Networked Communication for Decentralised Cooperative Agents in Mean-Field Control

Patrick Benjamin[1], **Alessandro Abate**[1]
{patrick.benjamin,alessandro.abate}@cs.ox.ac.uk

[1]**Department of Computer Science, University of Oxford, UK**

Abstract

We introduce networked communication to mean-field control (MFC) - the cooperative counterpart to mean-field games (MFGs) - and in particular to the setting where decentralised agents learn online from a single, non-episodic run of the empirical system. We adapt recent algorithms for MFGs 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. We show that the networked communication scheme allows agents to increase social welfare faster than under both the centralised and independent architectures, by computing a population of potential updates in parallel and then propagating the highest-performing ones through the population, via a method that can also be seen as tackling the credit-assignment problem. We prove this new result theoretically and provide experiments that support it across different classes of cooperative game (coordination and anti-coordination), as well as exploring the empirical finding that smaller communication radii can benefit convergence in anti-coordination games while still outperforming agents learning entirely independently. We provide numerous ablation studies and additional experiments on numbers of communication round and robustness to communication failures.

1 Introduction

Multi-agent reinforcement learning (MARL) can struggle to scale computationally as the number of agents N increases. The mean-field game (MFG) framework (Lasry & Lions, 2007; Huang et al., 2006) has been used to address this scaling difficulty; it 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 MFG is a non-cooperative scenario where each agent seeks to maximise its individual return, and the solution to the game is a mean-field Nash equilibrium (MFNE), which can be used as an approximation for the Nash equilibrium (NE) in a finite-agent game, with the error in the solution reducing as N tends to infinity (Yardim et al., 2024). Alternatively we can consider a cooperative scenario called a mean-field control (MFC) problem, where the population seeks to maximise a social welfare criterion such as the average return received by agents.

Since MFC problems can be interpreted as optimisation problems from the perspective of a social planner, classical approaches to MFC involve centralised methods whereby a central learner updates a policy that is assumed to be passed automatically to the population (Laurière et al., 2022a). Often the empirical mean field of the actual population is not used, with the central learner updating an estimate of the mean field based on its own policy (Carmona et al., 2019; Angiuli et al., 2022; 2023). However, recent works on MFGs, as in other areas of multi-agent research, have recognised that reliance on a central coordinator represents a bottleneck for computation and communication, and a vulnerable single point of failure of the system (Yardim et al., 2023; Benjamin & Abate, 2023; 2024; Horyna et al., 2025). These works also argue that, in order to be applicable to real-world, embodied problems, other desirable qualities for mean-field algorithms include: learning from the population's empirical mean field (i.e. this distribution is generated only by the agents' policies, rather than being updated by the algorithm itself or 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); model-free learning; and function approximation to allow high-dimensional observations.

Some recent MFC works have considered decentralisation, but Bayraktar & Kara (2024) requires that decentralised agents optimise for learnt models of the system dynamics (and 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). However, assuming decentralised agents have access to this global information is unrealistic, and Benjamin & Abate (2024) in the non-cooperative MFG setting has 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). We contribute theoretical proofs that decentralised policy communication allows networked populations to learn faster than both the independent *and* the centralised alternatives in the MFC setting in 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 for decentralised model-free training in MFC, as well as the first MFC algorithms for online learning from a single, non-episodic run of the empirical system.
- We prove theoretically that in this context, decentralised networked communication can improve learning speed over the independent *and* centralised alternatives.
- We further 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 provide extensive experiments supporting our theoretical results, and give ablation studies of various parts of our algorithms, as well as a study of robustness to communication failures.

We give preliminaries in Sec. 2, and our algorithms in Sec. 3. We present theoretical results in Sec. 4, and experiments in Sec. 5. A more detailed comparison to related work is in Appx. G.

2 Preliminaries

Solving the MFC problem involves finding the single policy that, when given to all agents in the infinite population, maximises the population's expected return. We give two ways to conceive of our work, illustrated in Fig. 2 (Appx. A), making 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). Firstly, we contribute algorithms that allow the solution to a MFC problem to be learnt using the empirical distribution of a decentralised finite population, without needing to make unrealistic assumptions about access to an oracle for the infinite population. Note that empirically it may be impractical to assume that the decentralised agents always follow a single identical policy.

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, 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 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); Benjamin & Abate (2023; 2024) follow a similar logic in MFGs).

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 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 2 (Individual expected discounted return)**.** *For all $i, j \in \{1,\dots,N\}$, i's return is* $V^i(\boldsymbol{\pi}, \mu_{\bar{t}}) = \mathbb{E}\left[\sum_{t=\bar{t}}^{\infty}\gamma^t R(s_t^i, a_t^i, \hat{\mu}_t)\middle|_{\substack{s_{\bar{t}}^j\sim\mu_{\bar{t}} \\ a_t^j\sim\pi^j(o_t^j) \\ s_{t+1}^j\sim P(\cdot|s_t^j,a_t^j,\hat{\mu}_t)}}\right]$.

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

Definition 3 (Population-average expected discounted return)**.** *For $i, j \in \{1,\dots,N\}$ the return is* $V^{pop}(\boldsymbol{\pi}, \mu_{\bar{t}}) = \frac{1}{N}\sum_i^N V^i(\boldsymbol{\pi}, \mu_{\bar{t}}) = \mathbb{E}\left[\frac{1}{N}\sum_{t=\bar{t}}^{\infty}\sum_i^N \gamma^t R(s_t^i, a_t^i, \hat{\mu}_t)\middle|_{\substack{s_{\bar{t}}^j\sim\mu_{\bar{t}} \\ a_t^j\sim\pi^j(o_t^j) \\ s_{t+1}^j\sim P(\cdot|s_t^j,a_t^j,\hat{\mu}_t)}}\right]$.

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 so-called '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 4 (Induced mean-field flow)**.** *We denote by $I(\pi)$ the mean-field flow $\boldsymbol{\mu}$ induced when all the agents follow π, where this is generated from π 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 5 (Social welfare)**.** *When all agents follow policy π giving mean-field flow $\boldsymbol{\mu} = I(\pi)$, π'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 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. 5, i.e. $\pi^* = \arg\max_{\pi \in \Pi} W(\pi; I(\pi))$.*

Remark 1. *Previous works showed that the MFC social optimum π^* gives a good approximation for the harder-to-solve finite-agent problem (i.e. if $\boldsymbol{\pi} = (\pi^*, \dots, \pi^*)$), 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 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, classical approaches to the MFC problem that conceive of a central planner trying to guide the population to a distribution that maximises the expected return might 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 use 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. 3 (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 $\mathcal{Q} = \{q : \mathcal{S} \times \Delta_{\mathcal{S}} \times \mathcal{A} \to \mathbb{R}\}$.[1] The communication graph of the decentralised population is given by:

Definition 7 (Time-varying communication network)**.** *The time-varying graph $(\mathcal{G}_t^{comm})_{t \geq 0}$ is given by $\mathcal{G}_t = (\mathcal{N}, \mathcal{E}_t)$, where $\mathcal{N}$ is the set of vertices each representing an agent $i = \{1, \dots, N\}$, 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 $d_{\mathcal{G}_t}$ is the maximum of the shortest path lengths between any pair of nodes.*

3 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 policy updates 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. 4, so we do not directly optimise social welfare from Def. 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. 3.1), and we describe our main learning method Alg. 2 in Sec. 3.2. Meanwhile Alg. 3 for estimating the mean field, which is taken from Alg. 3 in Benjamin & Abate (2024) for the MFG setting, is described in Appx. C.3. Our policy communication algorithm Alg. 4 is also based on that in Benjamin & Abate (2024) for the MFG setting, but since it is key to our novel theoretical results that we contribute for the MFC setting, we give a description of Alg. 4 in the main text in Sec. 3.3.

[1] When $o_t^i = s_t^i$, we instead have $\Pi = \{\pi : \mathcal{S} \to \Delta_{\mathcal{A}}\}$ and $\mathcal{Q} = \{q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$.

3.1 Sub-routine for networked estimation of global average reward

Our novel Alg. 1 (Appx. C.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 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).

3.2 Main learning algorithm

Our novel Alg. 2 (Appx. C.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. It is based on Munchausen Online Mirror Descent (for further background see Appx. B). Each agent i approximates its Q-function $\check{Q}_{\theta_k^i}(o, \cdot)$ with its own neural network parametrised by θ_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 $\left(o_t^i, a_t^i, \tilde{\hat{r}}_t^i, o_{t+1}^i\right)$, 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 M new transitions (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 $\check{Q}_{\theta_{k,l}^{i,\prime}}$ with the same architecture but parameters $\theta_{k,l}^{i,\prime}$ copied from $\theta_{k,l}^i$ less regularly than $\theta_{k,l}^i$ themselves are updated, i.e. only every ν 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 10, Appx. C.2 (Line 12). The trained Q-network determines i's updated policy (Line 16).

3.3 Sub-routine for communicating and refining policies

Alg. 4 (Appx. C.4, based on Alg. 1 in Benjamin & Abate (2024) for the MFG setting) 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 even centralised cases.

