*(s)_g, \{\pi^*(s)_j\}_{j \in \{i, \Delta^i\}}) \right| \right] \end{aligned}$$

Here, we have used the general property for a random variable  $X$  and constant  $c$  that  $\mathbb{E}[X] = \mathbb{E}[X \mathbb{1}\{X \leq c\}] + \mathbb{E}[(1 - \mathbb{1}\{X \leq c\})X]$ . Then,

$$\begin{aligned} & \mathbb{E}_{s \sim \mathcal{D}} \left| Q^*(s', \pi^*(s')) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_{i,\Delta^i}, \pi^*(s)_g, \pi^*(s)_{i,\Delta^i}) \right| \\ & \leq \frac{\tilde{r}}{1-\gamma} \sqrt{\frac{n-k+1}{8nk}} \sqrt{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}} + \epsilon_{k,m} + \frac{\tilde{r}}{1-\gamma} (1 - \mathbb{E}_{s' \sim \mathcal{D}} \mathcal{I}(s', k, \delta)) \\ & \leq \frac{\tilde{r}}{1-\gamma} \sqrt{\frac{n-k+1}{8nk}} \sqrt{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}} + \epsilon_{k,m} + \frac{\tilde{r}}{1-\gamma} |\mathcal{A}_g| k^{|\mathcal{A}_l|} \delta \\ & = \frac{\tilde{r}}{1-\gamma} \sqrt{\frac{n-k+1}{8nk}} \sqrt{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}} + \epsilon_{k,m} + \frac{\tilde{r}}{1-\gamma} |\mathcal{A}_g| k^{|\mathcal{A}_l|} \delta \end{aligned}$$

For the first term in the first inequality, we use  $\mathbb{E}[X \mathbb{1}\{X \leq c\}] \leq c$ . For the second term, we trivially bound  $Q^*(s', \pi^*(s')) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_{i,\Delta^i}, \pi^*(s)_g, \pi^*(s)_{i,\Delta^i})$  by the maximum value  $Q^*$  can take, which is  $\frac{\tilde{r}}{1-\gamma}$  by Theorem C.4. In the second inequality, we use the fact that the expectation of an indicator function is the conditional probability of the underlying event. The second inequality follows from Theorem G.5.

Since this is a uniform bound that is independent of  $\Delta^j$  for  $j \in [n]$  and  $i \in [n]$ , we have:

$$\begin{aligned} & \sum_{\Delta \in \binom{[n]}{k} \Delta^1, \dots, \Delta^n:} \sum_{\Delta^i \in \binom{[n]}{k-1} \setminus i} \frac{1/n}{\binom{n}{k} \binom{n-1}{k-1}^n} \sum_{i \in [n]} \mathbb{E}_{s \sim \mathcal{D}} \left| Q^*(s, \pi^*(s)) - \hat{Q}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta^i}}, \pi^*(s)_g, \{\pi^*(s)_j\}_{j \in \{i, \Delta^i\}}) \right| \\ & \leq \frac{\tilde{r}}{1-\gamma} \sqrt{\frac{n-k+1}{8nk}} \sqrt{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}} + \epsilon_{k,m} + \frac{\tilde{r}}{1-\gamma} |\mathcal{A}_g| k^{|\mathcal{A}_l|} \delta \end{aligned}$$

This proves the claim.  $\square$

**Corollary G.7.** *Changing the notation of the policy to  $\pi_{k,m}^{\text{est}}$  in the proofs of Theorems G.5 and G.6, such that  $a \sim \tilde{\pi}_{k,m}^{\text{est}}$  then yields that:*

$$\begin{aligned} & \mathbb{E}_{s \sim \mathcal{D}} \left[ \sum_{\Delta \in \binom{[n]}{k} \Delta^1, \dots, \Delta^n:} \sum_{\Delta^i \in \binom{[n]}{k-1} \setminus i} \frac{1}{\binom{n}{k}} \frac{1}{\binom{n-1}{k-1}^n} \sum_{i \in [n]} \frac{1}{n} \left| Q^*(s, a) - \hat{Q}_{k,m}^*(s_g, s_i, F_{s_{\Delta^i}}, a_g, a_{i,\Delta^i}) \right| \right] \\ & \leq \frac{\tilde{r}}{1-\gamma} \sqrt{\frac{n-k+1}{8nk}} \sqrt{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}} + \epsilon_{k,m} + \frac{\tilde{r}}{1-\gamma} |\mathcal{A}_g| k^{|\mathcal{A}_l|} \delta \end{aligned}$$

**Lemma G.8** (Optimizing Parameters). *With probability at least  $1 - \frac{1}{100e^k}$ ,*

$$V^{\pi^*}(s_0) - V^{\pi_{k,m}}(s_0) \leq \tilde{O} \left( \frac{1}{\sqrt{k}} \right)$$

*Proof.* Setting  $\delta = \frac{(1-\gamma)^2}{2\tilde{r}\varepsilon|\mathcal{A}_g|k^{|\mathcal{A}_l|}}$  in the bound for the optimality gap in Theorem G.3 gives:

$$\begin{aligned} V^{\pi^*}(s_0) - V^{\tilde{\pi}_{k,m}^{\text{est}}}(s_0) & \leq \frac{\tilde{r}}{(1-\gamma)^2} \sqrt{\frac{n-k+1}{2nk}} \sqrt{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}} + 2\epsilon_{k,m} + \frac{2\tilde{r}}{(1-\gamma)^2} |\mathcal{A}_g| k^{|\mathcal{A}_l|} \delta \\ & = \frac{\tilde{r}}{(1-\gamma)^2} \sqrt{\frac{n-k+1}{2nk}} \sqrt{\ln \frac{4\tilde{r}\varepsilon|\mathcal{S}_l||\mathcal{A}_l||\mathcal{A}_g|k^{|\mathcal{A}_l|}}{(1-\gamma)^2}} + \frac{1}{\varepsilon} + 2\epsilon_{k,m} \end{aligned}$$

Setting  $\varepsilon = 10\sqrt{k}$  for any  $c > 0$  recovers a decaying optimality gap of the order

$$V^{\pi^*}(s_0) - V^{\tilde{\pi}_{k,m}^{\text{est}}}(s_0) \leq \frac{\tilde{r}}{(1-\gamma)^2} \sqrt{\frac{n-k+1}{2nk}} \sqrt{\ln \frac{40\tilde{r}|\mathcal{S}_l||\mathcal{A}_l||\mathcal{A}_g|k^{|\mathcal{A}_l|+\frac{1}{2}}}{(1-\gamma)^2}} + \frac{1}{10\sqrt{k}} + 2\epsilon_{k,m}$$

Finally, using the probabilistic bound of  $\epsilon_{k,m} \leq \tilde{O}(\frac{1}{\sqrt{k}})$  from Theorem G.10 yields that with probability at least  $1 - \frac{1}{100e^k}$ ,

$$V^{\pi^*}(s_0) - V^{\tilde{\pi}_{k,m}}(s_0) \leq \tilde{O}\left(\frac{1}{\sqrt{k}}\right),$$

which proves the claim.  $\square$

## G.1 Bounding the Bellman Error

This section is devoted to the proof of Theorem 4.3.

