Mean-Field Control on Sparse Graphs: From Local Limits to GNNs via Neighborhood Distributions

Tobias Schmidt
Technische Universität Darmstadt, Germany
tobias.schmidt@tu-darmstadt.de

Kai Cui
Technische Universität Darmstadt, Germany
kai.cui@tu-darmstadt.de

Abstract

Mean-field control (MFC) offers a scalable solution to the curse of dimensionality in multi-agent systems but traditionally hinges on the restrictive assumption of exchangeability via dense, all-to-all interactions. In this work, we bridge the gap to real-world network structures by proposing a rigorous framework for MFC on large sparse graphs. We redefine the system state as a probability measure over *decorated rooted neighborhoods*, effectively capturing local heterogeneity. Our central contribution is a theoretical foundation for scalable reinforcement learning in this setting. We prove *horizon-dependent locality*: for finite-horizon problems, an agent's optimal policy at time t depends strictly on its $(T-t)$-hop neighborhood. This result renders the infinite-dimensional control problem tractable and underpins a novel Dynamic Programming Principle (DPP) on the lifted space of neighborhood distributions. Furthermore, we formally and experimentally justify the use of Graph Neural Networks (GNNs) for actor-critic algorithms in this context. Our framework naturally recovers classical MFC as a degenerate case while enabling efficient, theoretically grounded control on complex sparse topologies.

1 Introduction

The theory of mean-field games and control (Andersson & Djehiche, 2011; Bensoussan et al., 2013) has emerged as a cornerstone for the study of large-scale multi-agent systems (Lasry & Lions, 2007; Huang et al., 2006; Carmona et al., 2018; Laurière et al., 2022a). By considering the limit as the number of agents $N \to \infty$, these frameworks circumvent the curse of dimensionality inherent in multi-agent reinforcement learning (MARL) problems (Canese et al., 2021; Gronauer & Diepold, 2022). The core assumption is that agents are exchangeable and interact with the collective through a tractable aggregate statistic – the empirical distribution of states, or "mean field" (MF).

A critical implicit assumption is that the interaction graph is the complete graph. While all-to-all interaction is reasonable for systems with long-range, homogeneous interactions, it fails to capture a vast array of real-world systems with local interactions. Examples include social networks (Jackson, 2008), power grids (Dörfler et al., 2013), biological swarms (Vicsek et al., 1995), and robotic networks (Bullo et al., 2009). The challenge of sparse graphs is profound: The local environment of each agent is distinct and stochastic, breaking symmetry that underpins classical MF theory. Recent work has begun to tackle this challenge by studying MF learning on random graphs (Caines & Huang, 2021; Cui & Koeppl, 2022; Fabian et al., 2023; Zhang et al., 2024a; Fabian et al., 2024; Zhang et al., 2024b; Fabian et al., 2025), showing that a MF limit can still be obtained. However, the limiting MF is no longer a simple state distribution but is related to the structure of the graph itself.

This paper builds on these insights to construct a rigorous framework for reinforcement learning (RL) on sparse graphs. The central idea is to redefine the state of the system. Instead of a measure $\mu \in \mathcal{P}(\mathcal{X})$ on the agent's state space $\mathcal{X}$, the state is a measure $\boldsymbol{\mu} \in \mathcal{P}(\mathcal{G}_*^{\mathcal{X}})$ on the space of *decorated rooted graphs*, which we denote by $\mathcal{G}_*^{\mathcal{X}}$. This object, $\boldsymbol{\mu}$, describes the probability of observing any

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

given local neighborhood structure, with each node in that neighborhood "decorated" by its state. For a finite T-horizon problem, we obtain and apply a crucial locality result: an agent's optimal policy at time t depends only on the distribution of its $(T-t)$-hop neighborhood, providing a direct link between the problem horizon and the required locality of information. As a result, we have a Markov decision process (MDP) as a more tractable single-agent RL problem, and are able to design appropriate RL algorithms using GNNs with strong theoretical justification.

1.1 Related Work

A large body of work studies how to compute equilibria and optimal controls in MFG/MFC using RL and deep learning (Perrin et al., 2020; Gu et al., 2021; Laurière et al., 2022a). In the finite-state/discrete-time setting, foundational RL approaches for MFGs include Q-learning-style algorithms that couple best responses with the MF fixed point (Guo et al., 2019), and scalable deep RL variants that target large state spaces (Laurière et al., 2022b). Complementary learning paradigms include fictitious play and extensions (e.g., common noise), with convergence analyses and practical implementations (Perrin et al., 2020; Laurière et al., 2022b) and generalization viewpoints (Perrin et al., 2022). For mean-field *control* (cooperative) rather than games, various policy-gradient methods have also been developed (Gu et al., 2020, 2023; Cui et al., 2024; Guan et al., 2024).

A second related strand addresses *many-agent RL* by replacing interactions with a mean-field (average-neighbor) surrogate, yielding mean-field Q-learning / actor–critic algorithms with strong empirical performance in large populations (Yang et al., 2018; Gu et al., 2021). These methods typically assume that the relevant aggregate seen by an agent is an empirical average (often over neighbors), but they do not build a *graph-local* limiting distribution over rooted neighborhoods, and use a separate set of assumptions. There exist scalable actor-critic methods for networked systems exploit localized policies and local reward decomposition (Qu et al., 2020). The objectives we treat, however, may be non-linear and globally coupled (e.g., a penalty on the total infection rate in a SIR model).

Large graph theory and mean fields. To incorporate network heterogeneity, several works adopt *dense* graph limits via graphons, leading to graphon MFG/MFC and learning algorithms when interactions vary by latent graphon index (Cui & Koeppl, 2022; Hu et al., 2023; Zhang et al., 2024a,b; Zhou et al., 2024). More recently, various graphon variants have been proposed to better capture sparsity (e.g. L^p-graphons) beyond dense graphons (Fabian et al., 2023, 2024). Graphon convergence describes normalized edge density, whereas local weak convergence retains unnormalized finite neighborhoods. Therefore, both limit procedures are complementary rather than nested.

On the probability theory side, interacting processes on sparse graphs have a limiting description governed by local weak (Benjamini–Schramm) convergence, where the natural object is the law of a rooted neighborhood process; continuity of the dynamics with respect to local weak convergence is established in general frameworks (Lacker et al., 2023) and successfully to MFC via heuristic algorithms (Fabian et al., 2025). Our contribution differs in emphasis and structure: we turn the local-weak-limit viewpoint into a *dynamic programming* framework for control by taking the system state to be a distribution over decorated rooted neighborhoods, proving a *horizon-dependent locality* principle that identifies the minimal radius required at each time step, and theoretically justifies the use of Graph Neural Networks as policy approximators with approximate optimality. Our results show the limits of the heuristic but successful approach in Fabian et al. (2025), contributing a rigorous solution and theoretical basis for algorithms; Table 1 in the appendix gives a theorem-level comparison. Our work is complementary to mean-field subsampling (Anand et al., 2026) and graph-attention methods (Veličković et al., 2017), which address how local information is estimated rather than which local information is sufficient. Moreover, kernel mean-field methods (Tang et al., 2026) offer an alternative route to high-dimensional local policies.

MARL on graphs. Finally, our algorithmic blueprint connects to extensive MARL literature on graph-structured environments and GNN-based policies/value functions (Jiang et al., 2020; Sunehag et al., 2018; Rashid et al., 2020), but these works typically proceed from a finite-N training objective rather than from an explicit sparse-graph mean-field limit. Our work provides separate motivation for architecture choices such as GNN especially in large sparse graphs where neighborhoods are of interest up to isomorphism.

In contrast to classical MFC and learning, where the MF is a distribution over agent states and interaction structure is implicit (typically all-to-all), our framework replaces the mean field by a

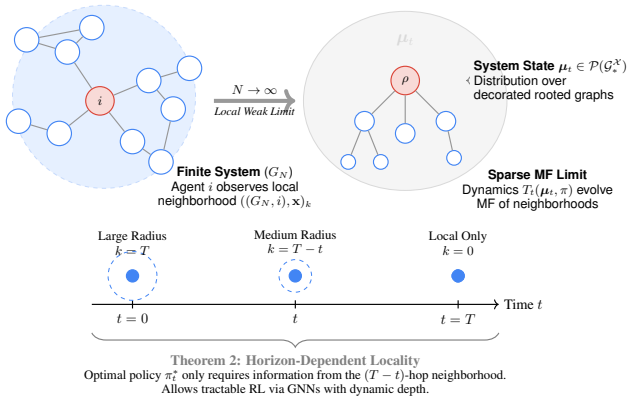

Figure 1: **Overview of the Sparse Mean-Field Control Framework.** (Top Left) The physical N-agent system on a sparse graph where agents interact locally. (Top Right) The rigorous mean-field limit where the state is redefined as a probability measure $\boldsymbol{\mu}_t$ over the space of decorated rooted graphs. (Bottom) Our Horizon-Dependent Locality result (Theorem 2) establishes the $(T - t)$-hop neighborhood at time t as sufficient statistic for optimal control.

distribution over decorated rooted neighborhoods arising from the local weak limit of sparse graphs as shown in Fig. 1. This shift preserves classical MFC as a special case while enabling a dynamic programming formulation that remains valid when interactions are local and the population is no longer exchangeable.

1.2 Main Contributions

Our main contributions are: (i) A novel mathematical framework with optimality characterized to horizon-dependent locality, using the distribution over decorated neighborhoods and deriving the corresponding state transition dynamics; (ii) a DPP and MDP, establishing a Bellman equation and proving the existence of an optimal local policy under standard continuity and compactness; (iii) an algorithmic blueprint where we propose practical RL algorithms on sparse large graphs, that are theoretically justified as the RL solution to a well-defined MDP; and (iv) theoretical guarantees that show approximate optimality of mean-field solutions through a "propagation of chaos" type result, a policy gradient theorem for sparse graphs, as well as a rigorous justification for GNNs in RL on large sparse graphs. This work rigorously bridges the gap between classical mean-field learning algorithms and the reality of highly sparse systems, providing both a rigorous foundation and a practical pathway for designing decentralized control strategies on large networks. Several of our proof ingredients such as Bellman recursion and the likelihood-ratio trick are well-known. Our core contribution is structural: identifying the correct framework for sparse, non-exchangeable systems and deriving its consequences.

2 Background and Preliminaries

In this section, we expand upon classical MFC and our proposed framework. We begin by explaining how classical MFC and our approach fit together.

2.1 From Classical MFC to Graph-Lifted MFC

In a standard N-agent MFC problem, each agent $i \in \{1, \dots, N\}$ has a state $x_t^i \in \mathcal{X}$ and chooses an action $u_t^i \in \mathcal{U}$. The dynamics and reward for agent i depend on its own state and action, and on the empirical distribution of all agent states, $\mu_t^N = \frac{1}{N} \sum_{j=1}^N \delta_{x_t^j}$. In the limit $N \to \infty$, the control problem simplifies to solving a Bellman equation on the space $\mathcal{P}(\mathcal{X})$ for the value function $V(t, \mu)$; see also Gu et al. (2023) for a more thorough introduction. This formulation relies on the all-to-all interaction structure, which is a strong symmetry assumption: agents are exchangeable and interact only through the empirical distribution of states $\mu_t \in \mathcal{P}(X)$. While this assumption is natural when interactions are dense or effectively all-to-all, it is unreasonable in systems where

interactions are mediated by a sparse graph. From a modeling perspective, the failure is structural rather than technical. In sparse networks, an agent does not interact with a representative sample of the population, but with a small, random neighborhood whose composition varies across agents. As a result, the empirical state distribution μ_t is no longer a closed sufficient statistic for the dynamics or rewards, even in the limit $N \to \infty$.

Our framework can be understood as a minimal extension of classical MFC by lifting the mean field from states to *local environments*. Concretely, instead of tracking only the marginal distribution of agent states, we consider the distribution of *decorated rooted neighborhoods*,

$$\boldsymbol{\mu}_t \ \in \ \mathcal{P}(\mathcal{G}_*^{\mathcal{X}}),$$

where each element describes both the local interaction structure and the states of agents within that neighborhood. This lift has three key consequences.

(i) Closure under Sparse Dynamics. Local transition kernels and rewards depend on bounded neighborhoods by assumption. By taking the system state to be a distribution over such neighborhoods, the limiting dynamics become deterministic/Markovian at the level of $\boldsymbol{\mu}_t$, exactly as in classical MFC, but now without requiring dense interactions.

(ii) Recovery of Classical MFC as a Special Case. When the interaction graph is complete, or when rewards and dynamics depend only on the root state, the neighborhood distribution $\boldsymbol{\mu}_t$ collapses to the marginal state distribution in $\mathcal{P}(\mathcal{X})$. In this sense, classical mean-field control is a degenerate case of the graph-lifted formulation, corresponding to zero-radius neighborhoods.

(iii) Compatibility with Dynamic Programming. Although $\mathcal{P}(\mathcal{G}_*^{\mathcal{X}})$ is formally infinite-dimensional, for finite-horizon problems the relevant information is localized. As shown in Section 4, the optimal policy at time t depends only on $(T-t)$-hop neighborhoods, yielding a tractable DPP on progressively smaller lifted state spaces.

Viewed this way, graph-lifted MFC as we propose it in this work does not replace classical MFC but extends it; the mean field is enriched just enough to encode the interaction structure required while preserving the conceptual simplicity of a single representative control problem.

