# Networked Communication for Decentralised Cooperative Agents in Mean-Field Control

**Anonymous authors**
**Paper under double-blind review**

## Abstract

The mean-field framework has been used to find approximate solutions to problems involving very large populations of symmetric, anonymous agents, which may be intractable by other methods. The cooperative mean-field control (MFC) problem has received less attention than the non-cooperative mean-field game (MFG), despite the former potentially being more useful as a tool for engineering large-scale collective behaviours. Decentralised communication algorithms have recently been introduced to MFGs, giving benefits to learning speed and robustness. Inspired by this, we introduce networked communication to MFC - where populations arguably have broader incentive to communicate - and in particular to the setting where decentralised agents learn online from a single, non-episodic run of the empirical system. We adapt recent MFG algorithms to this new setting, as well as contributing a novel sub-routine allowing networked agents to estimate the global average reward from their local neighbourhood. Previous theoretical analysis of decentralised communication in MFGs does not extend trivially to MFC. We therefore contribute new theory proving that in MFC the networked communication scheme allows agents to increase social welfare faster than under *both* of the two typical alternative architectures, namely independent and centralised learning. We provide experiments that support this new result across different classes of cooperative game, and also give numerous ablation studies and additional experiments concerning numbers of communication round and robustness to communication failures.

## 1 Introduction

The mean-field framework (Lasry & Lions, 2007; Huang et al., 2006) models a representative agent as interacting not with the rest of the population on a per-agent basis, but instead with a distribution over the other agents, known as the *mean field*. The framework analyses the limiting case when the population consists of an infinite number of symmetric and anonymous agents, that is, they have identical reward and transition functions which depend on the mean-field distribution rather than on the actions of specific other players. The mean-field *control* (MFC) problem is a cooperative scenario where the population seeks to maximise a social welfare criterion such as the average return received by the agents. Alternatively we can consider a non-cooperative scenario called a mean-field *game* (MFG), where each agent seeks to maximise its individual return, to which the solution is a MFG-Nash equilibrium (MFG-NE).

The MFC social optimum and the MFG-NE can respectively be used as approximate solutions to the associated finite-agent problem/game, with the error in the solution reducing as the number of agents $N$ tends to infinity (Saldi et al., 2018; Gu et al., 2021; Mondal et al., 2022; Cui et al., 2023b;c; Anahtarci et al., 2023; Yardim et al., 2024; Toumi et al., 2024; Hu & Zhang, 2024; Chen et al., 2024; Bayraktar & Kara, 2024). MFC and MFGs have therefore been used to address the difficulty faced by multi-agent reinforcement learning (MARL), which can struggle to scale computationally as $N$ increases (Yardim & He, 2024; Zeng et al., 2024). While MFGs have been well-studied and applied to a wide variety of real-world problems (Laurière et al., 2022a), MFC has received less attention, despite possibly being more useful for engineering collective behaviours to achieve global objectives, such as in consensus, synchronisation,

rendezvous, exploration, coverage or task allocation problems (Cui et al., 2023c). This paper seek to redress some of this imbalance by adapting recent developments in MFGs for MFC.

Since MFC problems can be interpreted as optimisation problems from the perspective of a social planner, classical approaches involve centralised methods (they also do so for reasons of simplicity, as in MFGs). In this context 'centralised' does not necessarily imply global observability of the whole population's actions - which could make computation infeasible given the complexity of the problem - but rather that learning is conducted from the samples of a single representative agent, whose policy updates are assumed to be automatically pushed to the rest of the population by the central node (Fornasier & Solombrino, 2014; Carmona et al., 2019; Ruthotto et al., 2020; Laurière et al., 2022a; Angiuli et al., 2022; 2023; Cui et al., 2023a; Lee et al., 2024; Denkert et al., 2024). For this reason, whilst 'centralised learning' is the term used in prior works, we generally refer to 'central-agent learning' to reduce confusion. Often the empirical mean field of the actual population is not even used to compute rewards or transitions, with the central learner instead updating an estimate of the mean field based only on its own policy, which is in turn used as input to its reward and transition functions (Carmona et al., 2019; Angiuli et al., 2022; 2023).

Recent works on MFGs, as in other areas of multi-agent research, have recognised that the existence of a central learner is a strong assumption in complex, real-world settings, as well as representing a bottleneck for computation and communication, and a vulnerable single point of failure of the system (Zhang et al., 2018; 2021a;b; Chen et al., 2021; Yardim et al., 2023; Benjamin & Abate, 2023; 2024; Jiang et al., 2024; Xu et al., 2025; Agyeman et al., 2025; Horyna et al., 2025). They advocate instead for the individual agents in the empirical population to learn policies for themselves without relying on a central node. Such works also argue that other strong classical assumptions should similarly be loosened in order to make MFGs applicable to real-world, embodied problems such as swarm robotics. They therefore contend that, aside from decentralised learning, desirable qualities for mean-field algorithms include: learning from the population's empirical mean field (i.e. the mean-field distribution is generated only by the agents' policies, rather than being manipulated by the algorithm itself or by an external oracle/simulator); learning online from a single, non-episodic system run (i.e. similar to above, the population is not arbitrarily reset by an external controller); learning without reliance on a model of the system; and using function approximation to allow scalability to high-dimensional observations (including the option to include the mean field in the input to policies).

Until now, no work on MFC has met all these criteria. Some recent works have considered decentralisation in MFC, but Bayraktar & Kara (2024) requires that decentralised agents optimise for learnt models of the system dynamics (and learning is only fully independent when the population is large but finite rather than infinite), while Cui et al. (2023c) presents a model-free deep learning algorithm that gives decentralised execution but requires centralised, episodic training. This latter work stipulates that decentralised training can be achieved if all agents can directly observe the mean-field distribution and use the same seed to correlate their actions, though they only provide empirical results for the centralised scenario, while Bayraktar & Kara (2024) provides no empirical results at all. However, assuming decentralised agents have access to this global information is unrealistic, and in the non-cooperative MFG setting Benjamin & Abate (2024) have shown that networked communication between decentralised agents allows agents to estimate the global mean field from a local neighbourhood. They also show that proliferating high-performing policies through the population via decentralised communication (in a manner reminiscent of distributed embodied evolutionary algorithms (Hart et al., 2015; Fernández Pérez et al., 2018; Fernández Pérez & Sanchez, 2019; Cazenille et al., 2025; Sissodia et al., 2025)) improves training time and avoidance of local optima, particularly over the case of agents learning entirely independently, but often also over populations with a single central learner.

Inspired by this non-cooperative MFG work, we introduce networked communication to MFC for the first time, where populations arguably have even more incentive to communicate. This allows us to present a model-free deep learning algorithm that fulfils all of the proposed desiderata, including learning online from a single non-episodic run of the empirical system, and decentralised training without needing to observe global information: we contribute a novel sub-routine for estimating the global average reward from local communication, in addition to the existing sub-routine for estimating the global mean field from Benjamin & Abate (2024). Previous theoretical analysis of networked communication in the non-cooperative MFG setting does not extend trivially to MFC, so we contribute new theoretical proofs showing that decentralised policy exchange allows networked populations to learn faster than both the independent *and* the central-

agent alternatives in the MFC setting, across different classes of cooperative game (coordination and anti-coordination). We also demonstrate this finding empirically in numerous games, as well as contributing an empirical study of the algorithms' robustness to communication failures, along with several ablation studies. In summary, our contributions include:

- We provide the first algorithms in MFC for model-free training without any central provision of information or coordination, as well as the first MFC algorithms for online learning from a single, non-episodic run of the empirical system.
  - We contribute a novel sub-routine allowing decentralised agents to estimate the global average reward via networked communication, and incorporate an existing sub-routine used in MFGs for estimating the global mean field via local communication.
- We prove theoretically that in this context, decentralised networked communication can improve learning speed over the independent *and* central-agent architectures.
- We provide extensive experiments supporting our theoretical results in numerous games, and give ablation studies of various parts of our algorithms, as well as a study of robustness to communication failures.

We provide further comparison with related work in Sec. 2, give preliminaries in Sec. 3, and our algorithms in Sec. 4. We present theoretical results in Sec. 5 and experiments in Sec. 6, before suggesting future work in Sec. 7.

## 2 Related work

We discuss here the research most closely related to our present work, focusing on decentralisation and networked communication, and clarifying the differences with prior methods and settings. We refer the reader to Laurière et al. (2022a) for a broader survey of MFC.

Numerous works claiming to study decentralisation in MFC take this to mean only that agents do not have access to the specific states of all other agents, and have policies depending on their local state and possibly the mean field, all of which we take as a given in our work. They nevertheless rely on a central learner or coordinator that provides global information to all agents, a dependence that we remove in our work. This applies, for example, to Grammatico et al. (2016), where a 'central population coordinator' broadcasts a common signal to all agents, and to Tajeddini et al. (2017), which presents a leader-follower setting where a 'central population coordinator' estimates the mean-field trajectory. Farzaneh et al. (2020) similarly requires a central coordinator, and also presents a non-cooperative scenario so does not actually fall under MFC despite being referred to as such.

Bayraktar & Kara (2024) considers independent, 'online' learning for MFC in a setting that is different from ours. Crucially, their method involves agents first estimating a model (reward and transition functions) of the system by conducting 'online' updates using samples collected while following exploration policies. Only once having done so do they compute execution policies that are optimal with respect to the estimated model. We argue that having a dedicated exploration phase is infeasible for many real-world applications, and instead present a fully model-free online learning algorithm. Moreover, their setting only permits independent learning if $N$ is large but finite. For infinite populations, a central coordinator is required to supply common noise to aid exploration during the initial phase, and if the optimal policy for the estimated model is not unique, centralised coordination is required to allow the agents to agree on which policy to execute. Our algorithm requires no such special considerations. Finally, their work is purely theoretical, whereas we provide extensive empirical results.

In Cui et al. (2023c), decentralisation applies only during execution, and they offer a centralised-training decentralised-execution method (as also in Cui et al. (2023a)). They say that decentralised training could be achieved if the global mean field is observable and all agents use the same seed to correlate their actions, whilst we do not require either assumption for our decentralised training algorithm. They also train episodically whereas we learn online from a single run of the system. Finally, their experiments focus only on

coordination games, whereas we additionally explore empirical effects resulting from decentralised training in anti-coordination games, where agents can gain higher rewards by diversifying their behaviour.

Angiuli et al. (2022) and Angiuli et al. (2023) provide algorithms for MFC learning from a single run, but there it is a single run only of a 'representative' player that is used to simulate the mean field, rather than a single run of the empirical population as in our work. Their algorithms are thus inherently centralised, as well as involving two timescales for updating the mean-field approximation, which we argue is unlikely to be a practical paradigm for training in complex real-world systems such as robotic swarms.

Our work is also closely related to Benjamin & Abate (2023) and Benjamin & Abate (2024), which introduce networked communication to the non-cooperative MFG setting. By adapting their communication scheme and learning algorithm, we introduce networked communication to the cooperative MFC setting, where it is arguably more applicable due to broader incentives for agents to communicate policies. Their works focus on coordination games to justify the sharing of policies (though Benjamin & Abate (2024) does demonstrate empirically that networked agents outperform independent agents in a non-cooperative anti-coordination game, indicating that self-interested agents do nevertheless have incentive to communicate), whilst we provide extensive theoretical and empirical results on the benefits of policy sharing in MFC for both coordination and anti-coordination games. We leverage Alg. 4 from Benjamin & Abate (2024) for estimating the global mean field from a local neighbourhood, but additionally contribute the novel Alg. 1 for estimating the global average reward from a local neighbourhood for the MFC setting.

## 3 Preliminaries

### 3.1 Mean-field control

We use the following notation. $N$ is the number of agents in a population, with $\mathcal{S}$ and $\mathcal{A}$ representing the finite state and common action spaces. The set of probability measures on a finite set $\mathcal{X}$ is denoted $\Delta_{\mathcal{X}}$, and $\mathbf{e}_x \in \Delta_{\mathcal{X}}$ for $x \in \mathcal{X}$ is a one-hot vector with only the entry corresponding to $x$ set to 1, and all others set to 0. For time $t \geq 0$, $\hat{\mu}_t = \frac{1}{N}\sum_{i=1}^{N}\sum_{s\in\mathcal{S}} \mathbb{1}_{s_t^i=s}\mathbf{e}_s \in \Delta_{\mathcal{S}}$ is a vector of length $|\mathcal{S}|$ denoting the empirical categorical state distribution of the $N$ agents at time $t$. For agent $i \in \{1 \dots N\}$, $i$'s policy $\pi^i \in \Pi$ depends on its observation $o_t^i$. We give different forms that this observation can take, and relatedly a more formal definition of the policy, after the following.

**Definition 3.1** ($N$-player stochastic cooperative control problem with symmetric, anonymous agents)**.** This is given by the tuple $\langle N, \mathcal{S}, \mathcal{A}, P, R, \gamma\rangle$, where $\mathcal{A}$ is the action space, identical for each agent, $\mathcal{S}$ is the identical state space of each agent, such that their initial states are $\{s_0^i\}_{i=1}^N \in \mathcal{S}^N$ sampled from some initial distribution $\mu_0 \in \Delta_{\mathcal{S}}$, and their policies are $\{\pi^i\}_{i=1}^N \in \Pi^N$. $P : \mathcal{S}\times\mathcal{A}\times\Delta_{\mathcal{S}} \to \Delta_{\mathcal{S}}$ is the transition function and $R : \mathcal{S}\times\mathcal{A}\times\Delta_{\mathcal{S}} \to [0,1]$ is the reward function, both identical to all agents, and which map each agent's local state and action and the population's empirical distribution to transition probabilities and bounded rewards, respectively, i.e. $\forall i \in \{1,\dots,N\}$: $s_{t+1}^i \sim P(\cdot|s_t^i, a_t^i, \hat{\mu}_t)$ and $r_t^i = R(s_t^i, a_t^i, \hat{\mu}_t)$.

For the joint policy $\boldsymbol{\pi} := (\pi^1, \dots, \pi^N) \in \Pi^N$, an individual agent's discounted return is given by:

**Definition 3.2** (Individual expected discounted return)**.** For all $i, j \in \{1,\dots,N\}$, $i$'s return is

$$V^i(\boldsymbol{\pi}, \mu_{\bar{t}}) = \mathbb{E}\left[\sum_{t=\bar{t}}^{\infty}\gamma^t R(s_t^i, a_t^i, \hat{\mu}_t)\Big|_{\substack{s_{\bar{t}}^j\sim\mu_{\bar{t}} \\ a_t^j\sim\pi^j(o_t^j) \\ s_{t+1}^j\sim P(\cdot|s_t^j,a_t^j,\hat{\mu}_t)}}\right].$$

However, the maximisation objective for this *cooperative* problem is:

**Definition 3.3** (Population-average expected discounted return)**.** For $i, j \in \{1,\dots,N\}$ the return is

$$V^{pop}(\boldsymbol{\pi}, \mu_{\bar{t}}) = \frac{1}{N}\sum_{i}^{N} V^i(\boldsymbol{\pi}, \mu_{\bar{t}}) = \mathbb{E}\left[\frac{1}{N}\sum_{t=\bar{t}}^{\infty}\sum_{i}^{N}\gamma^t R(s_t^i, a_t^i, \hat{\mu}_t)\Big|_{\substack{s_{\bar{t}}^j\sim\mu_{\bar{t}} \\ a_t^j\sim\pi^j(o_t^j) \\ s_{t+1}^j\sim P(\cdot|s_t^j,a_t^j,\hat{\mu}_t)}}\right].$$

That is, the solution to the control problem is $\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi}\in\Pi^N} V^{pop}(\boldsymbol{\pi}, \mu_{\bar{t}})$.

At the limit as $N \to \infty$, the infinite population of agents can be characterised as a limit distribution $\mu \in \Delta_{\mathcal{S}}$; the infinite-agent setting is termed a MFC problem. The *mean-field flow* $\boldsymbol{\mu}$ is given by the infinite sequence of mean-field distributions s.t. $\boldsymbol{\mu} = (\mu_t)_{t \geq 0}$.

**Definition 3.4** (Induced mean-field flow)**.** We denote by $I(\pi)$ the mean-field flow $\boldsymbol{\mu}$ induced when all the agents follow $\pi$, where this is generated from $\pi$ by

$$\mu_{t+1}(s') = \sum_{s,a} \mu_t(s)\pi(a|o_t)P(s'|s, a, \mu_t).$$

The snapshot of this induced flow at $t$ is given by $I(\pi)_t$.

**Definition 3.5** (Social welfare)**.** When all agents follow policy $\pi$ giving mean-field flow $\boldsymbol{\mu} = I(\pi)$, $\pi$'s social welfare is

$$W(\pi; I(\pi)) = \mathbb{E}\left[\sum_{t=\bar{t}}^{\infty} \gamma^t (R(s_t, a_t, I(\pi)_t)) \middle|_{\substack{s_{\bar{t}} \sim \mu_{\bar{t}} \\ a_t \sim \pi(\cdot|o_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t, I(\pi)_t)}}\right].$$

**Definition 3.6** (Social optimum)**.** The solution to the MFC problem is a social optimum policy $\pi^* \in \Pi$ that maximises the social welfare function in Def. 3.5, i.e. $\pi^* = \arg\max_{\pi \in \Pi} W(\pi; I(\pi))$.

**Remark 3.7.** Previous works showed that the MFC social optimum $\pi^*$ gives a good approximation for the harder-to-solve finite-agent problem (i.e. if $\boldsymbol{\pi} = (\pi^*, \ldots, \pi^*)$), with the error characterised by $\mathcal{O}(\frac{1}{\sqrt{N}})$ (Gu et al., 2021; Mondal et al., 2022; Cui et al., 2023b;c; Bayraktar & Kara, 2024).

When the distribution is the same for all $t$, i.e. $\mu_t = \mu_{t+1}\ \forall t \geq 0$, we say the mean-field flow is *stationary*, giving a stationary MFC problem. *Non-stationary* problems may require the policy to depend on the mean field such that $o_t^i = (s_t^i, \hat{\mu}_t)$, whereas the observation in the stationary case can be simplified to $o_t^i = s_t^i$. However, since classical approaches to the MFC problem often conceive of a central planner trying to guide the population to a distribution that maximises the expected return, they sometimes have policies that depend on the mean field even in the stationary case (Laurière et al., 2022a; Carmona et al., 2023; Cui et al., 2023c). Therefore *we permit mean field-dependent policies for the sake of generality, but show through our ablation studies that in practice our algorithms require only* $\pi^i(a|o_t^i) = \pi^i(a|s_t^i)$ in our experimental tasks, which have *stationary* solutions.

Furthermore, it is unrealistic to assume that decentralised agents with a possibly limited communication radius would have perfect observability of the global mean field $\hat{\mu}_t$. Therefore we allow agents to form a local estimate $\tilde{\hat{\mu}}_t^i$ which can be improved by communication with neighbours, using Alg. 4 (from Alg. 3 in Benjamin & Abate (2024) for the MFG setting). We thus have $o_t^i = (s_t^i, \tilde{\hat{\mu}}_t^i)$. Formally we can now say that when $o_t^i = (s_t^i, \hat{\mu}_t)$ or $(s_t^i, \tilde{\hat{\mu}}_t^i)$, we have the set of policies defined as $\Pi = \{\pi : \mathcal{S} \times \Delta_{\mathcal{S}} \to \Delta_{\mathcal{A}}\}$, and the set of Q-functions denoted $\mathcal{Q} = \{q : \mathcal{S} \times \Delta_{\mathcal{S}} \times \mathcal{A} \to \mathbb{R}\}$. (N.b. when $o_t^i = s_t^i$, we instead have $\Pi = \{\pi : \mathcal{S} \to \Delta_{\mathcal{A}}\}$ and $\mathcal{Q} = \{q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$.)