**Theorem G.9** (Theorem 2 of Li et al. [2022]). *If  $m \in \mathbb{N}$  is the number of samples in the Bellman update, there exists a universal constant  $0 < c_0 < 2$  and a Bellman noise  $0 < \epsilon_{k,m} \leq \frac{1}{1-\gamma}$  such that*

$$\|\hat{\mathcal{T}}_{k,m}\hat{Q}_{k,m}^{\text{est}} - \hat{\mathcal{T}}_k\hat{Q}_k^*\|_\infty = \|\hat{Q}_{k,m}^{\text{est}} - \hat{Q}_k^*\|_\infty \leq \epsilon_{k,m},$$

where  $\epsilon_{k,m}$  satisfies

$$m = \frac{c_0 \cdot \tilde{r} \cdot t_{\text{cover}}}{(1-\gamma)^5 \epsilon_{k,m}^2} \log^2\left(\frac{|\mathcal{S}||\mathcal{A}|}{\rho}\right) \log\left(\frac{1}{(1-\gamma)^2}\right), \quad (32)$$

with probability at least  $1 - \rho$ , for any  $\rho \in (0, 1)$ . Here,  $t_{\text{cover}}$  stands for the cover time, which is the time taken for the trajectory to visit all state-action pairs at least once. Formally,

$$t_{\text{cover}} := \min \left\{ t \mid \min_{(s_0, a_0) \in \mathcal{S} \times \mathcal{A}} \mathbb{P}(\mathcal{B}_t | s_0, a_0) \geq \frac{1}{2} \right\}, \quad (33)$$

where  $\mathcal{B}_t$  denotes the event that all  $(s, a) \in \mathcal{S} \times \mathcal{A}$  have been visited at least once between time 0 and time  $t$ , and  $\mathbb{P}(\mathcal{B}_t | s_0, a_0)$  denotes the probability of  $\mathcal{B}_t$  conditioned on the initial state  $(s_0, a_0)$ .

**Lemma G.10.** *If  $T = \frac{2}{1-\gamma} \log \frac{\tilde{r}\sqrt{k}}{1-\gamma}$ , SUB-SAMPLE-MFQ: Learning runs in time  $\tilde{O}(T|\mathcal{S}_g|^2|\mathcal{A}_g|^2|\mathcal{A}_l|^2|\mathcal{S}_l|^2 k^{3.5+2|\mathcal{S}_l||\mathcal{A}_l|}\tilde{r})$ , while accruing a Bellman noise  $\epsilon_{k,m} \leq \tilde{O}(1/\sqrt{k})$  with probability at least  $1 - \frac{1}{100e^k}$ .*

*Proof.* We first prove that  $\|\hat{Q}_k^T - \hat{Q}_k^*\|_\infty \leq \frac{1}{\sqrt{k}}$ .

For this, it suffices to show  $\gamma^T \frac{\tilde{r}}{1-\gamma} \leq \frac{1}{\sqrt{k}} \implies \gamma^T \leq \frac{1-\gamma}{\tilde{r}\sqrt{k}}$ . Then, using  $\gamma = 1 - (1-\gamma) \leq e^{-(1-\gamma)}$ , it again suffices to show  $e^{-(1-\gamma)T} \leq \frac{1-\gamma}{\tilde{r}\sqrt{k}}$ . Taking logarithms, we have

$$\begin{aligned} \exp(-T(1-\gamma)) &\leq \frac{1-\gamma}{\tilde{r}\sqrt{k}} \\ -T(1-\gamma) &\leq \log \frac{1-\gamma}{\tilde{r}\sqrt{k}} \\ T &\geq \frac{1}{1-\gamma} \log \frac{\tilde{r}\sqrt{k}}{1-\gamma} \end{aligned}$$

Since  $T = \frac{2}{1-\gamma} \log \frac{\tilde{r}\sqrt{k}}{1-\gamma} > \frac{1}{1-\gamma} \log \frac{\tilde{r}\sqrt{k}}{1-\gamma}$ , the condition holds and  $\|\hat{Q}_k^T - \hat{Q}_k^*\|_\infty \leq \frac{1}{\sqrt{k}}$ . Then, rearranging Equation (32) and incorporating the convergence error of the  $\hat{Q}_k$ -function, one has that with probability at least  $1 - \rho$ ,

$$\epsilon_{k,m} \leq \frac{1}{\sqrt{k}} + \frac{k\sqrt{2\tilde{r}t_{\text{cover}}}}{(1-\gamma)^{2.5}\sqrt{m}} \log\left(\frac{|\mathcal{S}_g||\mathcal{A}_g||\mathcal{A}_l||\mathcal{S}_l|}{\rho}\right) \log\left(\frac{1}{(1-\gamma)^2}\right) \quad (34)$$

Then, using the naïve bound  $t_{\text{cover}}$  (since we are doing offline learning), we have

$$t_{\text{cover}} \leq |\mathcal{S}_g||\mathcal{A}_g||\mathcal{S}_l||\mathcal{A}_l|k^{|\mathcal{S}_l||\mathcal{A}_l|}.$$

Substituting this in Equation (34) yields that with probability  $1 - \rho$ ,

$$\epsilon_{k,m} \leq \frac{1}{\sqrt{k}} + \frac{k\sqrt{2\tilde{r}|\mathcal{S}_g||\mathcal{A}_g||\mathcal{S}_l||\mathcal{A}_l|k^{|\mathcal{S}_l||\mathcal{A}_l|}}}{(1-\gamma)^{2.5}\sqrt{m}} \log\left(\frac{|\mathcal{S}_g||\mathcal{A}_g||\mathcal{A}_l||\mathcal{S}_l|}{\rho}\right) \log\left(\frac{1}{(1-\gamma)^2}\right) \quad (35)$$

Therefore, letting  $\rho = \frac{1}{100e^k}$ , setting

$$m = \frac{2\tilde{r}|\mathcal{S}_g||\mathcal{A}_g||\mathcal{S}_l||\mathcal{A}_l|k^{3.5+|\mathcal{S}_l||\mathcal{A}_l|}}{(1-\gamma)^5} \log(100|\mathcal{S}_g||\mathcal{A}_g||\mathcal{A}_l||\mathcal{S}_l|) \log\left(\frac{1}{(1-\gamma)^2}\right) \quad (36)$$

attains a Bellman error of  $\epsilon_{k,m} \leq \tilde{O}(1/\sqrt{k})$  with probability at least  $1 - \frac{1}{100e^k}$ . Finally, the runtime of our learning algorithm is

$$O(mT|\mathcal{S}_g||\mathcal{S}_l||\mathcal{A}_g||\mathcal{A}_l|k^{|\mathcal{S}_l||\mathcal{A}_l|}) = \tilde{O}(|\mathcal{S}_g|^2|\mathcal{A}_g|^2|\mathcal{A}_l|^2|\mathcal{S}_l|^2k^{3.5+2|\mathcal{S}_l||\mathcal{A}_l|}\tilde{r}),$$

which is still polynomial in  $k$ , proving the claim.  $\square$

## H Generalization to Stochastic Rewards

Suppose we are given two families of distributions,  $\{\mathcal{G}_{s_g, a_g}\}_{s_g, a_g \in \mathcal{S}_g \times \mathcal{A}_g}$  and  $\{\mathcal{L}_{s_i, s_g, a_i}\}_{s_i, s_g, a_i \in \mathcal{S}_l \times \mathcal{S}_g \times \mathcal{A}_l}$ . Let  $R(s, a)$  denote a stochastic reward of the form

$$R(s, a) = r_g(s_g, a_g) + \sum_{i \in [n]} r_l(s_i, s_g, a_i), \quad (37)$$

where the rewards of the global agent  $r_g$  emerge from a distribution  $r_g(s_g, a_g) \sim \mathcal{G}_{s_g, a_g}$ , and the rewards of the local agents  $r_l$  emerge from a distribution  $r_l(s_i, s_g, a_i) \sim \mathcal{L}_{s_i, s_g, a_i}$ . For  $\Delta \subseteq [n]$ , let  $R_\Delta(s, a)$  be defined as:

$$R_\Delta(s, a) = r_g(s_g, a_g) + \sum_{i \in \Delta} r_l(s_i, s_g, a_i) \quad (38)$$

We make some standard assumptions of boundedness on  $\mathcal{G}_{s_g, a_g}$  and  $\mathcal{L}_{s_i, s_g, a_i}$ .