2.2 Sparse Random Graphs and Local Weak Convergence

Our work focuses on systems where agents are nodes in a large sparse graph, contrasting against simpler dense graphs where a MF may be computed for neighbors of a node. A canonical example is the Erdős-Rényi random graph $G(N, d/N)$. As $N \to \infty$, such graphs are *locally tree-like*, formalized by the concept of **local weak convergence** or Benjamini-Schramm convergence (Benjamini & Schramm, 2001). By a k-neighborhood of a node ρ in $\mathcal{G}$ we mean the set $\{v \in \mathcal{G} : \|v - \rho\| \leq k\}$, where the norm is defined through the graph distance. We will also call a k-neighborhood a k-hop neighborhood occasionally and denote it by $(\mathcal{G}, \rho)_k$.

Definition 1. *Let $\mathcal{G}_*$ be the space of rooted graphs. A sequence of finite, random graphs $(G_N)_{N\geq 1}$ converges to a random rooted graph $(\mathcal{G}, \rho)$ if for every fixed radius $k \geq 1$, the distribution of the k-hop neighborhood of a uniformly chosen vertex in G_N converges to the distribution of the k-hop neighborhood of the root ρ in $\mathcal{G}$.*

This implies that for a sequence of large sparse graphs $(G_N)_{N\in\mathbb{N}}$ that converges weakly locally to $(\mathcal{G}, \rho)$, agents typically experience a *distribution* of local environments.

All continuity statements below refer to the *marked local topology* on decorated rooted graphs (two elements are close when sufficiently large rooted balls are isomorphic and decorations are close) and the weak topology on $\mathcal{P}(\mathcal{G}_*^{\mathcal{X}})$, metrized by the bounded-Lipschitz metric d_{BL}. Marked Benjamini-Schramm convergence of G_N is then exactly weak convergence of the empirical measures $\boldsymbol{\mu}_t^N$, meaning the notions are compatible by construction. Possible limiting objects that we can treat this way include trees, grids, lattices, clustered graphs, and typed or community-structured sparse graphs qualify. Only Theorem 6 additionally requires a uniform maximum degree Δ.

3 The Sparse Mean-Field Control Framework

We now formalize the control problem on a large sparse graph over a finite time horizon T. In all what follows we write $\mathcal{P}(\Omega)$ to denote the space of probability measures over the set Ω. We also

require the notion of a decorated graph. Throughout, all spaces considered will be assumed to be (at least) standard Borel spaces. Before we introduce the formal definition, let us discuss a simple example. Consider a path $v_1 - v_2 - v_3$ with SIR states (S, I, R) viewed from the root $\rho = v_2$. Here, *rooted* means viewed from the distinguished agent and *decorated* that every node carries a state from $\mathcal{X}$. The lifted mean field $\boldsymbol{\mu}_t$ is the distribution of such local views for a uniformly sampled agent. Projecting to radius zero recovers the classical mean field in $\mathcal{P}(\mathcal{X})$.

Definition 2 (Decorated Rooted Graph)**.** *A decorated rooted graph is a pair $((\mathcal{G}, \rho), \mathbf{x})$, where $(\mathcal{G}, \rho)$ is a rooted graph and $\mathbf{x} : V(\mathcal{G}) \to \mathcal{X}$ is a state decoration. Let $\mathcal{G}_*^{\mathcal{X}}$ be the space of all such objects. For a radius k, let $\mathcal{G}_{*,k}^{\mathcal{X}}$ be the space of decorated k-neighborhoods.*

3.1 The Finite N-Agent System

Let $G_N = (V_N, E_N)$ be a graph with $|V_N| = N$ agents. Let $((G_N, i), \mathbf{x}_t)_k$ denote the k-hop decorated neighborhood of agent i at time t. The system is modeled as follows.

- **State**: Each agent $i \in V_N$ has a state $\mathbf{x}_t^i \in \mathcal{X}$. The joint state is $\mathbf{x}_t = (\mathbf{x}_t^i)_{i \in V_N} \in \mathcal{X}^N$. The initial states are given as (random) boundary conditions $\mathbf{x}_0$.
- **Action**: Each agent i takes an action $u_t^i \in \mathcal{U}$.
- **Policy**: Agents follow a shared, local policy π. The action u_t^i depends on the agent's local decorated neighborhood. For simplicity of exposition, consider nearest-neighbor interactions ($k = 1$):
$$u_t^i \sim \pi_t(\mathrm{d}u_t^i \mid (((G_N, i), \mathbf{x}_t)_1). \tag{1}$$
More generally, a policy π_t is a stochastic kernel $\pi_t : \mathcal{G}_*^{\mathcal{X}} \to \mathcal{P}(\mathcal{U})$ that provides an action $u_t \sim \pi_t(\cdot \mid ((\mathcal{G}, \rho), \mathbf{x}))$ for node ρ.
- **Dynamics**: The state of each agent evolves based on its local neighborhood:
$$\mathbf{x}_{t+1}^i \sim P(\mathrm{d}\mathbf{x}_{t+1}^i \mid ((G_N, i), \mathbf{x}_t)_1, u_t^i). \tag{2}$$
- **Objective**: Maximize the total expected reward:
$$J^N(\pi) = \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^{N} r(((G_N, i), \mathbf{x}_t)_1, u_t^i)\right] \tag{3}$$
Moreover, we assume $r : \mathcal{G}_{*,k}^{\mathcal{X}} \times \mathcal{U} \to \mathbb{R}$ to be continuous and bounded.

The choice $h = 1$ is for notational convenience only. For general radius h, the sufficient radius at time t is $h(T - t)$ and the DPP maps $h(T - t)$ to $h(T - t - 1)$. This removes the dependence on N but not on hT, since a radius-r ball may contain $O(\Delta^r)$ vertices and needs r message-passing rounds. Theorem 6 extends with an h-hop influence coefficient in place of α; see also Proposition 4.

The above setting describes a fully observable MDP, i.e. a standard RL problem, when we optimize over all policies. This is reflected in the fact that the joint states and actions of agents may be equivalently considered, as the MDP's states and actions. For tractability however, in standard MFC it is fruitful to introduce a *system state* given by a probability measure $\mu \in \mathcal{P}(\mathcal{X})$, as well as to look for feedback policies shared by all agents (Saldi et al., 2018) that are optimal under a certain class of policies. However, such a simple system state is insufficient in the setting of sparse graphs. The above shared-policy setting is the one we will study for the remainder of this paper, developing the suitable sparse mean-field limit for solving the RL problem.

3.2 The Limiting Sparse-MFC System

As $N \to \infty$, we assume G_N converges to a random rooted graph $(\mathcal{G}, \rho)$ in the Benjamini–Schramm sense. The state of our Sparse-MFC (SMFC) system at time t is therefore naturally given by a probability measure $\boldsymbol{\mu}_t \in \mathcal{P}(\mathcal{G}_*^{\mathcal{X}})$. By formally taking $N \to \infty$ in (3), the normalized sum converges to an integral over the space of decorated graphs, distributed according to some law determined by the Benjamini–Schramm limit. The evolution of the state is governed by a transition operator $\boldsymbol{\mu}_{t+1} = T_t(\boldsymbol{\mu}_t, \pi_t)$. Indeed, it is just the result of an application of the two kernels according to (1) and (2). One averages these dynamics according to the current system distribution, to obtain

$$\boldsymbol{\mu}_{t+1}(\mathrm{d}S) = \int \boldsymbol{\mu}_t(\mathrm{d}y) \int \Pi_t(\mathrm{d}\overline{u}|y)\overline{P}(\mathrm{d}S|y, \overline{u}). \tag{4}$$

Here, $\overline{u} = (u^{v_1}, ...)$ for some enumeration of $V(\mathcal{G})$ and $\Pi_t(\mathrm{d}\overline{u}|S) = \prod_{v\in\mathcal{G}} \pi_t(\mathrm{d}u^v|(S,v)_1)$. Similarly, $\overline{P}$ uses the state and the local action to sample the next state for every agent $v \in V(\mathcal{G})$. Note that the evolution operators $(T_t)_{t\in\mathbb{N}}$ yield a deterministic flow of probability measures in case the initial state $\boldsymbol{\mu}_0$ and the policy π is specified. We will usually omit explicit dependence on π when clear from context.

The objective functional in the limit is given with $S = (\mathcal{G}, \rho)$ being a rooted graph, by

$$J(\pi) = \mathbb{E}\Big[\sum_{t=0}^{T-1} \int_{\mathcal{G}_*^{\mathcal{X}}} \int_{\mathcal{U}} r((S,\mathbf{x}_t)_1, u_t)\pi_t(\mathrm{d}u_t \mid S)\, \boldsymbol{\mu}_t(\mathrm{d}S)\Big]$$

where it is important that the reward only takes local neighborhoods, meaning that taking $N \to \infty$ indeed yields a sensible limit by assumption. The value function of a sequence of policies $\pi = (\pi_t)_{t\leq T}$ with boundary conditions $\mathbf{x}_0$ is defined, with $0 < \gamma \leq 1$, by

$$V_t^{\pi}(\boldsymbol{\mu}) = \mathbb{E}^{\pi}\Big[\sum_{t'=t}^{T-1} \gamma^{t'} \int \boldsymbol{\mu}_{t'}(\mathrm{d}(\mathcal{G},\rho)) r\big(((\mathcal{G},\rho),\mathrm{x}_{t'})_1, u_{t'}\big)\Big],$$

where the expectation w.r.t. π is defined such that

$$\boldsymbol{\mu}_{t+1} = T_t(\boldsymbol{\mu}_t), \quad u_t^i \sim \pi_t\big(\mathrm{d}u_t^i \mid \big((\mathcal{G},i),\mathbf{x}_t\big)_1\big), \quad \mathbf{x}_{t+1}^i \sim P(\mathrm{d}\mathbf{x}_{t+1}^i \mid \big((\mathcal{G},i),\mathbf{x}_t\big)_1, u_t^i).$$

Finally, we abbreviate $V(\boldsymbol{\mu}) = \sup_\pi V^\pi(\boldsymbol{\mu})$. For convenience only we will always choose $\gamma = 1$. It can be seen that all our results extend to the case of general γ without problems.

3.3 Approximation Guarantee

The practical value of this framework depends on its ability to approximate the finite N-agent system, which we obtain with rigorous guarantees in the following.

Theorem 1 (Convergence of Optimal Values)**.** *Let $(G_N)_{N\geq 1}$ be a sequence of graphs converging locally to $(\mathcal{G},\rho)$. Assume that the space of admissible policies Π is compact and that the family $(J^N)_{N\in\mathbb{N}}$ is equicontinuous (with respect to the topology of weak convergence of policies) on Π. Then*

$$\lim_{N\to\infty} \sup_\pi J^N(\pi) \;=\; \sup_\pi J(\pi). \tag{5}$$

Proof. Fix an admissible policy π. By Lacker et al. (2023, Theorem 3.2), we have $J^N(\pi) \to J(\pi)$ as $N \to \infty$. Next, let $V_N^* = \sup_\pi J^N(\pi)$ and $V^* = \sup_\pi J(\pi)$. By assumption, the maps $\pi \mapsto J^N(\pi)$ and $\pi \mapsto J(\pi)$ are continuous; hence maximizers exist on the compact set of admissible policies. Pick $\pi^* \in \arg\max_\pi J$ and $\pi_N \in \arg\max_\pi J^N$. Then $V_N^* \geq J^N(\pi^*) \to J(\pi^*) = V^*$, so $\liminf_N V_N^* \geq V^*$. By compactness, along a subsequence $\pi_N \Rightarrow \bar{\pi}$ and by continuity, $V_N^* = J^N(\pi_N) \to J(\bar{\pi}) \leq V^*$, hence $\limsup_N V_N^* \leq V^*$. Therefore $V_N^* \to V^*$. □

Appendix A.1 gives verifiable sufficient conditions for equicontinuity and settings in which it can fail.

Corollary 1 (Approximate Optimality)**.** *An optimal policy π^* for the SMFC problem is ϵ-optimal for the N-agent problem, with $\epsilon \to 0$ for sufficiently large N.*

We want to note that, during execution, agents do not need the full graph information. Instead of taking the empirical measure of all neighborhood distributions in the graph, the neighborhoods can be approximated (due to the assumed Benjamini-Schramm limit) by subsampling from a given graph distribution / pooling only parts of the graph. The same also goes for the initial distributions. Moreover, the meta-policy generates the parameters for a local policy (e.g., once per timestep or via a broadcasted signal). Therefore, for known initial conditions, agents can determine their action using only their 1-hop neighborhood. The MFC approach hence allows a "Centralized Training, Decentralized Execution" (CTDE) scheme; see also Cui et al. (2024) for more information.

4 Reinforcement Learning on Sparse Graphs

In what follows, we will develop an SMFC-based RL algorithm applicable to sparse graphs, that will naturally fit with GNNs, and is theoretically well-founded through solving a limiting MDP, whose foundation is proved in the following.

4.1 Horizon-Dependent Local Dynamic Programming

To begin, a key insight is that for a finite horizon, we do not need to consider the distribution over infinite graphs.