Alg. 4 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. 4, agents obtain an approximation of their *individual* discounted expected return $\{V^i(\boldsymbol{\pi}, \mu_t)\}_{i=1}^N$ (Def. 2, i.e. *not* the population-average return, which would not give differentiation between the different updated policies). They do so by collecting individual rewards for E steps, and calculating the discounted sum of rewards over these finite steps, setting this value to σ_{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 σ_{k+1}^i (Line 9). Receiving these from their neighbours J_t^i on the communication 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: $\text{adopted}^i \sim \Pr\big(\text{adopted}^i = j\big) = \frac{\exp(\sigma_{k+1}^j/\tau_k^{comm})}{\sum_{x \in J_t^i}\exp(\sigma_{k+1}^x/\tau_k^{comm})}$ (Lines 10-12). They can repeat this broadcast and adoption process for C_p rounds (distinct from the C_r and C_e communication rounds for the other sub-routines). We theoretically prove the benefits of this method in the following.

4 Theoretical results

To demonstrate the networked architecture's benefits, we compare it with the results of modified versions of our algorithm for centralised and independent learners. In the centralised case, as in similar MFG and MFC works, only arbitrary central agent $i = 1$ updates a Q-network and automatically pushes this to all other agents, and the true global mean-field distribution is always used in place of the local estimate i.e. $\tilde{\hat{\mu}}_t^i = \hat{\mu}_t$. In the independent case, there are no links in $\mathcal{G}_t^{comm}$, i.e. $\mathcal{E}_t^{comm} = \emptyset$. We now prove theoretically that the policy communication and adoption scheme allows networked agents to increase their returns faster than the centralised and independent architectures. Rem. 2 (Appx. D) suggests informal reasons for our formal results, to aid intuitive understanding.

We give the theoretical analysis separately for two important subclasses of cooperative game usually found in MFC, which have different reward structures and therefore require different population behaviour, namely: 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 policies, such as in exploration, coverage or task allocation games. The phenomenon of diversity being desirable in cooperative anti-coordination games is an artifact of having a finite, albeit large, population: the benefit of diversity will decrease as the empirical population tends to infinity, until the single social optimum policy must be followed by all agents. While it is intuitive that adopting policies from neighbours via the communication scheme would be beneficial in coordination games, we show theoretically and empirically that the scheme also benefits populations in anti-coordination games.

To define formally the two types of game, we first introduce the following two functions. $b : \Pi \rightarrow \mathbb{R}_{\geq 0}$ 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.[2] $\mathbb{I}[\cdot]$ is the indicator function, which equals 1 if the condition inside is true and 0 otherwise.

Definition 8 (Coordination game). *$f_c : \mathbb{N} \rightarrow \mathbb{R}_{>0}$ is a 'coordination scaling function'. It has minimum $f_c(1) > 0$, and increases monotonically with the number of agents whose policies match i's. A coordination game is one where the agents' return can be decomposed as follows, $\forall i, j \in \{1, \ldots, N\}$: $V^i(\boldsymbol{\pi}, \mu_{\bar{t}}) = h\left(b(\pi^i), f_c\left(\sum_{j\in\{1,\ldots,N\}} \mathbb{I}\left[\pi^i = \pi^j\right]\right)\right)$, where $h : \mathbb{R}_{\geq 0} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}_{\geq 0}$ 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, results in a higher return.*

Definition 9 (Anti-coordination game). *$f_d : \mathbb{N} \rightarrow \mathbb{R}_{>0}$ is an 'anti-coordination scaling function'. It has minimum $f_d(N) > 0$, and increases monotonically with the number of agents whose policies are different from that of i. An anti-coordination game is one where the agents' return can be decomposed as follows, $\forall i, j \in \{1, \ldots, N\}$: $V^i(\boldsymbol{\pi}, \mu_{\bar{t}}) = h\left(b(\pi^i), f_d\left(\sum_{j\in\{1,\ldots,N\}} \mathbb{I}\left[\pi^i = \pi^j\right]\right)\right)$, where $h : \mathbb{R}_{\geq 0} \times \mathbb{R}_{>0} \rightarrow \mathbb{R}_{\geq 0}$ 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, results in a higher return.*

For simplicity of the theory, we make several assumptions giving conditions under which networked agents *do* outperform centralised ones (for reasons of space these are detailed fully in Ass. 1-6 of Appx. E). These assumptions do not always hold in practice, which explains why networked agents may not always outperform centralised ones, though they do in the majority of our experiments.

Recall that at each iteration k of Alg. 2, after independently updating their policies in Line 16, the population has the policies $\{\pi_{k+1}^i\}_{i=1}^N$. Ass. 5 assumes that after the C_p policy exchange rounds in Lines 8-15 (Alg. 4), the networked population is left with a single policy. Call this consensus policy π_{k+1}^{net}. Recall that the centralised case is where the updated Q-network of arbitrary agent $i = 1$ is automatically pushed to all the others instead of the policy evaluation and exchange in Lines 1-15

[2] 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.

(Alg. 4); this is equivalent to a networked case where policy consensus is reached on a *random* one of the policies $\{\pi^i_{k+1}\}_{i=1}^N$. Call this policy *arbitrarily* given to the whole population π^{cent}_{k+1}.

Theorem 1. *In coordination and anti-coordination games where Ass. 1, 2, 3, 4, 5 and 6 (Appx. E) apply, we have* $\mathbb{E}[W(\pi^{\text{net}}_{k+1}, I(\pi^{\text{net}}_{k+1}))] > \mathbb{E}[W(\pi^{\text{cent}}_{k+1}, I(\pi^{\text{cent}}_{k+1}))]$ *(i.e. in expectation networked agents will increase their returns faster than centralised ones). Full proof in Appx. F.1.*

We now give results showing why learning can be faster in the networked than the independent case.[3] However, since we cannot expect independent agents to share a single policy π_{k+1} after the update in each iteration, it is not possible to extract a solution to the MFC problem from the independent policies (a further weakness of the independent case). We therefore give these results in terms of the population-average return (Def. 3) instead of the social welfare (Def. 5) as before. We say the joint policy in the networked case after communication round c is $\boldsymbol{\pi}^{\text{net}}_{k+1,c} = \left(\pi^{(1,\text{net})}_{k+1,c}, \ldots, \pi^{(N,\text{net})}_{k+1,c}\right)$, and the joint policy in the independent case is $\boldsymbol{\pi}^{\text{ind}}_{k+1} = \left(\pi^{(1,\text{ind})}_{k+1}, \ldots, \pi^{(N,\text{ind})}_{k+1}\right)$.

Theorem 2. *In a coordination game, given Ass. 2, 3, 4 and 7 (Appx. E), even a single round of communication in the networked case improves on the independent case, i.e. for* $c = 0$, $\mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{net}}_{k+1,c+1}, \mu_t)\right] > \mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{ind}}_{k+1}, \mu_t)\right]$. *Full proof in Appx. F.2.*

Theorem 3. *In an anti-coordination game, given Ass. 2, 3, 4, 7 and 8 (Appx. E), even a single round of communication in the networked case improves on the independent case, i.e. for* $c = 0$, $\mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{net}}_{k+1,c+1}, \mu_t)\right] > \mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{ind}}_{k+1}, \mu_t)\right]$. *Full proof in Appx. F.3.*

5 Experiments

5.1 Experimental setup

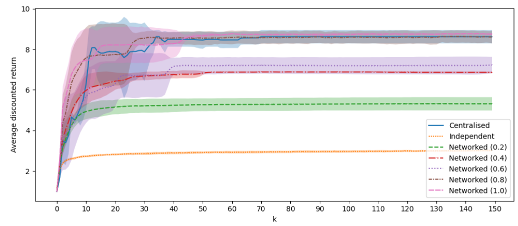
We present experiments from grid worlds, following the gold standard in similar works on MFGs (Laurière et al., 2022a). We give results from six tasks similar to those found in prior works, defined by the agents' reward/transition functions and relating to agents' positions relative to other agents. Two (*cluster*; *target selection*) are coordination games and four (*disperse*; *target coverage*; *beach bar*; *shape formation*) 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. 4). Appx. H.1 has full technical descriptions of the tasks. In these spatial environments, $\mathcal{G}^{comm}_t$ is determined by the physical distance from i; we show plots for various broadcast radii, given as fractions of the maximum distance in the grid. We evaluate our experiments according to a finite-step approximation of the discounted population-average return (Def. 3) over M steps within each outer k loop, i.e. $\hat{V}^{pop}(\boldsymbol{\pi}_k, \mu_t; M)$.

5.2 Results and discussion

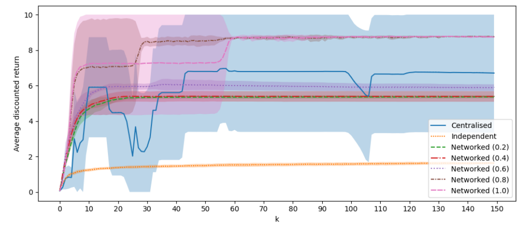
We present results in Fig. 1 for our standard experimental settings involving 500 agents each with their own Q-network. When networked agents communicate, they have only a *single* communication round. See Appx. H.3 for additional experiments with more communication rounds, a study of robustness to failures in the communication network, and ablation studies for our various sub-routines. The ablation studies of Algs. 1 (estimating global average reward) and 3 (estimating global empirical mean field) suggest that in our experimental settings the policy communication scheme (Alg. 4) is the dominant factor in the better performance of networked populations over the other architectures.