**Assumption H.1.** Define

$$\begin{aligned} \bar{\mathcal{G}} &= \bigcup_{(s_g, a_g) \in \mathcal{S}_g \times \mathcal{A}_g} \text{supp}(\mathcal{G}_{s_g, a_g}) \\ \bar{\mathcal{L}} &= \bigcup_{(s_i, s_g, a_i) \in \mathcal{S}_l \times \mathcal{S}_g \times \mathcal{A}_l} \text{supp}(\mathcal{L}_{s_i, s_g, a_i}) \end{aligned}$$

where for any distribution  $\mathcal{D}$ ,  $\text{supp}(\mathcal{D})$  is the support (set of all random variables  $\mathcal{D}$  that can be sampled with probability strictly larger than 0) of  $\mathcal{D}$ . Then, let  $\hat{\mathcal{G}} = \sup(\bar{\mathcal{G}})$ ,  $\hat{\mathcal{L}} = \sup(\bar{\mathcal{L}})$ ,  $\check{\mathcal{G}} = \inf(\bar{\mathcal{G}})$ , and  $\check{\mathcal{L}} = \inf(\bar{\mathcal{L}})$ . We assume that  $\hat{\mathcal{G}} < \infty$ ,  $\hat{\mathcal{L}} < \infty$ ,  $\check{\mathcal{G}} > -\infty$ ,  $\check{\mathcal{L}} > -\infty$ , and that  $\hat{\mathcal{G}}, \hat{\mathcal{L}}, \check{\mathcal{G}}, \check{\mathcal{L}}$  are all known in advance.

**Definition H.2.** Let the randomized empirical adapted Bellman operator be  $\hat{\mathcal{T}}_{k,m}^{\text{random}}$  such that:

$$\hat{\mathcal{T}}_{k,m}^{\text{random}} \hat{Q}_{k,m}^t(s_g, s_1, F_{z_\Delta}, a_g, a_1) = R_\Delta(s, a) + \frac{\gamma}{m} \sum_{\ell \in [m]} \max_{a' \in \mathcal{A}} \hat{Q}_{k,m}^t(s_g^\ell, s_j^\ell, F_{z_{\Delta \setminus j}^\ell}, a_j^\ell, a_g^\ell), \quad (39)$$

**SUBSAMPLE-MFQ: Learning with Stochastic Rewards.** The proposed algorithm averages  $\Xi$  samples of the adapted randomized empirical adapted Bellman operator,  $\mathcal{T}_{k,m}^{\text{random}}$  and updates the  $\hat{Q}_{k,m}$  function using with the average. One can show that  $\mathcal{T}_{k,m}^{\text{random}}$  is a contraction operator with module  $\gamma$ . By Banach's fixed point theorem,  $\mathcal{T}_{k,m}^{\text{random}}$  admits a unique fixed point  $\hat{Q}_{k,m}^{\text{random}}$ .

---

**Algorithm 4** SUBSAMPLE-MFQ: Learning with Stochastic Rewards

---

**Require:** A multi-agent system as described in Section 2. Parameter  $T$  for the number of iterations in the initial value iteration step. Sampling parameters  $k \in [n]$  and  $m \in \mathbb{N}$ . Discount parameter  $\gamma \in (0, 1)$ . Oracle  $\mathcal{O}$  to sample  $s'_g \sim P_g(\cdot | s_g, a_g)$  and  $s_i \sim P_l(\cdot | s_i, s_g, a_i)$  for all  $i \in [n]$ .

- 1: Set  $\tilde{\Delta} = \{2, \dots, k\}$ .
- 2: Set  $\mu_{k-1}(Z_l) = \{0, \frac{1}{k-1}, \frac{2}{k-1}, \dots, 1\}^{|\mathcal{S}_l| \times |\mathcal{A}_l|}$ .
- 3: Set  $\hat{Q}_{k,m}^0(s_g, s_1, F_{z_{\tilde{\Delta}}}, a_g, a_1) = 0$ , for  $(s_g, s_1, F_{z_{\tilde{\Delta}}}, a_1, a_g) \in \mathcal{S}_g \times \mathcal{S}_l \times \mu_{k-1}(\mathcal{Z}_l) \times \mathcal{A}_g \times \mathcal{A}_l$ .
- 4: **for**  $t = 1$  to  $T$  **do**
- 5:     **for**  $(s_g, s_1, F_{z_{\tilde{\Delta}}}, a_g, a_1) \in \mathcal{S}_g \times \mathcal{S}_l \times \mu_{k-1}(\mathcal{Z}_l) \times \mathcal{A}_g \times \mathcal{A}_l$  **do**
- 6:          $\rho = 0$
- 7:         **for**  $\xi \in \{1, \dots, \Xi\}$  **do**
- 8:              $\rho = \rho + \hat{\mathcal{T}}_{k,m}^{\text{random}} \hat{Q}_{k,m}^t(s_g, s_1, F_{z_{\tilde{\Delta}}}, a_g, a_1)$
- 9:          $\hat{Q}_{k,m}^{t+1}(s_g, s_1, F_{z_{\tilde{\Delta}}}, a_g, a_1) = \rho / \Xi$
- 10:    For all  $(s_g, s_i, F_{s_{\Delta \setminus i}}) \in \mathcal{S}_g \times \mathcal{S}_l \times \mu_{k-1}(\mathcal{S}_l)$ , let

$$\hat{\pi}_{k,m}^{\text{est}}(s_g, s_i, F_{s_{\Delta \setminus i}}) = \arg \max_{(a_g, a_i, F_{a_{\Delta \setminus i}}) \in \mathcal{A}_g \times \mathcal{A}_l \times \mu_{k-1}(\mathcal{Z}_l)} \hat{Q}_{k,m}^T(s_g, s_i, F_{s_{\Delta \setminus i}}, a_g, a_i).$$


---

**Theorem H.3** (Hoeffding's Theorem [Tsybakov, 2008]). *Let  $X_1, \dots, X_n$  be independent random variables such that  $a_i \leq X_i \leq b_i$  almost surely. Then, let  $S_n = X_1 + \dots + X_n$ . Then, for all  $\epsilon > 0$ ,*

$$\mathbb{P}[|S_n - \mathbb{E}[S_n]| \geq \epsilon] \leq 2 \exp \left( -\frac{2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2} \right) \quad (40)$$

**Lemma H.4.** *When  $\pi_{k,m}^{\text{est}}$  is derived from the randomized empirical value iteration operation and applying our online subsampling execution in Algorithm 2, we get*

$$\Pr \left[ V^{\pi^*}(s_0) - V^{\tilde{\pi}_{k,m}}(s_0) \leq \tilde{O} \left( \frac{1}{\sqrt{k}} \right) \right] \geq 1 - \frac{1}{100\sqrt{k}}. \quad (41)$$

*Proof.*

$$\begin{aligned} \Pr \left[ \left| \frac{\rho}{\Xi} - \mathbb{E}[R_{\Delta}(s, a)] \right| \geq \frac{\epsilon}{\Xi} \right] &\leq 2 \exp \left( -\frac{2\epsilon^2}{\sum_{i=1}^{\Xi} |\hat{\mathcal{G}} + \hat{\mathcal{L}} - \check{\mathcal{G}} - \check{\mathcal{L}}|^2} \right) \\ &= 2 \exp \left( -\frac{2\epsilon^2}{\Xi |\hat{\mathcal{G}} + \hat{\mathcal{L}} - \check{\mathcal{G}} - \check{\mathcal{L}}|^2} \right) \end{aligned}$$