Theorem 2 (Horizon-Dependent Locality)**.** *Consider the SMFC problem with horizon T. The value function V_t depends only on the distribution of $(T-t)$-hop decorated neighborhoods, i.e., $V_t : \mathcal{P}(\mathcal{G}^{\mathcal{X}}_{*,T-t}) \to \mathbb{R}$. In other words, V_t is measurable w.r.t. the information contained in $\mathcal{G}^{\mathcal{X}}_{*,T-t}$. Consequently, the optimal policy π^*_t at time t depends only on the $(T-t)$-hop decorated neighborhood of the root.*

Remark 1. *Theorem 2 is crucial for computation. It implies that for a T-horizon problem, the state space is not $\mathcal{P}(\mathcal{G}^{\mathcal{X}}_*)$ but $\mathcal{P}(\mathcal{G}^{\mathcal{X}}_{*,T})$. On a regular graph like a grid or a d-regular tree, the number of non-isomorphic k-hop neighborhoods is finite. This means the state space $\mathcal{P}(\mathcal{G}^{\mathcal{X}}_{*,k})$ becomes computationally tractable.*

Note that a k-layer GNN requires k communication rounds and $O(k|E|d_h)$ scalars per decision step, so the exact radius $T-t$ costs $O(T^2|E|d_h)$ per episode, and global pooling requires centralized aggregation or an aggregation-and-broadcast protocol (a CTDE scheme). Moreover, "tractable" means that the dependence on N is removed: $\mathcal{P}(\mathcal{G}^{\mathcal{X}}_{*,T-t})$ may still grow as $O(\Delta^{T-t})$ and is not finite-dimensional for continuous decorations.

This locality result leads to a well-defined DPP/MDP on a sequence of progressively smaller state spaces.

Theorem 3 (Dynamic Programming Principle)**.** *The value function $V_t : \mathcal{P}(\mathcal{G}^{\mathcal{X}}_{*,T-t}) \to \mathbb{R}$ satisfies the Bellman equation:*

$$\begin{cases} V_T(\cdot) = 0, \\ V_t(\boldsymbol{\mu}) \;=\; \sup\limits_{\pi_t} \Big\{ \mathbb{E}_{S\sim\boldsymbol{\mu}}\big[\mathbb{E}_{u\sim\pi_t(\cdot|S)}[r]\big] + V_{t+1}\big(T_t(\boldsymbol{\mu},\pi_t)\big) \Big\}, \qquad t < T. \end{cases} \tag{6}$$

where $\boldsymbol{\mu} \in \mathcal{P}(\mathcal{G}^{\mathcal{X}}_{*,T-t})$ and T_t maps to a distribution over $(T-t-1)$-hop neighborhoods. An optimal policy π^*_t exists.

This DPP gives rise to a natural solution as an MDP and via using RL, such as value-based or policy-based RL. In particular, note that the MDP is however at least continuous or in the worst case infinite-dimensional, since the action space is now given by the initial policy space of the finite graph problem. Therefore, in the following we will formulate a policy gradient algorithm based on this limiting SMFC problem, since the value-based approach is intractable.

4.2 Policy Gradients on Sparse Graphs

To provide a solid foundation for the proposed RL algorithm, we derive a policy gradient theorem for the sparse mean-field objective. We treat the problem as a "Lifted MDP" where the state is the neighborhood distribution $\boldsymbol{\mu}$.

To handle the complexity of the space, we employ a hierarchical MFC policy structure:

1. A **Meta-Policy** $\hat{\pi}_\theta(\psi \mid \boldsymbol{\mu})$ parameterized by θ, which observes the global mean-field state $\boldsymbol{\mu}$ and samples a local policy parameter $\psi \in \Psi$.
2. A **Local Policy** $\pi_\psi(u \mid s)$ parameterized by ψ, which agents use to select actions u based on their local neighborhood s.

Assumption 1 (Regularity)**.** *The meta-policy $\hat{\pi}_\theta(\psi \mid \boldsymbol{\mu})$ is differentiable with respect to θ and satisfies the standard score function requirements. Specifically, $\nabla_\theta \log \hat{\pi}_\theta(\psi \mid \boldsymbol{\mu})$ exists and is bounded. The reward r and transition dynamics are bounded.*

Fix the horizon T. Let $\boldsymbol{\mu}_0$ be the initial distribution. The evolution of the system is deterministic at the level of distributions given the chosen parameters ψ_t:

$$\boldsymbol{\mu}_{t+1} = T_t(\boldsymbol{\mu}_t, \pi_{\psi_t}).$$

However, since the meta-policy $\hat{\pi}_\theta$ is stochastic, the trajectory of parameters $\tau = (\psi_0, \dots, \psi_{T-1})$ is random, inducing a probability distribution over the trajectories of $\boldsymbol{\mu}_t$.

Theorem 4 (Policy Gradient for the Limiting SMFC)**.** *Consider the limiting SMFC problem formulated as a Lifted MDP on the state space $\mathcal{P}(\mathcal{G}_*^{\mathcal{X}})$. The gradient of the objective $J(\theta) = \mathbb{E}[\sum_{t=0}^{T-1} R(\boldsymbol{\mu}_t, \psi_t)]$ with respect to the meta-parameters θ is given by*

$$\nabla_\theta J(\theta) = \sum_{t=0}^{T-1} \mathbb{E}_{\boldsymbol{\mu}_t, \psi_t} \left[\nabla_\theta \log \hat{\pi}_\theta(\psi_t \mid \boldsymbol{\mu}_t) \cdot Q_t^{MF}(\boldsymbol{\mu}_t, \psi_t) \right], \tag{7}$$

where $Q_t^{MF}(\boldsymbol{\mu}, \psi)$ is the cumulative future reward starting from state $\boldsymbol{\mu}$ with action ψ:

$$Q_t^{MF}(\boldsymbol{\mu}, \psi) = R(\boldsymbol{\mu}, \psi) + \mathbb{E}_{\tau > t} \left[\sum_{k=t+1}^{T-1} R(\boldsymbol{\mu}_k, \psi_k) \right], \tag{8}$$

with $R(\boldsymbol{\mu}, \psi) \coloneqq \int r((s)_1, u) \pi_\psi(\mathrm{d}u|s) \boldsymbol{\mu}(\mathrm{d}s)$.

The proof is provided in Appendix C. We note that the form of this gradient is a standard result for MDPs; the novelty lies in the definition of the state space $\boldsymbol{\mu}$ and the rigorous limit justifying it.

Finite-Graph Policy Gradient Approximation. The core practical challenge is that we cannot compute the gradient on the infinite limit directly. Instead, we compute the gradient on a finite graph G_N. Explicitly, the gradient we compute in Algorithm 1 is:

$$\nabla_\theta J^N(\theta) = \mathbb{E} \left[\sum_{t=0}^{T-1} \left(\sum_{k=t}^{T-1} \bar{r}_k^N \right) \nabla_\theta \log \hat{\pi}_\theta(\psi_t \mid \boldsymbol{\mu}_t^N) \right], \tag{9}$$

where $\boldsymbol{\mu}_t^N$ is the empirical distribution of decorated neighborhoods in G_N, and $\bar{r}_k^N = \frac{1}{N} \sum_{i=1}^N r(s_k^i, u_k^i)$ is the average reward at time k.

We must establish that this finite computation approximates the true gradient of the underlying SMFC MDP.

Theorem 5 (Finite-Graph PG Approximation)**.** *Let $J^N(\theta)$ be the objective on the finite graph G_N and $J(\theta)$ be the limiting objective. Assume that (i) $\boldsymbol{\mu}_0^N \Rightarrow \boldsymbol{\mu}_0$ in the marked local topology; (ii) the finite transition and reward operators converge locally uniformly to their limits, uniformly over the compact policy-parameter set; (iii) the local transition kernel is Feller and Lipschitz with respect to d_{BL} and the action; and (iv) rewards and policy-score functions are bounded and continuous. Then centralized training on the finite graph asymptotically optimizes the infinite limit:*

$$\lim_{N \to \infty} \left\| \nabla_\theta J^N(\theta) - \nabla_\theta J(\theta) \right\| = 0.$$

Our proof uses forward induction to obtain $\boldsymbol{\mu}_t^N \Rightarrow \boldsymbol{\mu}_t$ for every $t \leq T$, backward induction in the finite-horizon Bellman recursion to obtain locally uniform convergence $Q_t^N \to Q_t$, and bounded convergence to pass to the policy gradient. For details see Appendix C.1. We want to point out that the previous result validates Alg. 1: we can use standard PPO on the global graph state to optimize the mean-field control problem. We have thus motivated a practical reinforcement learning approach based on local function approximation. The local policy π_t takes a decorated neighborhood as input. Using a GNN for $\hat{\pi}$, the resulting algorithm is Algorithm 1, with its application to graphs visualized in Figure 2. By design, the proposed algorithm is sensitive w.r.t. the local neighborhood distribution. On the other hand, if two different graphs have similar neighborhood distributions, SMFAC generalized its knowledge from one graph to the other.

4.3 Theoretical Analysis of GNN Usage

Our framework relies on two key approximations: (1) using a finite radius k (via GNN depth) to approximate the optimal policy which formally requires radius $T - t$, and (2) using a finite graph G_N to approximate the infinite limit. Here we provide theoretical bounds for both.

For a message-passing GNN, one layer propagates information across one edge; inductively, after k layers the embedding at a node depends exactly on its decorated k-hop neighborhood. Hence the radius-$(T-t)$ sufficient statistic of Theorem 2 corresponds to a $(T-t)$-layer receptive field, whereas Theorem 6 justifies replacing it by a shallower k-layer architecture under weak spatial influence. Using an r-hop message-passing kernel per layer grows the receptive field as rk, which can reduce the depth needed for a given accuracy; the framework is otherwise agnostic to the message-passing architecture.

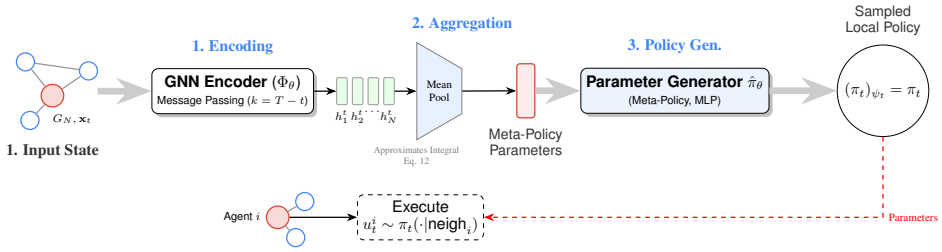

Figure 2: **The Sparse Mean-Field RL Architecture.** The computational flow of Algorithm 1. (1) The GNN encodes the local structure of the entire graph G_N. (2) A global pooling operation aggregates these embeddings to obtain a representation of the mean-field state $\boldsymbol{\mu}_t$. (3) The parameter generator (meta-policy actor $\hat{\pi}_\theta$) uses the mean-field state-dependent output to generate the parameters for the local policy π_t. In our implementation, π_t is parameterized by a set of parameters we denote by ψ_t. Individual agents i then execute actions u_t^i by conditioning the shared policy π_t on their specific local neighborhoods.

4.3.1 Approximation by Truncated Policies.

While Theorem 2 guarantees that the radius $T - t$ is sufficient, practical GNNs often use a fixed, smaller depth $k \ll T$. We show that if the dynamics are in some suitable sense not too much influenced by a single decoration in the 1-hop neighborhood, then the approximation error is exponentially small. For the exact assumptions, we refer the reader to the Appendix.

Theorem 6 (Truncation Error)**.** *Let Δ be the maximum degree and α the one-site influence from Assumption 2, so that $\alpha\Delta^2 < 1$. Choose any $\beta \in (\alpha\Delta, 1/\Delta)$ and set $\kappa = \alpha\Delta/\beta < 1$ and $q = \Delta\beta < 1$. Let π^* be the optimal policy (requiring radius T) and let $\pi^{(k)}$ be the optimal policy constrained to use only k-hop information. With C_r the reward bound,*

$$0 \leq J(\pi^*) - J(\pi^{(k)}) \leq \frac{2C_r}{\beta(1-\kappa)(1-q)}\, T\, q^{k+1}. \tag{10}$$

In particular, $k = O(\log(T/\varepsilon)/|\log q|)$ suffices for absolute error ε.

The constant we obtain is independent of T and k; the factor T in our bound corresponds to accumulated per-step truncation bias and seems hard to avoid for undiscounted cumulative reward. The condition $\alpha\Delta^2 < 1$ is sufficient, not necessary. Indeed, low-degree graphs tolerate strong local interactions, high-degree graphs require weaker per-edge influence, and near a phase transition $q \to 1$ the bound becomes vacuous. For weighted graphs, a Dobrushin condition on the influence matrix replaces the maximum-degree estimate (Appendix D). Shallow GNNs are therefore justified when interactions are not pathological. Our bound isolates truncation bias, while GNN approximation and optimization errors are separate (Appendix G).

4.3.2 GNNs as Universal Mean-Field Approximators

A central theoretical question is whether Graph Neural Networks are capable of representing the functions required for Sparse MFC, specifically the value function $V(\boldsymbol{\mu})$ and the meta-policy $\hat{\pi}(\boldsymbol{\mu})$. Since the state $\boldsymbol{\mu}$ is a probability measure over the space of decorated rooted graphs $\mathcal{G}_{*,k}^{\mathcal{X}}$, any approximator must satisfy two key geometric properties: (1) permutation invariant with respect to agent indices (measure property), and (2) isomorphism invariant with respect to the local neighborhood structure (the graph property). We analyze this through the lens of Geometric Deep Learning (Bronstein et al., 2021), decomposing the approximation into a global aggregation step and a local encoding step.