**Remark 3.8.** We have just seen that an optimal solution to the theoretical MFC problem is a single policy that, when followed by all agents in the infinite population, maximises the population's expected return. We give two ways to conceive of our work, illustrated in Fig. 1, which mirror and make more explicit the motivations underpinning other MFC works (Cui et al., 2023c; Dayanikli et al., 2024; Zaman et al., 2024; Bayraktar & Kara, 2024; Yang et al., 2025).

1. Firstly, while previous works might make unrealistic assumptions about access to an oracle for the infinite population, we contribute algorithms that allow the solution to a MFC problem to be learnt using the empirical distribution of a decentralised finite population. Note that it is unnecessary (and may be impractical) to assume the decentralised agents always follow a single identical policy throughout training.

2. Alternatively, we may have originally been interested in solving a cooperative problem for a large, finite population, but, due to the scalability issues of learning approaches like MARL, were forced to turn to the MFC framework to find a policy that gives an approximate solution to the finite-population problem. We contribute algorithms that allow the deployed finite population to find the

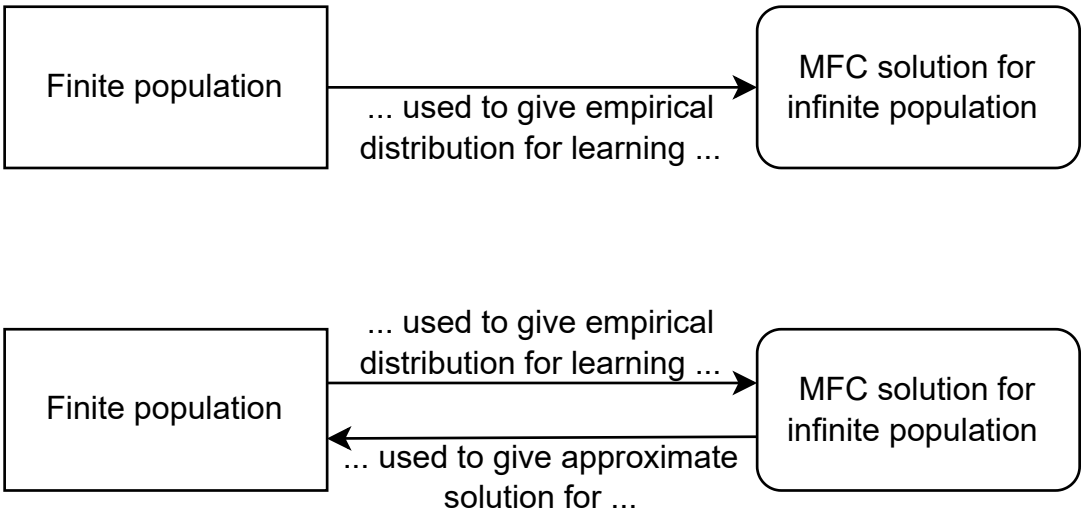



Figure 1: Two possible ways to conceive of our work regarding the relationship between the infinite- and finite-population control problems, described in Rem. 3.8. Note that using the finite empirical population to learn the single-policy MFC social optimum $\boldsymbol{\pi} = (\pi^*, \dots, \pi^*)$ for the infinite population (Def. 3.6) is *not* the same as directly finding $\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi} \in \Pi^N} V^{pop}(\boldsymbol{\pi}, \mu_{\bar{t}}) = (\pi^1, \dots, \pi^N)$, i.e. the tuple of *individual* policies that maximises the expected finite population-average return in Def. 3.3, a problem known to be hard (Cui et al., 2023c; Bernstein et al., 2002).

MFC solution that in turn approximately solves the original problem, without unrealistic assumptions about centralised training. Under this framing, it may matter less whether all agents follow a single policy in practice (Yardim et al. (2023) and Benjamin & Abate (2023; 2024) follow a similar logic in MFGs).

### 3.2 Munchausen Online Mirror Descent

Recent works have solved MFGs from non-episodic runs of the finite-population empirical system using a form of policy iteration called Online Mirror Descent (OMD) (Benjamin & Abate, 2024); we adapt this to learn a social optimum in the MFC setting. OMD involves beginning with an initial policy $\pi_0$, and then at each iteration $k$, evaluating the current policy $\pi_k$ with respect to its induced mean-field flow $\boldsymbol{\mu} = I(\pi_k)$ to compute its Q-function $Q_{k+1}$. To stabilise the learning process, we then use a weighted sum over this and past Q-functions, and set $\pi_{k+1}$ to be the softmax over this weighted sum, i.e. $\pi_{k+1}(\cdot|o) = softmax\left(\frac{1}{\tau_q}\sum_{\kappa=0}^{k+1} Q_\kappa(o, \cdot)\right)$. $\tau_q$ is a temperature parameter that scales the entropy in Munchausen RL (see Sec. 4.2) (Vieillard et al., 2020); this is a different temperature to the one agents use when communicating policies, denoted $\tau_k^{comm}$ and discussed in Sec. 4.3.

If the Q-function is approximated non-linearly, it is difficult to compute this weighted sum. The *Munchausen trick* addresses this by computing a single Q-function that mimics the weighted sum using implicit regularisation based on the Kullback-Leibler (KL) divergence between $\pi_k$ and $\pi_{k+1}$ (Vieillard et al., 2020). Using this reparametrisation gives Munchausen OMD (MOMD), detailed in Sec. 4.2 (Laurière et al., 2022b; Wu et al., 2024). MOMD does not bias policies, and has the same convergence guarantees as OMD (Hadikhanloo, 2017; Perolat et al., 2021; Wu et al., 2024).

### 3.3 Networks

Our decentralised population exhibits two time-varying graphs. The first is a communication network, by which agents can exchange information:

**Definition 3.9** (Time-varying communication network)**.** The time-varying graph $(\mathcal{G}_t^{comm})_{t\geq 0}$ is given by $\mathcal{G}_t^{comm} = (\mathcal{N}, \mathcal{E}_t^{comm})$, where $\mathcal{N}$ is the set of vertices each representing an agent $i \in \{1, \dots, N\}$, and the edge set $\mathcal{E}_t \subseteq \{(i,j) : i,j \in \mathcal{N}\}$ is the set of undirected links present at time $t$. A network's *diameter* $d_{\mathcal{G}_t^{comm}}$ is the maximum of the shortest path lengths between any pair of nodes.

In principle, agents can use this same communication network to receive information about others' state in order to estimate the mean field. However, we also define an alternative observation graph that is useful in a specific subclass of environments, which can most intuitively be thought of as those where agents' states are positions in physical space, which include those in our experiments. When this is the case, we usually think of agents' ability to observe each other as depending more abstractly on whether states are visible to each other. This visibility graph is:

**Definition 3.10** (Time-varying state-visibility graph)**.** The time-varying state visibility graph $(\mathcal{G}_t^{vis})_{t\geq 0}$ is given by $\mathcal{G}_t^{vis} = (\mathcal{S}', \mathcal{E}_t^{vis})$, where $\mathcal{S}'$ is the set of vertices representing the environment states $\mathcal{S}$, and the edge set $\mathcal{E}_t^{vis} \subseteq \{(m,n) : m,n \in \mathcal{S}'\}$ is the set of undirected links present at time $t$, indicating which states are visible to each other.

In Sec. 4.4 we present Alg. 4, which forms an initial estimate of the global empirical mean field (to serve as an observation input for agents' Q-/policy-networks) via the visibility graph $\mathcal{G}_t^{vis}$, before refining this estimate via the communication graph $\mathcal{G}_t^{comm}$. Benjamin & Abate (2024) discusses an algorithm for more general settings where the visibility graph $\mathcal{G}_t^{vis}$ does not apply.

## 4 Learning and estimation algorithms

We adapt recent algorithms for the MFG setting, where networked communication is used 1) to form local estimates of the global empirical mean field, and 2) to allow agents to adopt better-performing policies from neighbours to accelerate learning (Benjamin & Abate, 2024). We adapt these algorithms for cooperative MFC, where decentralised agents must optimise the population-average return instead of their individual one (the decentralised agents may not always follow a common policy while training unless we make strong assumptions on the communication network as in Sec. 5, so we do not directly optimise social welfare from Def. 3.5).

It is unrealistic to assume that decentralised agents have access to the global average reward, so we find a third use of the communication network in 3) allowing agents to estimate the global average reward $\hat{r}_t$ from a local neighbourhood. We contribute a novel algorithm Alg. 1 for this purpose (Sec. 4.1), and we describe our main learning method Alg. 2 in Sec. 4.2. Our policy communication algorithm Alg. 3, based on that in Benjamin & Abate (2024) for the MFG setting, is described in Sec. 4.3. Meanwhile Alg. 4 for estimating the mean field, which is taken from Alg. 3 in Benjamin & Abate (2024) for the MFG setting, is described in Sec. 4.4.

### 4.1 Sub-routine for networked estimation of global average reward

Our novel Alg. 1 involves agents using the communication network $\mathcal{G}_t^{comm}$ to locally estimate the global population-average reward received after a given step in the environment. Maximising the population-average reward ensures agents are solving the cooperative MFC problem instead of the non-cooperative MFG. Agents broadcast their received reward with a unique ID to ensure each reward is only counted once (Line 1). They collect those received from neighbours, and repeat the process of broadcasting and expanding their collections for a further $C_r - 1$ rounds, so as to receive rewards from agents more than one hop away on the network (Lines 2-6). They finally set their estimate of the global average to the average of the rewards they have collected (Line 7).

**Algorithm 1** Average reward estimation and communication

**Require:** Time-dependent communication graph $\mathcal{G}_t^{comm}$, rewards $\{r_t^i\}_{i=1}^N$, number of communication rounds $C_r$

1: $\forall i$ : Initialise reward sets $\hat{\mathcal{R}}_{t,1}^i \leftarrow \{(ID^i, r_t^i)\}$
2: **for** $c_r$ in $1, \ldots, C_r$ **do**
3: $\quad \forall i$ : Broadcast $\hat{\mathcal{R}}_{t,c_r}^i$
4: $\quad \forall i : J_t^i \leftarrow \{j \in \mathcal{N} : (i,j) \in \mathcal{E}_t^{comm}\}$
5: $\quad \forall i : \hat{\mathcal{R}}_{t,(c_r+1)}^i \leftarrow \hat{\mathcal{R}}_{t,c_r}^i \cup \bigcup_{j \in J_t^i} \hat{\mathcal{R}}_{t,c_r}^j$
6: **end for**
7: $\forall i : \tilde{\hat{r}}_t^i \leftarrow \frac{1}{|\hat{\mathcal{R}}_{t,C_r}^i|} \sum_{(ID,r) \in \hat{\mathcal{R}}_{t,C_r}^i} r$
8: **return** Estimates of average reward $\left\{\tilde{\hat{r}}_t^i\right\}_{i=1}^N$

**Algorithm 2** Decentralised MFC learning from non-episodic system run

**Require:** loop parameters $K, M, L, E, C_e, C_r, C_p$, learning parameters $\gamma, \tau_q, |B|, cl, \nu$, $\{\tau_k^{comm}\}_{k \in \{0,\ldots,K-1\}}$
**Require:** initial states $\{s_0^i\}_{i=1}^N$; $t \leftarrow 0$

1: $\forall i$ : Randomly initialise parameters $\theta_0^i$ of Q-networks $\check{Q}_{\theta_0^i}(o,\cdot)$, and set $\pi_0^i(a|o) = \text{softmax}\Big(\frac{1}{\tau_q}\check{Q}_{\theta_0^i}(o,\cdot)\Big)(a)$
2: **for** $k \in 0, \ldots, K-1$ **do**
3: $\quad \forall i$: Empty $i$'s buffer
4: $\quad$ **for** $m \in 0, \ldots, M-1$ **do**
5: $\quad\quad \{o_t^i\}_{i=1}^N \leftarrow$ **EstimateMeanFieldAlg. 4**$\big(\mathcal{G}_t^{vis}, \mathcal{G}_t^{comm}, \{s_t^i\}_{i=1}^N\big)$
6: $\quad\quad$ Take step $\forall i : a_t^i \sim \pi_k^i(\cdot|o_t^i), r_t^i = R(s_t^i, a_t^i, \hat{\mu}_t), s_{t+1}^i \sim P(\cdot|s_t^i, a_t^i, \hat{\mu}_t)$; $t \leftarrow t+1$
7: $\quad\quad \{\tilde{\hat{r}}_t^i\}_{i=1}^N \leftarrow$ **EstimateAverageRewardAlg. 1**$\big(\mathcal{G}_t^{comm}, \{r_t^i\}_{i=1}^N\big)$
8: $\quad\quad \forall i$: Add $\big(o_t^i, a_t^i, \tilde{\hat{r}}_t^i, o_{t+1}^i\big)$ to $i$'s buffer
9: $\quad$ **end for**
10: $\quad$ **for** $l \in 0, \ldots, L-1$ **do**
11: $\quad\quad \forall i$ : Sample batch $B_{k,l}^i$ from $i$'s buffer
12: $\quad\quad$ Update $\theta$ to minimise $\hat{\mathcal{L}}(\theta, \theta')$ as in Def. 4.1
13: $\quad\quad$ If $l \mod \nu = 0$, set $\theta' \leftarrow \theta$
14: $\quad$ **end for**
15: $\quad \check{Q}_{\theta_{k+1}^i}(o,\cdot) \leftarrow \check{Q}_{\theta_{k,L}^i}(o,\cdot)$
16: $\quad \forall i$ : $\pi_{k+1}^i(a|o) \leftarrow \text{softmax}\Big(\frac{1}{\tau_q}\check{Q}_{\theta_{k+1}^i}(o,\cdot)\Big)(a)$
17: $\quad \big(\{\pi_{k+1}^i\}_i, \{s_t^i\}_i, t\big) \leftarrow$ **CommunicatePolicyAlg. 3**$\big(\mathcal{G}_t^{comm}, \{\pi_{k+1}^i\}_i, \{s_t^i\}_i, t\big)$
18: **end for**
19: **return** policies $\{\pi_K^i\}_{i=1}^N$

### 4.2 Main learning algorithm for updating Q-networks and policies

Our novel Alg. 2, adapted from non-cooperative Alg. 1 in Benjamin & Abate (2024), contains the core method for online MFC learning using the empirical mean field in a non-episodic system run. Our MOMD-based method (Sec. 3.2) works as follows. Each agent $i$ approximates its Q-function $\check{Q}_{\theta_k^i}(o,\cdot)$ with its own neural network parametrised by $\theta_k^i$. Agent $i$'s policy is determined by

$$\pi_{\theta_k^i}(a|o) = \text{softmax}\left(\frac{1}{\tau_q}\check{Q}_{\theta_k^i}(o,\cdot)\right)(a)\,.$$

We denote this as $\pi_k^i(a|o)$ for simplicity when appropriate. Each agent maintains a buffer (with size $M$) of collected transitions of the form $\big(o_t^i, a_t^i, \tilde{\hat{r}}_t^i, o_{t+1}^i\big)$, where $\tilde{\hat{r}}_t^i$ is $i$'s local estimate of the global average reward obtained by running Alg. 1 (Line 7). At each iteration $k$, agents empty their buffer (Line 3) before collecting

**Algorithm 3** Policy communication and selection

**Require:** Time-dependent communication graph $\mathcal{G}_t^{comm}$, loop parameters $E, C_p$, learning parameters $\gamma$, $\{\tau_k^{comm}\}_{k\in\{0,\dots,K-1\}}$
**Require:** policies $\{\pi_{k+1}^i\}_{i=1}^N$; states $\{s_t^i\}_{i=1}^N$; $t$
1: $\forall i : \sigma_{k+1}^i \leftarrow 0$
2: **for** $e \in 0, \dots, E-1$ evaluation steps **do**
3: $\{o_t^i\}_{i=1}^N \leftarrow$ **EstimateMeanFieldAlg. 4**$\big(\mathcal{G}_t^{vis}, \mathcal{G}_t^{comm}, \{s_t^i\}_{i=1}^N\big)$
4: Take step $\forall i : a_t^i \sim \pi_k^i(\cdot|o_t^i), r_t^i = R(s_t^i, a_t^i, \hat{\mu}_t), s_{t+1}^i \sim P(\cdot|s_t^i, a_t^i, \hat{\mu}_t)$
5: $\forall i : \sigma_{k+1}^i \leftarrow \sigma_{k+1}^i + \gamma^e \cdot r_t^i$
6: $t \leftarrow t+1$
7: **end for**
8: **for** $C_p$ rounds **do**
9: $\forall i$ : Broadcast $\sigma_{k+1}^i, \pi_{k+1}^i$
10: $\forall i : J_t^i \leftarrow i \cup \{j \in \mathcal{N} : (i,j) \in \mathcal{E}_t^{comm}\}$
11: $\forall i$ : Select adopted$^i \sim \Pr\big(\text{adopted}^i = j\big) = \frac{\exp(\sigma_{k+1}^j / \tau_k^{comm})}{\sum_{x\in J_t^i} \exp(\sigma_{k+1}^x / \tau_k^{comm})}\ \forall j \in J_t^i$
12: $\forall i : \sigma_{k+1}^i \leftarrow \sigma_{k+1}^{\text{adopted}^i}, \pi_{k+1}^i \leftarrow \pi_{k+1}^{\text{adopted}^i}$
13: $\{o_t^i\}_{i=1}^N \leftarrow$ **EstimateMeanFieldAlg. 4**$\big(\mathcal{G}_t^{vis}, \mathcal{G}_t^{comm}, \{s_t^i\}_{i=1}^N\big)$
14: Take step $\forall i : a_t^i \sim \pi_k^i(\cdot|o_t^i), r_t^i = R(s_t^i, a_t^i, \hat{\mu}_t), s_{t+1}^i \sim P(\cdot|s_t^i, a_t^i, \hat{\mu}_t)$; $t \leftarrow t+1$
15: **end for**
16: **return** (policies $\{\pi_{k+1}^i\}_{i=1}^N$, states $\{s_t^i\}_{i=1}^N$, $t$)

$M$ new transitions in the environment (Lines 4-9). Each decentralised agent then trains its Q-network $\check{Q}_{\theta_k^i}$ via $L$ updates (Lines 10-14) as follows.

For stability, $i$ also maintains a target network $\check{Q}_{\theta_{k,l}^{i,'}}$ with the same architecture but parameters $\theta_{k,l}^{i,'}$ copied from $\theta_{k,l}^i$ less regularly than $\theta_{k,l}^i$ themselves are updated, i.e. only every $\nu$ learning iterations (Line 13). At each iteration $l$, the agent samples a random batch $B_{k,l}^i$ of $|B|$ transitions from its buffer (Line 11). It then trains its Q-network using stochastic gradient descent to minimise the loss in Def 4.1 (Line 12). The trained Q-network determines $i$'s updated policy (Line 16).

**Definition 4.1** (Q-network empirical loss)**.** The training loss to be minimised is given by

$$\hat{\mathcal{L}}(\theta, \theta') = \frac{1}{|B|} \sum_{transition \in B_{k,l}^i} \left| \check{Q}_{\theta_{k,l}^i}(o_t, a_t) - T \right|^2,$$

$$\text{where} \quad T = \tilde{\hat{r}}_t + \left[ \tau_q \ln \pi_{\theta_{k,l}^{i,'}}(a_t|o_t) \right]_{cl}^0 + \gamma \sum_{a\in\mathcal{A}} \pi_{\theta_{k,l}^{i,'}}(a|o_{t+1}) \left( \check{Q}_{\theta_{k,l}^{i,'}}(o_{t+1}, a) - \tau_q \ln \pi_{\theta_{k,l}^{i,'}}(a|o_{t+1}) \right).$$

For $cl < 0$, $[\cdot]_{cl}^0$ is a clipping function used in Munchausen RL to prevent numerical issues if the policy is too close to deterministic, as the log-policy term is otherwise unbounded (Vieillard et al., 2020; Wu et al., 2024).

### 4.3 Sub-routine for communicating and refining policies

Alg. 3 (based on Alg. 1 in Benjamin & Abate (2024) for MFGs) uses the communication network $\mathcal{G}_t^{comm}$ to spread policy updates that are estimated to be better performing through the population, allowing faster learning than in the independent and central-agent cases.