Rearranging this, we get:

$$\Pr \left[ \left| \frac{\rho}{\Xi} - \mathbb{E}[R_{\Delta}(s, a)] \right| \leq \sqrt{\ln \left( \frac{2}{\delta} \right) \frac{|\hat{\mathcal{G}} + \hat{\mathcal{L}} - \check{\mathcal{G}} - \check{\mathcal{L}}|^2}{2\Xi^2}} \right] \geq 1 - \delta \quad (42)$$

Then, setting  $\delta = \frac{1}{100\sqrt{k}}$ , and setting  $\Xi = 10|\hat{\mathcal{G}} + \hat{\mathcal{L}} - \check{\mathcal{G}} - \check{\mathcal{L}}|k^{1/4}\sqrt{\ln(200\sqrt{k})}$  gives:

$$\Pr \left[ \left| \frac{\rho}{\Xi} - \mathbb{E}[R_{\Delta}(s, a)] \right| \leq \frac{1}{\sqrt{200k}} \right] \geq 1 - \frac{1}{100\sqrt{k}}$$

Then, applying the triangle inequality to  $\epsilon_{k,m}$  allows us to derive a probabilistic bound on the optimality gap between  $\hat{Q}_{k,m}^{\text{est}}$  and  $Q^*$ , where the gap is increased by  $\frac{1}{\sqrt{200k}}$ , and where the randomness is over the stochasticity of the rewards. Then, the optimality gap between  $V^{\pi^*}$  and  $V^{\tilde{\pi}_{k,m}^{\text{est}}}$ , for when the policy  $\hat{\pi}_{k,m}^{\text{est}}$  is learned using Algorithm 4 in the presence of stochastic rewards obeying Theorem H.1 follows

$$\Pr \left[ V^{\pi^*}(s_0) - V^{\tilde{\pi}_{k,m}}(s_0) \leq \tilde{O} \left( \frac{1}{\sqrt{k}} \right) \right] \geq 1 - \frac{1}{100\sqrt{k}}, \quad (43)$$

which proves the lemma.  $\square$

**Remark H.5.** First, note that through the naïve averaging argument, the optimality gap above still decays to 0 with probability decaying to 1, as  $k$  increases.

**Remark H.6.** This method of analysis could be strengthened by obtaining estimates of  $\hat{\mathcal{G}}, \hat{\mathcal{L}}, \check{\mathcal{G}}, \check{\mathcal{L}}$  and using methods from order statistics to bound the errors of the estimates and use the deviations between estimates to determine an optimal online stopping time [Kleinberg, 2005] as is done in the online secretary problem. This, more sophisticated, argument would become an essential step in converting this to a truly online learning, with a stochastic approximation scheme. Furthermore, one could incorporate variance-based analysis in the algorithm, where we use information derived from higher-order moments to do a weighted update of the Bellman update [Jin et al., 2024], thereby taking advantage of the skewedness of the distribution, and allowing us the freedom to assign an optimism/pessimism score to our estimate of the reward.

## I Partially Relaxing the Offline Learning assumption: Off-Policy Learning

A limitation of the planning Algorithm 1 is that it learns  $\hat{Q}_k^*$  in an *offline* manner by assuming a generative oracle access to the transition functions  $\mathbb{P}_g, \mathbb{P}_l$ , and reward function  $r(\cdot, \cdot)$ . In certain realistic RL applications, such a generative oracle might not exist, and it is more desirable to perform off-policy learning where the agent continues to learn in an offline manner but from *historical data* Fujimoto et al. [2019]. In this setting, the agents learn the target policy  $\hat{\pi}_k^*$  using data generated by a different behavior policy (the strategy it uses to explore the environment). There is a significant body of work on the theoretical guarantees in off-policy learning Chen et al. [2021b], Chen and Maguluri [2022b], Chen et al. [2021a, 2025].

In fact, these previous results are amenable to transforming guarantees about offline  $Q$ -learning to off-policy  $Q$ -learning, at the cost of  $\log |\mathcal{S}_g||\mathcal{S}_l|^k|\mathcal{A}_g||\mathcal{A}_l|^k$  factors in the runtime. Therefore, this section is devoted to showing that our previous result satisfy the assumptions of transforming offline  $Q$ -learning to off-policy  $Q$ -learning for the subsampled  $\hat{Q}_k$ -function, and we further show that, in expectation, can maintain the decaying optimality gap of  $\tilde{O}(1/\sqrt{k})$  of the learned policy  $\pi_k$ , where the randomness is over the heuristic exploration policy  $\pi_b$ .

The off-policy  $\hat{Q}_k$ -learning algorithm is an iterative algorithm to estimate the optimal  $\hat{Q}_k$ -function as follows: first, a sample trajectory  $\{(s_g, s_\Delta, a_g, a_\Delta)\}$  is collected using a suitable behavior policy  $\pi_{k,b}$ . For simplicity of notation, let  $S_\Delta = (s_g, s_\Delta)$  and  $A_\Delta = (a_g, a_\Delta)$ . Then, initialize  $\hat{Q}_k^0 : |\mathcal{S}_g||\mathcal{S}_l|^k|\mathcal{A}_g||\mathcal{A}_l|^k \rightarrow \mathbb{R}$  and let  $\alpha > 0$  be determined later. For each  $t \geq 0$  and state-action pair  $(S_\Delta, A_\Delta)$  that is updated to  $S'_\Delta$ , the iterate  $\hat{Q}_k^t(S_\Delta, A_\Delta)$  is updated by

$$\hat{Q}_k^{t+1}(S_\Delta, A_\Delta) = (1 - \alpha)\hat{Q}_k^t(S_\Delta, A_\Delta) + \alpha_t \left( r(S_\Delta, A_\Delta) + \gamma \max_{A'_\Delta \in \mathcal{A}_g \times \mathcal{A}_l^k} \hat{Q}_k^t(S'_\Delta, A'_\Delta) \right). \quad (44)$$

Note that the update in Equation (44) does not include an expectation and can be computed in a single trajectory via historical data. We make the following ergodicity assumption:

**Assumption I.1.** The behavior policy  $\pi_b$  satisfies  $\pi_b(A_\Delta | S_\Delta) > 0$  for all  $(S_\Delta, A_\Delta) \in \mathcal{S}_g \times \mathcal{S}_l^k \times \mathcal{A}_g \times \mathcal{A}_l^k$  and the Markov chain  $\mathcal{M}_{S_\Delta} = \{S_\Delta^{(t)}\}_{t \geq 0}$  induced by  $\pi_b$  is irreducible and aperiodic with stationary distribution  $\mu$  and mixing time  $t_\delta(\mathcal{M}) = \min\{t \geq 0 : \max_{S_\Delta \in \mathcal{S}_g \times \mathcal{S}_l^k} \|P^t(S_\Delta, \cdot) - \mu(\cdot)\|_{TV} \leq \delta\}$ . There are many heuristics of such behavior policies [Fujimoto et al., 2019].

**Theorem I.2.** Let  $\pi_k$  be the policy learned through off-policy  $\hat{Q}_k$ -learning. Then, under Theorem I.1, we have that with probability at least  $1 - \frac{1}{100e^k}$ ,

$$\begin{aligned} \mathbb{E}[V^{\pi^*}(s_0) - V^{\pi_k}(s_0)] &\leq \frac{\tilde{r}}{(1-\gamma)^2} \sqrt{\frac{n-k+1}{2nk}} \sqrt{\ln \frac{40\tilde{r}|\mathcal{S}_l||\mathcal{A}_l||\mathcal{A}_g|k^{|\mathcal{A}_l|+\frac{1}{2}}}{(1-\gamma)^2}} + \frac{1}{10\sqrt{k}} + 2\epsilon_{k,m} \\ &= \tilde{O}(1/\sqrt{k}), \end{aligned}$$

where the randomness in the expectation is over the stochasticity of the exploration policy  $\pi_b$ .