Global Aggregation (Deep Sets). The value function $V : \mathcal{P}(\mathcal{G}_{*,k}^{\mathcal{X}}) \to \mathbb{R}$ is a function on a set (or measure). The Deep Sets theorem (Zaheer et al., 2017) establishes that any continuous permutation-invariant function on a set can be decomposed into the form $\rho(\sum \phi(x))$. In our mean-field context, this implies that the value function can be approximated as:

$$V(\boldsymbol{\mu}) \approx \rho\left(\int_{\mathcal{G}_{*,k}^{\mathcal{X}}} \phi(\mathfrak{g})\, \mathrm{d}\boldsymbol{\mu}(\mathfrak{g})\right), \tag{11}$$

where ϕ is a continuous embedding of a rooted graph, and ρ is a continuous function (e.g., an MLP).

Local Encoding (Weisfeiler-Leman). The function $\phi(\mathfrak{g})$ must encode the topology of the rooted neighborhood $\mathfrak{g}$. It is established that Message Passing Neural Networks (MPNNs) are as powerful as the 1-dimensional Weisfeiler-Leman (1-WL) graph isomorphism test (Xu et al., 2019; Morris et al., 2019; Maron et al., 2019). Therefore, parameterizing ϕ as a GNN allows us to distinguish all neighborhood structures distinguishable by the 1-WL test.

Proposition 1 (GNN Readout as Mean-Field Integration)**.** *Let $\Phi_\theta : \mathcal{G}^{\mathcal{X}}_{*,k} \to \mathbb{R}^d$ denote a k-layer MPNN parameterized by θ. The operation of applying the GNN to the global graph G_N and performing global mean pooling is mathematically equivalent to the Deep Sets decomposition over the empirical measure $\boldsymbol{\mu}_t^N$:*

$$\underbrace{\frac{1}{N}\sum_{i=1}^{N}\Phi_\theta\big((G_N,i)_k\big)}_{\textit{Global Average Pooling}} \;=\; \underbrace{\int_{\mathcal{G}^{\mathcal{X}}_{*,k}}\Phi_\theta(\mathfrak{g})\,\boldsymbol{\mu}_t^N(\mathrm{d}\mathfrak{g})}_{\textit{Expectation over Mean Field}}. \tag{12}$$

This identity provides the rigorous justification for our algorithmic choice. It confirms that the standard "GNN + Readout" architecture is not merely a heuristic, but a direct implementation of the function approximation required for the Mean-Field state space. Consequently, we approximate the global value function (Critic) and similarly the meta-policy as:

$$V_\phi(\boldsymbol{\mu}_t^N) \approx \rho_\phi\left(\frac{1}{N}\sum_{i=1}^{N}\Phi_{\theta_{\text{enc}}}\big((G_N,i)_k\big)\right). \tag{13}$$

Remark 2 (Expressivity Limitations)**.** *While this framework is robust, it inherits the limitations of the 1-WL test. If the optimal control depends on distinguishing local structures that are 1-WL equivalent (e.g., certain circular skip-connections vs. disconnected components), standard MPNNs may underfit. In such cases, the framework naturally extends to higher-order GNNs (k-GNNs) (Morris et al., 2019) without altering the mean-field logic.*

4.4 Empirical Summary

We validate SMFAC on two synthetic SIR tasks; the details are given in Appendix F.4. On a 20×20 grid, aggregate mean-field control tends to uniform intervention as it cannot distinguish neighborhoods. SMFAC, on the other hand, learns selective interventions, achieving lower loss (Fig. 3). On 20 disjoint triangles, concentrated and dispersed infections have identical global statistics but require different actions in our setup; a zero-message-passing controller cannot separate them, while a one-layer GNN can (Figs. 4 and 5).

5 Conclusion and Discussion

We have presented a rigorous framework for MFC on sparse graphs by lifting the system state to distributions over decorated neighborhoods. Our horizon-dependent locality theorem yields a tractable DPP and justifies GNNs as local policy approximators.

Future research could extend this to non-cooperative Sparse MFGs by coupling a local Hamilton-Jacobi-Bellman equation with a Fokker-Planck-Kolmogorov equation for the neighborhood evolution. For discounted infinite-horizon problems (meaning $\gamma < 1$), combining Bellman and spatial contraction gives a horizon-uniform bound of order $q^{k+1}/[\beta(1-\gamma\kappa)(1-q)(1-\gamma)]$, so depth grows only logarithmically in $1/(1-\gamma)$. The average-reward case requires recurrence or ergodicity and is left open. Limitations, including the bounded-degree requirement of Theorem 6, adversarial topologies, and broader societal considerations, are discussed in Appendix G.

Acknowledgements

This research has been partially funded by the German Federal Ministry of Education and Research (BMBF) through grant 01IS23067 Software Campus 3.0 Technische Universität Darmstadt as part of the Software Campus project.

References

Anand, E., Hoffmann, R., Liaw, S., and Wierman, A. Graphon mean-field subsampling for cooperative heterogeneous multi-agent reinforcement learning. *arXiv preprint arXiv:2602.16196*, 2026.

Andersson, D. and Djehiche, B. A maximum principle for sdes of mean-field type. *Applied Mathematics & Optimization*, 63:341–356, 2011.

Benjamini, I. and Schramm, O. Recurrence of distributional limits of finite planar graphs. *Electronic Journal of Probability*, 6, 2001.

Bensoussan, A., Frehse, J., Yam, P., et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

Bullo, F., Cortés, J., and Martinez, S. *Distributed Control of Robotic Networks: A Mathematical Approach to Motion Coordination Algorithms*. Applied Mathematics. Princeton University Press, Princeton, NJ, 2009.

Caines, P. E. and Huang, M. Graphon mean field games and their equations. *SIAM Journal on Control and Optimization*, 59(6):4373–4399, 2021.

Canese, L., Cardarilli, G. C., Di Nunzio, L., Fazzolari, R., Giardino, D., Re, M., and Spanò, S. Multi-agent reinforcement learning: A review of challenges and applications. *Applied Sciences*, 11 (11):4948, 2021.

Carmona, R., Delarue, F., et al. *Probabilistic Theory of Mean Field Games with Applications I-II*. Springer, 2018.

Cui, K. and Koeppl, H. Learning graphon mean field games and approximate nash equilibria. In *International Conference on Learning Representations*, 2022.

Cui, K., Hauck, S. H., Fabian, C., and Koeppl, H. Learning decentralized partially observable mean field control for artificial collective behavior. In *The Twelfth International Conference on Learning Representations*, 2024.

Dörfler, F., Chertkov, M., and Bullo, F. Synchronization in complex oscillator networks and smart grids. *Proceedings of the National Academy of Sciences*, 110(6):2005–2010, 2013.

Fabian, C., Cui, K., and Koeppl, H. Learning sparse graphon mean field games. In *International Conference on Artificial Intelligence and Statistics*, pp. 4486–4514. PMLR, 2023.

Fabian, C., Cui, K., and Koeppl, H. Learning mean field games on sparse graphs: A hybrid graphex approach. In *International Conference on Learning Representations (ICLR)*, 2024.

Fabian, C., Cui, K., and Koeppl, H. Learning mean field control on sparse graphs. In *Forty-second International Conference on Machine Learning*, 2025.

Gronauer, S. and Diepold, K. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 55(2):895–943, 2022.

Gu, H., Guo, X., Wei, X., and Xu, R. Mean-field controls with q-learning for cooperative marl: Convergence and complexity analysis. *arXiv preprint arXiv:2002.04131*, 2020.

Gu, H., Guo, X., Wei, X., and Xu, R. Mean-field multi-agent reinforcement learning: A decentralized network approach. *arXiv preprint arXiv:2108.02731*, 2021.

Gu, H., Guo, X., Wei, X., and Xu, R. Dynamic programming principles for mean-field controls with learning. *Operations Research*, 71(4):1040–1054, 2023.

Guan, Y., Afshari, M., and Tsiotras, P. Zero-sum games between mean-field teams: Reachability-based analysis under mean-field sharing. In *Proc. AAAI*, volume 38, pp. 9731–9739, 2024.

Guo, X., Hu, A., Xu, R., and Zhang, J. Learning mean-field games. In *Advances in Neural Information Processing Systems*, pp. 4966–4976, 2019.

Hagberg, A. A., Schult, D. A., and Swart, P. J. Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J. (eds.), *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, pp. 11–15, Pasadena, CA, USA, August 2008. doi: 10.25080/TCWV9851.

Hu, Y., Wei, X., Yan, J., and Zhang, H. Graphon mean-field control for cooperative multi-agent reinforcement learning. *Journal of the Franklin Institute*, 360(18):14783–14805, 2023.

Huang, M., Malhamé, R. P., Caines, P. E., et al. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. *Communications in Information & Systems*, 6(3):221–252, 2006.

Jackson, M. O. *Social and Economic Networks*. Princeton University Press, Princeton, NJ, 2008.

Jiang, J., Dun, C., Huang, T., and Lu, Z. Graph convolutional reinforcement learning. In *International Conference on Learning Representations*, 2020.

Kallenberg, O. *Foundations of modern probability*. Springer, 1997.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization, 2014. Published as a conference paper at ICLR 2015.

Lacker, D., Ramanan, K., and Wu, R. Local weak convergence for sparse networks of interacting processes. *The Annals of Applied Probability*, 33(2):843–888, 2023.

Lasry, J.-M. and Lions, P.-L. Mean field games. *Japanese journal of mathematics*, 2(1):229–260, 2007.

Laurière, M., Perrin, S., Geist, M., and Pietquin, O. Learning mean field games: A survey. *arXiv preprint arXiv:2205.12944*, 2022a.

Laurière, M., Perrin, S., Girgin, S., Muller, P., Jain, A., Cabannes, T., Piliouras, G., Pérolat, J., Elie, R., Pietquin, O., et al. Scalable deep reinforcement learning algorithms for mean field games. In *International Conference on Machine Learning*, pp. 12078–12095. PMLR, 2022b.

Maron, H., Ben-Hamu, H., Serviansky, H., and Lipman, Y. Provably powerful graph networks. *Advances in neural information processing systems*, 32, 2019.

Morris, C., Ritzert, M., Fey, M., Hamilton, W. L., Lenssen, J. E., Rattan, G., and Grohe, M. Weisfeiler and leman go neural: Higher-order graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 4602–4609, 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8026–8037, 2019.

Perrin, S., Perolat, J., Laurière, M., Geist, M., Elie, R., and Pietquin, O. Fictitious play for mean field games: Continuous time analysis and applications. *arXiv preprint arXiv:2007.03458*, 2020.

Perrin, S., Laurière, M., Pérolat, J., Élie, R., Geist, M., and Pietquin, O. Generalization in mean field games by learning master policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9413–9421, 2022.

Qu, G., Lin, Y., Wierman, A., and Li, N. Scalable multi-agent reinforcement learning for networked systems with average reward. *Advances in Neural Information Processing Systems*, 33:2074–2086, 2020.

Rashid, T., Samvelyan, M., De Witt, C. S., Farquhar, G., Foerster, J., and Whiteson, S. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

Saldi, N., Basar, T., and Raginsky, M. Markov–nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K., et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2085–2087, 2018.

Tang, H., Zhang, Y., Wang, C., Shi, Z., Zhang, E., Zeng, B., and Ding, W. SKARL: Provably scalable kernel mean field reinforcement learning for variable-size multi-agent systems, 2026. URL `https://openreview.net/forum?id=VJZ56DDkin`.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Vicsek, T., Czirók, A., Ben-Jacob, E., Cohen, I., and Shochet, O. Novel type of phase transition in a system of self-driven particles. *Physical Review Letters*, 75(6):1226, 1995.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., and Wang, J. Mean field multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 5571–5580. PMLR, 2018.

Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R., and Smola, A. Deep sets. In *Advances in neural information processing systems*, volume 30, 2017.

Zhang, F., Tan, V., Wang, Z., and Yang, Z. Learning regularized monotone graphon mean-field games. *Advances in Neural Information Processing Systems*, 36, 2024a.

Zhang, F., Tan, V. Y., Wang, Z., and Yang, Z. Learning regularized graphon mean-field games with unknown graphons. *Journal of Machine Learning Research*, 25(372):1–95, 2024b.

Zhou, F., Zhang, C., Chen, X., and Di, X. Graphon mean field games with a representative player: Analysis and learning algorithm. In *International Conference on Machine Learning (ICML)*, 2024.

Overview of the proofs. The common thread in our proofs is finite-speed propagation. The horizon-dependent locality theorem follows from a direct sigma-algebra induction: at each time step, the state and action of a node can depend only on information from one graph hop further away, so after $T - t$ steps only the $(T - t)$-ball can influence the root. Closure of the lifted state follows because the one-hop transition kernel maps a decorated rooted neighborhood to a new decorated rooted neighborhood. Averaging over the local action kernel and the population law yields the deterministic transition T_t on $\mathcal{P}(\mathcal{G}_*^{\mathcal{X}})$. The DPP is then the standard finite-horizon Bellman recursion on this closed state space, with compactness and continuity ensuring that the supremum is attained. Finally, the truncation theorem uses spatial contraction. Under $\alpha\Delta^2 < 1$, a disagreement at one decoration contracts geometrically as it propagates, so decorations outside the k-ball contribute at most $q^{k+1}/(1-q)$ to the weighted disagreement distance. Summing the resulting per-step bias over the horizon gives the $O(Tq^k)$ bound.

A Proof of Theorem 2

We start with the following observation.