Alg. 3 is run after agents have independently updated their policies according to their newly trained Q-networks at each iteration $k$ of the main learning algorithm (Line 17, Alg. 2). In Alg. 3, agents obtain an approximation of their *individual* discounted expected return $\{V^i(\boldsymbol{\pi}, \mu_t)\}_{i=1}^N$ (Def. 3.2, i.e. *not* the

**Algorithm 4** Mean-field estimation and communication for environments with $\mathcal{G}_t^{vis}$

**Require:** Time-dependent visibility graph $\mathcal{G}_t^{vis}$, time-dependent communication graph $\mathcal{G}_t^{comm}$, states $\{s_t^i\}_{i=1}^N$, number of communication rounds $C_e$

1: $\forall i, s$ : Initialise count vector $\hat{v}_{t,1}^i[s]$ with $\emptyset$
2: $\forall i$, $\forall s' \in \mathcal{S}' : (s_t^i, s') \in \mathcal{E}_t^{vis}$ : $\hat{v}_{t,1}^i[s'] \leftarrow \sum_{j \in 1,\dots,N : s_t^j = s'} 1$
3: **for** $c_e \in 1, \dots, C_e$ **do**
4: $\quad \forall i$ : Broadcast $\hat{v}_{t,c_e}^i$
5: $\quad \forall i : J_t^i \leftarrow i \cup \{j \in \mathcal{N} : (i,j) \in \mathcal{E}_t^{comm}\}$
6: $\quad \forall i, s$ : Initialise new count vector $\hat{v}_{t,(c_e+1)}^i[s]$ with $\emptyset$
7: $\quad \forall i, s$ and $\forall j \in J_t^i : \hat{v}_{t,(c_e+1)}^i[s] \leftarrow \hat{v}_{t,c_e}^j[s]$ if $\hat{v}_{t,c_e}^j[s] \neq \emptyset$
8: **end for**
9: $\forall i : counted\_agents_t^i \leftarrow \sum_{s \in \mathcal{S} : \hat{v}_t^i[s] \neq \emptyset} \hat{v}_t^i[s]$
10: $\forall i : uncounted\_agents_t^i \leftarrow N - counted\_agents_t^i$
11: $\forall i : unseen\_states_t^i \leftarrow \sum_{s \in \mathcal{S} : \hat{v}_t^i[s] = \emptyset} 1$
12: $\forall i, s$ where $\hat{v}_t^i[s]$ is not $\emptyset$ : $\tilde{\hat{\mu}}_t^i[s] \leftarrow \frac{\hat{v}_t^i[s]}{N}$
13: $\forall i, s$ where $\hat{v}_t^i[s]$ is $\emptyset$ : $\tilde{\hat{\mu}}_t^i[s] \leftarrow \frac{uncounted\_agents_t^i}{N \times unobserved\_states_t^i}$
14: **return** $\{$(states $s_t^i$, mean-field estimates $\tilde{\hat{\mu}}_t^i$)$\}_{i=1}^N$

population-average return, which would not give differentiation between the different updated policies). They do so by collecting individual rewards for $E$ steps (not added to the training buffer), and calculating the discounted sum of rewards over these finite steps, setting this value to $\sigma_{k+1}^i$ (Lines 1-7). We can characterise this approximation of the infinite-step return as $\{\sigma_{k+1}^i\}_{i=1}^N = \{\hat{V}^i(\boldsymbol{\pi}_{k+1}, \mu_t; E)\}_{i=1}^N$.

They then broadcast their Q-network parameters along with $\sigma_{k+1}^i$ (Line 9). Receiving these from their neighbours $J_t^i$ on the network, agents select which set of parameters to adopt by taking a softmax over their own and the received estimate values $\sigma_{k+1}^j \; \forall j \in J_t^i$, defined as follows (Lines 10-12):

$$\text{adopted}^i \sim \Pr\left(\text{adopted}^i = j\right) = \frac{\exp\left(\sigma_{k+1}^j / \tau_k^{comm}\right)}{\sum_{x \in J_t^i} \exp\left(\sigma_{k+1}^x / \tau_k^{comm}\right)}.$$

They repeat this broadcast and adoption process for $C_p$ rounds (distinct from the $C_r/C_e$ communication rounds for the other sub-routines).

### 4.4 Sub-routine for networked estimation of global empirical mean-field

Networked agents use Alg. 4 (this is Alg. 3 from Benjamin & Abate (2024) for the MFG setting) to locally estimate the global empirical mean field, to serve as an observation input for their Q-/policy-networks. Recall that we include this added observation and sub-routine for generality, especially for non-stationary problems. However it is often not necessary, particularly in stationary problems like those in our experiments, where agents can find the social optimum while only observing $o_t^i = s_t^i$, and therefore would not need to estimate the mean field.

Alg. 4 involves agents using the visibility graph $\mathcal{G}_t^{vis}$ to count the number of agents in locations that fall within the visibility radius (Line 2). For $C_e$ communication rounds, agents can supplement this local count with those received from neighbours over the communication network $\mathcal{G}_t^{comm}$, in order to count agents that do not fall within the visibility radius (Lines 3-8). We assume agents know the population's total size $N$, and therefore can distribute the uncounted agents uniformly over the states that remain unaccounted for after the communication rounds (Lines 9-11). Agents now have a vector containing a true or estimated count for every state; this is converted to an estimated empirical mean field by dividing all counts by $N$ (Lines 12-13).

# 5 Theoretical results

## 5.1 Introduction

We follow the definitions of the central-agent and independent-learning architectures from closely related works that learn MFGs online from a non-episodic run of the empirical system (Yardim et al., 2023; Benjamin & Abate, 2023; 2024); both architectures can each be seen as special cases of our networked algorithm:

- In the **central-agent** case, only arbitrary central agent $i = 1$ updates a Q-network and automatically pushes this to all other agents in place of the decentralised policy communication in Line 17 of Alg. 2. Additionally, the true global mean-field distribution and average reward are always used in place of the local estimates, i.e. $\tilde{\hat{\mu}}_t^i = \hat{\mu}_t$ and $\tilde{\hat{r}}_t^i = \hat{r}$.
- In the **independent** case, there are never any links in $\mathcal{G}_t^{comm}$ or $\mathcal{G}_t^{vis}$, i.e. $\mathcal{E}_t^{comm} = \mathcal{E}_t^{vis} = \emptyset$.

We prove theoretically that the policy communication and adoption scheme allows networked agents to increase their returns faster than these alternatives (with the central-agent paradigm being potentially unrealistic and vulnerable in any case). Rem. 5.1 suggests informal reasons for our formal results to aid intuitive understanding.

**Remark 5.1.** Like many cooperative learning paradigms, both the independent and central-agent alternatives to our networked architecture may suffer from the credit-assignment problem, in that it is not clear how agents' local state $s_t^i$ and local action $a_t^i$ contributed to the (locally estimated) *average* reward $\tilde{\hat{r}}_t^i$ (Li & Li, 2024; Cazenille et al., 2025). Agents may receive low individual reward $r_t^i$ by taking action $a_t^i$ given $o_t^i$, but would nevertheless learn that doing so was 'good' if the rest of the population took highly rewarded actions at the same step giving high average reward $\tilde{\hat{r}}_t^i$. By drawing spurious relations, an agent's updated policy $\pi_{k+1}^i(a|o)$ may negatively impact (or simply not advance) the goal of maximising social welfare. Including the (estimated) empirical mean field in the observation $o_t^i = (s_t^i, \tilde{\hat{\mu}}_t^i)$ might mitigate this slightly by indicating which mean fields gave high average rewards. However, this does not solve the issue of allowing learners to distinguish between helpful or unhelpful local actions $a_t^i$, whether those learners are centralised or not, since actions can affect rewards in ways other than simply by helping to reach a certain mean field. By updating policies with respect to average return but then spreading updates through the population which are estimated to give a higher *individual* return, despite this being a cooperative problem, we reduce the credit-assignment problem by replicating updated policies that should contribute positively to the population-average return, and filtering out those that do not.

Moreover, even if we assumed credit assignment were not a problem, there is randomness in the Q-network update: agents have stochastic policies and thus may collect a wide variety of transitions to add to their individual buffers, from which they sample randomly when training Q-networks. There may therefore be considerable variance in the quality of their estimated Q-functions, leading in turn to variance in the quality of policy updates. At each iteration of the central-agent algorithm, in *expectation* the central learner will by definition have an average-quality update, and its updated policy will be pushed to the entire population whether or not it performs well, also giving large variance in the quality of updates. Our decentralised networked approach permits beneficial parallelisation in place of this single-learner method, by generating a whole population of possible updates, from which the one(s) estimated to be best-performing can be selected via a process akin to the comparison of fitness functions in evolutionary algorithms. These are then spread around the population, biasing networked populations towards better performing updates.

We give the theoretical analysis separately for two important subclasses of cooperative game usually found in MFC, which have different reward structures and therefore can incentivise different population behaviour:

1. *coordination games*, where the social welfare is increased by agents aligning their strategies, such as in consensus/synchronisation/rendezvous tasks;
2. *anti-coordination games*, where the social welfare is increased by the population exhibiting diverse strategies, such as in exploration, coverage or task allocation games.

These subclasses cover a large proportion of cooperative objectives in anonymous, symmetric settings with large populations. We emphasise that the fact that agents would in principle benefit from having diverse policies in anti-coordination games does not contradict the classical MFC framework that simplifies the infinite population problem by finding the single policy to be shared by all agents. In the symmetric (i.e. identical reward and transition functions) MFC limit, an optimal solution can be realised by having the infinite agents all follow the single socially optimal policy, even for reward functions that favour diversity. A very large number of works on both MFC and MFGs conduct experiments on anti-coordination games, particularly dispersal and exploration tasks, despite assuming that the population follows a shared single policy learnt by a central node (Ruthotto et al., 2020; Laurière et al., 2022a; Lee et al., 2024). We make the distinction between coordination and anti-coordination games to aid theoretical analysis of our decentralised policy adoption scheme compared with entirely independent learning: while it is intuitive that adopting independently-updated policies from neighbours via the communication scheme would be beneficial in coordination games, we also show theoretically and empirically that the adoption scheme provides a benefit in anti-coordination games, though this requires separate analysis.

To define the two types of game, we first introduce the following functions. $\mathbb{I}[\cdot]$ is the indicator function, which equals 1 if the condition inside is true and 0 otherwise. $b : \Pi \to \mathbb{R}_{\geq 0}$ is a *base return function* that quantifies a policy's inherent ability to receive rewards regardless of how many other agents follow the same strategy. For example, if agents are rewarded for agreeing on one of a number of targets at which to meet, then policies that visit none of the designated targets will have lower returns than those that do, whether agents are aligned or not. $f_c : \mathbb{N} \to \mathbb{R}_{>0}$ (resp. $f_d : \mathbb{N} \to \mathbb{R}_{>0}$) is a *coordination (resp. anti-coordination) scaling function*. It has minimum $f_c(1) > 0$ (resp. $f_d(0) > 0$), and increases monotonically with the number of agents whose policies match (resp. are different from) $i$'s.

**Definition 5.2** (Coordination game)**.** The agents' return can be decomposed as follows, $\forall i, j \in \{1, \dots, N\}$: $V^i(\boldsymbol{\pi}, \mu_{\bar{t}}) = h\left(b(\pi^i), f_c\left(\sum_{j\in\{1,\dots,N\}} \mathbb{I}\left[\pi^i = \pi^j\right]\right)\right)$, where $h : \mathbb{R}_{\geq 0} \times \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ is a function that composes $b(\cdot)$ and $f_c(\cdot)$ and is monotonic in both arguments, i.e. an increase in either the policy's intrinsic ability to attain rewards, or the extent to which it is aligned with other agents' policies, gives a higher return.

**Definition 5.3** (Anti-coordination game)**.** The agents' return can be decomposed as follows, $\forall i, j \in \{1, \dots, N\}$: $V^i(\boldsymbol{\pi}, \mu_{\bar{t}}) = h\left(b(\pi^i), f_d\left(N - \sum_{j\in\{1,\dots,N\}} \mathbb{I}\left[\pi^i = \pi^j\right]\right)\right)$, where $h : \mathbb{R}_{\geq 0} \times \mathbb{R}_{>0} \to \mathbb{R}_{\geq 0}$ is a function that composes $b(\cdot)$ and $f_d(\cdot)$ and is monotonic in both arguments, i.e. an increase in either the policy's intrinsic ability to attain rewards, or the extent to which it is different from other agents' policies, gives a higher return.

Note that in our setting, where policy parameters are directly communicated and adopted among the population, we focus on exact equality of policies for simplicity of the theory. However, these definitions could be made more general and inclusive by instead considering similarity kernels or label mappings of strategically relevant parts of policies.

### 5.2 Analysis

Our sub-routines involve time-varying networks sharing different types of information at different points in the algorithm, meaning that theoretical analysis can potentially grow complicated. We seek to simplify this analysis to make it more intuitive and useful by focusing on the benefit of the decentralised policy exchange scheme in Alg. 3. This is because our ablation studies of Algs. 1 (average reward estimation) and 4 (mean field estimation) in Sec. 6.3 indicate that the policy exchange scheme is the dominant factor in driving the benefit of the networked paradigm in our experimental settings. Moreover, recall that Alg. 4 is only necessary when we allow population-dependent policies i.e. $o_t^i = (s_t^i, \tilde{\hat{\mu}}_t^i)$, whereas for stationary problems, including all those in our experiments and many others, using a mean field observation or estimation is not actually required for finding the optimal policy.

For simplicity of the theory, we make several assumptions. We explore the conditions under which these assumptions apply in practice, and discuss how even when loosening the assumptions, they still provide useful heuristic insight as to how our networked communication scheme affords benefits over the central-agent and independent-learning architectures. We do not enforce the assumptions in our experiments, and

our empirical results nevertheless follow our theoretical theorems in all but some specific instances that we discuss.

The first assumption simplifies the theory by presuming that it is only the decentralised policy communication scheme that creates a difference in learning between the networked and central-agent cases, by assuming that the estimated mean fields and average rewards are equivalent to the true ones used in the central-agent case. Note that populations with fully connected networks will in any case always be able to accurately estimate $\hat{r}$ and $\hat{\mu}_t$ by Algs. 1 and 4, even for $C_r = 1$ and $C_e = 0$. This is may apply reasonably commonly in practice depending on the scenario; for example, if the network is defined by a broadcast radius (as in our experiments), then the network will be fully connected whenever that radius is at least large enough to cover the area that all the agents fall within. Moreover, as just mentioned, our ablation studies suggest that the policy communication scheme is the dominant factor in our experimental settings anyway, with the estimated mean field not required at all in the broad class of stationary problems. We leave analysis of the theoretical impact of worsening mean-field and average-reward estimations for future work.

**Assumption 5.4.** Assume that Algs. 1 and 4 allow networked agents to obtain accurate estimations of the true population-average rewards and global empirical mean field respectively, i.e. $\forall i\ \tilde{\hat{\mu}}_t^i = \hat{\mu}_t$ and $\tilde{\hat{r}}_t^i = \hat{r}$.

Recall that at each iteration $k$ of Alg. 2, after individually updating their policies in Line 16, the population has the policies $\{\pi_{k+1}^i\}_{i=1}^N$. There is randomness in these individual policy updates, stemming from the random sampling of each agent's individually collected buffer. In Lines 1-7 of Alg. 3, agents estimate the individual infinite-step discounted returns $\{V^i(\boldsymbol{\pi}, \mu_0)\}_{i=1}^N$ (Def. 3.2) of their updated policies by computing $\{\sigma_{k+1}^i\}_{i=1}^N$: the $E$-step discounted return with respect to the empirical mean field generated when agents follow policies $\{\pi_{k+1}^i\}_{i=1}^N$.

We next assume that the populations' policies are all pair-wise distinct after the updates in Line 16 and before the policy communication. This ensures that policies that are estimated to receive higher returns (and are thus adopted) are being evaluated as higher-performing due to receiving higher base returns, rather than simply because of how aligned or distinct they already happen to be with regard to other policies. This avoids scenarios where, for example, significantly suboptimal policies that are shared across multiple agents after the update (in the case of a coordination game) end up spreading through the population by communication at the expense of a more promising but less common policy, decelerating rather than accelerating improvement. In practice, this assumption is highly likely to apply in most situations in any case. Even if agents start a given iteration with identical policies, their different random seeds are likely to mean that they collect different sample transitions to add to their reinitialised buffers. Even if their buffers end up containing identical transitions, their different random seeds are likely to mean that they sample differently from their buffers, leading to slightly different updates to their policy networks.

**Assumption 5.5.** Assume that directly after the policy updates in Line 16 (Alg. 2), before any policy transfer as in the networked or central-agent algorithms, all policies are pair-wise distinct due to the randomness in these updates, i.e. $\forall i, j \in \{1, \dots, N\}\ \pi_{k+1}^i \neq \pi_{k+1}^j$. This means the function $f_c$ attains its minimum $f_c(1)$, and $f_d$ attains its maximum $f_d(N-1)$.

We now assume that the finite-step estimations of the returns give sufficiently accurate comparisons between policies, so that better policies are indeed the ones that get adopted in expectation.

**Assumption 5.6.** Assume that $\{\sigma_{k+1}^i\}_{i=1}^N$ are sufficiently good estimations so as to respect the ordering of the true infinite discounted individual returns $\{V^i(\boldsymbol{\pi}_{k+1}, \mu_0)\}_{i=1}^N$, i.e.

$$V^i(\boldsymbol{\pi}_{k+1}, \mu_0) > V^j(\boldsymbol{\pi}_{k+1}, \mu_0) \iff \sigma_{k+1}^i > \sigma_{k+1}^j \qquad \forall i, j \in \{1, \dots, N\}.$$

In practice, even if Assumption 5.6 does not strictly hold, the softmax parameter $\tau_k^{comm}$ allows a smooth degradation as the ordering of the approximations worsens with respect to the ordering of the true values. That is, if instead of the exact correct policy ordering we have that better policies are simply *more likely* to be given higher estimated evaluations, then the softmax means that these policies remain *more likely* to spread, and a better policy may still be adopted even if it is not evaluated as being better.

The next assumption presumes that the networked population reaches consensus on a single policy within each $k$ iteration. We use it only in Thm. 5.9, and we do so to give general and intuitive comparison

with the central-agent population which always shares a single policy. Incomplete consensus would give different levels of alignment/diversity, such that the relative performance of the central-agent and networked architectures might otherwise depend on the specific reward function of the task, and whether base return or alignment/diversity is more important in that reward function.

**Assumption 5.7.** Assume that after the $C_p$ rounds in Lines 8-15 (Alg. 3), in which agents exchange and adopt policies from neighbours, the networked population is left with a single policy such that $\forall i, j \in \{1, \dots, N\}$ $\pi^i_{k+1} = \pi^j_{k+1}$.

While this may sound like a strong assumption, we phrase it like this so as not to make overly strong restrictions on the communication network instead - we intentionally leave it so that Assumption 5.7 can be fulfilled in numerous ways. Most simply we can think of Assumption 5.7 holding if:

1. we set $\tau^{comm}_k$ close to 0 for all $k$, such that the softmax essentially becomes a max function; and
2. the communication network $\mathcal{G}^{comm}_t$ is static and connected during the $C_p$ communication rounds, where $C_p$ is at least as large as the network diameter $d_{\mathcal{G}^{comm}_t}$.

Under these conditions, previous results on max-consensus algorithms show that all agents in the network will converge on the highest value $\sigma^{max}_{k+1}$ (and hence the unique associated $\pi^{max}_{k+1}$) within a number of rounds equal to the diameter $d_{\mathcal{G}^{comm}_t}$ (Nejad et al., 2009; Benjamin & Abate, 2023). If we assumed more strongly that the network was always *fully* connected, policy consensus would be achieved within a single communication round.