**Proposition I.3.** Recall that the following are true:

1.  $\|\hat{Q}_k(F_{S_\Delta}, F_{A_\Delta}) - \hat{Q}_k(F_{S'_\Delta}, F_{A_\Delta})\| \leq \frac{1}{1-\gamma} \|r_l(\cdot, \cdot)\|_\infty \cdot \|F_{S'_\Delta} - F_{S_\Delta}\|_1$ , for any  $S_\Delta, S'_\Delta \in \mathcal{S}_g \times \mathcal{A}_l^k$  and  $A_\Delta \in \mathcal{A}_g \times \mathcal{A}_l^k$  by Theorem E.3,
2.  $\|\hat{Q}_k\| \leq \frac{\tilde{r}}{1-\gamma}$  by Theorem C.4,
3.  $\|\hat{Q}_k^{t+1}(S_\Delta, A_\Delta) - \tilde{Q}_k^{t+1}(S_\Delta, A_\Delta)\|_\infty \leq \gamma \|\hat{Q}_k^t - \tilde{Q}_k^t(S_\Delta, A_\Delta)\|_\infty$  by Theorem C.6,
4. The Markov chain  $\mathcal{M}_S$  enjoys a rapid mixing property from Theorem I.1,

Then, by treating the single trajectory update of the  $\hat{Q}_k$ -function as a noisy addition to the expected update from the ideal Bellman operator, Chen et al. [2021b] uses Markovian stochastic approximation to bound  $\mathbb{E}[\|\hat{Q}_k^T - \hat{Q}_k^*\|_\infty^2]$ . We restate their result:

**Theorem I.4** (Theorem 3.1 in Chen et al. [2021b] adapted to our setting). Suppose  $\alpha_t = \alpha$  for all  $t \geq 0$ , where  $\alpha$  is chosen such that  $\alpha t_\alpha(\mathcal{M}_{S_\Delta}) \leq c_{Q,0} \frac{(1-\beta_1)^2}{\log |\mathcal{S}_g||\mathcal{S}_l^k||\mathcal{A}_g||\mathcal{A}_l^k|}$ , where  $c_{Q,0}$  is a numerical constant. Then, under Theorem I.3, for all  $t \geq t_\alpha(\mathcal{M}_{S_\Delta})$ , we have

$$\mathbb{E}[\|\hat{Q}_k^t - \hat{Q}_k^*\|_\infty^2] \leq c_{Q,1} \left(1 - \frac{(1-\gamma)\alpha}{2}\right)^{t-t_\alpha(\mathcal{M}_{S_\Delta})} + c_{Q,2} \frac{\log |\mathcal{S}_g||\mathcal{S}_l^k||\mathcal{A}_g||\mathcal{A}_l^k|}{(1-\gamma)^2} \alpha t_\alpha(\mathcal{M}_{S_\Delta}),$$

where  $c_{Q,1} = 3(\frac{\tilde{r}}{1-\gamma} + 1)^2$ ,  $c_{Q,2} = 912e(\frac{3\tilde{r}}{1-\gamma} + 1)^2$ , and where the randomness in the expectation is over the randomness of the stochasticity of the behavior/exploration policy  $\pi_b$ .

**Corollary I.5** (Corollary 3.2 in Chen et al. [2021b] adapted to our setting). To make  $\mathbb{E}[\|\hat{Q}_k^t - \hat{Q}_k^*\|_\infty \leq \frac{1}{100\sqrt{k}}]$  for  $\epsilon > 0$ , we need

$$t > \tilde{O}\left(\frac{10000k \log^2(100\sqrt{k}) |\mathcal{S}_g||\mathcal{S}_l^k||\mathcal{A}_g||\mathcal{A}_l^k| \log |\mathcal{S}_g||\mathcal{A}_g||\mathcal{S}_l^k||\mathcal{A}_l^k|}{(1-\gamma)^5}\right).$$

With this sample complexity, by the triangle inequality we also recover an expected-value analog of Theorem F.8.

**Corollary I.6.** For  $\delta \in (0, 1)^2$ , with probability at least  $1 - \delta$ , we have

$$\mathbb{E}[\hat{Q}_k^*(s_g, F_{s_\Delta}, a_g, F_{a_\Delta}) - Q_n^*(s_g, F_{s_{[n]}, a_g, F_{a_{[n]}}})] \leq \frac{\ln \frac{2|\mathcal{S}_l||\mathcal{A}_l|}{\delta}}{1-\gamma} \sqrt{\frac{n-k+1}{8kn}} \|r_l\|_\infty,$$

where the randomness in the expectation is over the stochasticity of the exploration policy  $\pi_b$ .

In turn, following the argument in the proof of Theorem G.3, it is straightforward to verify that this leads to a result on the expected performance difference using off-policy learning:

**Corollary I.7.** With probability at least  $1 - \frac{1}{100e^k}$ ,

$$\begin{aligned} \mathbb{E}[V^{\pi^*}(s_0) - V^{\pi_k}(s_0)] &\leq \frac{\tilde{r}}{(1-\gamma)^2} \sqrt{\frac{n-k+1}{2nk}} \sqrt{\ln \frac{40\tilde{r}|\mathcal{S}_l||\mathcal{A}_l||\mathcal{A}_g|k^{|\mathcal{A}_l|+\frac{1}{2}}}{(1-\gamma)^2}} + \frac{1}{10\sqrt{k}} + 2\epsilon_{k,m} \\ &= \tilde{O}(1/\sqrt{k}), \end{aligned}$$

where the randomness in the expectation is over the stochasticity of the exploration policy  $\pi_b$ .

## J Extension to Continuous State/Action Spaces

Multi-agent settings where each agent handles a continuous state and action space find many applications in optimization, control, and synchronization.

**Example J.1** (Quadcopter Swarm Drone [Preiss et al., 2017]). Consider a system of drones with a global controller, where each drone has to chase the controller, and the controller is designed to follow a bounded trajectory. Here, the state of each local agent  $i \in [n]$  is its position and velocity in the bounded region, and the state of the global agent  $g$  is a signal on its position and direction. The action of each local agent is a velocity vector  $a_l \in \mathcal{A}_l \subset \mathbb{R}^3$  which is a bounded subset of  $\mathbb{R}^3$ , and the action of the global agent is a vector  $a_g \in \mathcal{A}_g \subset \mathbb{R}^2$  which is a bounded subset of  $\mathbb{R}^2$ .

Hence, this section is devoted to extending our algorithm and theoretical results to the case where the state and action spaces can be continuous (and therefore have infinite cardinality).

**Preliminaries.** For a measurable space  $(\mathcal{S}, \mathcal{B})$ , where  $\mathcal{B}$  is a  $\sigma$ -algebra on  $\mathcal{S}$ , let  $\mathbb{R}^{\mathcal{S}}$  denote the set of all real-valued  $\mathcal{B}$ -measurable functions on  $\mathcal{S}$ . Let  $(X, d)$  be a metric space, where  $X$  is a set and  $d$  is a metric on  $X$ . A set  $S \subseteq X$  is *dense* in  $X$  if every element of  $X$  is either in  $S$  or a limit point of  $S$ . A set is *nowhere dense* in  $X$  if the interior of its closure in  $X$  is empty. A set  $S \subseteq X$  is of Baire first category if  $S$  is a union of countably many nowhere dense sets. A set  $S \subseteq X$  is of Baire second category if  $X \setminus S$  is of first category.

**Theorem J.2** (Baire Category Theorem [Jin et al., 2020]). Let  $(X, d)$  be a complete metric space. Then any countable intersection of dense open subsets of  $X$  is dense.