Fig. 1 shows that in all of our games, networked populations of *all* broadcast radii significantly outperform independent (orange) agents, which hardly appear to increase their returns, if at all. Networked populations of *all* broadcast radii also significantly outperform the centralised (blue) agents in all but the two coordination games, where only networked agents of the smaller radii

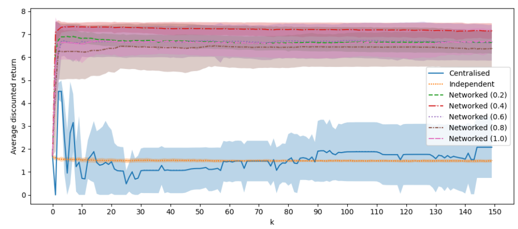
[3] Here we replace Ass. 1 with Ass. 7, Appx. E.2. To prove the benefit of the networked case over the independent case in anti-coordination games, we use an additional Ass. 8, Appx E.2.1.



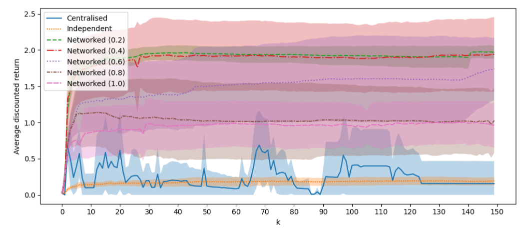
(a) ‘Cluster’ game.



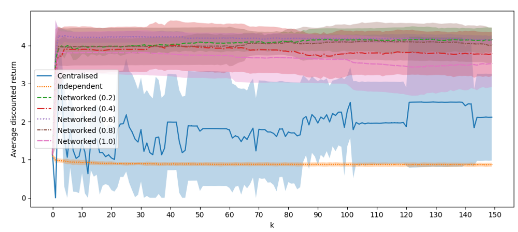
(b) ‘Target selection’ game.



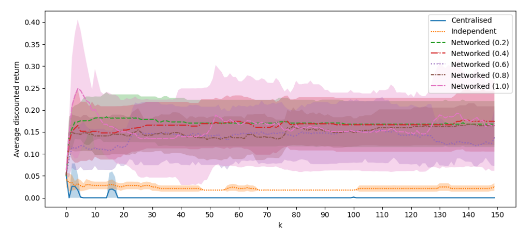
(c) ‘Disperse’ game.



(d) ‘Target coverage’ game.



(e) ‘Beach bar’ game.



(f) ‘Shape formation’ game.

Figure 1: Standard settings with $C_e = C_r = C_p = 1$. In almost all games networked agents of all broadcast radii significantly outperform the centralised (blue) and independent (orange) agents.

(green, 0.2; red, 0.4; purple, 0.6) underperform them. Indeed, in the anti-coordination games the centralised populations perform similarly to purely independent ones in hardly appearing to increase their returns, performing even worse than independent ones in the ‘shape formation’ game. The centralised populations also have markedly higher variance than networked ones in several games (‘target selection’, ‘disperse’, ‘beach bar’). This reflects our theoretical analysis in Sec. 4 that the centralised 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 four anti-coordination games, and most notably in the ‘target coverage’ game, networked agents of smaller broadcast radii often outperform those of larger radii, i.e. the ordering is reversed from that of the coordination games. This reflects the fact that our strong theoretical Ass. 8 (namely that an increase in the base return function must outweigh a decrease in the population’s policy diversity in anti-coordination games) only applies to a certain extent in practice, explained below.

The fact that a single round of communication improves return over the independent case in anti-coordination games reflects Ass. 8 holding for Thm. 3, in that for all networked populations the increase in average base policy quality outweighs the decrease in diversity. However, the different communication radii lead to different degrees of consensus after a single round, and hence different decreases in diversity. Beyond a certain point, maintaining some diversity does in fact outweigh the benefit of all agents using the policy that has the best base quality for a given iteration. Some policy sharing is better than none, but too much may be a disadvantage in anti-coordination games. 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. 6), where a single policy

π^* 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) 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.

6 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 independent and centralised architectures. We supported this with extensive numerical results, accompanied by ablation studies and discussion of the empirical effects of communication radii. For future work, see Appx. I.

References

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.

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.

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.

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.

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.

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 *Proceed-*

ings of Machine Learning Research, pp. 12078–12095. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/lauriere22a.html.

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.

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.

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.

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, 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.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Diagram for two possible conceptions of our work

See Fig. 2.

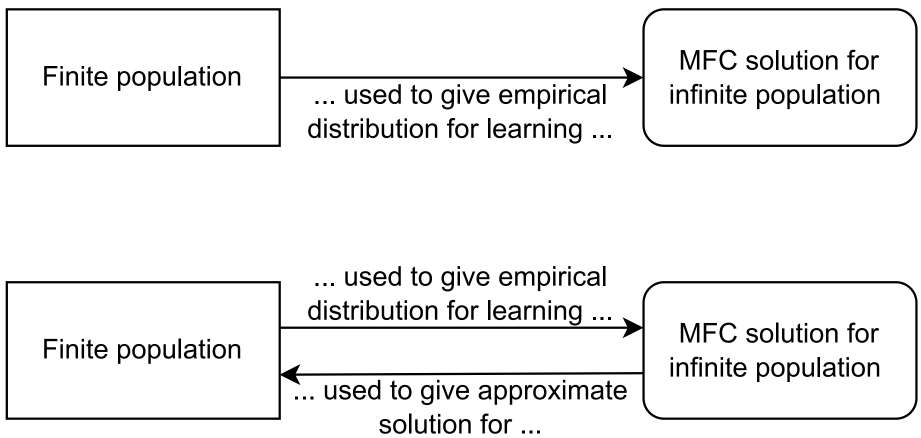


Figure 2: Two possible ways to conceive of our work regarding the relationship between the infinite- and finite-population control problems, described in Sec. 2. Note that using the finite empirical population to learn the single-policy MFC social optimum $\boldsymbol{\pi} = (\pi^*, \ldots, \pi^*)$ for the infinite population (Def. 6) is *not* the same as directly finding $\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi} \in \Pi^N} V^{pop}(\boldsymbol{\pi}, \mu_{\bar{t}}) = (\pi^1, \ldots, \pi^N)$, i.e. the tuple of *individual* policies that maximises the expected finite population-average return in Def. 3, a problem known to be hard (Cui et al., 2023c; Bernstein et al., 2002).

B Preliminaries on Munchausen Online Mirror Descent

Recent works have solved MFGs from non-episodic runs of the finite 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 π_0, and then at each iteration k, evaluating the current policy π_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 π_{k+1} to be the softmax over this weighted sum, i.e. $\pi_{k+1}(\cdot|o) = softmax\left(\frac{1}{\tau_q}\sum_{\kappa=0}^{k+1} Q_\kappa(o, \cdot)\right)$. τ_q is a temperature parameter that scales the entropy in Munchausen RL (see below) (Vieillard et al., 2020); this is a different temperature to the one agents use when communicating policies, denoted τ_k^{comm} and discussed in Sec. 3.3.

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 π_k and π_{k+1} (Vieillard et al., 2020). Using this reparametrisation gives Munchausen OMD (MOMD), detailed in Sec. 3.2 (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).

C Algorithms

C.1 Sub-routine for networked estimation of global average reward

See Alg. 1, discussed in Sec. 3.1.

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

$\forall i$: Initialise reward sets $\hat{\mathcal{R}}_t^i \leftarrow \{(ID^i, r_t^i)\}$

for c_r in $1, \ldots, C_r$ **do**

  $\forall i$: Broadcast $\hat{\mathcal{R}}_{t,c_r}^i$

  $\forall i$: $J_t^i \leftarrow \{j \in \mathcal{N} : (i,j) \in \mathcal{E}_t^{comm}\}$

  $\forall i$: $\hat{\mathcal{R}}_{t,(c_r+1)}^i \leftarrow \hat{\mathcal{R}}_{t,c_r}^i \cup \bigcup_{j \in J_t^i} \hat{\mathcal{R}}_{t,c_r}^j$

end for

$\forall i$: $\tilde{\hat{r}}_t^i \leftarrow \frac{1}{|\hat{\mathcal{R}}_{t,C_r}^i|} \sum_{(ID,r) \in \hat{\mathcal{R}}_{t,C_r}^i} r$

return Estimates of average reward $\left\{\tilde{\hat{r}}_t^i\right\}_{i=1}^N$

C.2 Main learning algorithm

See Alg. 2, discussed in Sec. 3.2.

Definition 10 (Q-network empirical loss)**.** *The training loss to be minimised is given by* $\hat{\mathcal{L}}(\theta, \theta') = \frac{1}{|B|} \sum_{transition \in B_{k,l}^i} \left| \check{Q}_{\theta_{k,l}^i}(o_t, a_t) - T \right|^2$, *where*

$$T = \tilde{\hat{r}}_t + \left[\tau_q \ln \pi_{\theta_{k,l}^{i,\prime}}(a_t|o_t) \right]_{cl}^0 + \gamma \sum_{a \in \mathcal{A}} \pi_{\theta_{k,l}^{i,\prime}}(a|o_{t+1}) \left(\check{Q}_{\theta_{k,l}^{i,\prime}}(o_{t+1}, a) - \tau_q \ln \pi_{\theta_{k,l}^{i,\prime}}(a|o_{t+1}) \right).$$

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).