Policy consensus can be achieved even outside of these conditions, including if the network is dynamic and not connected at every step. The *union* of a collection of graphs $\{\mathcal{G}_t, \mathcal{G}_{t+1}, \cdots, \mathcal{G}_{t+\omega}\}$ ($\omega \in \mathbb{N}$) is the graph with vertices and edge set equalling the union of the vertices and edge sets of the graphs in the collection (Jadbabaie et al., 2003). A collection is *jointly connected* if its members' union is connected. Now, instead of assuming that the communication network is static and connected, we assume instead only that the sequence of networks contains one or more jointly connected collections. Then max-consensus is reached within $C_p$ if $C_p$ is large enough that the number of jointly connected collections occurring within $C_p$ is equal to the largest diameter of the union of any such collection.

Thus Assumption 5.7 may not hold if $C_p$ is not large enough or if parts of the population remain isolated. However, we do not enforce this assumption in our experiments, where we use $C_p = 1$ to show the benefit of even just one communication round, yet we still see networked populations significantly outperforming central-agent populations across anti-coordination games. In coordination games, while networked populations that are more connected (due to having larger communication radii) usually perform similarly to or better than central-agent populations, those that are less connected occasionally perform less well than the central-agent populations. This is probably due to Assumption 5.7 being empirically more likely to be violated in less connected populations, which in turn is more of an issue in coordination games (where consensus is more likely to be beneficial) than in anti-coordination games (where some lack of consensus does not prevent, or even helps, networked populations to outperform central-agent ones in practice).

The next assumption presumes that if a certain policy, when followed by all members of a finite population, is better than another policy when the latter is followed by all members of a finite population, then the same quality ordering will apply when members of infinite populations follow each policy. We require this in order to relate our analysis of learning in the empirical finite population back to the mean field limit when comparing with central-agent learning. Since the finite population can be arbitrarily large, and in many environments when all agents follow the same policy the finite population-average return will converge smoothly to the infinite population social welfare, this assumption will naturally hold in many scenarios. For example, a policy that is better than another at getting a population of 500 or 5,000,000 agents to cluster in a particular location will also be better than the other policy at getting an infinite population to gather at the location. Nevertheless this order preservation is not a completely general phenomenon, and strict inequalities can vanish or reverse in the limit, especially in models with thresholds or discontinuities in the dependence of rewards or transitions on the mean field, so we state it as an explicit condition.

**Assumption 5.8.** Say we have two different policies that could be shared by the whole population such that $\boldsymbol{\pi}^x = (\pi^x, \dots, \pi^x)$ and $\boldsymbol{\pi}^y = (\pi^y, \dots, \pi^y)$. We assume that:

$$V^{pop}(\boldsymbol{\pi}^x, \mu_0) > V^{pop}(\boldsymbol{\pi}^y, \mu_0) \iff W(\pi^x, I(\pi^x)) > W(\pi^y, I(\pi^y)).$$

We have now given all the assumptions for our first theorem. Assumption 5.7 assumes that after the $C_p$ policy exchange rounds in Lines 8-15 of Alg. 3, the networked population is left with a single policy. Call this consensus policy $\pi^{\text{net}}_{k+1}$, and its associated finitely approximated return $\sigma^{\text{net}}_{k+1}$. Recall that the central-agent case is where the Q-network update of arbitrary agent $i = 1$ is automatically pushed to all the others instead of the policy evaluation and exchange in Line 17 of Alg. 2; this is equivalent to a networked case where policy consensus is reached on a *random* one of the policies $\{\pi^i_{k+1}\}^N_{i=1}$. Call this policy *arbitrarily* given to the whole population $\pi^{\text{cent}}_{k+1}$, and its associated finitely approximated return $\sigma^{\text{cent}}_{k+1}$.

We can now give our first theorem, namely that in expectation networked populations will increase their returns at least as fast as central-agent ones.

**Theorem 5.9.** *Let us set $\tau^{comm}_k \in \mathbb{R}_{>0}$. In coordination and anti-coordination games where Assumptions 5.4-5.8 apply, we have $\mathbb{E}[W(\pi^{\text{net}}_{k+1}, I(\pi^{\text{net}}_{k+1}))] \geq \mathbb{E}[W(\pi^{\text{cent}}_{k+1}, I(\pi^{\text{cent}}_{k+1}))]$.*

**Remark 5.10.** Assumption 5.5 presumes that all policies are pairwise distinct after the updates, but does not restrict their returns in the same way. If we additionally make the very weak assumption that at least one of these distinct policies in each $k$ iteration has a base return that is distinct from the others (which is likely to hold in all but the most trivial environments), the inequality in the theorem above will be strict, i.e. *networked learning will always be faster in expectation.*

*Proof.* Recall that before the communication rounds in Line 8 (Alg. 3), the randomly updated policies $\{\pi^i_{k+1}\}^N_{i=1}$ have associated estimated returns $\{\sigma^i_{k+1}\}^N_{i=1}$. Denote the mean and maximum of this set $\sigma^{\text{mean}}_{k+1}$ and $\sigma^{\max}_{k+1}$ respectively. Since $\pi^{\text{cent}}_{k+1}$ is chosen arbitrarily from $\{\pi^i_{k+1}\}^N_{i=1}$, it will obey $\mathbb{E}[\sigma^{\text{cent}}_{k+1}] = \sigma^{\text{mean}}_{k+1}$ $\forall k$, though there will be high variance. Conversely, for the networked case the softmax adoption scheme (Line 11, Alg. 3), which for $\tau^{comm}_k \in \mathbb{R}_{>0}$ gives non-uniform adoption probabilities for distinct $\sigma$ values, means by definition that some communicated policies are more likely to be adopted than others if they have distinct finitely estimated returns (those with higher $\sigma^i_{k+1}$ are more likely to be adopted at each communication round). Thus the consensus $\pi^{\text{net}}_{k+1}$ that gets adopted by the whole networked population will obey $\mathbb{E}[\sigma^{\text{net}}_{k+1}] > \sigma^{\text{mean}}_{k+1}$ if at least one policy receives a distinct return from the others, or $\mathbb{E}[\sigma^{\text{net}}_{k+1}] \geq \sigma^{\text{mean}}_{k+1}$ in the rare circumstance that all policies receive the same return. If $\tau^{comm}_{k+1} \to 0$, it will obey $\mathbb{E}[\sigma^{\text{net}}_{k+1}] = \sigma^{\max}_{k+1}$ $\forall k$. As such:

$$\mathbb{E}[\sigma^{\text{net}}_{k+1}] \geq \mathbb{E}[\sigma^{\text{cent}}_{k+1}]. \tag{1}$$

In Eq. 1 and the remaining equations of the proof, bear in mind that the equality will be strict if at least one policy receives a distinct return from the others.

Refer to the agent whose update originally gave rise to $\pi^{\text{net}}_{k+1}$ and $\sigma^{\text{net}}_{k+1}$ as agent $(i, \text{net})$; we equivalently also have the arbitrary agent $(j, \text{cent})$. Prior to consensus being attained in each case, the joint policy can be written as $\boldsymbol{\pi}^{(i,\text{net};j,\text{cent})} := (\pi^1, \dots, \pi^{i-1}, \pi^{(i,\text{net})}, \pi^{i+1}, \dots, \pi^{j-1}, \pi^{(j,\text{cent})}, \pi^{j+1}, \dots, \pi^N)$.

Given Eq. 1, and by Assumption 5.6 on the quality of finite-step estimations, we know that directly after the policy update in Line 16 (Alg. 2), *prior to the consensus being reached*, we have:

$$\mathbb{E}\left[V^{(i,\text{net})}(\boldsymbol{\pi}^{(i,\text{net};j,\text{cent})}_{k+1}, \mu_t)\right] \geq \mathbb{E}\left[V^{(j,\text{cent})}(\boldsymbol{\pi}^{(i,\text{net};j,\text{cent})}_{k+1}, \mu_t)\right]. \tag{2}$$

We now need to show that this ordering is maintained in the case that each policy is given to the whole population.

By Assumption 5.5 we know that straight after the random policy updates there is no alignment among policies, i.e. in a coordination game we have $f_c^{(i,\text{net})} = f_c^{(j,\text{cent})} = \min f_c$, and in an anti-coordination game we have $f_d^{(i,\text{net})} = f_d^{(j,\text{cent})} = \max f_d$. Therefore if Eq. 2 pertains, by Def. 5.2 it must be because:

$$\mathbb{E}[b(\pi^{(i,\text{net})})] \geq \mathbb{E}[b(\pi^{(j,\text{cent})})], \tag{3}$$

i.e. because the base policy quality is higher for $\pi^{(i,\text{net})}$ than for $\pi^{(j,\text{cent})}$.

By Assumption 5.7 on policy consensus, we know that in the networked and central-agent cases the joint policies respectively become $\boldsymbol{\pi}^{\text{net}} := (\pi^{\text{net}}, \pi^{\text{net}}, \pi^{\text{net}}, \dots)$ and $\boldsymbol{\pi}^{\text{cent}} := (\pi^{\text{cent}}, \pi^{\text{cent}}, \pi^{\text{cent}}, \dots)$. We therefore end up with maximum alignment in both cases, such that $f_c^{\text{net}} = f_c^{\text{cent}} = \max f_c$ in a coordination game, and $f_d^{\text{net}} = f_d^{\text{cent}} = \min f_d$ in an anti-coordination game. Due to this, along with Eqs. 2 and 3, we have

$$\mathbb{E}\left[V^i(\boldsymbol{\pi}^{\text{net}}_{k+1}, \mu_t)\right] \geq \mathbb{E}\left[V^j(\boldsymbol{\pi}^{\text{cent}}_{k+1}, \mu_t)\right]. \tag{4}$$

In turn we have:

$$\mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{net}}_{k+1}, \mu_t)\right] \geq \mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{cent}}_{k+1}, \mu_t)\right], \tag{5}$$

which by Assumption 5.8 gives

$$\mathbb{E}[W(\pi^{\text{net}}_{k+1}, I(\pi^{\text{net}}_{k+1}))] \geq \mathbb{E}[W(\pi^{\text{cent}}_{k+1}, I(\pi^{\text{cent}}_{k+1}))],$$

namely the result. □

We now give results showing that learning is at least as fast in the networked case than in the independent case - empirically we find networked learning always to be strictly faster. We give separate theorems for coordination and anti-coordination games. Since we cannot necessarily expect the independent agents to share a single policy $\pi_{k+1}$ after the update in each iteration of learning, we give these results in terms of the population-average return (Def. 3.3) instead of the single-policy social welfare (Def. 3.5) as before.

Again, we assume for simplicity of the theory that it is only the policy communication scheme that creates a difference in learning between the networked and independent cases, i.e. we assume that networked agents receive the same estimates of the mean field and average reward as independent agents. As mentioned above, our ablation studies suggest this is the dominant factor in our experimental settings anyway, with the estimated mean field not required at all in the broad class of stationary problems. Nevertheless, in practice the networked estimates of the (mean field and) average reward will be better than the independent ones, giving an additional performance increase over the independent case. Thus loosening this assumption is likely to actually enhance the effects identified in the theorems.

**Assumption 5.11.** Assume that the estimated global mean field and average reward in the networked case are the same as the independent case, i.e. $\forall i, j\ \tilde{\hat{\mu}}_t^{(i,net)} = \tilde{\hat{\mu}}_t^{(j,ind)}$ and $\tilde{\hat{r}}_t^{(i,net)} = r_t^i$.

We refer to the joint policy in the networked case after communication round $c$ as $\boldsymbol{\pi}^{\text{net}}_{k+1,c} = \left(\pi^{(1,\text{net})}_{k+1,c}, \dots, \pi^{(N,\text{net})}_{k+1,c}\right)$, and the joint policy in the independent case as $\boldsymbol{\pi}^{\text{ind}}_{k+1} = \left(\pi^{(1,\text{ind})}_{k+1}, \dots, \pi^{(N,\text{ind})}_{k+1}\right)$.

We can now give our second theorem, namely that in expectation networked populations will increase their returns at least as fast as independent ones in coordination games with only a single round of communication in each iteration.

**Theorem 5.12.** *Let us again set $\tau_k^{comm} \in \mathbb{R}_{>0}$. In a coordination game, given Assumptions 5.5, 5.6 and 5.11, for $c = 0$, $\mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{net}}_{k+1,c+1}, \mu_t)\right] \geq \mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{ind}}_{k+1}, \mu_t)\right]$.*

**Remark 5.13.** Assumption 5.5 presumes that all policies are pairwise distinct after the updates, but does not restrict their returns in the same way. If we additionally make the very weak assumption that at least one of the distinct policies in each $k$ iteration has a distinct base return from the others (as is generally likely to be the case), the inequality in the theorem above will be strict, i.e. *networked learning will always be faster in expectation.*

*Proof.* Let us consider two scenarios. Firstly let us imagine that within the communication round, agents swap policies, but no policy drops out of the population, such that if agent $i$ adopts policy $\pi^j_{k+1}$, there exists an agent $i'$ that adopts policy $\pi^i_{k+1}$, and so on. That way we end up with the same policies in the population as before the change, but with each one possibly carried by different arbitrary agents. This is equivalent to if no communication had taken place, meaning that in this scenario $V^{pop}(\boldsymbol{\pi}^{\text{net}}_{k+1,c+1}, \mu_t) = V^{pop}(\boldsymbol{\pi}^{\text{ind}}_{k+1}, \mu_t)$.

Let us now consider an alternative scenario. The softmax adoption scheme (Line 11, Alg. 3), which for $\tau_k^{comm} \in \mathbb{R}_{>0}$ gives non-uniform adoption probabilities for distinct $\sigma$ values, means by definition that some communicated policies are more likely to be adopted than others if they have distinct finitely estimated returns. Thus in expectation the number of distinct policies in the population will decrease if at least one policy has a distinct return from the others (of course, there is a possibility of this still happening even if no policy has a distinct return from the others). Let us start by saying for simplicity that during the first communication round a single $\pi_{k+1,c}^{(j,\text{net})}$ is replaced by $\pi_{k+1,c}^{(i,\text{net})}$, such that for $c = 0$

$$\boldsymbol{\pi}_{k+1,c}^{\text{net}} = \left(\pi_{k+1,c}^{(1,\text{net})}, \ldots, \pi_{k+1,c}^{(\mathbf{i},\text{net})}, \ldots, \pi_{k+1,c}^{(\mathbf{j},\text{net})}, \ldots \pi_{k+1,c}^{(N,\text{net})}\right),$$
$$\text{and} \quad \boldsymbol{\pi}_{k+1,c+1}^{\text{net}} = \left(\pi_{k+1,c+1}^{(1,\text{net})}, \ldots, \pi_{k+1,c+1}^{(\mathbf{i},\text{net})}, \ldots, \pi_{k+1,c+1}^{(\mathbf{i},\text{net})}, \ldots \pi_{k+1,c+1}^{(N,\text{net})}\right).$$

For this to have occurred, we know that

$$\mathbb{E}[\sigma_{k+1,c}^{(i,\text{net})}] > \mathbb{E}[\sigma_{k+1,c}^{(j,\text{net})}],$$

and therefore by Assumption 5.6 that

$$\mathbb{E}\left[V^{(i,\text{net})}(\boldsymbol{\pi}_{k+1,c}^{\text{net}}, \mu_t)\right] > \mathbb{E}\left[V^{(j,\text{net})}(\boldsymbol{\pi}_{k+1,c}^{\text{net}}, \mu_t)\right]. \tag{6}$$

By Assumption 5.5 we know that straight after the random policy updates there is no alignment among policies, i.e. in a coordination game we have $f_c^{(i,\text{net})} = f_c^{(j,\text{net})} = \min f_c$. Therefore if Eq. 9 pertains, by Def. 5.2 it must be because:

$$\mathbb{E}[b(\pi^{(i,\text{net})})] > \mathbb{E}[b(\pi^{(j,\text{net})})], \tag{7}$$

i.e. because the base policy quality is higher for $\pi^{(i,\text{net})}$ than for $\pi^{(j,\text{net})}$. For this reason we have, for $c = 0$:

$$\mathbb{E}\left[V^{pop}(\boldsymbol{\pi}_{k+1,c+1}^{\text{net}}, \mu_t)\right] > \mathbb{E}\left[V^{pop}(\boldsymbol{\pi}_{k+1,c}^{\text{net}}, \mu_t)\right]. \tag{8}$$

Additionally, replacing $\pi_{k+1,c}^{(j,\text{net})}$ with a second copy of $\pi_{k+1,c}^{(i,\text{net})}$ will increase the alignment ($f_c$) of $\pi_{k+1,c}^{(i,\text{net})}$ such that $\mathbb{E}\left[V^{(i,\text{net})}(\boldsymbol{\pi}_{k+1,c+1}^{\text{net}}, \mu_t)\right] > \mathbb{E}\left[V^{(i,\text{net})}(\boldsymbol{\pi}_{k+1,c}^{\text{net}}, \mu_t)\right]$, increasing the improvement even further. This effect is even greater if more than one policy is replaced.

Since the independent case is equivalent to the networked case when $C_p = 0$, we can say that $\boldsymbol{\pi}_{k+1}^{\text{ind}} = \boldsymbol{\pi}_{k+1,0}^{\text{net}}$. This gives the result, i.e.

$$\mathbb{E}\left[V^{pop}(\boldsymbol{\pi}_{k+1,c+1}^{\text{net}}, \mu_t)\right] \geq \mathbb{E}\left[V^{pop}(\boldsymbol{\pi}_{k+1}^{\text{ind}}, \mu_t)\right],$$

where this inequality is only not strict if the first scenario always applies rather than the second. □

To prove the benefit of the networked case over the independent case in anti-coordination games, we use a final additional assumption. This presumes that the base return is not yet fully maximised, and that the benefit to an agent's overall return of increasing its base return by adopting a neighbour's better-performing policy outweighs the resulting decrease in diversity. This establishes the conditions under which our policy adoption scheme is able to advantage networked agents over those whose policies are always independent. This assumption applies in most non-trivial scenarios (at least at the beginning of training), namely where the goal of the task is not simply for agents to have distinct policies that are otherwise inconsequential, and thus where the benefit of diverse behaviour can only be fully felt once agents have a certain level of aptitude at accomplishing the given task. For example, in all of the anti-coordination games in our experiments, agents are always penalised for moving, and only start to receive higher rewards if they are stationary. Therefore in these anti-coordination games agents will receive higher returns by *aligning* on policies that prioritise stationarity, than by maintaining diverse policies that have high levels of movement. Of course once base return is maximised and the assumption no longer holds, one can consider terminating policy communication and adoption to avoid decreases in diversity (one may also be ready to stop training entirely at this point, as the population is likely to be reaching the optimal average return). Please see Sec. 6.3 for further discussion of the applicability of this assumption in practice.

**Assumption 5.14.** Assume that an increase in the base return function outweighs a decrease in the policy diversity, namely $h(b + \Delta b, f_d - \Delta f_d) > h(b, f_d),\ \forall \Delta b > 0, \Delta f_d > 0$, and that the agents have not yet maximised their base return function i.e. $b(\pi^i_{k+1}) < \sup_{\pi \in \Pi} b(\pi)\ \ \forall i \in \{1, \dots, N\}$.

We now give our final theorem, namely that in anti-coordination games, in expectation networked populations will increase their returns at least as fast as independent ones with only a single round of communication in each iteration.

**Theorem 5.15.** *Let us once again set $\tau^{comm}_k \in \mathbb{R}_{>0}$. In an anti-coordination game, given Assumptions 5.5, 5.6, 5.11and 5.14, for $c = 0$, $\mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{net}}_{k+1,c+1}, \mu_t)\right] \geq \mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{ind}}_{k+1}, \mu_t)\right]$.*

*Proof.* The proof begins similarly to that for a coordination game. Let us consider two scenarios. Firstly let us imagine that within the communication round, agents swap policies, but no policy drops out of the population, such that if agent $i$ adopts policy $\pi^j_{k+1}$, there exists an agent $i'$ that adopts policy $\pi^i_{k+1}$, and so on. That way we end up with the same policies in the population as before the change, but with each one possibly carried by different arbitrary agents. This is equivalent to if no communication had taken place, meaning that in this scenario $V^{pop}(\boldsymbol{\pi}^{\text{net}}_{k+1,c+1}, \mu_t) = V^{pop}(\boldsymbol{\pi}^{\text{ind}}_{k+1}, \mu_t)$.