**Definition J.3** (Linear MDP). MDP  $(S, A, \mathbb{P}, r)$  is a linear MDP with feature map  $\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$  if there exist  $d$  unknown (signed) measures  $\mu = (\mu^1, \dots, \mu^d)$  over  $\mathcal{S}$  and a vector  $\theta \in \mathbb{R}^d$  such that for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , we have

$$\begin{aligned}\mathbb{P}(\cdot | s, a) &= \langle \phi(s, a), \mu(\cdot) \rangle, \\ r(s, a) &= \langle \phi(s, a), \theta \rangle\end{aligned}$$

**Assumption J.4** (Bounded features). Without loss of generality, we assume boundedness of the features:  $\|\phi(s, a)\| \leq 1$ , for all  $(s, a) \in \mathcal{S} \times \mathcal{A}$  and  $\max\{\|\mu(S)\|, \|\theta\|\} \leq \sqrt{d}$ .

We motivate our analysis by reviewing representation learning in RL via spectral decompositions. For instance, if  $P(s'|s, a)$  admits a linear decomposition in terms of some spectral features  $\phi(s, a)$  and  $\mu(s')$ , then the  $Q(s, a)$ -function can be linearly represented in terms of the spectral features  $\phi(s, a)$ .

Then, through a reduction from Ren et al. [2024] that uses function approximation to learn the spectral features  $\phi_k$  for  $\hat{Q}_k$ , we derive a performance guarantee for the learned policy  $\pi_k^{\text{est}}$ , where the optimality gap decays with  $k$ :

**Theorem J.5.** When  $\pi_k^{\text{est}}$  is derived from the spectral features  $\phi_k$  learned in  $\hat{Q}_k$ , and  $M$  is the number of samples used in the function approximation, then

$$\Pr \left[ V^{\pi^*}(s) - V^{\pi_k^{\text{est}}}(s) \leq \tilde{O} \left( \frac{1}{\sqrt{k}} + \frac{\|\phi_k\|^5 \log 2k^2}{\sqrt{M}} + \frac{2\gamma\tilde{r}}{(1-\gamma)\sqrt{k}} \|\phi_k\| \right) \right] \geq 1 + \frac{1}{50k} - \frac{201}{100\sqrt{k}}$$

We assume the system is a linear MDP, where  $\mathcal{S}_g$  and  $\mathcal{S}_l$  are infinite compact sets. By a reduction from Ren et al. [2024] and using function approximation to learn the spectral features  $\phi_k$  for  $\hat{Q}_k$ , we derive a performance guarantee for the learned policy  $\pi_k^{\text{est}}$ , where the optimality gap decays with  $k$ .

**Assumption J.6.** Suppose  $\mathcal{S}_g \subset \mathbb{R}^{\sigma_g}, \mathcal{S}_l \subset \mathbb{R}^{\sigma_l}, \mathcal{A}_g \subset \mathbb{R}^{\alpha_g}, \mathcal{A}_l \subset \mathbb{R}^{\alpha_l}$  are bounded compact sets. From the Baire category theorem, the underlying field  $\mathbb{R}$  can be replaced to any set of Baire first category, which satisfies the property that there exists a dense open subset. In particular, replace Theorem 2.1 with a boundedness assumption.

Multi-agent settings where each agent handles a continuous state/action space find many applications in optimization, control, and synchronization.

**Example J.7** (Federated Learning). Consider a peer-to-peer learning setting [Chaudhari et al., 2024] with  $n$  agents, where each agent possesses a common neural network architecture  $f(\mathbf{x}; \theta) : \mathbb{R}^N \rightarrow \mathbb{R}^M$ , parameterized by  $\theta \in \mathbb{R}^P$ . Here, each local agent has a common local objective function  $L_q(\theta) : \mathbb{R}^P \rightarrow \mathbb{R}$  defined using a local dataset  $\mathcal{D}_q \triangleq \{x_{q,i}, y_{q,i}\}_{i=1}^D$ . For instance,  $L_q(\theta)$  can be the local empirical risk function that evaluates the performance of  $f(\mathbf{x}; \theta)$  on a task involving the local dataset  $\mathcal{D}_q$ , where the agents cooperate to learn the optimal  $\theta^*$  that minimizes the global objective, which is a weighted average of all local objectives. Formally, the agents collaboratively aim to find  $\theta^*$  satisfying:

$$\begin{aligned}\theta^* &= \arg \min_{\theta \in \mathbb{R}^P} L(\theta) \\ L(\theta) &= \frac{1}{Q} \sum_{q=1}^Q L_q(\theta) \triangleq \frac{1}{Q} \sum_{q=1}^Q \frac{1}{D} \sum_{i=1}^D \ell(f(x_{q,i}; \theta), y_{q,i})\end{aligned}$$

Here, the global agent may act to balance the loss among agents, by ensuring that the variance on the losses across the agents is small. If the variance on the losses of the agents is large, it could assign a large cost to the system, and signal that the policies of the local agents must favor the mean state, promoting convergence.

**Lemma J.8** (Proposition 2.3 in Jin et al. [2020]). For any linear MDP, for any policy  $\pi$ , there exist weights  $\{w^\pi\}$  such that for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , we have  $Q^\pi(s, a) = \langle \phi(s, a), \mathbf{w}^\pi \rangle$ .

**Lemma J.9** (Proposition A.1 in Jin et al. [2020]). For any linear MDP, for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$ , and for any measurable subset  $\mathcal{B} \subseteq \mathcal{S}$ , we have that

$$\begin{aligned}\phi(s, a)^\top \mu(\mathcal{S}) &= 1, \\ \phi(s, a)^\top \mu(\mathcal{B}) &\geq 0.\end{aligned}$$

**Property J.10.** Suppose that there exist spectral representation  $\mu_g(s'_g) \in \mathbb{R}^d$  and  $\mu_l(s'_i) \in \mathbb{R}^d$  such that the probability transitions  $P_g(s'_g | s_g, a_g)$  and  $P_l(s'_i | s_i, s_g, a_i)$  can be linearly decomposed by

$$\begin{aligned}P_g(s'_g | s_g, a_g) &= \phi_g(s_g, a_g)^\top \mu_g(s'_g) \\ P_l(s'_i | s_i, s_g, a_i) &= \phi_l(s_i, s_g, a_i)^\top \mu_l(s'_i)\end{aligned}$$

for some features  $\phi_g(s_g, a_g) \in \mathbb{R}^d$  and  $\phi_l(s_i, s_g, a_i) \in \mathbb{R}^d$ . Then, the dynamics are amenable to the following spectral factorization, as in Ren et al. [2024].