C.3 Sub-routine for networked estimation of global empirical mean-field

Networked agents use Alg. 3 (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. To do so, we say that the population exhibits the following visibility graph, in addition to its communication network.

Definition 11 (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.*

This graph applies in the 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. Benjamin & Abate (2024) additionally contains a graph and algorithm for more general settings.

In our experiments in spatial environments, the visibility graph $\mathcal{G}_t^{vis}$ is determined by the physical distance from agent i, as with the communication network $\mathcal{G}_t^{comm}$. In the independent architecture, we assume there are no links in $\mathcal{G}_t^{vis}$, i.e. $\mathcal{E}_t^{vis} = \emptyset$.

Alg. 3 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

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 $\gamma, \tau_q, |B|, cl, \nu, \{\tau_k^{comm}\}_{k \in \{0,\dots,K-1\}}$

Require: initial states $\{s_0^i\}_{i=1}^N$; $t \leftarrow 0$

$\forall i$: Randomly initialise parameters θ_0^i of Q-networks $\check{Q}_{\theta_0^i}(o,\cdot)$, and set $\pi_0^i(a|o) = \text{softmax}\left(\frac{1}{\tau_q}\check{Q}_{\theta_0^i}(o,\cdot)\right)(a)$

for $k = 0, \dots, K-1$ **do**

  $\forall i$: Empty i's buffer

  for $m = 0, \dots, M-1$ **do**

    $\{o_t^i\}_{i=1}^N \leftarrow$ **EstimateMeanFieldAlg. 3**$\left(\mathcal{G}_t^{vis}, \mathcal{G}_t^{comm}, \{s_t^i\}_{i=1}^N\right)$

    Take step $\forall i : a_t^i \sim \pi_k^i(\cdot|o_t^i), r_t^i = R(s_t^i, a_t^i, \hat{\mu}_t), s_{t+1}^i \sim P(\cdot|s_t^i, a_t^i, \hat{\mu}_t); t \leftarrow t+1$

    $\{\tilde{\hat{r}}_t^i\}_{i=1}^N \leftarrow$ **EstimateAverageRewardAlg. 1**$\left(\mathcal{G}_t^{comm}, \{r_t^i\}_{i=1}^N\right)$

    $\forall i$: Add $\left(o_t^i, a_t^i, \tilde{\hat{r}}_t^i, o_{t+1}^i\right)$ to i's buffer

  end for

  for $l = 0, \dots, L-1$ **do**

    $\forall i$: Sample batch $B_{k,l}^i$ from i's buffer

    Update θ to minimise $\hat{\mathcal{L}}(\theta, \theta')$ as in Def. 10

    If $l \mod \nu = 0$, set $\theta' \leftarrow \theta$

  end for

  $\check{Q}_{\theta_{k+1}^i}(o,\cdot) \leftarrow \check{Q}_{\theta_{k,L}^i}(o,\cdot)$

  $\forall i : \pi_{k+1}^i(a|o) \leftarrow \text{softmax}\left(\frac{1}{\tau_q}\check{Q}_{\theta_{k+1}^i}(o,\cdot)\right)(a)$

  $\left(\{\pi_{k+1}^i\}_i, \{s_t^i\}_i, t\right) \leftarrow$ **CommunicatePolicyAlg. 4**$\left(\mathcal{G}_t^{comm}, \{\pi_{k+1}^i\}_i, \{s_t^i\}_i, t\right)$

end for

return policies $\{\pi_K^i\}_{i=1}^N$

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 3-7). 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 8-10). 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 11-12).

C.4 Sub-routine for communicating and refining policies

See Alg. 4, described in Sec. 3.3.

D Informal intuitions regarding formal results

Remark 2. *Like many cooperative learning paradigms, both the independent and centralised versions of our core learning algorithm (Alg. 2) 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. While 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, this does not solve the issue of allowing learners to distinguish between helpful or unhelpful local actions a_t^i, whether centralised or not. By spreading policies through the population which are expected to have a higher individual return, despite this being a cooperative*

Algorithm 3 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

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

Algorithm 4 Policy communication and selection

Require: Time-dependent communication graph $\mathcal{G}_t^{comm}$, loop parameters E, C_p, learning parameters γ, $\{\tau_k^{comm}\}_{k \in \{0,\dots,K-1\}}$
Require: policies $\{\pi_{k+1}^i\}_{i=1}^N$; states $\{s_t^i\}_{i=1}^N$; t

$\forall i : \sigma_{k+1}^i \leftarrow 0$
for $e = 0, \dots, E-1$ evaluation steps **do**
  $\{o_t^i\}_{i=1}^N \leftarrow$ **EstimateMeanFieldAlg. 3**$\big(\mathcal{G}_t^{vis}, \mathcal{G}_t^{comm}, \{s_t^i\}_{i=1}^N\big)$
  Take step $\forall i : a_t^i \sim \pi_k^i(\cdot|o_t^i), r_t^i = R(s_t^i, a_t^i, \hat{\mu}_t), s_{t+1}^i \sim P(\cdot|s_t^i, a_t^i, \hat{\mu}_t)$
  $\forall i : \sigma_{k+1}^i \leftarrow \sigma_{k+1}^i + \gamma^e \cdot r_t^i$
  $t \leftarrow t+1$
end for
for C_p rounds **do**
  $\forall i$: Broadcast $\sigma_{k+1}^i, \pi_{k+1}^i$
  $\forall i : J_t^i \leftarrow \{j \in \mathcal{N} : (i,j) \in \mathcal{E}_t^{comm}\}$
  $\forall i$: Select $\text{adopted}^i \sim \Pr\big(\text{adopted}^i = j\big) = \frac{\exp(\sigma_{k+1}^j / \tau_k^{comm})}{\sum_{x \in J_t^i} \exp(\sigma_{k+1}^x / \tau_k^{comm})} \; \forall j \in J_t^i$
  $\forall i : \sigma_{k+1}^i \leftarrow \sigma_{k+1}^{\text{adopted}^i}, \pi_{k+1}^i \leftarrow \pi_{k+1}^{\text{adopted}^i}$
  $\{o_t^i\}_{i=1}^N \leftarrow$ **EstimateMeanFieldAlg. 3**$\big(\mathcal{G}_t^{vis}, \mathcal{G}_t^{comm}, \{s_t^i\}_{i=1}^N\big)$
  Take step $\forall i : a_t^i \sim \pi_k^i(\cdot|o_t^i), r_t^i = R(s_t^i, a_t^i, \hat{\mu}_t), s_{t+1}^i \sim P(\cdot|s_t^i, a_t^i, \hat{\mu}_t)$; $t \leftarrow t+1$
end for
return (policies $\{\pi_{k+1}^i\}_{i=1}^N$, states $\{s_t^i\}_{i=1}^N$, t)

problem, we reduce the credit-assignment problem by replicating 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 centralised 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. Our decentralised networked approach permits beneficial parallelisation in place of this centralised method, by gener-

ating 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.

E Theoretical assumptions

Recall that at each iteration k of Alg. 2, after independently updating their policies in Line 16, the population has the policies $\{\pi^i_{k+1}\}^N_{i=1}$. There is randomness in these independent policy updates, stemming from the random sampling of each agent's independently collected buffer. In Lines 1-7 of Alg. 4, agents approximate their infinite discounted individual returns $\{V^i(\boldsymbol{\pi}, \mu_t)\}^N_{i=1}$ (Def. 2) of their updated policies by computing $\{\sigma^i_{k+1}\}^N_{i=1}$: the E-step discounted return with respect to the *empirical* mean field generated when agents follow policies $\{\pi^i_{k+1}\}^N_{i=1}$ (i.e. they do not at this stage all follow a single identical policy).

E.1 General assumptions

Assumption 1. *Assume that Algs. 1 and 3 allow networked agents to obtain accurate estimations of the true population-average rewards and global empirical mean field respectively, i.e.* $\forall i\ \tilde{\hat{\mu}}^i_t = \hat{\mu}_t$ *and* $\tilde{\hat{r}}^i_t = \hat{r}$.[4]

Assumption 2. *Assume that* $\{\sigma^i_{k+1}\}^N_{i=1}$ *are sufficiently good approximations so as to respect the ordering of the true infinite discounted individual returns* $\{V^i(\boldsymbol{\pi}_{k+1}, \mu_t)\}^N_{i=1}$,*i.e.* $\forall i, j \in \{1, \ldots, N\}$ $\sigma^i_{k+1} > \sigma^j_{k+1} \iff V^i(\boldsymbol{\pi}_{k+1}, \mu_t) > V^j(\boldsymbol{\pi}_{k+1}, \mu_t)$.

Assumption 3. *Assume that directly after the policy updates in Line 16 (Alg. 2), before any policy transfer as in the networked or centralised algorithms, all policies are different (due to the randomness in these updates). This means we have the minimum possible value of the* f_c *function and the maximum possible value of the* f_d *function. Assume also that each policy has a distinct return, such that* $\forall i, j \in \{1, \ldots, N\}$ $\pi^i_{k+1} \neq \pi^j_{k+1}, V^i(\boldsymbol{\pi}_{k+1}, \mu_t) \neq V^j(\boldsymbol{\pi}_{k+1}, \mu_t), \sigma^i_{k+1} \neq \sigma^j_{k+1}$.