Let us now consider an alternative scenario. The softmax adoption scheme (Line 11, Alg. 3), which for $\tau^{comm}_k \in \mathbb{R}_{>0}$ gives non-uniform adoption probabilities for distinct $\sigma$ values, means by definition that some communicated policies are more likely to be adopted than others if they have distinct finitely estimated returns. Thus in expectation the number of distinct policies in the population will decrease if at least one policy has a distinct return from the others. Say for simplicity that during the first communication round a $\pi^{(j,\text{net})}_{k+1,c}$ is replaced by $\pi^{(i,\text{net})}_{k+1,c}$, such that for $c = 0$

$$\boldsymbol{\pi}^{\text{net}}_{k+1,c} = \left(\pi^{(1,\text{net})}_{k+1,c}, \dots, \pi^{(\mathbf{i},\text{net})}_{k+1,c}, \dots, \pi^{(\mathbf{j},\text{net})}_{k+1,c}, \dots \pi^{(N,\text{net})}_{k+1,c}\right),$$
$$\text{and} \quad \boldsymbol{\pi}^{\text{net}}_{k+1,c+1} = \left(\pi^{(1,\text{net})}_{k+1,c+1}, \dots, \pi^{(\mathbf{i},\text{net})}_{k+1,c+1}, \dots, \pi^{(\mathbf{i},\text{net})}_{k+1,c+1}, \dots \pi^{(N,\text{net})}_{k+1,c+1}\right).$$

For this to have occurred, we know that

$$\mathbb{E}[\sigma^{(i,\text{net})}_{k+1,c}] > \mathbb{E}[\sigma^{(j,\text{net})}_{k+1,c}],$$

and therefore by Assumption 5.6 that

$$\mathbb{E}\left[V^{(i,\text{net})}(\boldsymbol{\pi}^{\text{net}}_{k+1,c}, \mu_t)\right] > \mathbb{E}\left[V^{(j,\text{net})}(\boldsymbol{\pi}^{\text{net}}_{k+1,c}, \mu_t)\right]. \tag{9}$$

By Assumption 5.5 we know that straight after the random policy updates there is no alignment among policies, i.e. in the anti-coordination game we have $f^{(i,\text{net})}_d = f^{(j,\text{net})}_d = \max f_d$, while by Assumption 5.14 we know that the agents have not yet maximised their base return function. Therefore if Eq. 9 pertains, by Def. 5.2 it must be because:

$$\mathbb{E}[b(\pi^{(i,\text{net})})] > \mathbb{E}[b(\pi^{(j,\text{net})})], \tag{10}$$

i.e. because the base policy quality is higher for $\pi^{(i,\text{net})}$ than for $\pi^{(j,\text{net})}$.

Assumption 5.14 assumes that any increase in the base quality of the policy will outweigh the decrease in diversity that will come from having more than one agent following $\pi^{(i,\text{net})}_{k+1,c+1}$. Therefore we have, for $c = 0$:

$$\mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{net}}_{k+1,c+1}, \mu_t)\right] > \mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{net}}_{k+1,c}, \mu_t)\right].$$

These steps apply similarly if more than one policy is replaced.

Since the independent case is equivalent to the networked case when $C_p = 0$, we can say that $\boldsymbol{\pi}^{\text{ind}}_{k+1} = \boldsymbol{\pi}^{\text{net}}_{k+1,0}$. This gives the result, i.e.

$$\mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{net}}_{k+1,c+1}, \mu_t)\right] \geq \mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{ind}}_{k+1}, \mu_t)\right],$$

where this inequality is only not strict if the first scenario always applies rather than the second. □

# 6 Experiments

## 6.1 Experimental setup

We present experiments from grid worlds, following the gold standard in similar works on MFGs and MFC (Laurière et al., 2022a). We give results from six tasks similar to those found in prior works, defined by the agents' reward/transition functions and relating to agents' positions relative to other agents. Two are coordination games and four are anti-coordination games, where in each case the reward function reflects a coordination/anti-coordination ($f_c/f_d$) element alongside other elements that may be crucial for receiving reward, reflected in the policies' base quality $b(\pi)$ (Sec. 5). In all cases, rewards are normalised in [0,1] after they are computed.

The two coordination games are:

- **Cluster.** This game is also used in Benjamin & Abate (2023; 2024). Agents are encouraged to gather together by the reward function $R(s_t^i, a_t^i, \hat{\mu}_t) = \log(\hat{\mu}_t(s_t^i))$. That is, agent $i$ receives a reward that is logarithmically proportional to the fraction of the population that is co-located with it at time $t$. We give the population no indication where they should cluster, agreeing this themselves over time.
- **Target selection.** This game is also used in Benjamin & Abate (2023; 2024). Unlike in the above 'cluster' game, the agents are given options of locations at which to gather, and they must reach consensus among themselves. If the agents are co-located with one of a number of specified targets $\phi \in \Phi$ (in our experiments we place one target in each of the four corners of the grid), and other agents are also at that target, they get a reward proportional to the fraction of the population found there; otherwise they receive a penalty of -1. In other words, the agents must coordinate on which of a number of mutually beneficial points will be their single gathering place. Define the magnitude of the distances between $x, y$ at $t$ as $dist_t(x, y)$. The reward function is given by $R(s_t^i, a_t^i, \hat{\mu}_t) = r_{targ}(r_{coord}(\hat{\mu}_t(s_t^i)))$, where

$$r_{targ}(x) = \begin{cases} x & \text{if } \exists \phi \in \Phi \text{ s.t. } dist_t(s_t^i, \phi) = 0 \\ -1 & \text{otherwise,} \end{cases}$$

$$r_{coord}(x) = \begin{cases} x & \text{if } \hat{\mu}_t(s_t^i) > 1/N \\ -1 & \text{otherwise.} \end{cases}$$

The anti-coordination games are:

- **Disperse.** This game is also used in Benjamin & Abate (2024) and is similar to the 'exploration' tasks in Laurière et al. (2022b); Wu et al. (2024) and other MFG works. In our version agents are rewarded for being located in more sparsely populated areas but only if they are stationary, to avoid trivial random policies. The reward function is given by $R(s_t^i, a_t^i, \hat{\mu}_t) = r_{stationary}(-\log(\hat{\mu}_t(s_t^i)))$, where

$$r_{stationary}(x) = \begin{cases} x & \text{if } a_t^i \text{ is 'remain stationary'} \\ -1 & \text{otherwise.} \end{cases}$$

- **Target coverage.** The population is rewarded for spreading across a certain number of targets, as long as agents are stationary at the target. As in the 'target selection' game, we have targets $\phi \in \Phi$, where in our experiments we place one target in each of the four corners of the grid. Again define the magnitude of the distances between $x, y$ at $t$ as $dist_t(x, y)$. The reward function is given by

$$R(s_t^i, a_t^i, \hat{\mu}_t) = r_{stationary}\left(r_{targ}\left(-\log(\hat{\mu}_t(s_t^i))\right)\right),$$

  where $r_{stationary}$ and $r_{targ}$ are as defined above.

- **Beach bar.** Such games are very common in MFG works (Perrin et al., 2020; Laurière et al., 2022a; Cui et al., 2023a; Wu et al., 2024). In our version agents are rewarded for being stationary in sparsely populated locations as close as possible to a target $\phi_b$, located in the centre of the grid. The maximum possible distance from the target is denoted *maxDist*. The reward is given by

$$R(s_t^i, a_t^i, \hat{\mu}_t) = r_{stationary}\left(maxDist - dist_t(s_t^i, \phi_b) - \log(\hat{\mu}_t(s_t^i))\right),$$

  where $r_{stationary}$ is as defined above.

- **Shape formation.** The population is rewarded for spreading around a ring shape, accomplished by encouraging agents to be a distance of 3 (chosen arbitrarily to fit the grid) from a centre point $\phi_c$. The reward is given by

$$R(s_t^i, a_t^i, \hat{\mu}_t) = r_{stationary}\left(r_{ring}\left(-\log(\hat{\mu}_t(s_t^i))\right)\right),$$

  where $r_{stationary}$ is as defined above, and

$$r_{ring}(x) = \begin{cases} x & \text{if } dist_t(s_t^i, \phi_c) = 3 \\ -1 & \text{otherwise.} \end{cases}$$

In these spatial environments, we define both the communication network $\mathcal{G}_t^{comm}$ and the visibility graph $\mathcal{G}_t^{vis}$ by the physical distance from $i$. We show plots for various transmission radii, given as fractions of the maximum distance in the grid. Note that the networked population with the largest radius is always fully connected, and therefore these agents are always able to accurately estimate $\hat{r}$ and $\hat{\mu}_t$ even for $C_r = 1$ and $C_e = 0$. That is, when we set $C_r = C_e > 0$ their observations are equivalent to those that the central-agent population would receive, albeit that policies are updated and spread differently.

We evaluate our experiments according to a finite-step estimation of the population-average discounted return (Def. 3.3) over $M$ steps within each outer $k$ loop, i.e. $\hat{V}^{pop}(\boldsymbol{\pi}_k, \mu_t; M)$. Experiments were conducted on a Linux-based machine with 2 x Intel Xeon Gold 6248 CPUs (40 physical cores, 80 threads total, 55 MiB L3 cache). We use the JAX framework to accelerate and vectorise our code. We run five trials with different random seeds for each experiment, and plot the mean and standard deviation of the mean across the seeds. Random seeds are set in our code in a fixed way dependent on the trial number to allow easy replication of experiments. Our code is included in the publicly available supplementary material for reproducibility.

### 6.2 Hyperparameters

See Table 1 for our hyperparameter choices. We can group our hyperparameters into those controlling the size of the experiment, those controlling the size of the Q-network, those controlling the number of iterations of each loop in the algorithms and those affecting the learning/policy updates or policy adoption.

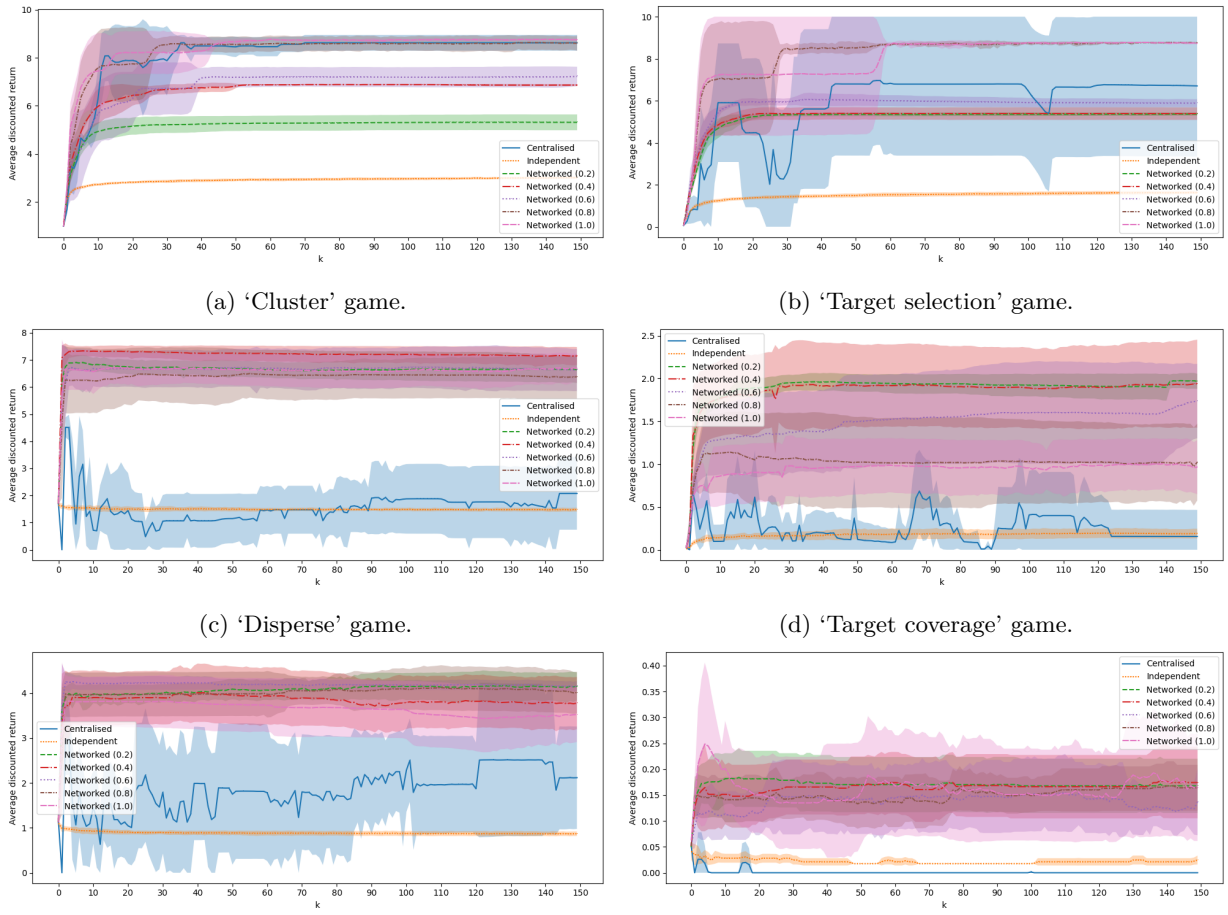
In our experiments we generally want to demonstrate that our communication-based algorithm learns faster than the central-agent and independent architectures, even when the Q-function / mean field / average reward are poorly estimated as is likely to be the case in complex real-world scenarios. There is a similar motivation in the related works on networked communication in the MFG setting by Benjamin & Abate (2023; 2024). Moreover we want to show that there is a benefit even to a small amount of communication, so that communication rounds themselves do not excessively add to time complexity. As such, we generally select hyperparameters at the lowest end of those we tested during development, to show that our algorithms are particularly successful and robust given what might otherwise be considered 'undesirable' hyperparameter choices.

### 6.3 Results and discussion

Fig. 2 gives results for our standard experimental settings involving 500 agents, each with their own Q-network. When networked agents communicate, they have only a *single* communication round. Fig. 2 shows that in all of our games, networked populations of all broadcast radii significantly outperform independent

Table 1: Hyperparameters

| Hyperparam. | Value | Comment |
|---|---|---|
| Trials | 5 | We run 5 trials with different random seeds for each experiment. We plot the mean and standard deviation of the mean for each metric across the seeds. |
| Gridsize | 20x20 | - |
| Population | 500 | We chose 500 for our demonstrations to show that our algorithm can handle large populations, indeed often larger than those demonstrated in other mean-field works, especially for grid-world environments, while also being feasible to simulate with respect to time and computation constraints (Yang et al., 2018; Subramanian & Mahajan, 2019; Ganapathi Subramanian et al., 2020; 2021; Cui & Koeppl, 2021; Yongacoglu et al., 2024; Subramanian et al., 2022; Cui et al., 2023a; Guo et al., 2023; Benjamin & Abate, 2023; 2024; Wu et al., 2024). For example, the MFC work in Carmona et al. (2019) uses 10 agents; the work on decentralised execution for MFC by Cui et al. (2023c) uses 200 agents. |
| Number of neurons in input layer | 440 | The agent's position is represented by two concatenated one-hot vectors, indicating the agent's row and column. The mean-field distribution is a flattened vector of the same size as the grid. As such, the input size is $[(2 \times \text{dimension}) + (\text{dimension}^2)]$. |
| Neurons per hidden layer | 256 | We draw inspiration from common rules of thumb when selecting the number of neurons in hidden layers, e.g. it should be between the number of input neurons and output neurons / it should be 2/3 the size of the input layer plus the size of the output layer / it should be a power of 2 for computational efficiency. Using these rules of thumb as rough heuristics, we select the number of neurons per hidden layer by rounding the size of the input layer down to the nearest power of 2. The layers are all fully connected. |
| Hidden layers | 2 | We achieved sufficient learning speed with just 2 hidden layers, but further optimising the number of layers may lead to better results. |
| Activation function | ReLU | This is a common choice in deep RL. |
| $K$ | 150 | $K$ is chosen to be large enough to see convergence in most networked cases. |
| $M$ | 20 | We tested $M$ in {20,50,100} and found that the lowest value was sufficient to achieve convergence while minimising training time. It may be possible to converge with even smaller choices of $M$. |
| $L$ | 20 | We tested $L$ in {20,50,100} and found that the lowest value was sufficient to achieve convergence while minimising training time. It may be possible to converge with even smaller choices of $L$. |
| $E$ | 20 | We tested $E$ in {20,50,100}, and choose the lowest value to show the benefit to convergence even from very few evaluation steps. It may be possible to reduce this value further and still achieve similar results. |
| $C_p$ | 1 (10/50) | As in Benjamin & Abate (2023; 2024), we choose a value of 1 for most experiments to show the convergence benefits brought by even a single communication round, even in networks that may have limited connectivity. We also conduct additional studies to show the effect of additional rounds in Figs. 4 and 5. |
| $C_r$ | 1 (10/50) | Similar to $C_p$, we choose this value to show our algorithm's ability to appropriately estimate the average reward even with only a single round, even in networks that may have limited connectivity. We conduct additional studies to show the effect of additional rounds in Figs. 4 and 5. |
| $C_e$ | 1 (10/50) | Similar to $C_p$, we choose this value to show the ability of our algorithm to appropriately estimate the mean field even with only a single communication round, even in networks that may have limited connectivity. We also conduct additional studies to show the effect of additional rounds in Figs. 4 and 5. |
| $\gamma$ | 0.9 | Standard choice across RL literature. |
| $\tau_q$ | 0.03 | We follow Vieillard et al. (2020) and Benjamin & Abate (2024), which tested a range of values. |
| $\|B\|$ | 32 | This is a common choice of batch size that trades off noisy updates and computational efficiency. |
| $cl$ | -1 | We use the same value as in Vieillard et al. (2020) and Benjamin & Abate (2024). |
| $\nu$ | $L-1$ | We follow Benjamin & Abate (2024), which is similar to Laurière et al. (2022b). |
| Optimiser | Adam | As in Vieillard et al. (2020), we use the Adam optimiser with initial learning rate 0.01. |
| $\tau_k^{comm}$ | cf. comment | We follow Benjamin & Abate (2024), where $\tau_k^{comm}$ increases linearly from 0.001 to 1 across the $K$ iterations. Further optimising this inverse annealing process may lead to better results; we provide an ablation study in Fig. 10. |




(a) 'Cluster' game.

(b) 'Target selection' game.

(c) 'Disperse' game.

(d) 'Target coverage' game.

(e) 'Beach bar' game.

(f) 'Shape formation' game.

Figure 2: Standard settings with $C_e = C_r = C_p = 1$. In all games networked agents of all broadcast radii significantly outperform the independent (orange) populations, and in most games they also outperform the central-agent (blue) populations, reflecting our theoretical results. The central-agent populations also have markedly higher variance than networked ones in several games, since the central learner pushes an arbitrary updated policy to the whole population regardless of its quality, leading to large fluctuations in performance, whereas our communication scheme biases networked populations towards better performing updates.