Proposition 2. *Let $(E, \mathcal{E})$ be a measurable space and let $(S, \mathcal{S})$ be a standard Borel space. Let*

$$\pi : E \times \mathcal{S} \to [0, 1]$$

be a Markov kernel, i.e. for each $x \in E$, $\pi(x, \cdot)$ is a probability measure on $(S, \mathcal{S})$, and for each $A \in \mathcal{S}$, the map $x \mapsto \pi(x, A)$ is $\mathcal{E}$-measurable.

Then there exist a measurable function

$$f : E \times [0, 1] \to S$$

and a random variable $U \sim \mathrm{Unif}[0, 1]$ such that for every fixed $x \in E$,

$$f(x, U) \sim \pi(x, \cdot).$$

Equivalently, for all $A \in \mathcal{S}$,

$$\mathbb{P}\big(f(x, U) \in A\big) = \pi(x, A).$$

The proof of this Proposition is standard and can be found in slightly different phrasing in e.g. Kallenberg (1997). In order to formalize what we mean by "the value function only depending on the local neighborhood", we use the basic concept of measurability. By what we have just shown, every Markov kernel π can be identified with a function f_π and a random variable U_π, effectively factoring out the randomness into U_π. We can therefore associate a random variable U_t^v and W_t^v to the evolution of agent v at time t, where U_t^v corresponds to the action choice of agent sitting at node v at time t (i.e., using π_t). Similarly, W_t^v describes the state evolution of agent v at time t [1].

For any rooted graph $(\mathcal{G}, \rho)$ we define the σ-algebra

$$\mathcal{F}_k^\rho := \sigma\left(\{\mathbf{x}_0^v, U_t^v, W_t^v : \|v - \rho\| \le k, t < k\}\right).$$

The definition of $\mathcal{F}_k$ tracks all probabilistic inputs that are needed in order to determine the state of the root node at time k (and hence its action at time k).

Lemma 1. *For any $i \in V(\mathcal{G})$, $\mathbf{x}_t^i$ is measurable with respect to $\mathcal{F}_t^i$.*

Proof. The statement is verified by induction. For $t = 0$ and arbitrary $i \in V(\mathcal{G})$, $\mathbf{x}_0^i$ is measurable with respect to $\mathcal{F}_0^i$ by definition. For the induction step, assume $\mathbf{x}_t^v$ is measurable with respect to $\mathcal{F}_t^v$ for every node v in the graph. By locality of the dynamics, $\mathbf{x}_{t+1}^i$ is a measurable function of the 1–hop neighborhood states $\{\mathbf{x}_t^v : v \in (\mathcal{G}, i)_1\}$ together with the random variables U_t^i and W_t^i. Since for each $v \in (\mathcal{G}, i)_1$ we have $\|v - i\| \le 1$, the sigma-algebra generated by $\bigcup_{v \in (\mathcal{G}, i)_1} \mathcal{F}_t^v$ is contained in $\mathcal{F}_{t+1}^i$. Moreover, U_t^i and W_t^i are contained in $\mathcal{F}_{t+1}^i$ by definition. Hence $\mathbf{x}_{t+1}^i$ is measurable with respect to $\mathcal{F}_{t+1}^i$, which closes the induction. □

[1] Note that while here the transition kernel stays the same, we still need a new random variable for each agent at each time. The difference to the policy is that these random variables all parameterize the same function, where as in the policy case, for each time t, different functions are parameterized.

We allow the underlying graph structure to be chosen randomly. In order to obtain full information on the graph distribution, we not only need information about the current states of the agents, but also about the graph structure. We therefore introduce the σ-algebra

$$\mathcal{K}_k^\rho := \sigma\left(\mathcal{F}_k^\rho, \mathcal{J}\right),$$

where $\mathcal{J}$ is the canonical σ-algebra on the space of graphs with at most countably many nodes.

Proof of Theorem 2. Recall the definition of V_t from (3.2). Fix $t \in \{0, \dots, T-1\}$ and set $k := T-t$. By Lemma 1, the random variables $((\mathcal{G}, \rho), \mathbf{x}_s)_1$ for $s = t, t+1, \dots, T$ are measurable with respect to $\mathcal{K}_k^\rho$ (since information propagates at unit speed and $k = T-t$). Since instantaneous rewards depend only on the 1–hop neighborhood, the reward-to-go from time t to $T-1$ (and hence the value V_t) is $\mathcal{K}_k^\rho$-measurable. This shows that V_t depends only on the k-neighborhood information, i.e. on the local neighborhood. □

A.1 Verification of Equicontinuity for Theorem 1

The following sufficient conditions cover the finite-state epidemic models used in our experiments and are representative of the bounded-degree setting.

Proposition 3. *Suppose that (i) $\mathcal{X}$ and $\mathcal{U}$ are finite; (ii) $\sup_N \max_{v \in G_N} \deg(v) \le \Delta < \infty$; (iii) rewards are bounded and depend on a fixed-radius neighborhood; (iv) transition kernels are uniformly Lipschitz in their local state-action inputs; and (v) the policy class is parameterized by a compact metric space (Θ, d_Θ) with*

$$\sup_{t,S} \left\| \pi_t^\theta(\cdot \mid S) - \pi_t^{\theta'}(\cdot \mid S) \right\|_{\mathrm{TV}} \le L_\pi \, d_\Theta(\theta, \theta').$$

Then the family (J^N) is equicontinuous and the conclusion of Theorem 1 holds.

Proof sketch. Couple the actions under π^θ and $\pi^{\theta'}$ maximally and use the same transition randomness whenever the local inputs agree. Since the horizon and the interaction radius are finite and local neighborhood growth is uniformly bounded in N, a policy disagreement can affect only a uniformly bounded number of space-time locations before time T. A finite-horizon coupling therefore gives

$$|J^N(\pi^\theta) - J^N(\pi^{\theta'})| \le C_{T,\Delta} \, d_\Theta(\theta, \theta'),$$

where $C_{T,\Delta}$ may grow with T and Δ but is independent of N. This is equicontinuity. The same argument extends to compact metric state-action spaces using bounded-Lipschitz or Wasserstein estimates instead of total variation. □

Uniform equicontinuity can fail for example when degrees or local influence grow with N without a compensating moment bound or the interaction radius or horizon grows with N The finite-state SIR model on a bounded-degree graph satisfies the assumptions above, since its local transition probabilities are Lipschitz in the local infection configuration, transmission parameters, and randomized control probabilities.

A.2 General Interaction Radius

Proposition 4. *Suppose that rewards and transition kernels depend on at most h-hop neighborhoods. Then the value function at time t depends only on the decorated $h(T-t)$-hop neighborhood, i.e.*

$$V_t : \mathcal{P}\big(\mathcal{G}^{\mathcal{X}}_{*,h(T-t)}\big) \to \mathbb{R},$$

and the DPP transition maps radius $h(T-t)$ to radius $h(T-t-1)$. A direct implementation requires $h(T-t)$ message-passing rounds at time t, and on a bounded-degree graph the input may contain $O(\Delta^{h(T-t)})$ vertices.

Proof sketch. Replace the one-hop sigma-algebra in the proof of Theorem 2 by the h-hop sigma-algebra and repeat the induction: each time step expands the region that can influence the root by at most h hops, so after $T-t$ steps only the $h(T-t)$-ball matters. The complexity statement counts the vertices in that ball and the number of sequential communication rounds. □

For Theorem 6, the one-hop influence coefficient α is replaced by an h-hop Dobrushin coefficient that accounts for the number and strength of dependencies within an h-ball; the exponential truncation conclusion persists only when this enlarged coefficient is strictly subcritical.

B Proof of Theorem 3

Recall that the Bellman equation simply states that the value of a state $\boldsymbol{\mu}$ is simply given as the maximum of the sum of the immediate best reward and the best possible future value. In order to prove (6) we follow the ideas from Gu et al. (2023). For any action kernel h and sequence of controls $\pi = (\pi_t)_{t \leq T}$, define the IQ function

$$Q_t(\boldsymbol{\mu}, h) := \sup_{\pi \in \Pi_1} Q_t^{\pi}(\boldsymbol{\mu}, h).$$

where

$$Q_t^{\pi}(\boldsymbol{\mu}, h) := \mathbb{E}^{\pi}\left[\sum_{\ell=t}^{T-1} \gamma^{\ell-t} r(((\mathcal{G}, \rho), \mathbf{x}_\ell)_1, u_\ell)\right],$$

with $0 < \gamma \leq 1$. We abbreviate $Q(\boldsymbol{\mu}, h) = Q_0(\boldsymbol{\mu}, h)$. The expectation on the rhs is taken according to $((\mathcal{G}, \rho), \mathbf{x}_t)_1 \sim \boldsymbol{\mu}$ and $u_t \sim h$.

Proof of Theorem 3. In order to simplify notation we will abbreviate a state of the 1–hop neighborhood of ρ by s. Similarly, we abbreviate $s_t = ((\mathcal{G}, \rho), \mathbf{x}_t)_1$. Define the (average) immediate reward as

$$\hat{r}(\boldsymbol{\mu}, h) := \int \boldsymbol{\mu}(\mathrm{d}s) \int h(\mathrm{d}u \mid s) r(s, u).$$

By definition we have with $\pi^{1:} := (\pi_t)_{1 \leq t \leq T}$ that

$$V_0^{\pi}(\boldsymbol{\mu}) = \hat{r}(\boldsymbol{\mu}, \pi_0) + \gamma V_1^{\pi^{1:}}(T_1 \boldsymbol{\mu})$$

holds, and similarly for $t > 0$. Recall that

$$V(\boldsymbol{\mu}) = \sup_{\pi \in \Pi_1} V_0^{\pi}(\boldsymbol{\mu})$$

is defined to be the value of state $\boldsymbol{\mu}$. We claim

$$V_t(\boldsymbol{\mu}) = \sup_{h \in \mathcal{H}} Q_t(\boldsymbol{\mu}, h) \tag{14}$$

for any state distribution $\boldsymbol{\mu}$. It suffices to show the claim for $t = 0$. To show "$\leq$", note that by definition

$$V_0^{\pi}(\boldsymbol{\mu}) = Q_0^{\pi^{1:}}(\boldsymbol{\mu}, \pi_0),$$

which immediately shows this inequality by taking the sup over all policies and next–step policies and then taking the sup over the policies on the left–hand side. For the other direction, fix any start distribution π_0 and choose π^{ϵ} such that

$$Q_0^{\pi^{\epsilon}}(\boldsymbol{\mu}, \pi_0) \geq Q_0(\boldsymbol{\mu}, \pi_0) - \epsilon.$$

The existence of π^{ϵ} is guaranteed by the definition of a supremum. Defining the "glued policy"

$$\tilde{\pi} := (\pi_0, \pi_1^{\epsilon}, ..., \pi_{T-1}^{\epsilon})$$

and inserting yields the inequality

$$V(\boldsymbol{\mu}) \geq V_0^{\tilde{\pi}}(\boldsymbol{\mu}) = Q_0^{\pi^{\epsilon}}(\boldsymbol{\mu}, \pi_0) \geq Q_0(\boldsymbol{\mu}, \pi_0) - \epsilon.$$

Since $\epsilon > 0$ was arbitrary, equation (14) is shown. The DPP follows now from the calculation

$$Q(\boldsymbol{\mu}, h) = \sup_{\pi \in \Pi_1} \hat{r}(\boldsymbol{\mu}, h) + \gamma V_1^{\pi}(T_1 \boldsymbol{\mu}) = \hat{r}(\boldsymbol{\mu}, h) + \gamma \sup_{h' \in \mathcal{H}} Q_1(T_1 \boldsymbol{\mu}, h')$$

and writing out definitions. □

C Proof of Theorem 4

We provide a detailed proof for the Policy Gradient Theorem applied to the Sparse Mean-Field Control system.

The problem is a Markov Decision Process (MDP) defined by the tuple $(\mathcal{S}, \Psi, \mathcal{T}, R)$, where:

- State space $\mathcal{S} = \mathcal{P}(\mathcal{G}_*^{\mathcal{X}})$ (measures over decorated neighborhoods).
- Action space Ψ (parameters for the local policy).
- Dynamics $\boldsymbol{\mu}_{t+1} = \mathcal{T}(\boldsymbol{\mu}_t, \psi_t)$. Note that given $\boldsymbol{\mu}$ and ψ, the next state is deterministic.
- Reward $R(\boldsymbol{\mu}, \psi)$.