Assumption 4. *Say that* $\tau^{comm}_k \in \mathbb{R}_{\geq 0}$, *such that the softmax adoption scheme (Line 11, Alg. 4) gives non-uniform probabilities of policies being adopted as they are exchanged among neighbours.*

Assumption 5. *Assume that after the* C_p *rounds in Lines 8-15 (Alg. 4), in which agents exchange and adopt policies from neighbours, the networked population is left with a single policy such that* $\forall i, j \in \{1, \ldots, N\}$ $\pi^i_{k+1} = \pi^j_{k+1}$.[5]

Assumption 6. *We have two different policies that could be shared by the whole population such that* $\boldsymbol{\pi}^x = (\pi^x, \ldots, \pi^x)$ *and* $\boldsymbol{\pi}^y = (\pi^y, \ldots, \pi^y)$. *We assume that:*

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

E.2 Assumptions for outperformance of the independent case

Assumption 7. *Assume the estimated global mean field and average reward are the same in the networked and independent cases, i.e.* $\forall i, j\ \tilde{\hat{\mu}}^{(i,net)}_t = \tilde{\hat{\mu}}^{(j,ind)}_t$ *and* $\tilde{\hat{r}}^{(i,net)}_t = r^i_t$.[6]

[4] In other words, we assume for simplicity that the only difference between the networked and centralised cases is the networked policy communication scheme. In practice, our ablation studies indicate that this is empirically the dominant factor in our experimental settings anyway.

[5] Most simply we can think of Ass. 5 holding if 1) $\tau^{comm}_k \to 0\ \forall k$, 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 equal to 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 σ^{max}_{k+1} value (and hence the unique associated π^{max}_{k+1}) within a number of rounds equal to the diameter $d_{\mathcal{G}^{comm}_t}$ (Nejad et al., 2009; Benjamin & Abate, 2023). However, policy consensus as in Ass. 5 might be achieved even outside of these conditions, including if the network is dynamic and not connected at every step, given appropriate values for C_p and $\tau^{comm}_k \in \mathbb{R}_{>0}$.

[6] In other words, we assume for simplicity that the only difference between the networked and independent cases is the networked policy communication scheme. *In practice, the networked estimates will be better due to communication, giving an additional performance increase over the independent case.*

E.2.1 Assumption for outperformance of the independent case in anti-coordination games

Assumption 8. *Assume that an increase in the base return function outweighs a decrease in the population's policy diversity, namely* $h(b + \Delta b, f_d - \Delta f_d) > h(b, f_d), \ \forall \Delta b > 0, \Delta f_d > 0$.

F Proofs

Ass. 5 assumes that after the C_p policy exchange rounds in Lines 8-15 (Alg. 4), the networked population is left with a single policy. Call this consensus policy π^{net}_{k+1}, and its associated finitely approximated return $\sigma^{\text{net}}_{k+1}$. Recall that the centralised 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 Lines 1-15 (Alg. 4); 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 π^{cent}_{k+1}, and its associated finitely approximated return $\sigma^{\text{cent}}_{k+1}$.

F.1 Proof of Thm. 1

Theorem 1. *In coordination games and anti-coordination games where Ass. 1, 2, 3, 4, 5 and 6 apply, we have* $\mathbb{E}[W(\pi^{\text{net}}_{k+1}, I(\pi^{\text{net}}_{k+1}))] > \mathbb{E}[W(\pi^{\text{cent}}_{k+1}, I(\pi^{\text{cent}}_{k+1}))]$ *(i.e. in expectation networked agents will increase their returns faster than centralised ones).*

Proof. Recall that before the communication rounds in Line 8 (Alg. 4), the randomly updated policies $\{\pi^i_{k+1}\}^N_{i=1}$ have associated approximated returns $\{\sigma^i_{k+1}\}^N_{i=1}$. Denote the mean and maximum of this set $\sigma^{\text{mean}}_{k+1}$ and $\sigma^{\text{max}}_{k+1}$ respectively. Since π^{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, the softmax adoption probability (Line 11, Alg. 4) for the networked case means by definition that policies with higher σ^i_{k+1} are more likely to be adopted at each communication round. Thus the consensus π^{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 $\tau^{comm}_{k+1} \to 0$, it will obey $\mathbb{E}[\sigma^{\text{net}}_{k+1}] = \sigma^{\text{max}}_{k+1} \ \forall k$). As such:

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

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

Given Eq. 1, and by Ass. 2, we know that directly after the policy update in Line 16 (Alg. 2), *prior to the consensus being reached*, we have:

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

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

By Ass. 3 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. 8 it must be because:

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

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

By Ass. 5 know that in the networked and centralised cases the joint policies respectively become $\boldsymbol{\pi}^{\text{net}} := (\pi^{\text{net}}, \pi^{\text{net}}, \pi^{\text{net}}, \ldots)$ and $\boldsymbol{\pi}^{\text{cent}} := (\pi^{\text{cent}}, \pi^{\text{cent}}, \pi^{\text{cent}}, \ldots)$. We therefore end up with maximum alignment in both cases, such that f_c^{net} = $f_c^{\text{cent}} = \max f_c$ in a coordination game, and f_d^{net} = $f_d^{\text{cent}} = \min f_d$ in an anti-coordination game. Due to this, along with Eqs. 2 and 3, we have

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

In turn we have:

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

which by Ass. 6 gives

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

namely the result. □

F.2 Proof of Thm. 2

Theorem 2. *In a coordination game, given Ass. 2, 3, 4 and 7, even a single round of communication in the networked case improves on the independent case, i.e. for* $c = 0$, $\mathbb{E}\left[V^{pop}(\boldsymbol{\pi}_{k+1,c+1}^{\text{net}}, \mu_t)\right] > \mathbb{E}\left[V^{pop}(\boldsymbol{\pi}_{k+1}^{\text{ind}}, \mu_t)\right]$.

Proof. The softmax adoption scheme (Line 11, Alg. 4), which according to Ass. 3 and 4 gives non-uniform adoption probabilities, is such that some policies are more likely to be adopted than others. Thus the number of distinct policies in the population is expected to decrease. Say for simplicity that during the first communication round a $\pi_{k+1,c}^{(j,\text{net})}$ 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 Ass. 2 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 Ass. 3 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. 8 pertains, by Def. 8 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 $c = 0$: $\mathbb{E}\left[V^{pop}(\boldsymbol{\pi}_{k+1,c+1}^{\text{net}}, \mu_t)\right] > \mathbb{E}\left[V^{pop}(\boldsymbol{\pi}_{k+1,c}^{\text{net}}, \mu_t)\right]$. 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 $\mathbb{E}\left[V^{(i,\text{net})}(\boldsymbol{\pi}_{k+1,c+1}^{\text{net}}, \mu_t)\right] > \mathbb{E}\left[V^{(i,\text{net})}(\boldsymbol{\pi}_{k+1,c}^{\text{net}}, \mu_t)\right]$, accelerating the improvement even further. 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 $\boldsymbol{\pi}_{k+1}^{\text{ind}} = \boldsymbol{\pi}_{k+1,0}^{\text{net}}$. This gives the result, i.e.

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

□

F.3 Proof of Thm. 3

Theorem 3. *In an anti-coordination game, given Ass. 2, 3, 4, 7 and 8, even a single round of communication in the networked case improves on the independent case, i.e. for* $c = 0$, $\mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{net}}_{k+1,c+1}, \mu_t)\right] > \mathbb{E}\left[V^{pop}(\boldsymbol{\pi}^{\text{ind}}_{k+1}, \mu_t)\right]$.

Proof. The proof begins similarly to that for a coordination game. The softmax adoption scheme (Line 11, Alg. 4), which according to Ass. 3 and 4 gives non-uniform adoption probabilities, is such that some policies are more likely to be adopted than others. Thus the number of distinct policies in the population is expected to decrease. 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 $c = 0$

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

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

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 Ass. 2 that

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

By Ass. 3 we know that straight after the random policy updates there is no alignment among policies, i.e. in the anti-coordination game we have $f^{(i,\text{net})}_d = f^{(j,\text{net})}_d = \max f_d$. Therefore if Eq. 8 pertains, by Def. 8 it must be because:

$$\mathbb{E}[b(\pi^{(i,\text{net})})] > \mathbb{E}[b(\pi^{(j,\text{net})})], \tag{9}$$

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

Ass. 8 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.

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

□

G Extended comparison with related work

We discuss here the works 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, 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 reliance which 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 Tajeddini et al. (2017), which presents a leader-follower setting where a 'central population coordinator' estimates the mean-field trajectory. Farzaneh et al. (2020) requires a central coordinator, and 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 to 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 algorithms require 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 - 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 gain higher rewards by diversifying their behaviour.

Angiuli et al. (2022; 2023) provide algorithms for MFC learning from a single run, but here it is a single run only of a 'representative' player that estimates 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 time-scales for updating the mean-field estimate, 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 communication of 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 integrate Alg. 3 from Benjamin & Abate (2024) for estimating the global mean field from a local neighbourhood, but additionally contribute novel Alg. 1 for estimating the global average reward from a local neighbourhood for the MFC setting.

H Experiments

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. We discuss hyperparameters in Appx. H.2.

H.1 Games

We conduct numerical tests with six games. In all cases, rewards are normalised in [0,1] after they are computed.

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.