(orange) agents, which hardly appear to increase their returns, if at all. Networked populations of all broadcast radii also significantly outperform the central-agent (blue) populations in all but the two coordination games, where only networked agents of the smaller radii (green, 0.2; red, 0.4; purple, 0.6) underperform them (due to these less connected populations being more likely to experience violations of Assumption 5.7 on policy consensus, which is a disadvantage in scenarios where alignment is beneficial). Indeed, in the anti-coordination games the central-agent populations perform similarly to purely independent ones in hardly appearing to increase their returns, performing even worse than independent agents in the 'shape formation' game. The central-agent populations also have markedly higher variance than networked ones in several games ('target selection', 'disperse', 'beach bar'). This reflects our theoretical analysis in Sec. 5 that the central learner pushes an arbitrary updated policy to the whole population regardless of its quality, leading to large fluctuations in performance, whereas our communication scheme biases networked populations towards better performing updates.

In the 'target coverage' game, and sometimes the other anti-coordination games to a lesser extent, networked agents of smaller broadcast radii appear to outperform those of larger radii, i.e. the ordering is reversed from that of the coordination games, albeit not necessarily significantly so. This reflects the point up to which

our Assumption 5.14 (increase in base return outweighs decrease in diversity in anti-coordination games, and base return is not yet maximised), holds in practice, which we discuss in the following.

The first part of Assumption 5.14 strictly holds throughout the 'disperse', 'target coverage' and 'shape formation' anti-coordination games: agents get no reward for diversity unless they are stationary (and also unless they are in one of the correct locations in the latter two cases). This means that any increase in base return (likelihood of being stationary or in the right location) achieved by policy adoption does indeed outweigh the loss of diversity. The first part of Assumption 5.14 mostly holds in the 'beach bar' game, apart from in a small window for agents that are stationary close to the bar target, with the window defined by the size of the population and hence the potential magnitude of the $\log(\hat{\mu}_t(s_t^i)$ term in the reward function. Inside this window, increasing base return by moving even closer to the target, at the cost of being in a more crowded area, would not necessarily be beneficial. Regardless, in all these games the networked populations of all broadcast radii significantly outperform the independent agents, which do not appear to be able to learn at all without the helpful bias towards policies with better base returns enabled by the communication scheme.

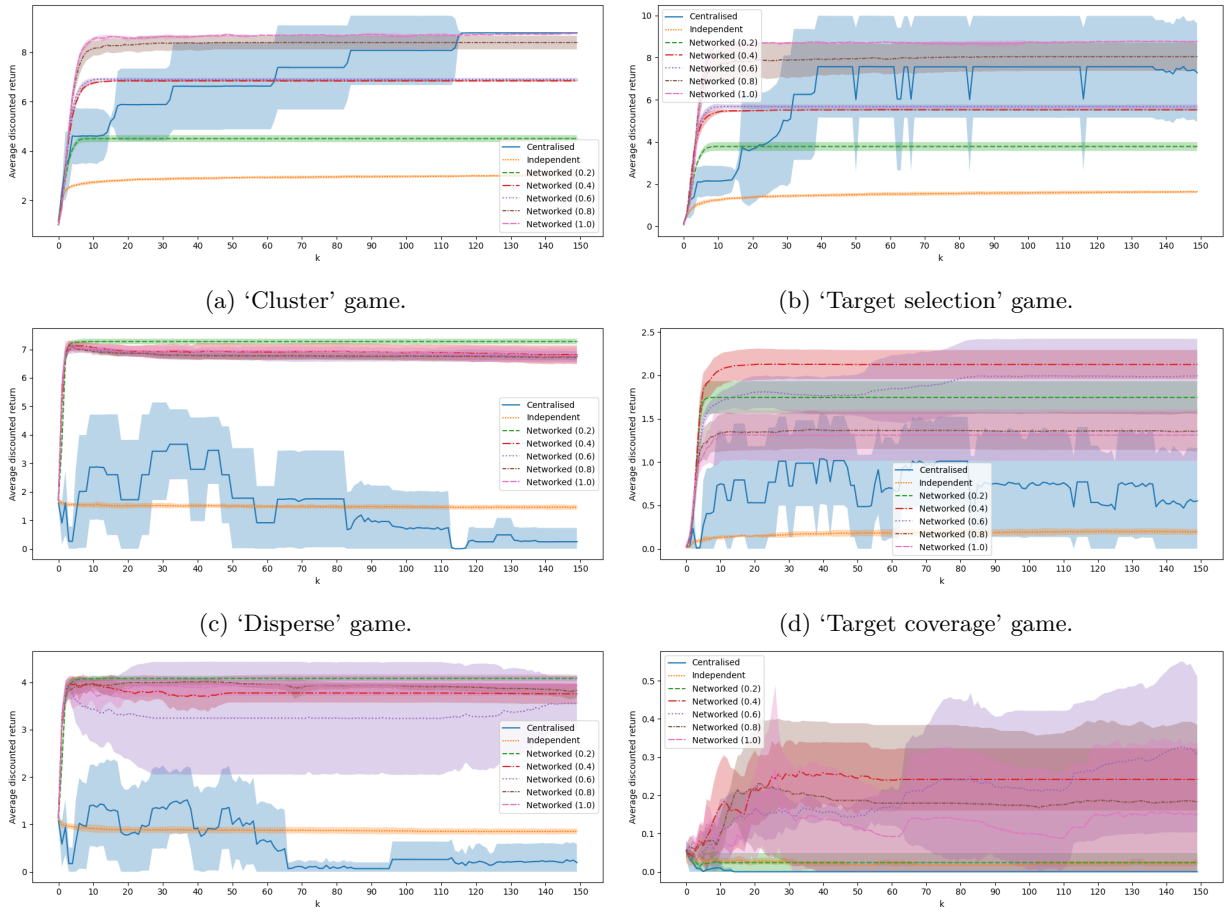
However, among these networked populations, the base return quickly reaches its capacity, i.e. agents learn to be primarily stationary in one of the right locations, such that the second part of Assumption 5.14 no longer holds. This is not an issue when comparing with the independent populations, which have not maximised their base returns and therefore perform worse, but it does give rise to the reverse ordering of returns which we see among networked populations of different radii and hence connectivities. Once base return is maximised, policies that are estimated to receive higher returns may be less aligned with other policies than those other policies are with each other (at least regarding the strategically relevant parts of policies which are rewarded for greater diversity, e.g. these policies visit the less congested locations), or they simply visited the less congested locations by chance during the finite evaluation steps. Either way, more adoption of policies now becomes a disadvantage, since it reduces diversity without an additional positive impact on base return. Therefore architectures that give less communication now perform better by preserving diversity. Populations with lower broadcast radii usually have less connected networks, especially if sub-populations become isolated from each other, which is more likely in our 'target coverage' game than the others since the target locations are as far apart as possible from each other. Therefore these populations have less communication than those with larger broadcast radii and so may perform better, even while all networked populations outperform the independent agents that have not maximised their base returns.

This intuition also helps to understand why networked populations outperform central-agent populations in these anti-coordination games, especially when policy consensus is not enforced for the networked populations. The ultimate choice of consensus level might depend on whether one is using the empirical population as a practical way of learning the social optimum for a MFC problem (Def. 3.6), where a single policy $\pi^*$ is desired to be given to an infinite population, or whether one is solving the MFC problem to approximate the solution to a finite-agent control problem (Def. 3.3) involving the same number of agents as the empirical population from which one is learning. In the latter case some policy diversity may be accepted/desired if it affords a better approximation to the $N$-agent solution.

We provide numerous additional experiments and ablation studies. We list these below, but please find the full discussion of results in the caption for each figure. Of particular note, the ablation studies of Algs. 1 (estimating global average reward) and 4 (estimating global empirical mean field) suggest that in our experimental settings the policy communication scheme (Alg. 3) is the dominant factor in the better performance of networked populations over the other architectures.

- Robustness to communication failure - Fig. 3.
- Increased communication rounds - Figs. 4 and 5.
- Ablation study with population-independent policies - Fig. 6.
- Ablation study of Alg. 4 for estimating the empirical mean field - Fig. 7.
- Ablation study for observation of true/estimated average reward (agents only see their individual reward) - Fig. 8.

- Ablation study for Alg. 1 for estimating the true global average reward (all agents receive true global average reward) - Fig. 9.
- Ablation study of the choice of $\tau_k^{comm}$ - Fig. 10.




(a) 'Cluster' game.

(b) 'Target selection' game.

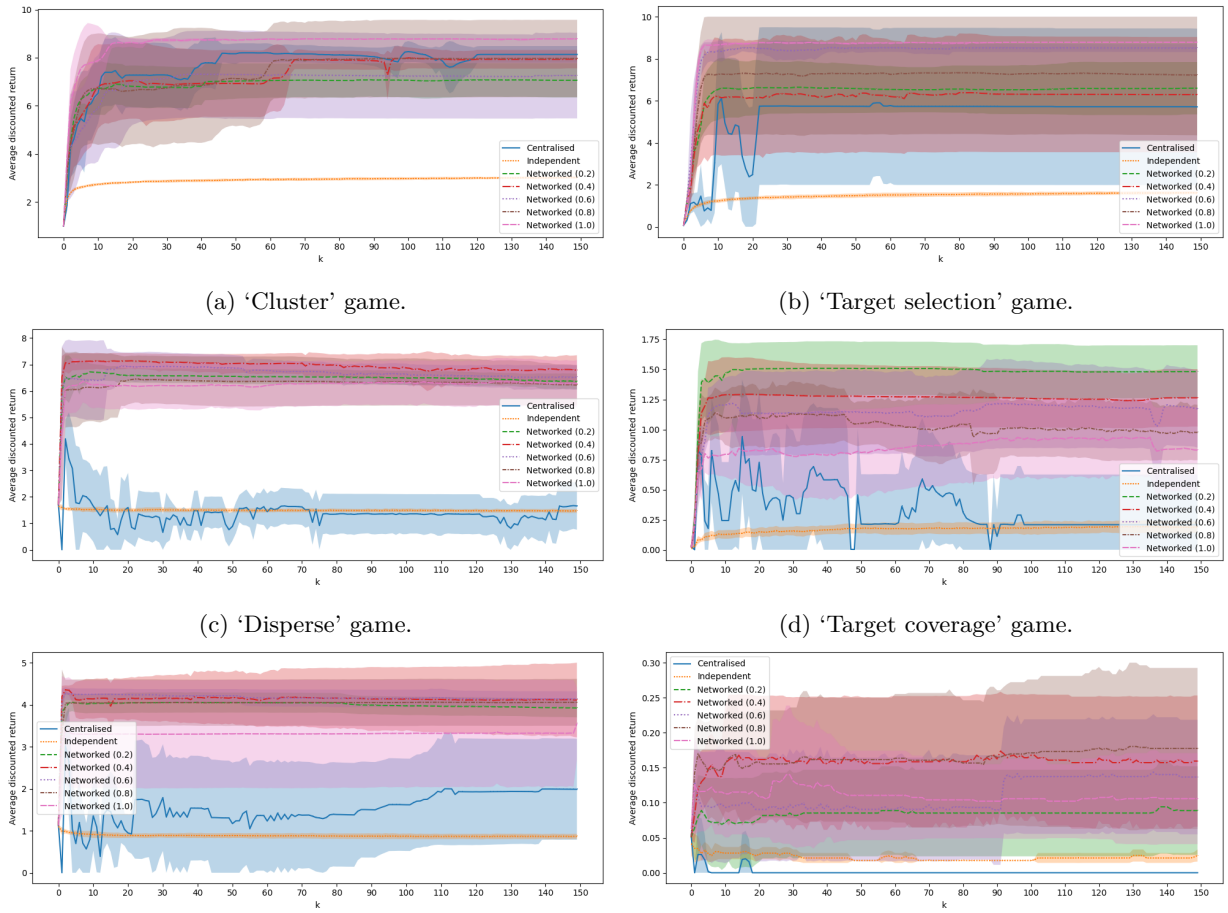
(c) 'Disperse' game.

(d) 'Target coverage' game.

(e) 'Beach bar' game.

(f) 'Shape formation' game.

Figure 3: All communication links suffer a 90% probability of failure, including in the central-agent case, where the link between the central learner and the rest of the population may fail. $C_e = C_r = C_p = 1$. The central-agent population, which in the standard setting matched networked performance only in the 'cluster' game, now learns slower even in this game, due to suffering from the single point of failure. Our networked scheme appears robust to the failures in all games, with only small differences compared to performance in the standard setting. In fact, several broadcast radii appear to perform better in the 'shape formation' game with these failures than without (though not significantly so), probably because they permit greater diversity policies while still having an advantage over purely independent learners (as discussed in the body of Sec. 6.3). However, the smallest broadcast radius (green, 0.2) does drop in performance in this game, which might be expected given it now acts similarly to the independent case. Networked populations appear to have less variance in this setting than in the standard setting, at least in the first four games. This is possibly because the communication failures prevent both particularly high and particularly low performing policies from spreading fast in the population, preventing large performance fluctuations and smoothing learning progress. Meanwhile a central-agent population still has large variance even with communication failures, due to enforcing the adoption of an arbitrarily-chosen consensus policy - in some games variance is higher in this setting (though in some it may be marginally lower). This points to an additional benefit of our networked scheme over the central-agent case.




(a) 'Cluster' game.

(b) 'Target selection' game.

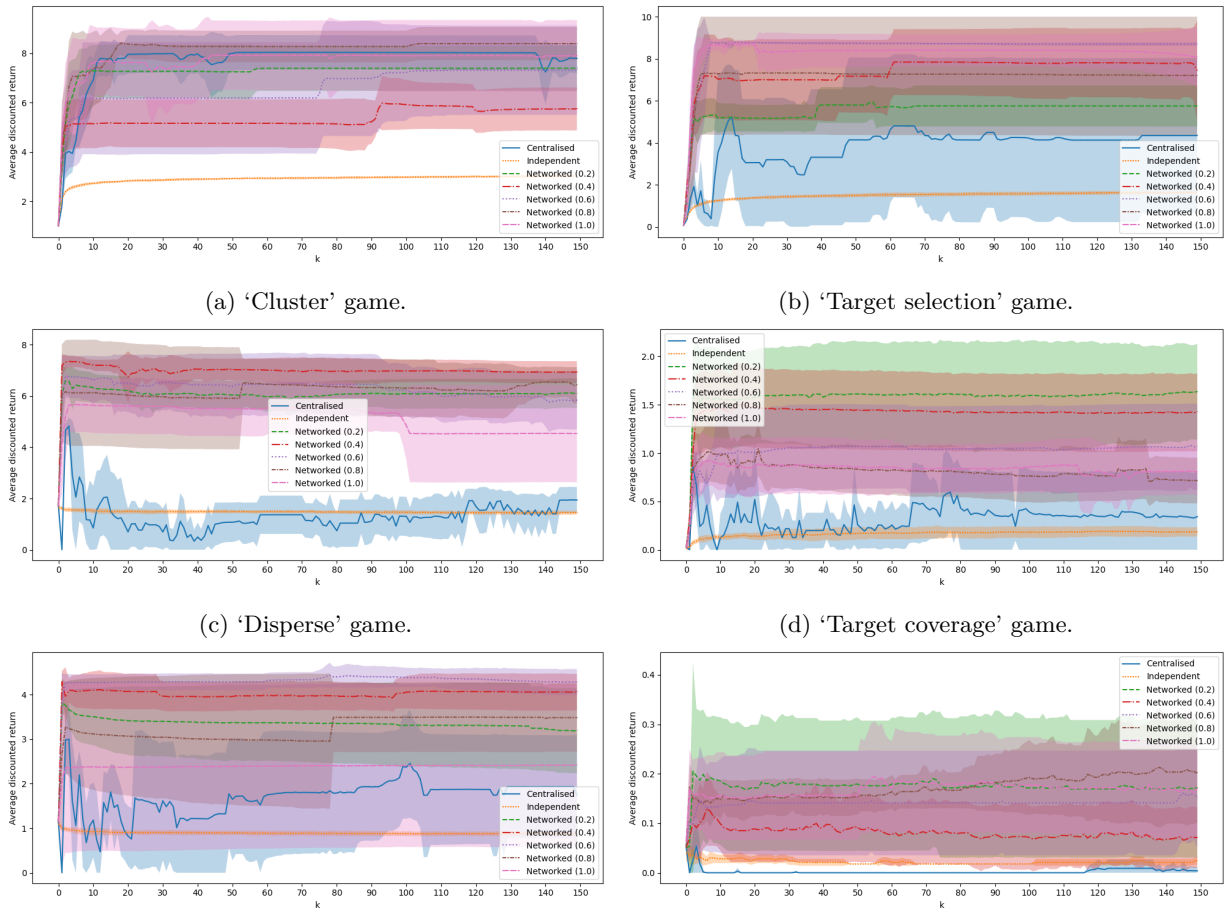
(c) 'Disperse' game.

(d) 'Target coverage' game.

(e) 'Beach bar' game.

(f) 'Shape formation' game.

Figure 4: Standard algorithms but $C_e = C_r = C_p = \mathbf{10}$. As is expected, in the coordination games the networked agents with lower broadcast radii now receive returns almost as high as those with larger radii, albeit at the cost of greater variance (having more communication rounds leads to greater policy consensus in the population at each iteration of the outer loop, and there may be some noise in the quality of these consensus policies). In the 'target selection' game, now all networked populations appear to outperform the central-agent (orange) population, though again with high variance. In the anti-coordination 'target coverage' game, the smaller broadcast radii (green, 0.2; red, 0.4; purple, 0.6) receive slightly lower returns than before, since the additional communication rounds now make policy alignment more likely, reducing $f_d$ as per Def 5.3. The same is true of the smallest radius population (green, 0.2) in the 'shape formation' game, which receives a lower return than before. This reflects the discussion in Sec. 6.3 regarding the detrimental effect of additional policy adoption once the maximum base return has been achieved in anti-coordination games. Nevertheless, *all* networked populations receive higher returns than the independent agents in all games, and also than the central-agent population in all but the 'cluster' game. This shows that in our experimental settings there is a very large benefit to a single communication round, with limited benefit to increasing the algorithms' time complexity with additional communication rounds.




(a) 'Cluster' game.

(b) 'Target selection' game.

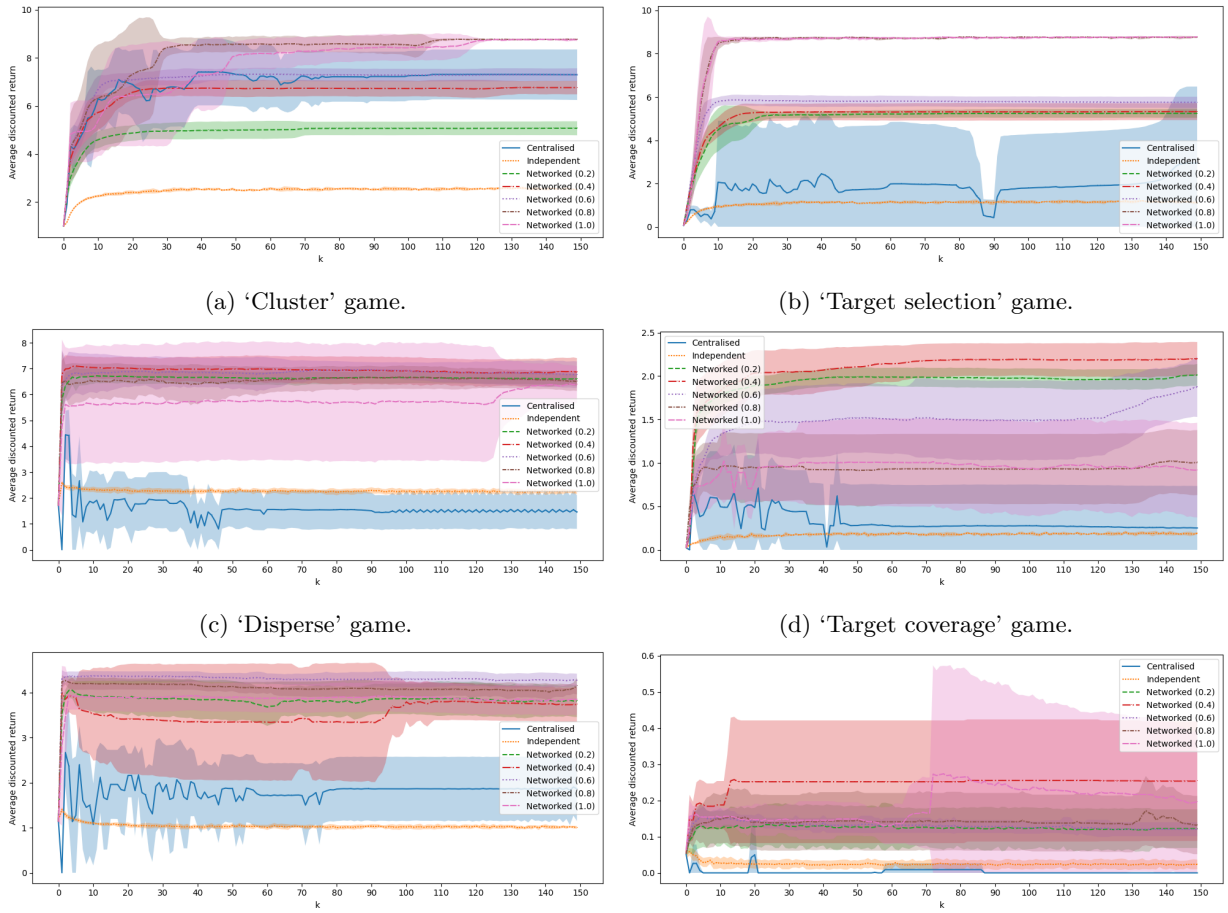
(c) 'Disperse' game.