The agent (meta-policy) $\hat{\pi}_\theta(\psi \mid \boldsymbol{\mu})$ is stochastic. Let $V_t(\boldsymbol{\mu})$ be the value function at time t under policy $\hat{\pi}_\theta$, as

$$V_t(\boldsymbol{\mu}) = \int_\Psi \hat{\pi}_\theta(\psi \mid \boldsymbol{\mu}) \left[R(\boldsymbol{\mu}, \psi) + V_{t+1}(\mathcal{T}(\boldsymbol{\mu}, \psi))\right] \mathrm{d}\psi.$$

We aim to find $\nabla_\theta V_0(\boldsymbol{\mu}_0)$. Differentiating V_t with respect to θ gives

$$\begin{aligned}\nabla_\theta V_t(\boldsymbol{\mu}) &= \nabla_\theta \int_\Psi \hat{\pi}_\theta(\psi \mid \boldsymbol{\mu}) Q_t(\boldsymbol{\mu}, \psi) \mathrm{d}\psi \\ &= \int_\Psi \Big[\nabla_\theta \hat{\pi}_\theta(\psi \mid \boldsymbol{\mu}) Q_t(\boldsymbol{\mu}, \psi) + \hat{\pi}_\theta(\psi \mid \boldsymbol{\mu}) \nabla_\theta Q_t(\boldsymbol{\mu}, \psi)\Big] \mathrm{d}\psi.\end{aligned}$$

Note that this differentiation step is standard in the Policy Gradient literature (e.g., Sutton et al., 1999). Using the log-derivative trick $\nabla \hat{\pi} = \hat{\pi} \nabla \log \hat{\pi}$,

$$\text{Term 1} = \mathbb{E}_{\psi \sim \hat{\pi}_\theta} \left[\nabla_\theta \log \hat{\pi}_\theta(\psi \mid \boldsymbol{\mu}) Q_t(\boldsymbol{\mu}, \psi)\right]. \tag{15}$$

Now consider **Term 2**, which accounts for the dependency of the future value on θ. Since R and $\mathcal{T}$ do not depend directly on θ (only via ψ), we have

$$\begin{aligned}\nabla_\theta Q_t(\boldsymbol{\mu}, \psi) &= \nabla_\theta \left[R(\boldsymbol{\mu}, \psi) + V_{t+1}(\mathcal{T}(\boldsymbol{\mu}, \psi))\right] \\ &= \nabla_\theta V_{t+1}(\boldsymbol{\mu}_{t+1}) \quad \text{where } \boldsymbol{\mu}_{t+1} = \mathcal{T}(\boldsymbol{\mu}, \psi).\end{aligned} \tag{16}$$

Substituting back into Eq. (15), we obtain

$$\nabla_\theta V_t(\boldsymbol{\mu}_t) = \mathbb{E}_{\psi_t} \left[\nabla_\theta \log \hat{\pi}_\theta(\psi_t \mid \boldsymbol{\mu}_t) Q_t(\boldsymbol{\mu}_t, \psi_t)\right] + \mathbb{E}_{\psi_t} \left[\nabla_\theta V_{t+1}(\boldsymbol{\mu}_{t+1})\right].$$

This is a recursive relationship. The differentiation of the dynamics is implicitly contained in the term $\nabla_\theta V_{t+1}$. By unrolling this recurrence from $t = 0$ to $T - 1$, we have

$$\begin{aligned}\nabla_\theta J(\theta) &= \nabla_\theta V_0(\boldsymbol{\mu}_0) \\ &= \mathbb{E}_{\psi_0}[\nabla \log \hat{\pi}_0 \cdot Q_0 + \nabla V_1] \\ &= \mathbb{E}_{\psi_0}[\nabla \log \hat{\pi}_0 \cdot Q_0 + \mathbb{E}_{\psi_1}[\nabla \log \hat{\pi}_1 \cdot Q_1 + \nabla V_2]] \\ &= \sum_{t=0}^{T-1} \mathbb{E}_\tau \left[\nabla_\theta \log \hat{\pi}_\theta(\psi_t \mid \boldsymbol{\mu}_t) Q_t(\boldsymbol{\mu}_t, \psi_t)\right].\end{aligned} \tag{17}$$

The term ∇V_T vanishes because $V_T(\cdot) = 0$.

C.1 Proof of Theorem 5

Proof. Let $\boldsymbol{\mu}_t^N$ be the empirical measure state on graph G_N, and $\boldsymbol{\mu}_t$ be the limiting measure. Let the gradient on the finite system be

$$g^N(\theta) = \sum_{t=0}^{T-1} \mathbb{E}\left[Q_t^N(\boldsymbol{\mu}_t^N, \psi_t) \nabla_\theta \log \hat{\pi}_\theta(\psi_t \mid \boldsymbol{\mu}_t^N)\right].$$

Let the gradient on the limit system be

$$g^\infty(\theta) = \sum_{t=0}^{T-1} \mathbb{E}\left[Q_t^\infty(\boldsymbol{\mu}_t, \psi_t) \nabla_\theta \log \hat{\pi}_\theta(\psi_t \mid \boldsymbol{\mu}_t)\right].$$

We assume $\hat{\pi}_\theta$ is Lipschitz continuous and bounded, and R is bounded. Our objective is now to show $\|g^N - g^\infty\| \to 0$. By Theorem 5(i)-(iii), forward induction gives $\boldsymbol{\mu}_t^N \Rightarrow \boldsymbol{\mu}_t$ in the marked local topology for every $t \leq T$: the base case is assumption (i), and the induction step uses local uniform convergence of the transition operators together with the Feller/Lipschitz property of the local kernel. Backward induction in the finite-horizon Bellman recursion, using local uniform convergence of the reward operators, gives locally uniform convergence $Q_t^N \to Q_t$ on compact sets of measures and policies. Assumption 1 makes the score function bounded and continuous, so bounded convergence yields the claim.

Consider now the difference for a fixed time step t and use the triangle inequality, which yields

$$\begin{aligned}\Delta_t &= \left\|\mathbb{E}[Q_t^N(\boldsymbol{\mu}_t^N)\nabla \log \pi(\boldsymbol{\mu}_t^N)] - \mathbb{E}[Q_t^\infty(\boldsymbol{\mu}_t)\nabla \log \pi(\boldsymbol{\mu}_t)]\right\| \\ &\leq \underbrace{\left\|\mathbb{E}[(Q_t^N(\boldsymbol{\mu}_t^N) - Q_t^\infty(\boldsymbol{\mu}_t^N))\nabla \log \pi(\boldsymbol{\mu}_t^N)]\right\|}_{\text{(I)}} \\ &\quad + \underbrace{\left\|\mathbb{E}[Q_t^\infty(\boldsymbol{\mu}_t^N)\nabla \log \pi(\boldsymbol{\mu}_t^N) - Q_t^\infty(\boldsymbol{\mu}_t)\nabla \log \pi(\boldsymbol{\mu}_t)]\right\|}_{\text{(II)}}.\end{aligned}$$

Term (I): Since $Q^N \to Q^\infty$ uniformly and $\nabla \log \pi$ is bounded, this term vanishes as $N \to \infty$.

Term (II): Let $f(\mu, \psi) = Q_t^\infty(\mu, \psi)\nabla \log \hat{\pi}(\psi \mid \mu)$. This function is bounded and continuous. Since $\boldsymbol{\mu}_t^N \to \boldsymbol{\mu}_t$ weakly, by the definition of weak convergence of measures (and the Portmanteau theorem), $\mathbb{E}[f(\boldsymbol{\mu}_t^N)] \to \mathbb{E}[f(\boldsymbol{\mu}_t)]$.

Thus, $\lim_{N\to\infty} \|g^N - g^\infty\| = 0$. Optimizing the finite graph objective asymptotically optimizes the mean-field objective. □

D Proof of Theorem 6

In the following we provide a proof for Theorem 6. In order to do this, some assumptions are necessary.

Assumption 2.

- *There exists* $\Delta \in \mathbb{N}$ *such that* $\deg(\rho) \leq \Delta$ *for all* $\rho \in \mathcal{G}$*;*
- *Fix any vertex* $\rho \in \mathcal{G}$*. Let* $\mathbf{x}$ *and* $\mathbf{x}'$ *be two graph decorations that differ at most by on one node in the* 1*-hop neighborhood around* ρ*. I.e.,*

  $$|\{v \in (\mathcal{G}, \rho)_1 : \mathbf{x}^v \neq (\mathbf{x}')^v\}| \leq 1;$$

  Then uniformly in $u \in \mathcal{U}$*,*

  $$\left\|P\big(\,\cdot\,|((\mathcal{G}, \rho), \mathbf{x})_1, u\big) - P\big(\,\cdot\,|((\mathcal{G}, \rho), \mathbf{x}')_1, u\big)\right\|_{\mathrm{TV}} \leq \alpha$$

  for $\alpha < 1$*;*
- *Assume* $\alpha\Delta^2 < 1$.

Intuitively α is the worst-case total-variation effect on one node's next-state distribution when one input decoration changes, while Δ controls the number of ways along which such disagreements can propagate. The condition $\alpha\Delta^2 < 1$ does not require either quantity to be small independently. In order to avoid confusion with the mean-field value function $V_t(\boldsymbol{\mu})$, we denote the value function for a specific local state S by (the slightly untypical) $W_t(S)$.

Definition 3 (Local Value Function)**.** *Let* $S = ((\mathcal{G}, \rho), \mathbf{x}) \in \mathcal{G}_*^{\mathcal{X}}$ *be a decorated rooted graph state at time* t*. For a policy* π*, the local value function* $W_t^\pi(S)$ *is the expected cumulative reward collected by the agent at the root* ρ*:*

$$W_t^\pi(S) := \mathbb{E}\left[\sum_{s=t}^{T-1} r\big(((\mathcal{G}, \rho), \mathbf{x}_s)_1, u_s^\rho\big) \;\middle|\; ((\mathcal{G}, \rho), \mathbf{x}_t) = S\right].$$

The expectation is taken over the action sequence $u_s^\rho \sim \pi_s$ *and the state dynamics* $\mathbf{x}_{s+1} \sim P$*, holding the graph topology* $(\mathcal{G}, \rho)$ *fixed. Let* $W_t^*(S) = \sup_\pi W_t^\pi(S)$ *be the optimal local value. We define similarly the local Q-function and the optimal (local) action value* Q^**.*

Note that in the previous display we set the discount factor $\gamma = 1$ for simplicity. Next, given a fixed rooted graph $(\mathcal{G}, \rho)$ and a spatial weight $\beta \in (0, 1)$, define the weighted distance by

$$d_\beta(S, S') := \sum_{v \in V(G): \mathbf{x}(v) \neq \mathbf{x}'(v)} \beta^{\|\rho - v\|}.$$

Proposition 5. *Let $(\mathcal{G}, \rho)$ be a fixed rooted graph. Let $S = ((\mathcal{G}, \rho), \mathbf{x})$ and $S' = ((\mathcal{G}, \rho), \mathbf{x}')$ be two states with possibly different node decorations $\mathbf{x}, \mathbf{x}'$. Fix any $\beta \in (\alpha\Delta, 1/\Delta)$, which exists by Assumption 2, and set $\kappa := \alpha\Delta/\beta < 1$. Then the optimal local value function $W_t^*(S)$ satisfies:*

$$|W_t^*(S) - W_t^*(S')| \leq L_t \, d_\beta(S, S'),$$

where $L_t \leq C_r/(\beta(1-\kappa))$ is uniform in t, T, and $(\mathcal{G}, \rho)$.

Proof. We proceed by backward induction on t. For $t = T$, the inequality holds trivially. Next, assume the bound holds for the function W_{t+1}^*. That is, we assume W_{t+1}^* is Lipschitz continuous with constant L_{t+1}. The Bellman equation for W_t^* reads

$$W_t^*(S) = \sup_{u \in \mathcal{U}} \left\{ r((S)_1, u) + \mathbb{E}[W_{t+1}^*(S_{\text{next}}) \mid S, u] \right\}.$$

Fix states S, S'. Let $D = \{v \in V(\mathcal{G}) : \mathbf{x}^v \neq (\mathbf{x}')^v\}$ be the (deterministic) set of disagreeing nodes. Using $|\sup f - \sup g| \leq \sup |f - g|$, we have the estimate

$$|W_t^*(S) - W_t^*(S')| \leq \sup_{u \in \mathcal{U}} \Bigg(\underbrace{|r((S)_1, u) - r((S')_1, u)|}_{\textbf{(I)}} + \underbrace{\left| \mathbb{E}[W_{t+1}^*(S_{\text{next}}) \mid S, u] - \mathbb{E}[W_{t+1}^*(S'_{\text{next}}) \mid S', u] \right|}_{\textbf{(II)}} \Bigg).$$

We now bound both terms separately, starting with $(\mathbf{I})$. Since r is local (radius 1) and bounded, $|r(S) - r(S')| \leq C_r \sum_{v \in D \cap B(\rho, 1)} 1$. In particular, $1 = \beta^{-1}\beta^{\|\rho - v\|}$ for ρ, v being neighbors. Thus

$$\text{(I)} \leq (C_r/\beta) d_\beta(S, S').$$

For $(\mathbf{II})$, we apply the induction hypothesis to every realization, which yields

$$\text{(II)} \leq \mathbb{E}\left[|W_{t+1}^*(S_{\text{next}}) - W_{t+1}^*(S'_{\text{next}})| \right] \leq L_{t+1} \mathbb{E}\left[d_\beta(S_{\text{next}}, S'_{\text{next}}) \right].$$

Now, expand the expectation of the distance. Let $\mathbf{x}_{\text{next}}$ and $\mathbf{x}'_{\text{next}}$ be the random decorations at time $t+1$. By Assumption 2, for any node w it follows from the triangle inequality that

$$\mathbb{P}\left(\mathbf{x}_{\text{next}}^w \neq (\mathbf{x}'_{\text{next}})^w \right) \leq \alpha \sum_{v \sim w} \mathbf{1}_{v \in D}.$$

By substituting this probability bound into the expectation one obtains

$$\mathbb{E}[d_\beta] = \sum_{w \in V(\mathcal{G})} \beta^{\|\rho - w\|} \mathbb{P}(\mathbf{x}_{\text{next}}^w \neq (\mathbf{x}'_{\text{next}})^w) \leq \sum_w \beta^{\|\rho - w\|} \left(\alpha \sum_{v \sim w} \mathbf{1}_{v \in D} \right).$$

Swapping sums thus gives the estimate

$$\mathbb{E}[d_\beta] \leq \alpha \sum_{v \in D} \sum_{w \sim v} \beta^{\|\rho - w\|}.$$

Using the triangle inequality $\|\rho - w\| \geq \|\rho - v\| - 1$, we have (as $\beta < 1$) $\beta^{\|\rho - w\|} \leq \beta^{-1}\beta^{\|\rho - v\|}$. With max degree Δ this means that

$$\mathbb{E}[d_\beta] \leq \alpha\Delta\beta^{-1} \sum_{v \in D} \beta^{\|\rho - v\|} = \kappa \, d_\beta(S, S').$$

Thus (II) $\leq \kappa L_{t+1} d_\beta(S, S')$. Combining terms gives the recursion $L_t = C_r/\beta + \kappa L_{t+1}$ with $L_T = 0$. Since $\kappa < 1$, unrolling yields

$$L_t \leq \frac{C_r}{\beta} \sum_{j=0}^{T-t-1} \kappa^j \leq \frac{C_r}{\beta(1-\kappa)},$$

which is uniform in the horizon T. □

Lemma 2. *For any two Markov kernels $\pi, \tilde{\pi}$ and any finite horizon $T \in \mathbb{N}$ we have*

$$J(\pi) - J(\tilde{\pi}) = \sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\tilde{\pi}} \left[\mathbb{E}_{u\sim\pi(\cdot|S_t)}[Q_t^\pi(S_t, u)] - \mathbb{E}_{\tilde{u}\sim\tilde{\pi}(\cdot|S_t)}[Q_t^\pi(S_t, \tilde{u})] \right].$$

Here $\tau = (S_0, \tilde{u}_0, S_1, \ldots, \tilde{u}_{T-1}, S_T)$ is the state–action trajectory generated by $\tilde{\pi}$, and $\mathbb{E}_{\tau\sim\tilde{\pi}}$ integrates over $S_t \sim d_t^{\tilde{\pi}}$, the state occupancy induced by $\tilde{\pi}$. The inner expectation over $u \sim \pi$ is conditional on the state visited under $\tilde{\pi}$; it is not part of the trajectory-generating policy.