Agree on a single target. 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

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

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

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). Agents are rewarded for being stationary in sparsely populated locations as close as possible to a target ϕ_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 ϕ_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

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

H.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.

As in the related works on networked communication in the MFG setting by Benjamin & Abate (2023; 2024), in our experiments we generally want to demonstrate that our communication-based algorithms outperform the centralised and independent architectures by allowing policies that are estimated to be better performing to proliferate through the population, such that convergence occurs within fewer iterations and computationally faster, even when the Q-function is poorly approximated and/or the mean field is poorly estimated, as is likely to be the case in real-world scenarios. 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 given what might otherwise be considered ‘undesirable’ hyperparameter choices.

H.3 Additional experiments and ablations

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.

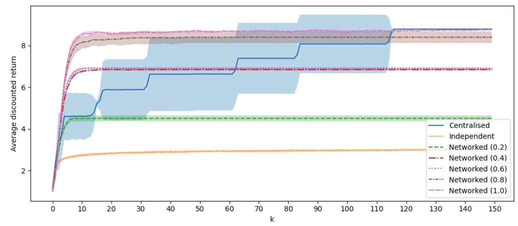
- Robustness to communication failure - Fig. 3.
- Increased communication rounds - Figs. 4 and 5.
- Ablation study with population-independent policies - Fig. 6.
- Ablation study of Alg. 3 for estimating the empirical mean field - Fig. 7.
- Ablation study for observation of true/estimated average reward (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 τ_k^{comm} - Fig. 10.

I Future work

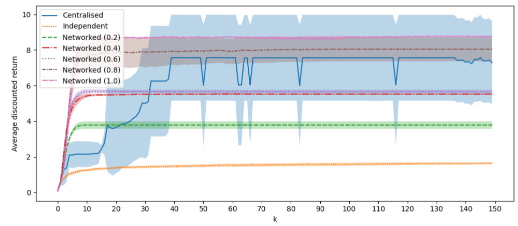
We leave more general theoretical results, such as proofs of convergence and sample complexity, for future work. Future work also includes experiments in other types of game, including more realistic environments and ones where the transition function also depends on the mean field. Our algorithms contain 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

Table 1: Hyperparameters

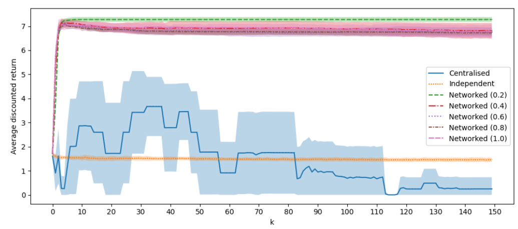
Hyperparam.	Value	Comment
Trials	5	We run 5 trials with different random seeds for each experiment. We plot the mean and standard deviation of the mean for each metric across the seeds.
Gridsize	20x20	-
Population	500	We chose 500 for our demonstrations to show that our algorithm can handle large populations, indeed often larger than those demonstrated in other mean-field works, especially for grid-world environments, while also being feasible to simulate wrt. 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 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 (Sec. H.3).
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 (Sec. H.3).
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 (Sec. H.3).
γ	0.9	Standard choice across RL literature.
τ_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); Benjamin & Abate (2024).
ν	$L-1$	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.
τ_k^{comm}	cf. comment	We follow Benjamin & Abate (2024), where τ_k^{comm} increases linearly from 0.001 to 1 across the K iterations. Further optimising the annealing process may lead to better results; we provide an ablation study in Appx. H.3.



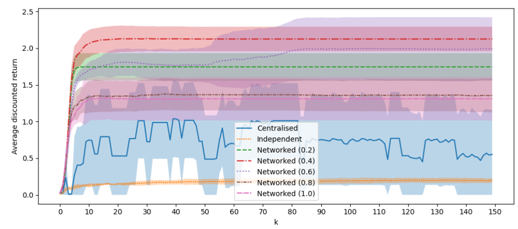
(a) 'Cluster' game.



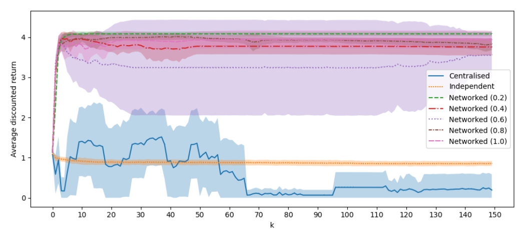
(b) 'Target selection' game.



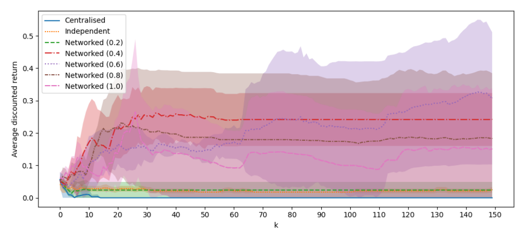
(c) 'Disperse' game.



(d) 'Target coverage' game.



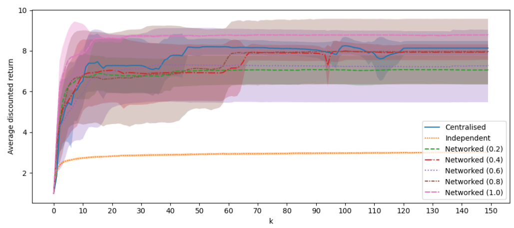
(e) 'Beach bar' game.



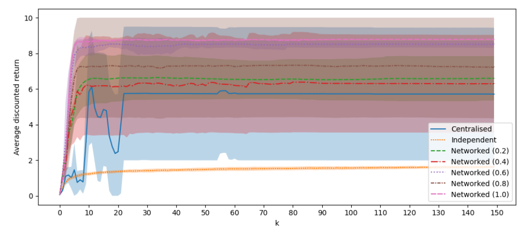
(f) 'Shape formation' game.

Figure 3: All communication links suffer a 90% probability of failure, including in the centralised case, where the link between the central learner and the rest of the population may fail. $C_e = C_r = C_p = 1$. The centralised 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 to performance in the standard setting. In fact, several broadcast radii perform better in the 'shape formation' game with these failures than without, probably because they permit greater diversity policies while still having an advantage over purely independent learners (as discussed in Sec. 5.2). 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 likely 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 centralised 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 centralised case.

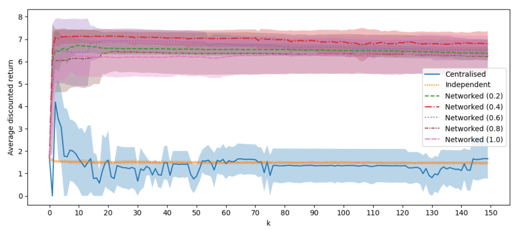
(Fig. 3) indicate that this is not necessarily a problem in practice, but future work nevertheless lies in simplifying the nested loops of our algorithms.



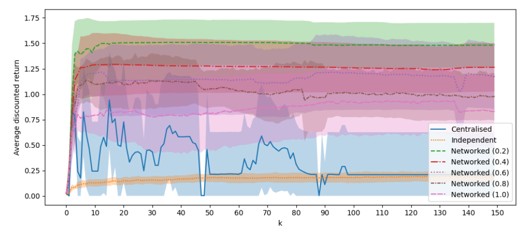
(a) ‘Cluster’ game.



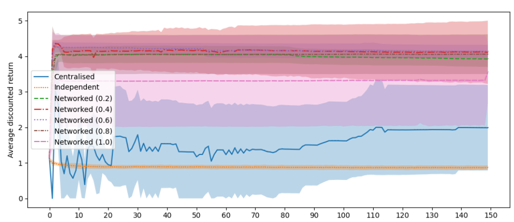
(b) ‘Target selection’ game.



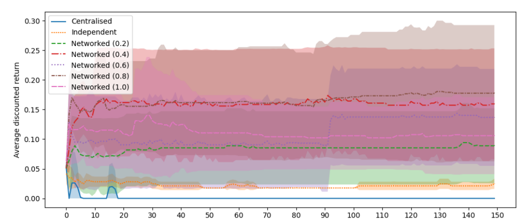
(c) ‘Disperse’ game.



(d) ‘Target coverage’ game.

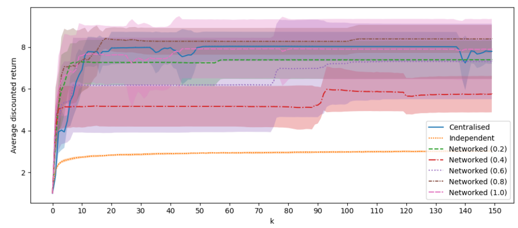


(e) ‘Beach bar’ game.

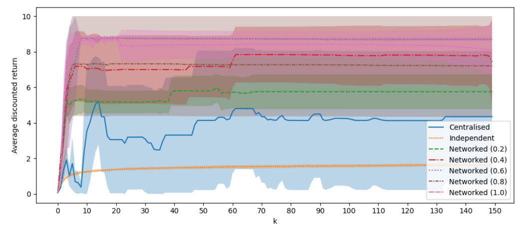


(f) ‘Shape formation’ game.