(d) 'Target coverage' game.

(e) 'Beach bar' game.

(f) 'Shape formation' game.

Figure 5: Standard algorithms but $C_e = C_r = C_p = \mathbf{50}$. Having 50 communication rounds does not appear to significantly change networked performance compared to 10 rounds (Fig. 4), with most increases or decreases in average return appearing within the margin of error. Most notably, the largest broadcast radius (pink, 1.0) receives slightly lower return now than with 10 rounds in the 'disperse' game, while pink (1.0), brown (0.8) and green (0.2) receive lower returns and have higher variance now in the 'beach bar' game. As in the case of $C_e = C_r = C_p = 10$, additional communication rounds make policy alignment more likely, reducing $f_d$ as per Def 5.3. This reflects the discussion in Sec. 6.3 regarding the detrimental effect of additional policy adoption once the maximum base return has been achieved in anti-coordination games. Nevertheless, *all* networked populations receive higher returns than the independent agents in all games, and also than the central-agent population in all but the 'cluster' game. This shows that in our experimental settings there is a very large benefit to a single communication round, with limited benefit to increasing the algorithms' time complexity with additional communication rounds.




(a) 'Cluster' game.

(b) 'Target selection' game.

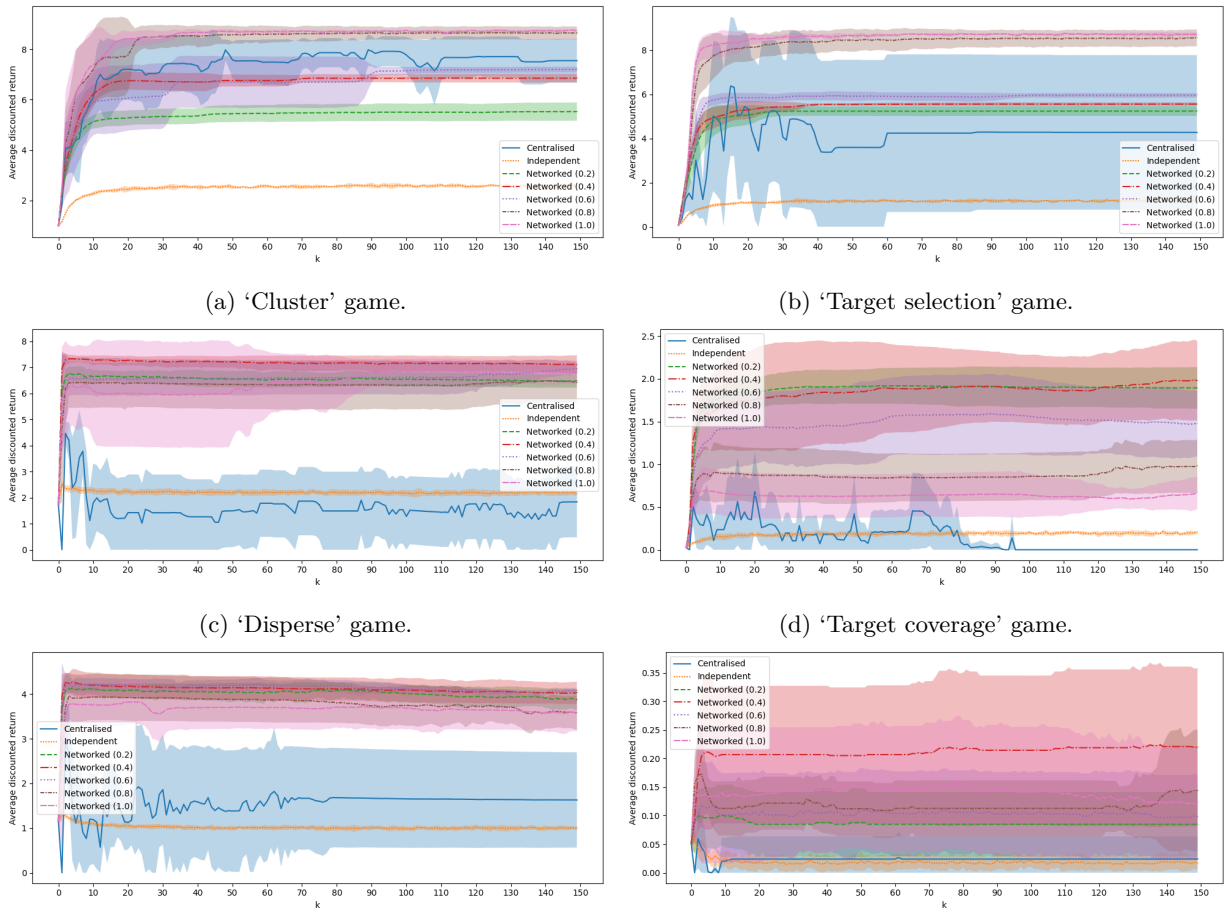
(c) 'Disperse' game.

(d) 'Target coverage' game.

(e) 'Beach bar' game.

(f) 'Shape formation' game.

Figure 6: Ablation study on population-*independent* policies. No agents, including centralised and networked ones, observe the empirical mean field, and all receive a vector of zeros in its place (so as to keep the neural networks the same size as in the standard setting). $C_r = C_p = 1$. Networked populations do not appear to perform substantially differently to the standard population-dependent setting, though some radii (red, 0.4; pink, 1.0) appear to perform slightly better in the 'shape formation' game. This is likely because all of our games have stationary solutions, such that observing the mean field is not actually necessary, even if it could potentially be useful (see Sec. 3.1 for discussion of the conception of MFC as a central planner trying to guide the population to a distribution that maximises the expected return). Indeed, in the coordination games, and particularly the 'target selection' game, the central-agent population receives a lower return in this setting, whereas our networked populations are robust to this change.




(a) 'Cluster' game.

(b) 'Target selection' game.

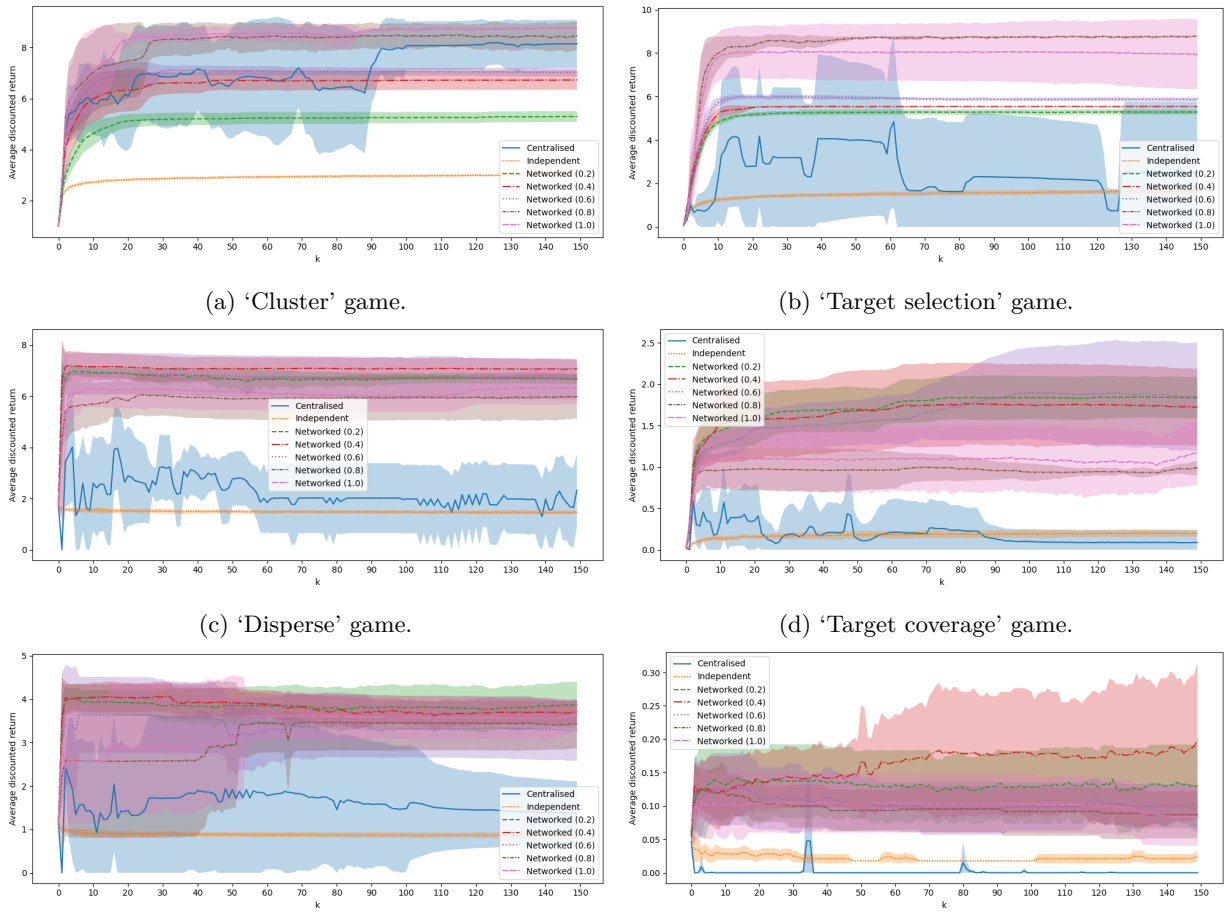
(c) 'Disperse' game.

(d) 'Target coverage' game.

(e) 'Beach bar' game.

(f) 'Shape formation' game.

Figure 7: Ablation study of Alg. 4 for estimating the empirical mean field - all agents, including independent ones, directly receive the true global empirical mean field. $C_r = C_p = 1$. This does not appear to change performance in the networked populations (apart from greater variance here in the 'shape formation' game), nor does it help independent agents. This may be evidence that Alg. 4 enables networked agents to accurately estimate the global mean field from local observations. However, our ablation study on population-independent policies (Fig. 6) suggests that not observing the mean field does not markedly disadvantage agents in our experimental settings in any case (apart from for the central-agent populations in the coordination games). This is likely because all of our games have stationary solutions, such that observing the mean field is not necessary. Therefore, in order to confirm the efficacy of Alg. 4 for estimating the mean field, further evidence is perhaps needed in MFC settings that require population-dependent policies, though Benjamin & Abate (2024) has already confirmed this for non-stationary games in the non-cooperative MFG setting.




(a) 'Cluster' game.

(b) 'Target selection' game.

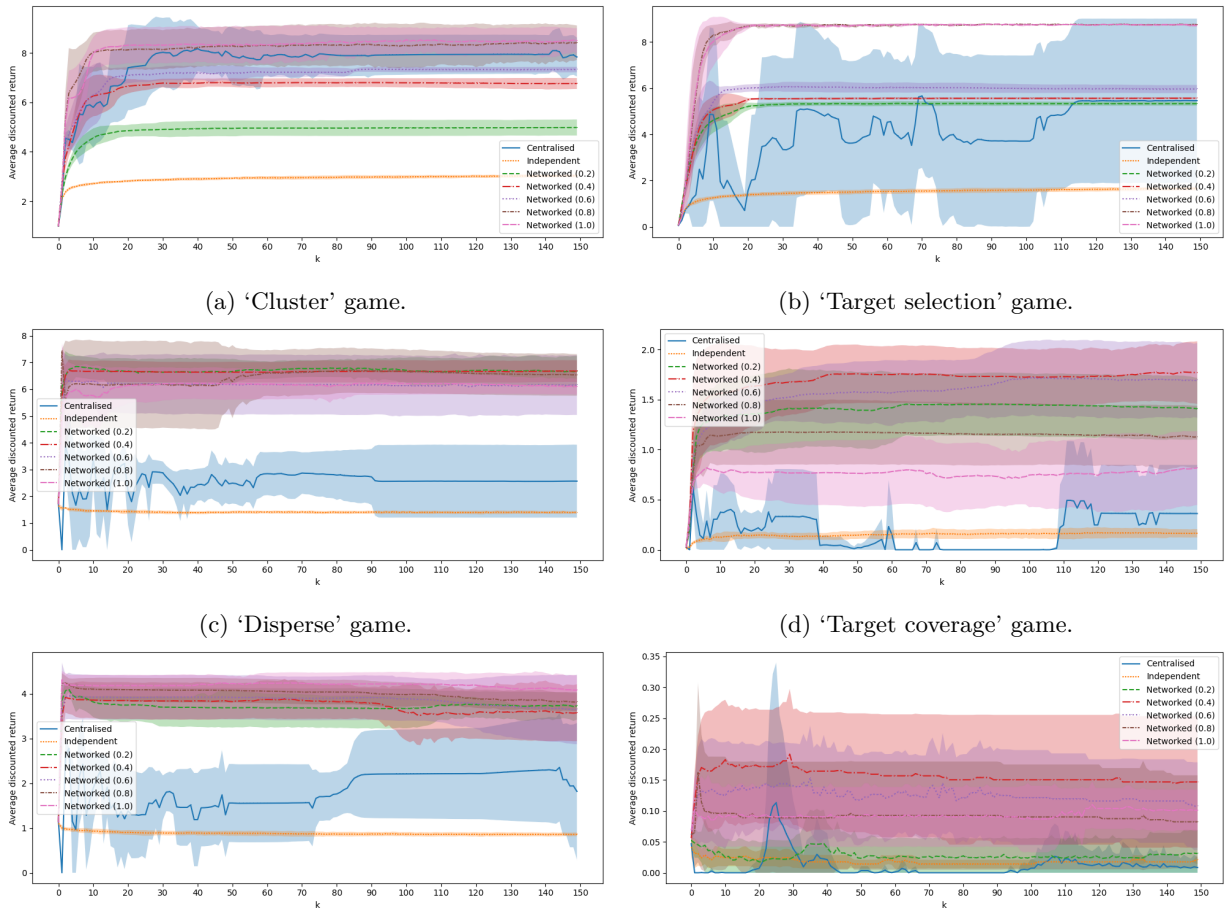
(c) 'Disperse' game.

(d) 'Target coverage' game.

(e) 'Beach bar' game.

(f) 'Shape formation' game.

Figure 8: Ablation study for observation of true/estimated global average reward $\hat{r}/\tilde{\hat{r}}_t^i$, where all agents, including centralised ones, only have access to $r_t^i$, where in the central-agent case $i = 1$. $C_e = C_p = 1$. The greatest effect of this is on the central-agent (blue) populations, which perform much worse in the 'target selection' game, and with higher variance in the 'cluster' and 'beach bar' games, i.e. they suffer without access to the global average reward. The networked agents appear more robust to the loss of the (estimated) average reward, pointing to an additional benefit of the policy communication scheme, though do experience a slight performance decrease, mostly among populations with the largest broadcast radii (pink, 1.0; brown, 0.8), i.e. those most similar to the central-agent case in terms of $\tilde{\hat{r}}_t^i$, as might be expected. In particular, note the greater variance of pink (1.0) in the 'target selection' game; slower learning and higher variance of pink (1.0) and brown (0.8) in the 'beach bar' game; lower returns for pink (1.0) and brown (0.8) in the 'shape formation' game; and slower learning and convergence of the smallest radii (green, 0.2; red, 0.4) in the 'target coverage' game. This all demonstrates the usefulness and efficacy of our novel Alg. 1 for decentralised estimation of the global average reward.




(a) 'Cluster' game.

(b) 'Target selection' game.

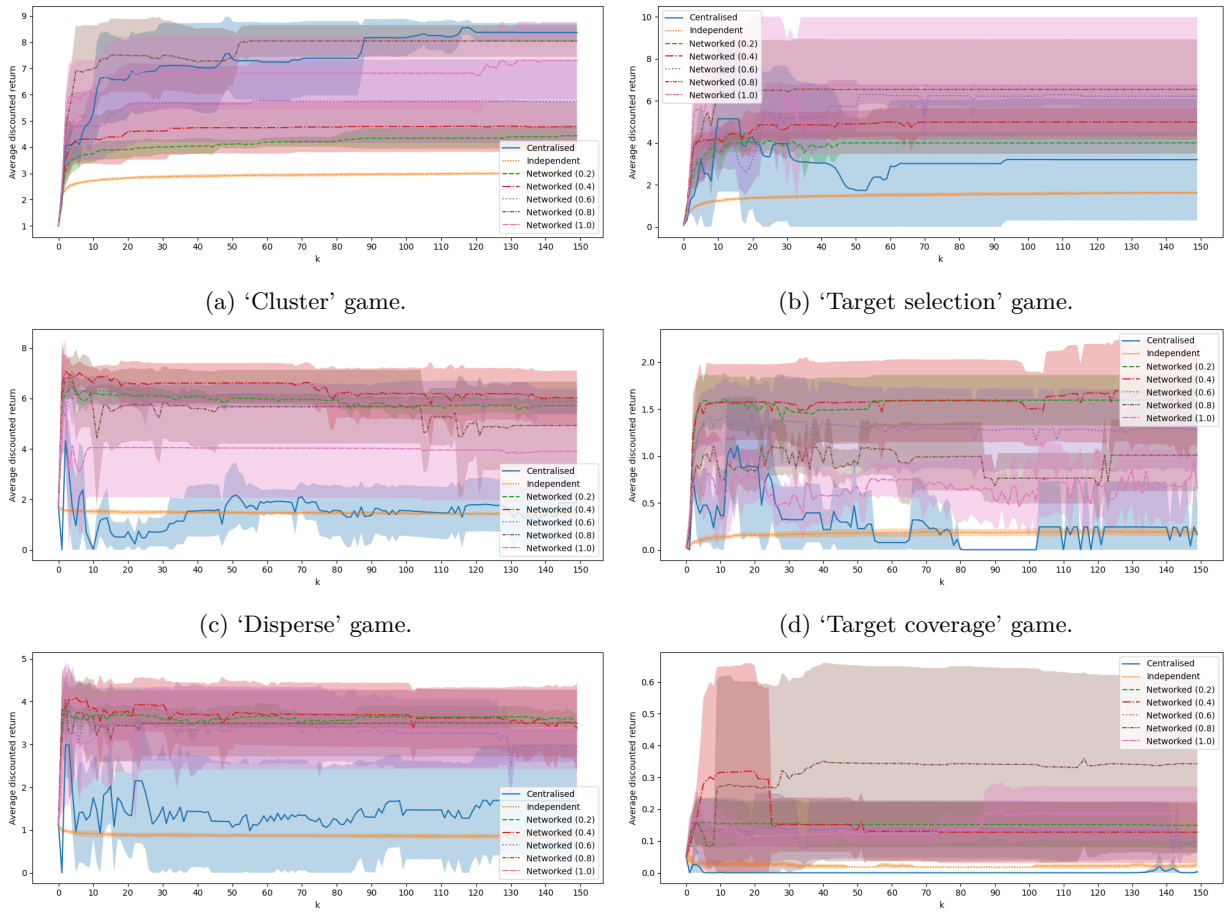
(c) 'Disperse' game.