Proof. Recall that $W_t^\pi(S) = \mathbb{E}_{u\sim\pi(\cdot|S)}[Q_t^\pi(S, u)]$ is the value function of state S with policy π. By definition, the Q-function satisfies the Bellman recursion

$$Q_t^\pi(S, \tilde{u}) = r(S, \tilde{u}) + \mathbb{E}_{S'\sim P(\cdot|S,\tilde{u})}[V_{t+1}^\pi(S')],$$

with $W_T^\pi(S) = 0$. Consider the term inside the summation on the RHS. Let $\tau \sim \tilde{\pi}$ denote a trajectory generated by $\tilde{\pi}$. At time t we define the advantage

$$\begin{aligned} A_t &:= \mathbb{E}_{u\sim\pi}[Q_t^\pi(S_t, u)] - \mathbb{E}_{\tilde{u}\sim\tilde{\pi}}[Q_t^\pi(S_t, \tilde{u})] \\ &= W_t^\pi(S_t) - \mathbb{E}_{\tilde{u}\sim\tilde{\pi}} \left[r(S_t, \tilde{u}) + \mathbb{E}_{S_{t+1}\sim P(\cdot|S_t,\tilde{u})}[W_{t+1}^\pi(S_{t+1})] \right]. \end{aligned}$$

A_t essentially measures the "loss" one gets when running $\tilde{\pi}$ instead of π for a single time step. Sample S_t according to the dynamics with respect to $\tilde{\pi}$, which yields

$$\mathbb{E}_{\tau\sim\tilde{\pi}}[A_t] = \mathbb{E}_{\tau\sim\tilde{\pi}}[W_t^\pi(S_t)] - \mathbb{E}_{\tau\sim\tilde{\pi}}[r(S_t, \tilde{u})] - \mathbb{E}_{\tau\sim\tilde{\pi}}[W_{t+1}^\pi(S_{t+1})].$$

Summing this over $t = 0$ to $T - 1$, one therefore obtains

$$\sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\tilde{\pi}}[A_t] = \sum_{t=0}^{T-1} \left(\mathbb{E}_{\tau\sim\tilde{\pi}}[W_t^\pi(S_t) - W_{t+1}^\pi(S_{t+1})] \right) - \underbrace{\sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\tilde{\pi}}[r(S_t, \tilde{u})]}_{J(\tilde{\pi})}.$$

Note that the first sum telescopes. Since $W_T^\pi \equiv 0$ and $\mathbb{E}[W_0^\pi(S_0)] = J(\pi)$, we obtain

$$\sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\tilde{\pi}}[A_t] = J(\pi) - J(\tilde{\pi}).$$ □

Proof of Theorem 6. Let π^* be the optimal policy. Fix $\beta \in (\alpha\Delta, 1/\Delta)$, set $\kappa = \alpha\Delta/\beta$, $q = \Delta\beta$ and let $L_Q := C_r/(\beta(1-\kappa))$ be the uniform Lipschitz constant from Proposition 5. Let $\mathcal{T}_k(S)$ be the state that has every node decoration at distance $\|\rho - v\| > k$ set to some fixed state, say 0, and that agrees with S on the k-ball around ρ. Next, define the truncated policy $\tilde{\pi}$ by $\tilde{\pi}(\cdot|S) := \pi^*(\cdot|\mathcal{T}_k(S))$. Since $\pi^{(k)}$ denotes the optimal k-local policy, we have $J(\pi^*) - J(\pi^{(k)}) \leq J(\pi^*) - J(\tilde{\pi})$.

Our plan is to apply Lemma 2. Let S_t be the state at time t and $S'_t = \mathcal{T}_k(S_t)$ be its truncation. The loss term at time t is

$$\Delta_t = \mathbb{E}_{u\sim\pi^*(\cdot|S_t)}[Q_t^*(S_t, u)] - \mathbb{E}_{\tilde{u}\sim\pi^*(\cdot|S'_t)}[Q_t^*(S_t, \tilde{u})],$$

because $\pi^*(\cdot \mid S'_t) = \tilde{\pi}(\cdot \mid S_t)$. Note that the first term is simply the optimal value function $V_t^*(S_t)$. The second term involves actions $\tilde{u}$ chosen according to the truncated state S'_t, but evaluated on the real state S_t.

We add and subtract the optimal value of the truncated state, $W_t^*(S'_t) = \mathbb{E}_{\tilde{u}\sim\pi^*(\cdot|S'_t)}[Q_t^*(S'_t, \tilde{u})]$ so that we obtain

$$\begin{aligned} \Delta_t &= W_t^*(S_t) - W_t^*(S'_t) + W_t^*(S'_t) - \mathbb{E}_{\tilde{u}\sim\pi^*(\cdot|S'_t)}[Q_t^*(S_t, \tilde{u})] \\ &= \underbrace{(W_t^*(S_t) - W_t^*(S'_t))}_{\text{(I)}} + \underbrace{\mathbb{E}_{\tilde{u}\sim\pi^*(\cdot|S'_t)}[Q_t^*(S'_t, \tilde{u}) - Q_t^*(S_t, \tilde{u})]}_{\text{(II)}}. \end{aligned}$$

We bound both terms separately. For the first term (I), observe that by the Lipschitz continuity of the optimal value function, we have

$$|W_t^*(S_t) - W_t^*(S'_t)| \leq L_Q d_\beta(S_t, S'_t).$$

For the second term (II), note that this is the expectation of the difference in Q-values for the *same* action $\tilde{u}$ but different states. By the Lipschitz continuity of Q^* (which follows from the Lipschitz estimate on W_t^*)

$$|Q_t^*(S_t', \tilde{u}) - Q_t^*(S_t, \tilde{u})| \leq L_Q d_\beta(S_t, S_t').$$

Taking the expectation over $\tilde{u}$ does not change this bound. Combining these estimates, we get $\Delta_t \leq 2L_Q d_\beta(S_t, \mathcal{T}_k(S_t))$. Finally, we bound the distance. Since S_t and $\mathcal{T}_k(S_t)$ agree on the k-hop neighborhood, disagreements occur only at distance $j > k$, and there are at most Δ^j nodes at distance j. Hence

$$d_\beta(S_t, \mathcal{T}_k(S_t)) \leq \sum_{j=k+1}^{\infty} \Delta^j \beta^j = \frac{q^{k+1}}{1-q}.$$

Summing over $t = 0 \dots T-1$ yields

$$J(\pi^*) - J(\pi^{(k)}) \leq \sum_{t=0}^{T-1} 2L_Q \frac{q^{k+1}}{1-q} \leq \frac{2C_r}{\beta(1-\kappa)(1-q)}\, T\, q^{k+1}. \qquad \square$$

D.1 Weighted graphs

We briefly discuss the extension of our results to weighted graphs. Indeed, the maximum-degree estimate can be sharpened for edge-weighted graphs. Let $P_{t,w}(\cdot \mid z)$ be the next-state kernel at vertex w and define the edgewise influence

$$c_{wv}^{(t)} := \sup_{z, z'} \big\| P_{t,w}(\cdot \mid z) - P_{t,w}(\cdot \mid z') \big\|_{\mathrm{TV}},$$

where the supremum ranges over decorations z, z' that agree everywhere except possibly at v. For the weighted disagreement distance, the relevant condition becomes

$$\kappa_\beta := \sup_{t,o,v} \sum_w \beta^{\mathrm{dist}(o,w) - \mathrm{dist}(o,v)}\, c_{wv}^{(t)} < 1.$$

If $c_{wv}^{(t)} \leq L a_{wv}$ for edge weights a_{wv} and $A = \sup_v \sum_w a_{wv}$, then $\kappa_\beta \leq LA/\beta$; together with the summability condition $\Delta\beta < 1$, it suffices that $LA < \beta < 1/\Delta$, which is possible whenever $LA\Delta < 1$. In the unweighted case this recovers $\alpha\Delta^2 < 1$. This formulation permits high degree provided the total incoming influence remains small.

E Theorem-Level Comparison with Fabian et al. (2025)

Table 1 makes precise where the present framework goes beyond the rigorous parts of Fabian et al. (2025) without understating their contribution in detail. Fabian et al. provide a rigorous sparse-graph mean-field model based on degree-indexed state distributions and prove convergence for fixed policy ensembles. Their optimal-policy transfer result is formulated for a finite candidate set of policies with a strict value gap, and their scalability approximation aggregates high-degree agents while invoking a heuristic approximation of neighbor-degree distributions.

F Experiments

Our algorithm (meaning Algorithm 1) can be seen as an extension of the strong but heuristic algorithm LWMFMARL (Local Weak Mean Field MARL) as introduced in Fabian et al. (2025), which conditions actions on global population statistics and node degrees. While already very effective in a variety of settings with millions of nodes, this algorithm is heuristic and relies on aggregated representations, discarding fine-grained, structural information. Because there are already plenty of experiments comparing LWMFMARL with more classical baseline algorithms (see in particular Section 7 in Fabian et al. (2025)), we refrain from running experiments comparing our algorithm with said baselines. Instead, the experiments presented in this section isolate two complementary failure modes of mean-field control on sparse graphs and demonstrate how local, graph-aware policies resolve them. Details of our setup can be found in Section F.4.

Aspect	Fabian et al. (2025)	This work
Mean-field state	degree-indexed state marginals and neighborhood-state counts	full law of decorated rooted neighborhoods
Value convergence	fixed policies; optimal-policy transfer for a finite candidate set with a strict value gap	optimized value over a compact policy class (Theorem 1)
DPP	not established	DPP on lifted neighborhood spaces (Theorem 3)
Policy gradient	not established	PG identity and finite-graph approximation (Theorems 4, 5)
GNN justification	heuristic	GNN/mean-field integration correspondence (Proposition 1)
Locality	not established	horizon-dependent radius and quantitative truncation (Theorems 2, 6)
Computational approximation	aggregates high-degree agents; heuristic neighbor-degree approximation	subsampling justified by local weak convergence; depth–accuracy tradeoff

Table 1: Theorem-level comparison with the sparse-graph MFC approach of Fabian et al. (2025).

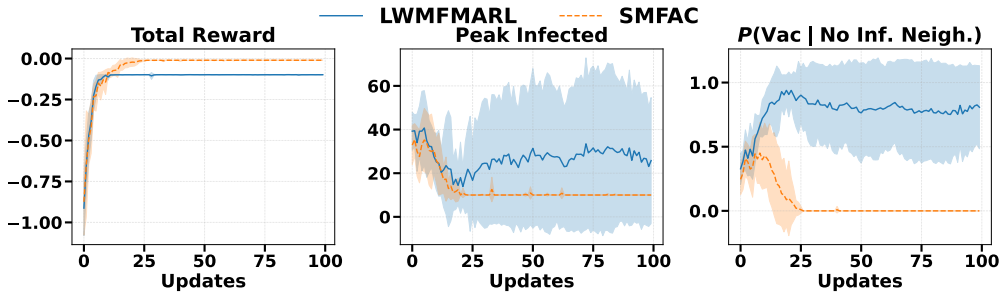

Figure 3: Additional information at the local node makes it easier for the agents to decide when to vaccinate, yielding a better policy compared to only local information.

F.1 Resolving Local Heterogeneity

In supercritical infection regimes where infection costs dominate, purely mean-field/LWMFMARL controllers converge to a degenerate *vaccinate-all* policy. More generally, because they only observe aggregated statistics, they cannot distinguish locally exposed nodes from spatially isolated ones. On sparse graphs, this behavior is highly suboptimal, as vaccinating a small number of strategically placed agents is often sufficient to block transmission.