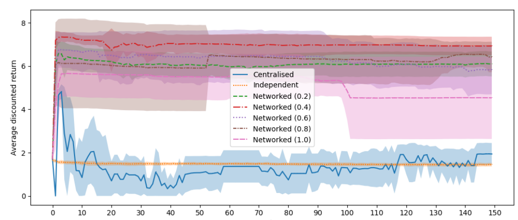
Figure 4: Standard algorithms but $C_e = C_r = C_p = \mathbf{10}$. As is expected, in the coordination games the networked agents with lower broadcast radii now receive returns almost as high as those with larger radii, albeit at the cost of greater variance (as there may be some noise in the quality of the policy that gets spread to the whole population as a result of more communication rounds). In the ‘target selection’ game, now all networked populations outperform the centralised agents. 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 9. The same is true of the smallest radius population (green, 0.2) in the ‘shape formation’ game, which receives a lower return than before. Nevertheless, *all* networked populations receive higher returns than the independent agents in all games, and also than the centralised 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.



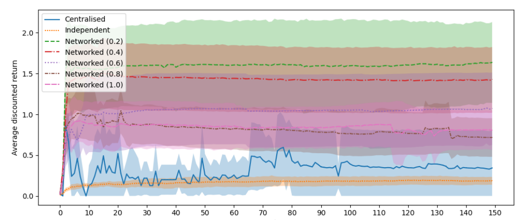
(a) 'Cluster' game.



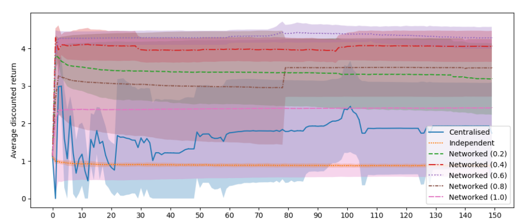
(b) 'Target selection' game.



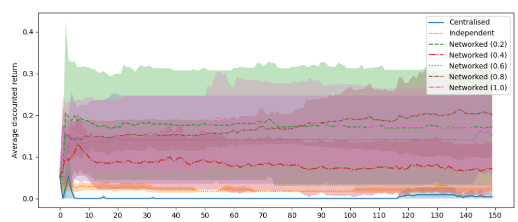
(c) 'Disperse' game.



(d) 'Target coverage' game.

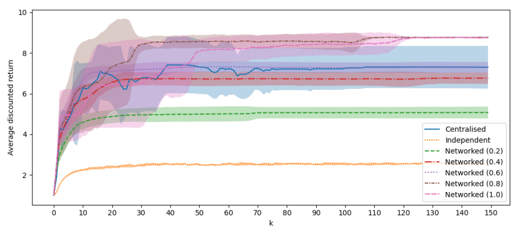


(e) 'Beach bar' game.

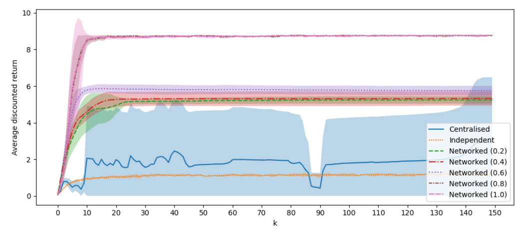


(f) 'Shape formation' game.

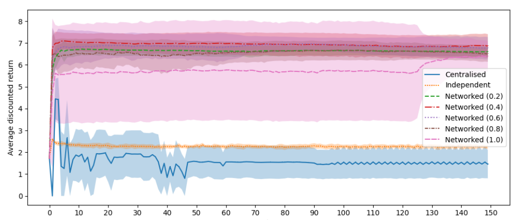
Figure 5: Standard algorithms but $C_e = C_r = C_p = \mathbf{50}$. Having 50 communication rounds does not appear to significantly change networked performance compared to 10 rounds (Fig. 4), with most increases or decreases in average return appearing within the margin of error. Most notably, the largest 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 9. Nevertheless, *all* networked populations receive higher returns than the independent agents in all games, and also than the centralised 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.



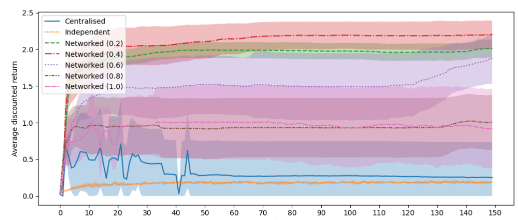
(a) ‘Cluster’ game.



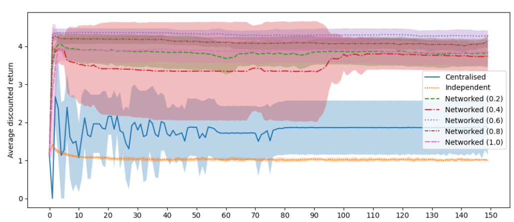
(b) ‘Target selection’ game.



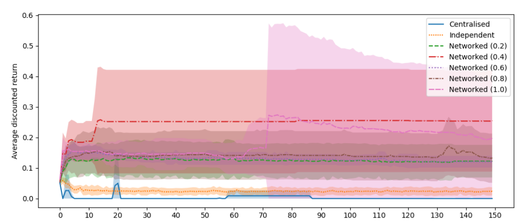
(c) ‘Disperse’ game.



(d) ‘Target coverage’ game.

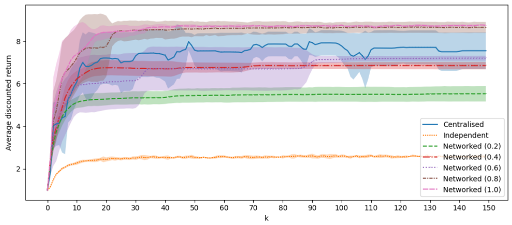


(e) ‘Beach bar’ game.

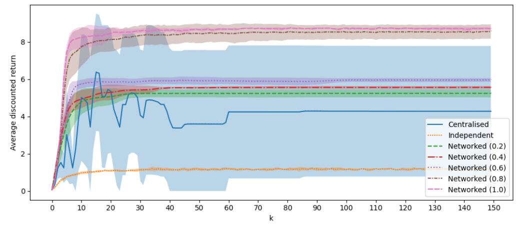


(f) ‘Shape formation’ game.

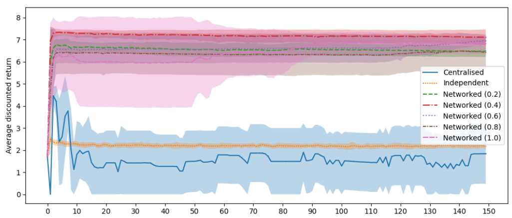
Figure 6: Ablation study on population-*independent* policies. No agents, including centralised and networked ones, observe the empirical mean field, and all receive a vector of zeros in its place (so as to keep the neural networks the same size as in the standard setting). $C_r = C_p = 1$. In our stationary games, 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. On the other hand, in the coordination games, and particularly the ‘target selection’ game, the centralised population receives a significantly lower return in this setting.



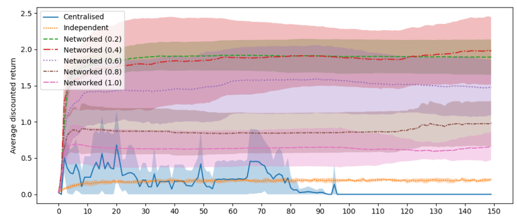
(a) ‘Cluster’ game.



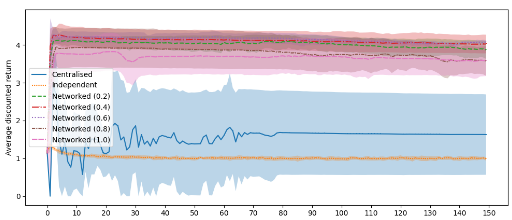
(b) ‘Target selection’ game.



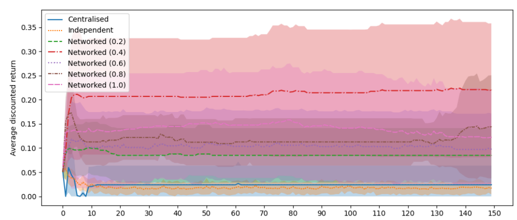
(c) ‘Disperse’ game.



(d) ‘Target coverage’ game.

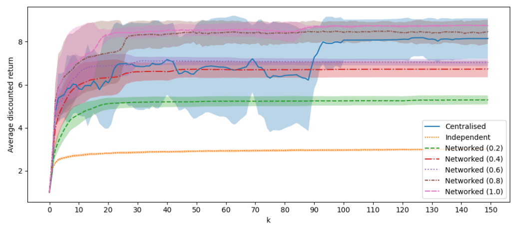


(e) ‘Beach bar’ game.

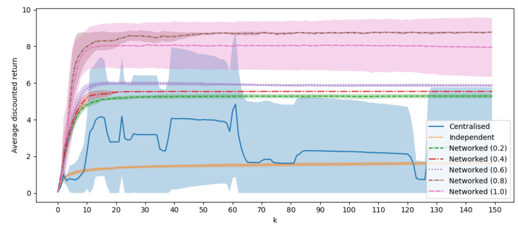


(f) ‘Shape formation’ game.