(d) 'Target coverage' game.

(e) 'Beach bar' game.

(f) 'Shape formation' game.

Figure 9: Ablation study for Alg. 1 for estimating the true global average reward. All agents, including both networked and independent ones, directly receive the true global average reward such that $\tilde{\hat{r}}_t^i = \hat{r}$. $C_e = C_p = 1$. Access to the true average reward does not help networked agents to improve their returns, demonstrating that our novel Alg. 1 already affords networked populations robustness against the lack of access to this global information (having this global information would be an unrealistic assumption in practice). Access to the true average reward also does not help independent agents to improve their returns, suggesting the *policy* communication scheme is the dominant factor in improving the performance of decentralised agents.




(a) 'Cluster' game.

(b) 'Target selection' game.

(c) 'Disperse' game.

(d) 'Target coverage' game.

(e) 'Beach bar' game.

(f) 'Shape formation' game.

Figure 10: Ablation study of the choice of $\tau_k^{comm}$. Here $\forall k\ \tau_k^{comm}$ = 1e-18 (i.e. $\tau_k^{comm}$ is close to 0, turning the softmax into a max function), rather than linearly increasing from 0.001 to 1 across the $K$ iterations as in all other experiments (see Table 1). $C_e = C_r = C_p = 1$. In this setting, networked agents continue to outperform the central-agent (blue) and independent (orange) populations in all games except the 'cluster' game, but otherwise generally appear to receive lower average returns and with greater variance. This is because Assumption 5.6 on the quality of the finite-step approximations $\{\sigma_{k+1}^i\}_{i=1}^N = \{\hat{V}^i(\boldsymbol{\pi}_{k+1}, \mu_t; E)\}_{i=1}^N$ may not always apply in practice, especially as policies get more deterministic closer to convergence. This means the policy estimated to perform the best may not actually be among the best updates, such that enforcing the adoption of this policy can lead to noisy, unstable learning. Using a higher temperature value smooths out this noise (while having a lower temperature at the beginning of training ensures faster learning when the difference in the quality of nascent policies is likely to be more stark, hence our inverse annealing scheme). Moreover, using $\tau_k^{comm}$ close to 0 more effectively enforces consensus on a single policy in the networked case, which in anti-coordination games may also reduce the average return (see Sec. 6.3). This all provides empirical evidence for our scheme for $\tau_k^{comm}$, but further optimising the choice might lead to additional performance increase.

## 7 Conclusion

We provided the first algorithms for decentralised training in MFC, as well as the first for online learning in MFC from a single non-episodic run of the empirical system. We did so by modifying existing algorithms for the MFG setting, and contributing a novel algorithm for estimating the global average reward via local communication. We proved theoretically that networked communication accelerates learning over both the independent and central-agent architectures. We supported this with extensive numerical results, accompanied by ablation studies and discussion of the empirical effects of communication radii.

Our work follows the gold standard in MFGs by presenting experiments on grid world toy environments. Nevertheless future work includes experiments in other games, including non-stationary games, more realistic environments and ones where both the transition function and the reward function depend on the mean field. Please note, however, that Benjamin & Abate (2024) already demonstrated in the MFG setting that the communication scheme affords faster learning when both the transition and reward functions depend on the mean field.

In Sec. 5 we give theoretical results showing that our networked algorithm can outperform the central-agent and independent alternative. We leave more general theoretical results, such as proofs of convergence and sample complexity, for future work.

Our algorithms contain numerous inner loops and thus require synchronisation between communicating agents. Our ablation studies of the sub-routines and our experiment on robustness to communication failures (Fig. 3) indicate that synchronisation failure is not necessarily a problem in practice, but future work nevertheless lies in simplifying the nested loops of our algorithms.

In grid-world settings such as those in our experiments, passing the (estimated or true global) mean-field distribution as a flat vector to the Q-network ignores the geometric structure of the problem. Perrin et al. (2022) therefore proposes to create an embedding of the distribution by first passing the vector to a convolutional neural network, essentially treating the categorical distribution as an image. This technique is also followed in Wu et al. (2024) (for their additional experiments, but not in the main body of their paper). As future work, we can test whether such a method increases the usefulness of observing the mean field in population-dependent policies, and therefore increases the importance of be able to accurately estimate the global mean field via Alg. 4.

### Broader Impact Statement

We identified no specific ethical concerns regarding our work, which explores new game theoretical and machine learning algorithms in general settings.

## References

Bernard T. Agyeman, Benjamin Decardi-Nelson, Jinfeng Liu, and Sirish L. Shah. A semi-centralized multi-agent RL framework for efficient irrigation scheduling. *Control Engineering Practice*, 155:106183, 2025. ISSN 0967-0661. doi: https://doi.org/10.1016/j.conengprac.2024.106183. URL `https://www.sciencedirect.com/science/article/pii/S0967066124003423`.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, 13(1):89–117, 2023.

Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. Unified reinforcement Q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, 34(2):217–271, 2022.

Andrea Angiuli, Jean-Pierre Fouque, Mathieu Laurière, and Mengrui Zhang. Convergence of Multi-Scale Reinforcement Q-Learning Algorithms for Mean Field Game and Control Problems. *arXiv preprint arXiv:2312.06659*, 2023.

Erhan Bayraktar and Ali D Kara. Learning with Linear Function Approximations in Mean-Field Control. *arXiv preprint arXiv:2408.00991*, 2024.

Patrick Benjamin and Alessandro Abate. Networked Communication for Decentralised Agents in Mean-Field Games. *arXiv preprint arXiv:2306.02766*, 2023.

Patrick Benjamin and Alessandro Abate. Networked Communication for Mean-Field Games with Function Approximation and Empirical Mean-Field Estimation. *arXiv preprint arXiv:2408.11607*, 2024.

Daniel S. Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The Complexity of Decentralized Control of Markov Decision Processes. *Mathematics of Operations Research*, 27(4):819–840, 2002. ISSN 0364765X, 15265471. URL `http://www.jstor.org/stable/3690469`.

René Carmona, Mathieu Laurière, and Zongjun Tan. Linear-quadratic mean-field reinforcement learning: convergence of policy gradient methods. *arXiv preprint arXiv:1910.04295*, 2019.

René Carmona, Mathieu Laurière, and Zongjun Tan. Model-Free Mean-Field Reinforcement Learning: Mean-Field MDP and Mean-Field Q-Learning. *The Annals of Applied Probability*, 33(6B):5334–5381, 2023.

Leo Cazenille, Maxime Toquebiau, Nicolas Lobato-Dauzier, Alessia Loi, Loona Macabre, Nathanaël Aubert-Kato, Anthony J Genot, and Nicolas Bredeche. Signalling and social learning in swarms of robots. *Philosophical Transactions A*, 383(2289):20240148, 2025.

Mingzhe Chen, Deniz Gündüz, Kaibin Huang, Walid Saad, Mehdi Bennis, Aneta Vulgarakis Feljan, and H Vincent Poor. Distributed Learning in Wireless Networks: Recent Progress and Future Challenges. *IEEE Journal on Selected Areas in Communications*, 39(12):3579–3605, 2021.

Yufan Chen, Lan Wu, Renyuan Xu, and Ruixun Zhang. Periodic Trading Activities in Financial Markets: Mean-field Liquidation Game with Major-Minor Players. *arXiv preprint arXiv:2408.09505*, 2024.

Kai Cui and Heinz Koeppl. Approximately Solving Mean Field Games via Entropy-Regularized Deep Reinforcement Learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1909–1917. PMLR, 2021.

Kai Cui, Christian Fabian, and Heinz Koeppl. Multi-Agent Reinforcement Learning via Mean Field Control: Common Noise, Major Agents and Approximation Properties. *arXiv preprint arXiv:2303.10665*, 2023a.

Kai Cui, Christian Fabian, and Heinz Koeppl. Multi-agent reinforcement learning via mean field control: Common noise, major agents and approximation properties. 03 2023b. doi: 10.48550/arXiv.2303.10665.

Kai Cui, Sascha Hauck, Christian Fabian, and Heinz Koeppl. Learning Decentralized Partially Observable Mean Field Control for Artificial Collective Behavior. *arXiv preprint arXiv:2307.06175*, 2023c.

Gökçe Dayanikli, Mathieu Laurière, and Jiacheng Zhang. Deep Learning for Population-Dependent Controls in Mean Field Control Problems with Common Noise. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, pp. 2231–2233, 2024.

Robert Denkert, Idris Kharroubi, and Huyên Pham. A randomisation method for mean-field control problems with common noise. *arXiv preprint arXiv:2412.20782*, 2024.

Hesam Farzaneh, Mohammad Shokri, Hamed Kebriaei, and Farrokh Aminifar. Robust Energy Management of Residential Nanogrids via Decentralized Mean Field Control. *IEEE Transactions on Sustainable Energy*, 11(3):1995–2002, 2020. doi: 10.1109/TSTE.2019.2949016.

Iñaki Fernández Pérez and Stéphane Sanchez. Influence of Local Selection and Robot Swarm Density on the Distributed Evolution of GRNs. In Paul Kaufmann and Pedro A. Castillo (eds.), *Applications of Evolutionary Computation*, pp. 567–582, Cham, 2019. Springer International Publishing. ISBN 978-3-030-16692-2.

Iñaki Fernández Pérez, Amine Boumaza, and François Charpillet. Maintaining Diversity in Robot Swarms with Distributed Embodied Evolution. In Marco Dorigo, Mauro Birattari, Christian Blum, Anders L. Christensen, Andreagiovanni Reina, and Vito Trianni (eds.), *Swarm Intelligence*, pp. 395–402, Cham, 2018. Springer International Publishing. ISBN 978-3-030-00533-7.

Massimo Fornasier and Francesco Solombrino. Mean-Field Optimal Control. *ESAIM: Control, Optimisation and Calculus of Variations*, 20(4):1123–1152, 2014. doi: 10.1051/cocv/2014009.

Sriram Ganapathi Subramanian, Pascal Poupart, Matthew E Taylor, and Nidhi Hegde. Multi Type Mean Field Reinforcement Learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 411–419, 2020.

Sriram Ganapathi Subramanian, Matthew E Taylor, Mark Crowley, and Pascal Poupart. Partially Observable Mean Field Reinforcement Learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 537–545, 2021.

Sergio Grammatico, Francesca Parise, Marcello Colombino, and John Lygeros. Decentralized Convergence to Nash Equilibria in Constrained Deterministic Mean Field Control. *IEEE Transactions on Automatic Control*, 61(11):3315–3329, 2016. doi: 10.1109/TAC.2015.2513368.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-Field Controls with Q-Learning for Cooperative MARL: Convergence and Complexity Analysis. *SIAM Journal on Mathematics of Data Science*, 3(4): 1168–1196, 2021. doi: 10.1137/20M1360700. URL `https://doi.org/10.1137/20M1360700`.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. A General Framework for Learning Mean-Field Games. *Mathematics of Operations Research*, 48(2):656–686, 2023.

Saeed Hadikhanloo. Learning in anonymous nonatomic games with applications to first-order mean field games. *arXiv preprint arXiv:1704.00378*, 2017.

Emma Hart, Andreas Steyven, and Ben Paechter. Improving Survivability in Environment-Driven Distributed Evolutionary Algorithms through Explicit Relative Fitness and Fitness Proportionate Communication. In *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation*, GECCO '15, pp. 169–176, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450334723. doi: 10.1145/2739480.2754688. URL `https://doi.org/10.1145/2739480.2754688`.

Jiří Horyna, Roland Jung, Stephan Weiss, Eliseo Ferrante, and Martin Saska. Swarming Without an Anchor (SWA): Robot Swarms Adapt Better to Localization Dropouts Then a Single Robot. *IEEE Robotics and Automation Letters*, 10(6):6207–6214, 2025. doi: 10.1109/LRA.2025.3562786.

Anran Hu and Junzi Zhang. MF-OML: Online Mean-Field Reinforcement Learning with Occupation Measures for Large Population Games. *arXiv preprint arXiv:2405.00282*, 2024. URL `https://arxiv.org/abs/2405.00282`.

Minyi Huang, Roland P. Malhamé, and Peter E. Caines. 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.

A. Jadbabaie, Jie Lin, and A.S. Morse. Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, 48(6):988–1001, 2003. doi: 10.1109/TAC.2003.812781.

Jiechuan Jiang, Kefan Su, and Zongqing Lu. Fully Decentralized Cooperative Multi-Agent Reinforcement Learning: A Survey. *arXiv preprint arXiv:2401.04934*, 2024.

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

Mathieu Laurière, Sarah Perrin, Matthieu Geist, and Olivier Pietquin. Learning Mean Field Games: A Survey. *arXiv preprint arXiv:2205.12944*, 2022a.

Mathieu Laurière, Sarah Perrin, Sertan Girgin, Paul Muller, Ayush Jain, Theophile Cabannes, Georgios Piliouras, Julien Perolat, Romuald Elie, Olivier Pietquin, and Matthieu Geist. Scalable Deep Reinforcement Learning Algorithms for Mean Field Games. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song,

Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12078–12095. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/lauriere22a.html.

Taeyoung Lee et al. Mean Field Game and Control for Switching Hybrid Systems. *arXiv preprint arXiv:2412.10522*, 2024.

Changling Li and Ying Li. Scaling up Energy-Aware Multi-Agent Reinforcement Learning for Mission-Oriented Drone Networks With Individual Reward. *IEEE Internet of Things Journal*, pp. 1–1, 2024. doi: 10.1109/JIOT.2024.3511253.

Washim Uddin Mondal, Mridul Agarwal, Vaneet Aggarwal, and Satish V. Ukkusuri. On the approximation of cooperative heterogeneous multi-agent reinforcement learning (MARL) using Mean Field Control (MFC). *J. Mach. Learn. Res.*, 23(1), January 2022. ISSN 1532-4435.

Behrang Monajemi Nejad, Sid Ahmed Attia, and Jorg Raisch. Max-consensus in a max-plus algebraic setting: The case of fixed communication topologies. In *2009 XXII International Symposium on Information, Communication and Automation Technologies*, pp. 1–7, 2009. doi: 10.1109/ICAT.2009.5348437.

Julien Perolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. Scaling up Mean Field Games with Online Mirror Descent. *arXiv preprint arXiv:2103.00623*, 2021.

Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious Play for Mean Field Games: Continuous Time Analysis and Applications. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Sarah Perrin, Mathieu Laurière, Julien Pérolat, Romuald Élie, Matthieu Geist, and Olivier Pietquin. Generalization in mean field games by learning master policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9413–9421, 2022.

Lars Ruthotto, Stanley J. Osher, Wuchen Li, Levon Nurbekyan, and Samy Wu Fung. A machine learning framework for solving high-dimensional mean field game and mean field control problems. *Proceedings of the National Academy of Sciences*, 117(17):9183–9193, 2020. doi: 10.1073/pnas.1922204117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1922204117.

Naci Saldi, Tamer Başar, and Maxim Raginsky. Markov–Nash Equilibria in Mean-Field Games with Discounted Cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018. doi: 10.1137/17M1112583. URL https://doi.org/10.1137/17M1112583.

Rajeshwari Sissodia, ManMohan Singh Rauthan, Varun Barthwal, and Vinay Dwivedi. Evolutionary Algorithms for Optimization and Swarm Intelligence-Based Optimization. In *Optimization Tools and Techniques for Enhanced Computational Efficiency*, pp. 17–42. IGI Global Scientific Publishing, 2025.

Jayakumar Subramanian and Aditya Mahajan. Reinforcement Learning in Stationary Mean-Field Games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '19, pp. 251–259, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450363099.

Sriram Ganapathi Subramanian, Matthew E Taylor, Mark Crowley, and Pascal Poupart. Decentralized Mean Field Games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9439–9447, 2022.

Mohammad Amin Tajeddini, Hamed Kebriaei, and Luigi Glielmo. Robust decentralised mean field control in leader following multi-agent systems. *IET Control Theory & Applications*, 11(16):2707–2715, 2017.

Noureddine Toumi, Roland Malhame, and Jerome Le Ny. A mean field game approach for a class of linear quadratic discrete choice problems with congestion avoidance. *Automatica*, 160:111420, 2024. ISSN 0005-1098. doi: https://doi.org/10.1016/j.automatica.2023.111420. URL https://www.sciencedirect.com/science/article/pii/S0005109823005873.

Nino Vieillard, Olivier Pietquin, and Matthieu Geist. Munchausen Reinforcement Learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 4235–4246. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/2c6a0bae0f071cbbf0bb3d5b11d90a82-Paper.pdf.

Zida Wu, Mathieu Laurière, Samuel Jia Cong Chua, Matthieu Geist, Olivier Pietquin, and Ankur Mehta. Population-aware Online Mirror Descent for Mean-Field Games by Deep Reinforcement Learning. *arXiv preprint arXiv:2403.03552*, 2024.

Kun Xu, Yue Li, Jun Sun, Shuyuan Du, Xinpeng Di, Yuguang Yang, and Bo Li. Targets capture by distributed active swarms via bio-inspired reinforcement learning. *Science China Physics, Mechanics & Astronomy*, 68(1):1–12, 2025.

Hongyi Yang, Jingzhi Liu, Geng Li, Jianming Zhang, Ling Jiang, and Shoulian Yang. Distributed Intelligent Power Distribution Optimization Method Based on Mean Field Game Theory. In *2025 IEEE 5th International Conference on Power, Electronics and Computer Applications (ICPECA)*, pp. 818–822, 2025. doi: 10.1109/ICPECA63937.2025.10928821.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean Field Multi-Agent Reinforcement Learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5571–5580. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/yang18d.html.

Batuhan Yardim and Niao He. Exploiting Approximate Symmetry for Efficient Multi-Agent Reinforcement Learning. *arXiv preprint arXiv:2408.15173*, 2024.

Batuhan Yardim, Semih Cayci, Matthieu Geist, and Niao He. Policy Mirror Ascent for Efficient and Independent Learning in Mean Field Games. In *International Conference on Machine Learning*, pp. 39722–39754. PMLR, 2023.

Batuhan Yardim, Artur Goldman, and Niao He. When is Mean-Field Reinforcement Learning Tractable and Relevant? *arXiv preprint arXiv:2402.05757*, 2024.

Bora Yongacoglu, Gürdal Arslan, and Serdar Yüksel. Mean-field games with finitely many players: independent learning and subjectivity. *Journal of Machine Learning Research*, 25(419):1–69, 2024.

Muhammad Aneeq Uz Zaman, Mathieu Lauriere, Alec Koppel, and Tamer Başar. Robust cooperative multi-agent reinforcement learning: A mean-field type game perspective. In *6th Annual Learning for Dynamics & Control Conference*, pp. 770–783. PMLR, 2024.

Sihan Zeng, Sujay Bhatt, Alec Koppel, and Sumitra Ganesh. A Single-Loop Finite-Time Convergent Policy Optimization Algorithm for Mean Field Games (and Average-Reward Markov Decision Processes). *arXiv e-prints*, pp. arXiv–2408, 2024.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. Fully Decentralized Multi-Agent Reinforcement Learning with Networked Agents. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5872–5881. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/zhang18n.html.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Decentralized Multi-Agent Reinforcement Learning with Networked Agents: Recent Advances. *Frontiers of Information Technology & Electronic Engineering*, 22 (6):802–814, 2021a.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. *"Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms"*, pp. 321–384. Springer International Publishing, Cham, 2021b. ISBN 978-3-030-60990-0. doi: 10.1007/978-3-030-60990-0_12. URL https://doi.org/10.1007/978-3-030-60990-0_12.