**Lemma J.11.**  $\hat{Q}_k$  admits a linear representation.

*Proof.* Given the factorization of the dynamics of the multi-agent system, we have:

$$\begin{aligned}\mathbb{P}(s' | s, a) &= P_g(s'_g | s_g, a_g) \cdot \prod_{i=1}^n P_l(s'_i | s_i, s_g, a_i) \\ &= \langle \phi_g(s_g, a_g), \mu_g(s'_g) \rangle \cdot \prod_{i=1}^n \langle \phi_l(s_i, s_g, a_i), \mu_l(s'_i) \rangle \\ &= \phi_g(s_g, a_g), \mu_g(s'_g) \rangle \cdot \langle \otimes_{i=1}^n \phi_l(s_i, s_g, a_i), \otimes_{i=1}^n \mu_l(s'_i) \rangle \\ &:= \langle \bar{\phi}_n(s, a), \bar{\mu}_n(s') \rangle\end{aligned}$$

Similarly, for any  $\Delta \subseteq [n]$  where  $|\Delta| = k$ , the subsystem consisting of  $k$  local agents  $\Delta$  has a subsystem dynamics given by

$$\mathbb{P}(s'_\Delta, s_\Delta | s_\Delta, s_g, a_g, a_\Delta) = \langle \bar{\phi}_k(s_g, s_\Delta, a_g, a_\Delta), \bar{\mu}_k(s'_\Delta) \rangle.$$

Therefore,  $\hat{Q}_k$  admits the linear representation:

$$\begin{aligned}\hat{Q}_k^{\pi_k}(s_g, F_{z_\Delta}, a_g) &= r_g(s_g, a_g) + \frac{1}{k} \sum_{i \in \Delta} r_l(s_i, s_g, a_i) + \gamma \mathbb{E}_{s'_g, s_{\Delta'}, a'_{\Delta'}} \max_{a'_g, a'_{\Delta'}} \hat{Q}_k^{\pi_k}(s'_g, F_{z'_{\Delta'}}, a'_g) \\ &= r_g(s_g, a_g) + \frac{1}{k} \sum_{i \in \Delta} r_l(s_i, s_g, a_i) + (\mathbb{P}V^{\pi_k}(s_g, F_{s_\Delta}, a_g)) \\ &= \left[ \frac{r_\Delta(s, a)}{\bar{\phi}_k(s_g, s_\Delta, a_g, a_\Delta)} \right] \left[ 1 - \gamma \int_{s'_\Delta} \mu_k(s'_\Delta) V^{\pi_k}(s'_\Delta) ds'_\Delta \right],\end{aligned}$$

proving the claim.  $\square$

Therefore,  $\bar{\phi}_k$  serves as a good representation for the  $\hat{Q}_k$ -function making the problem amenable to the classic linear functional approximation algorithms, consisting of *feature generation* and *policy gradients*.

In feature generation [Ren et al., 2024], we generate the appropriate features  $\phi$ , comprising of the local reward functions and the spectral features coming from the factorization of the dynamics. In

applying the policy gradient, we perform a gradient step to update the local policy weights  $\theta_i$  and update the new policy.

For this, we update the weight  $w$  via the TD(0) target semi-gradient method (with step size  $\alpha$ ), to get

$$w_{t+1} = w_t + \alpha(r_\Delta(s^t, a^t) + \gamma\bar{\phi}_k(s_\Delta^{t+1})^\top w_t - \bar{\phi}_k(s_\Delta^t)^\top w_t)\bar{\phi}_k(s_\Delta^t) \quad (45)$$

**Definition J.12.** Let the complete feature map be denoted by  $\Phi : \mathcal{S} \times d \rightarrow \mathbb{R}$ , and let the subsampled feature map be denoted by  $\hat{\Phi}_k : \mathcal{S}_g \times \mathcal{S}_l^k \times d \in \mathbb{R}$ , where we let

$$\hat{\Phi}_k = \begin{bmatrix} \bar{\Phi}_k(1) \\ \bar{\Phi}_k(2) \\ \vdots \\ \bar{\Phi}_k(|\mathcal{S}_g| \times |\mu_{k-1}(\mathcal{S}_l)|) \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}_g| \times |\mu_{k-1}(\mathcal{S}_l)| \times d} \quad (46)$$

Here, the system's stage rewards are given by

$$r = [r(1) \quad \dots \quad r(S)] \in \mathbb{R}^S$$

and

$$r_k = [r_k(1), \dots, r_k(|\mathcal{S}_g| \times |\mu_{k-1}(\mathcal{S}_l)|)] \in \mathbb{R}^{|\mathcal{S}_g| \times |\mu_{k-1}(\mathcal{S}_l)|},$$

where  $d$  is a low-dimensional embedding we wish to learn.

Agnostically, the goal is to approximate the value function through the approximation

$$V_w \approx \hat{\Phi}_k w = \begin{bmatrix} \hat{\Phi}_k(1) \\ \vdots \\ \hat{\Phi}_k(|\mathcal{S}_g| \times |\mu_{k-1}(\mathcal{S}_l)|) \end{bmatrix} w \in \text{span}(\hat{\Phi}_k)$$

This manner of updating the weights can be considered via the projected Bellman equation  $\hat{\Phi}_k w = \Pi_\mu \mathcal{T}(\hat{\Phi}_k w)$ , where  $\Pi_\mu(v) = \arg \min_{z \in \hat{\Phi}_k w} \|z - v\|_\mu^2$ . Notably, the fixed point of the projected Bellman equation satisfies

$$w = (\hat{\Phi}_k^\top D_\mu \hat{\Phi}_k)^{-1} \hat{\Phi}_k^\top D_\mu (r + \gamma P^\pi \hat{\Phi}_k w),$$

where  $D_\mu$  is a diagonal matrix comprising of  $\mu$ 's. In other words,  $D_\mu = \text{diag}(\mu_1, \dots, \mu_n)$ . Then,

$$(\hat{\Phi}_k^\top D_\mu \hat{\Phi}_k)w = \hat{\Phi}_k^\top D_\mu (r + \gamma P^\pi \hat{\Phi}_k w)$$

In turn, this implies

$$\hat{\Phi}_k^\top D_\mu (I - \gamma P^\pi) \Phi w = \hat{\Phi}_k^\top D_\mu r.$$

Therefore, the problem is amenable to Algorithm 1 in Ren et al. [2024]. To bound the error of using linear function approximation to learn  $\hat{Q}_k$ , we use a result from Ren et al. [2024].

**Lemma J.13** (Policy Evaluation Error, Theorem 6 of Ren et al. [2024]). *Suppose the sample size  $M \geq \log\left(\frac{4n}{\delta}\right)$ , where  $n$  is the number of agents and  $\delta \in (0, 1)$  is an error parameter. Then, with probability at least  $1 - 2\delta$ , the ground truth  $\hat{Q}_k^\pi(s, a)$  function and the approximated  $\hat{Q}_k$ -function  $\hat{Q}_k^{\text{LFA}}(s, a)$  satisfies, for any  $(s, a) \in \mathcal{S} \times \mathcal{A}$ ,*

$$\mathbb{E} \left[ \left| \hat{Q}_k^\pi(s, a) - \hat{Q}_k^{\text{LFA}}(s, a) \right| \right] \leq O \left( \underbrace{\log\left(\frac{2k}{\delta}\right) \frac{\|\bar{\phi}_k\|^5}{\sqrt{M}}}_{\text{statistical error}} + \underbrace{2\epsilon_P \gamma \cdot \frac{\tilde{r}}{1-\gamma} \cdot \|\bar{\phi}_k\|}_{\text{approximation error}} \right),$$

where  $\epsilon_P$  is the error in approximating the spectral features  $\phi_g, \phi_l$ .

---

**Algorithm 5** SUB-SAMPLE-MFQ: Execution with weakly shared randomness

---

**Require:** A multi-agent system as described in Section 2. A distribution  $s_0$  on the initial global state  $s_0 = (s_g, s_{[n]})$ . Parameter  $T'$  for the number of iterations for the decision-making sequence. Hyperparameter  $k \in [n]$ . Discount parameter  $\gamma$ . Policy  $\hat{\pi}_{k,m}^{\text{est}}(s_g, s_\Delta)$ .