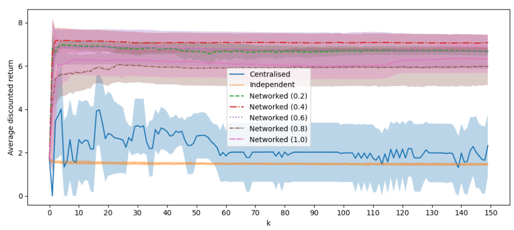
Figure 7: Ablation study of Alg. 3 for estimating the empirical mean field - all agents, including independent ones, directly receive the true global empirical mean field. $C_r = C_p = 1$. This does not appear to change performance in the networked populations (apart from greater variance here in the ‘shape formation’ game), nor does it help independent agents. This may be evidence that Alg. 3 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 centralised populations in the coordination games). Therefore further evidence is required in settings that require population-dependent policies to confirm the efficacy of Alg. 3.



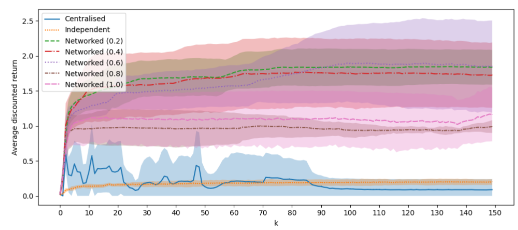
(a) ‘Cluster’ game.



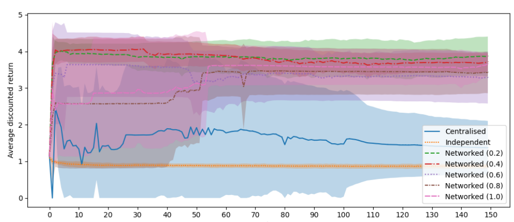
(b) ‘Target selection’ game.



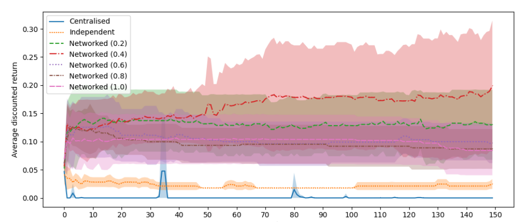
(c) ‘Disperse’ game.



(d) ‘Target coverage’ game.

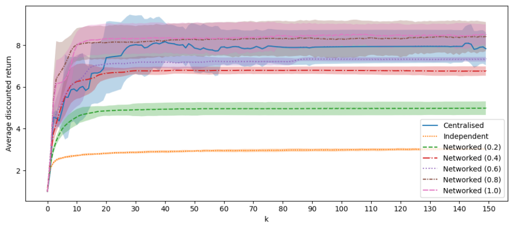


(e) ‘Beach bar’ game.

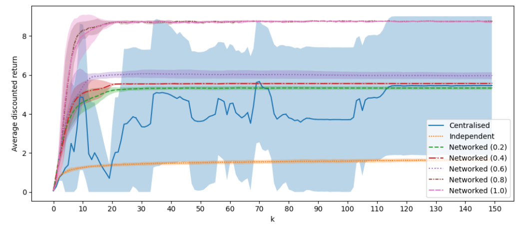


(f) ‘Shape formation’ game.

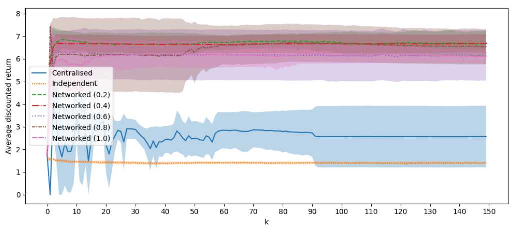
Figure 8: Ablation study for observation of true/estimated global average reward $\hat{r}/\tilde{\hat{r}}^i_t$, where all agents, including centralised ones, only have access to r^i_t, where in the centralised case $i = 1$. $C_e = C_p = 1$. The greatest effect of this is on the *centralised* (blue) case, which performs much worse in the ‘target selection’ game, and with higher variance in the ‘cluster’ and ‘beach bar’ games. The networked agents appear more robust, 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 centralised case in terms of $\tilde{\hat{r}}^i_t$, 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.



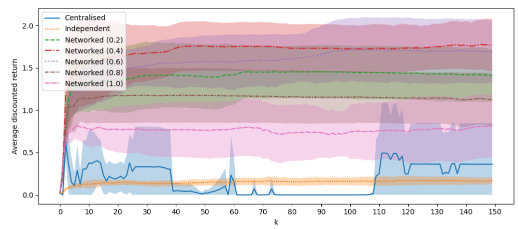
(a) 'Cluster' game.



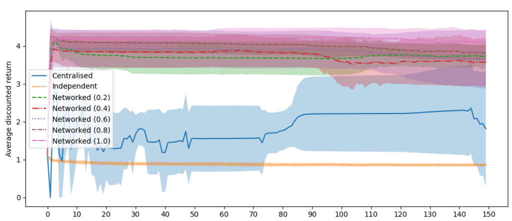
(b) 'Target selection' game.



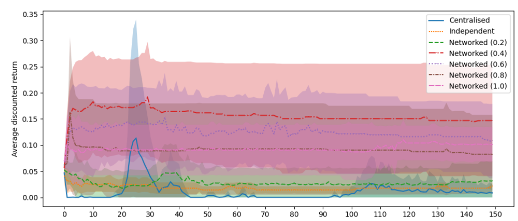
(c) 'Disperse' game.



(d) 'Target coverage' game.

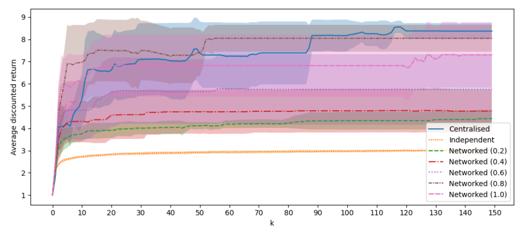


(e) 'Beach bar' game.

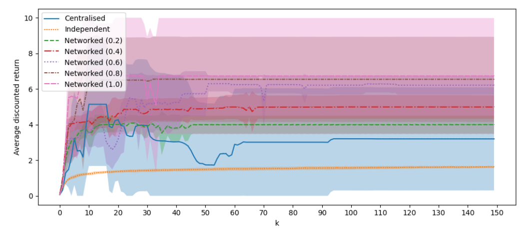


(f) 'Shape formation' game.

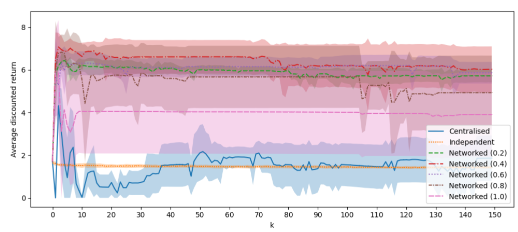
Figure 9: Ablation study for Alg. 1 for estimating the true global average reward. All agents, including both networked and independent ones, directly receive the true global average reward such that $\tilde{\hat{r}}^i_t = \hat{r}$. Access to the true average reward does not help networked (or independent) 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). In fact, networked populations' performance actually seems to be worse with this global information in the 'shape formation' game, particularly in the case with the smallest broadcast radius (green, 0.2), but perhaps not by a statistically significant amount.



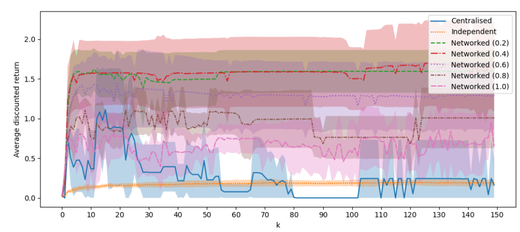
(a) ‘Cluster’ game.



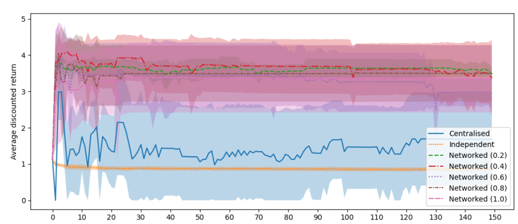
(b) ‘Target selection’ game.



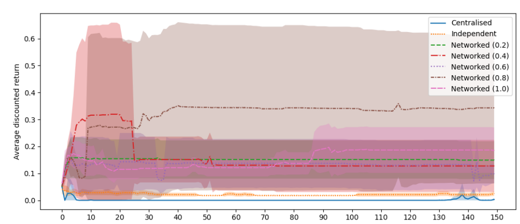
(c) ‘Disperse’ game.



(d) ‘Target coverage’ game.



(e) ‘Beach bar’ game.



(f) ‘Shape formation’ game.

Figure 10: Ablation study of the choice of τ_k^{comm}. Here $\forall k$ τ_k^{comm} = 1e-18 (i.e. $\tau_k^{comm} \rightarrow 0$), 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 centralised (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 Ass. 2 on the quality of the finite-step approximations $\{\sigma_{k+1}^i\}_{i=1}^N = \{\hat{V}^i(\boldsymbol{\pi}_{k+1}, \mu_t; E)\}_{i=1}^N$ does not always apply in practice, meaning the policy estimated to perform the best may not actually be a good update, such that enforcing the adoption of this policy can lead to noisy, unstable learning. Using a higher temperature value smooths out this noise. Moreover, using $\tau_k^{comm} \rightarrow 0$ effectively enforces consensus on a single policy for the finite population in the networked case, which in anti-coordination games may also reduce the average return. This all provides empirical evidence for our scheme for τ_k^{comm}, but further optimising the choice might lead to additional performance increase.