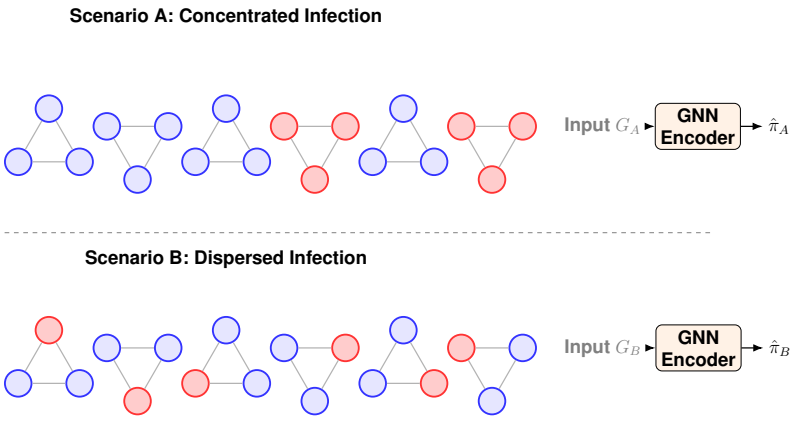

Figure 4: Two infection scenarios with identical global counts but different local structure.

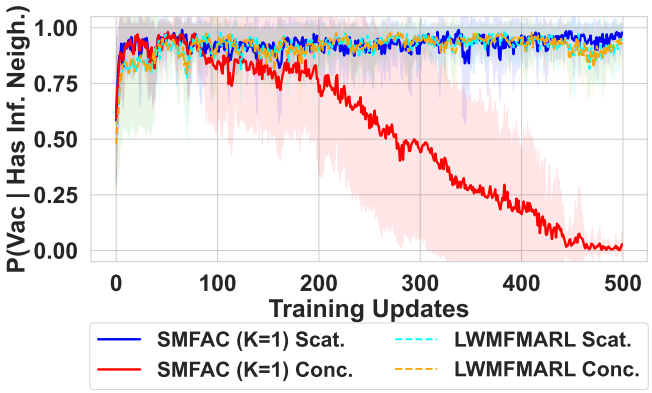

Figure 5: Applying a GNN encoder on the underlying graph can be essential in finding the optimal policy. Observe that our algorithm is able to find the optimal policy (not vaccinating nodes with infected neighbor if disease is already contained), where as the LWMFMARL is not able to distinguish between the concentrated and scattered initial state.

We illustrate this effect on a two-dimensional grid. Figure 3 shows that incorporating local neighborhood information via a lifted state space yields substantially lower loss and more selective interventions. Agents learn to condition actions on nearby infection states rather than reacting uniformly to global infection levels.

F.2 Disambiguating Global Structure

We next expose a more fundamental limitation of MFC and LWMFMARL: distinct local configurations can induce identical global statistics. Any controller that conditions solely on aggregated information is thus unable to distinguish such states, leading to suboptimal actions.

For simplicity, consider $N = 20$ disjoint triangles with 5 initially infected nodes. With equal probability, infections are either concentrated within a few triangles (Scenario A in Figure 4) or dispersed across many (Scenario B), while global infection counts remain identical. A global penalty is incurred if more than a fixed number of triangles contain multiple infected nodes. When infections are dispersed, local containment is required; when they are concentrated, non-intervention is optimal. Since both cases induce the same MF statistics, the MF baseline must select the same action and is therefore suboptimal in at least one regime.

Algorithm 1 Sparse Mean-Field Actor–Critic (SMFAC)

Initialize actor GNN π_θ and critic V_ϕ.
for each training episode **do**
  Initialize N-agent system $(G_N, \mathbf{x}_0)$.
  for time $t = 0, \dots, T-1$ **do**
    Apply GNN to G_N with $k = T - t$ message passes.
    Sample $\pi_t \sim \text{MLP}\left(\frac{1}{N}\sum_{i=1}^{N} \text{GNN}_k^i(G_N)\right)$.
    for agent $i \in G_N$ **do**
      Sample $u_t^i \sim \pi_t(\cdot \mid \left((G_N, i), \mathbf{x}_t\right)_1)$.
    end for
    Execute all actions, observe $\mathbf{x}_{t+1}$ and rewards $\mathbf{r}_t$.
    Store transitions in a replay buffer.
  end for
  Compute advantages; update π_θ and V_ϕ using PPO.
end for

Figure 5 shows that the GNN-based policy successfully conditions its behavior on the underlying graph structure and separates the two regimes, whereas the baseline (equivalently, a GNN with zero message-passing steps) collapses to a single compromised policy.

F.3 Experimental Takeaways

Across both experiments, MF aggregation fails either by obscuring localized infection pressure or by collapsing distinct configurations into identical global statistics. Neighborhood-aware policies resolve both issues by conditioning actions on local graph structure, improving MFC for sparse graphs.

F.4 Experimental Setup

F.4.1 SIR Model Dynamics

The experiments are conducted on a graph $\mathcal{G} = (V, E)$ where each node represents an agent. The system follows a discrete-time **SIR (Susceptible-Infected-Recovered)** process with interventions.

- **States:** each node i is in one of three states at time t: $S_i^t \in \{S, I, R\}$.
  - S (Susceptible): can be infected.
  - I (Infected): can infect neighbors.
  - R (Recovered/Removed): immune, cannot infect or be infected (includes vaccinated nodes).
- **Infection Dynamics:** a susceptible node i becomes infected at time $t+1$ with probability P_{inf}. If node i decides to *isolate*, the susceptibility is reduced by a factor $f_{\text{iso}} \in [0, 1]$,

$$P(S_i^{t+1} = I | S_i^t = S) = 1 - \prod_{j \in \mathcal{N}(i): S_j^t = I} (1 - \beta \cdot (1 - 1_{a_i=2} \cdot f_{\text{iso}}))$$

  where β is the base transmission probability per edge.
- **Recovery Dynamics:** an infected node recovers naturally with probability γ at each step, meaning

$$P(S_i^{t+1} = R | S_i^t = I) = \gamma$$

- **Vaccination:** an explicit action ($a_i = 1$) transitions a susceptible node directly to the recovered state ($S \to R$), granting immediate immunity.

The reward for every agent i at time t reads

$$r_i^t = -\left[c_I \cdot 1(S_i^t = I) + c_V \cdot 1(S_i^t \in \mathcal{V}) + c_Q \cdot 1(a_i^t = \text{Isolate}) + C_{\text{global}} \cdot 1\left(\frac{1}{N} \sum_{j=1}^{N} 1(S_j^t = I) > \tau \right) \right], \tag{18}$$

where

- c_I, c_V, c_Q are the costs for infection, vaccination, and isolation respectively;
- $1(\cdot)$ is the indicator function;
- $\mathcal{V}$ denotes the set of vaccinated states (distinct from naturally recovered);
- C_{global} is the penalty applied when the global infection rate exceeds the threshold τ.

F.4.2 Setup

Both our experiments utilize the architecture that is described in Figure 2. A GNN (parameterized by ϕ) predicts the means μ of a policy parameter distribution based on the global state S_{global}. The actual policy parameters θ for an episode are sampled from a Gaussian distribution centered at $\mu = \text{GNN}(S_{\text{global}})$ with a fixed standard deviation ($\sigma = 3.0$), meaning

$$\theta \sim \mathcal{N}(\text{GNN}(S_{\text{global}}); \phi), \sigma^2 I).$$

The local agents operate using a decentralized policy parameterized by θ. The policy then maps a discrete local observation state s_{local} to action probabilities. The state space is characterized as follows. The local state space is defined by the agent's own state $s_i \in \{S, I, R\}$ and binary flags indicating the presence of neighbors in each state. To decrease the dimensions of the learning problem, the policy is restricted to detect only the presence of *infected* neighbors, having no information about S and R neighbors. This effectively reduces the decision space to $3 \times 2 = 6$ functional states: $\{S, I, R\} \times \{0, 1\}$. All experiments were run on a Macbook Pro (2024) and across 10 random seeds. We use NetworkX (Hagberg et al., 2008) to generate our graphs and the Adam optimizer (Kingma & Ba, 2014) to optimize the learnable parameters. Everything is implemented in PyTorch (Paszke et al., 2019).

Experiment 1: Graph-Lifted State Space

This setup focuses on controlling an outbreak on a lattice using isolation/vaccination. As graph we chose a 20×20 grid, meaning $N = 400$ nodes. The environment parameters are chosen so that the infection would spread supercritically in case no controller is present. In particular, we choose the (average) cost of being infected greater than the cost of vaccination ($c_v = 1, c_I = 0.5, c_Q = 0.2$). By choosing the recovery rate to be suitably low ($\gamma = 0.2$) and the infection probability suitably high ($\beta = 0.25$), the optimal policy is to immediately stop the infection from the get go and not hope for it to die out on its own. We use a time horizon of $T = 100$; the infection reaches the grid boundary much earlier, and increasing T did not change the results qualitatively.

We found that using the aggregated mean-field (i.e. $k > 0$ message passes for the GNN encoder) does not improve performance for the model. This is to be expected; the optimal policy can be inferred from local knowledge alone.

Experiment 2: Sensitivity to Initial Conditions

This setup tests the ability to manage infection in a graph of disjoint cliques under a hard vaccination budget constraint. The graph topology is now $N = 20$ disjoint triangles. For simplicity we set $\beta = 1.0, \gamma = 0.0$, meaning no natural recovery and guaranteed infection spread. We now start in one of the following initial scenarios.

- *Scattered:* 5 infected nodes, each in a distinct triangle.
- *Concentrated:* 5 infected nodes concentrated in 4 triangles.

Vaccination is now very expensive ($c_v = 50$) and being in state I is cheap ($c_I = 0$, again choosing $c_Q = 0.2$). A global penalty is added to the reward if more than 25 percent of all triangles have at least two infected. The meta policy has thus to distinguish the scattered state and the concentrated state: in the scattered state, vaccination must happen immediately for all nodes with infected neighbors; otherwise, the reward becomes very small/negative. In the concentrated state, however, the threshold for the global punishment term of the reward can never be achieved, meaning that it is cheaper to let agents become infected in the local triangles rather than to vaccinate. Here, we find that the GNN can distinguish both initial conditions, where as the mean field can not (note that the mean field is constant across both different initial scenarios). This validates that local neighborhood information on the mean-field level can be important to find optimal strategies.

G Limitations and Broader Impact

Our guarantees assume a fixed, trustworthy graph with bounded local growth and, for Theorem 6, bounded degree. We do not cover adversarial edge manipulation, corrupted messages, or topology attacks. Those mechanisms could invalidate both the neighborhood representation and the learned policy. Genuinely unbounded-degree power-law networks are also excluded from the quantitative truncation guarantee. In such highly heterogeneous graphs, high-degree nodes make local policies difficult to parameterize, while low-degree nodes limit graphon-based approximations, which is why we treat this regime as out of scope for the quantitative bound. Moreover, uniform equicontinuity may fail in certain cases, for example when degrees or local influence grow with N without a compensating moment bound or when perturbations are amplified near a critical regime. The finite-horizon theory does not directly cover undiscounted average-reward control, which requires recurrence or ergodicity

of the lifted process. The hierarchical policy structure is also necessary: without a meta-policy that senses the global mean-field state, the lifted MDP reduces to optimizing an action sequence, for which classical methods such as relative-entropy policy search exist. Their extension to this infinite-dimensional state is non-trivial as far as we can tell and therefore out of scope.

We also need to mention that the truncation bound in Theorem 6 controls only the bias from restricting the information radius to k hops. Practical performance additionally depends on the approximation capacity of the GNN, optimization error, and finite-sample or statistical error. Standard message-passing networks inherit the expressivity limitations of the 1-WL test and may suffer from oversquashing or oversmoothing at large depth; the developed theory in this paper does not remove these architectural limitations, and they could in principle become more pronounced as T grows. Finally, experiments we provide use only synthetic graphs and simulated SIR states, and the paper makes no deployment or policy recommendation. Applications involving epidemics or social networks may use sensitive relational data and may allocate intervention costs unequally. Local message passing need not centralize raw neighborhoods, but learned messages may still leak information. Moreover, optimizing a global average reward can disproportionately burden particular communities or peripheral nodes. Any human-facing application would require privacy-preserving communication, access controls, fairness or group-specific constraints as well as domain-specific oversight.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We included all key results of the paper in the abstract, and not more, either.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We clearly state that our experimental section only discusses scenarios in which a published algorithm can be outperformed. Moreover, we also argue that finite time horizon, as well as discrete time, are key ingredients for the approach as presented in this work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: To the best of our knowledge, all proofs presented in this work are correct, and assumptions clearly stated.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided the algorithmic blueprint that was used to implement our approach. Moreover, we also disclosed the libraries used. As far as we are aware, the experiments in this paper can be reproduced with the information provided.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may

be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: At this point, we did not make our code available to the public.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provided all hyperparameters (and experiment setup) that we found necessary to reproduce our results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: While we did run our training over multiple random seeds, we did not carefully track quantities such as standard deviations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Because the experiments are not exhaustive and on small graph topologies, there is no requirement for any special computing hardware.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Yes, it does.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [N/A]

Justification: Our contribution is of theoretical nature, and therefore as potentially beneficial/harmful as any other theoretical machine learning paper. Hence, we did not feel the need for further discussion.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: By virtue of subject studied.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: We are the original owners of all the material used in this work.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: There are no new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We did not make use of crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core of our methods did not involve LLMs in a key component.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.