- 1: Sample  $(s_g(0), s_{[n]}(0)) \sim s_0$ .
- 2: Define groups  $h_1, \dots, h_x$  of agents where  $x := \lceil \frac{n}{k} \rceil$  and  $|h_1| = |h_2| = \dots = |h_{x-1}| = k$  and  $|h_x| = n \bmod k$ .
- 3: **for**  $t = 0$  to  $T' - 1$  **do**
- 4:   **for**  $i \in [x]$  **do**
- 5:     Let  $\Delta_i$  be a uniform sample of  $\binom{[n] \setminus g_i}{k-1}$ , and let  $a_{g_i}(t) = [\hat{\pi}_{k,m}^{\text{est}}(s_g(t), s_{\Delta_i}(t))]_{1:k}$ .
- 6:     Let  $a_g(t) = \text{majority} \left( \left\{ [\hat{\pi}_{k,m}^{\text{est}}(s_g(t), s_{\Delta_i}(t))]_g \right\}_{i \in [x]} \right)$ .
- 7:     Let  $s_g(t+1) \sim P_g(\cdot | s_g(t), a_g(t))$
- 8:     Let  $s_i(t+1) \sim P_l(\cdot | s_i(t))$
- 9:      $s_g(t), a_i(t)$ , for all  $i \in [n]$ .
- 10:  Get reward  $r(s, a) = r_g(s_g, a_g) + \frac{1}{n} \sum_{i \in [n]} r_l(s_i, a_i, s_g)$

---

**Corollary J.14.** Therefore, when  $\pi_{k,m}$  is derived from the spectral features learned in  $\hat{Q}_k^{\text{LFA}}$ , applying the triangle inequality on the Bellman noise and setting  $\delta = \epsilon_P = \frac{1}{\sqrt{k}}$  yields<sup>5</sup>

$$\Pr \left[ V^{\pi^*}(s_0) - V^{\pi_{k,m}}(s_0) \leq \tilde{O} \left( \frac{1}{\sqrt{k}} + \log(2k^2) \frac{\|\bar{\phi}_k\|^5}{\sqrt{M}} + \frac{2}{\sqrt{k}} \cdot \frac{\gamma \tilde{r}}{1-\gamma} \|\bar{\phi}_k\| \right) \right] \geq 1 + \frac{1}{50k} - \frac{201}{100\sqrt{k}}$$

Using  $\|\tilde{\phi}_k\| \leq 1$  and simplifying gives

$$\Pr \left[ V^{\pi^*}(s_0) - V^{\pi_{k,m}}(s_0) \leq \tilde{O} \left( \frac{1}{\sqrt{k}} + \frac{\log(2k^2)}{\sqrt{M}} + \frac{2\gamma \tilde{r}}{\sqrt{k}(1-\gamma)} \right) \right] \geq 1 - \frac{201}{100\sqrt{k}}$$

**Remark J.15.** Hence, in the continuous state/action space setting, as  $k \rightarrow n$  and  $M \rightarrow \infty$ ,  $V^{\tilde{\pi}_k}(s_0) \rightarrow V^{\tilde{\pi}_n}(s_0) = V^{\pi^*}(s_0)$ . Intuitively, as  $k \rightarrow n$ , the optimality gap diminishes following the arguments in Theorem G.7. As  $M \rightarrow \infty$ , the number of samples obtained allows for more effective learning of the spectral features.

## K Towards Deterministic Algorithms: Sharing Randomness

In large distributed systems, the random sampling in our communication network may be a bottleneck for “decentralized execution”. In light of this, we have provided a practical derandomized heuristic where the agents can share some randomness by only sampling within pre-defined fixed blocks of size. In this section, we propose algorithms RANDOM-SUB-SAMPLE-MFQ and RANDOM-SUB-SAMPLE-MFQ+, which shares randomness between agents through various careful groupings and using shared randomness within each groups. By implementing these algorithms, we derive simulation results, and establish progress towards derandomizing inherently stochastic approximately-optimal policies.

**Algorithm 5** (Execution with weakly shared randomness). The local agents are heuristically divided into arbitrary groups of size  $k$ . For each group  $h_i$ , the  $k-1$  local agents sampled are the same at each iteration. The global agent’s action is then the majority global agent action determined by each group of local agents. At each round, this requires  $O(\lceil n^2/k \rceil (n-k))$  random numbers.

**Algorithm 6** (Execution with strongly shared randomness). The local agents are randomly divided in to groups of size  $k$ . For each group, each agent uses the other agents in the group as the  $k-1$  other sampled states. Similarly, the global agent’s action is then the majority global agent action determined by each group of local agents. Here, at each round, this requires  $O(\lceil n^2/k \rceil)$  random numbers.

---

<sup>5</sup>Following Ren et al. [2024], the result easily generalizes to any positive-definite transition Kernel noise (e.g. Laplacian, Cauchy, Matérn, etc.

---

**Algorithm 6** SUB-SAMPLE-MFQ: Execution with strongly shared randomness

---

**Require:** A multi-agent system as described in Section 2. A distribution  $s_0$  on the initial global state  $s_0 = (s_g, s_{[n]})$ . Parameter  $T'$  for the number of iterations for the decision-making sequence. Hyperparameter  $k \in [n]$ . Discount parameter  $\gamma$ . Policy  $\hat{\pi}_{k,m}^{\text{est}}(s_g, s_\Delta)$ .

- 1: Sample  $(s_g(0), s_{[n]}(0)) \sim s_0$ .
- 2: Define groups  $h_1, \dots, h_x$  of agents where  $x := \lceil \frac{n}{k} \rceil$  and  $|h_1| = |h_2| = \dots = |h_{x-1}| = k$  and  $|h_x| = n \bmod k$ .
- 3: **for**  $t = 0$  to  $T' - 1$  **do**
- 4:   **for**  $i \in [x - 1]$  **do**
- 5:     Let  $a_{h_i}(t) = [\hat{\pi}_{k,m}^{\text{est}}(s_g(t), s_{h_i}(t))]_{1:k}$ .
- 6:     Let  $a_{h_i}^g(t) = [\hat{\pi}_{k,m}^{\text{est}}(s_g(t), s_{h_i}(t))]_g$ .
- 7:     Let  $\Delta_{\text{residual}} := \binom{[n] \setminus h_x}{k - (n - \lceil n/k \rceil)k}$ .
- 8:     Let  $a_{h_x}(t) = [\hat{\pi}_{k,m}^{\text{est}}(s_g(t), s_{h_x} \cup \Delta_{\text{residual}}(t))]_{1:n - \lceil n/k \rceil k}$ .
- 9:     Let  $a_{h_x}^g(t) = [\hat{\pi}_{k,m}^{\text{est}}(s_g(t), s_{h_x} \cup \Delta_{\text{residual}}(t))]_g$ .
- 10:    Let  $a_g(t) = \text{majority}(\{a_{h_i}^g(t) : i \in [x]\})$ .
- 11:    Let  $s_g(t+1) \sim P_g(\cdot | s_g(t), a_g(t))$ ,  $s_i(t+1) \sim P_l(\cdot | s_i(t), s_g(t), a_i(t))$ , for all  $i \in [n]$ .
- 12:    Get reward  $r(s, a) = r_g(s_g, a_g) + \frac{1}{n} \sum_{i \in [n]} r_l(s_i, a_i, s_g)$

---

The probability of a bad event (the policy causing the  $Q$ -function to exceed the Lipschitz bound) scales with the  $O(n^k)$  independent random variables, which is polynomially large. Agnostically, some randomness is always needed to avoid periodic dynamics that may occur when the underlying Markov chain is not irreducible. In this case, an adversarial reward function could be designed such that the system does not minimize the optimality gap, thereby penalizing excessive determinism.

This ties into an open problem of interest, which has been recently explored [Larsen et al., 2024]. What is the minimum amount of randomness required for SUBSAMPLE-MFQ (or general mean-field or multi-agent RL algorithms) to perform well? Can we derive a theoretical result that demonstrates and balances the tradeoff between the amount of random bits and the performance of the subsampled policy when using the greedy action from the  $k$ -agent subsystem derived from  $\hat{Q}_k^*$ ? We leave this problem as an exciting direction for future research.