Coordinating the Crowd: Inducing Desirable Equilibria in Non-Cooperative Systems

David Mguni
PROWLER.io
Cambridge, UK
davidmg@prowler.io

Emilio Sison
Department of Mechanical
Engineering, MIT
Cambridge, MA, USA
esison@mit.edu

Joel Jennings
PROWLER.io
Cambridge, UK
joel@prowler.io

Sofia Ceppi
PROWLER.io
Cambridge, UK
sofia@prowler.io

Enrique Munoz de Cote
PROWLER.io
Cambridge, UK
enrique@prowler.io

Sergio Valcarcel Macua
PROWLER.io
Cambridge, UK
sergio@prowler.io

ABSTRACT
Many real-world systems such as taxi systems, traffic networks and smart grids involve self-interested actors that perform individual tasks in a shared environment. However, in such systems, the self-interested behaviour of agents produces welfare inefficient and globally suboptimal outcomes that are detrimental to all — some common examples are congestion in traffic networks, demand spikes for resources in electricity grids and over-extraction of environmental resources such as fisheries. We propose an incentive-design method which modifies agents’ rewards in non-cooperative multi-agent systems that results in independent, self-interested agents choosing actions that produce optimal system outcomes in strategic settings. Our framework combines multi-agent reinforcement learning to simulate (real-world) agent behaviour and black-box optimisation to determine the optimal modifications to the agents’ rewards or incentives given some fixed budget that results in optimal system performance. By modifying the reward functions and generating agents’ equilibrium responses within a sequence of offline Markov games, our method enables optimal incentive structures to be determined offline through iterative updates of the reward functions of a simulated game. Our theoretical results show that our method converges to reward modifications that induce system optimality. We demonstrate the applications of our framework by tackling a challenging problem within economics that involves thousands of selfish agents and tackle a traffic congestion problem.

1 INTRODUCTION
Complex systems such as traffic networks, smart grids and fleet networks involve autonomous agents that each seek to perform individual tasks. One such example is a ride-sharing network such as an Uber fleet which involves many self-interested (freelance) drivers that each use the same road network and have access to a common supply of customers. Other examples are road traffic networks used by commuters, electricity grids with households drawing from the network and smart grids. In each of these settings, agents utilise a shared resource to maximise their individual objectives.

Multi-agent systems (MASs) in which agents act non-cooperatively to maximise their own interests are modelled by Markov game (MGs). In MGs, although each agent acts rationally, that is, to maximise its own interests, the lack of coordination produces stable outcomes or Nash equilibria (NE) that are vastly suboptimal from a system perspective and undermine firm efficiency [7].

In the case of ride-sharing networks, drivers’ self-interested behaviour and their preference to locate at certain regions results in inefficient clustering that produces a distribution of taxis that does not match customer locations [16]. This results in a market inefficiency and prevents firms from maximising output. In electricity networks, excessive demand at specific periods leads to demand spikes that overwhelm electrical supply; in traffic networks the actions of self-interested commuters leads to heavy congestion and traffic delays resulting in poor network outcomes.

To alleviate these problems, network designers can employ incentives to modify the strategic behaviour of the self-interested agents. However, in an MAS, these incentives must be carefully calibrated to induce desirable outcomes from the joint behaviour of selfish actors in dynamic environments and often, with (budgetary) constraints on the size of incentives or penalties. Additionally, in settings such as smart grids and traffic networks, the design of incentives must also account for adjustments in the system state such as changes in customer demand for taxis; consequently, designing incentives is a formidable challenge [22].

Although in many MAS, the agents’ reward functions are known (e.g. minimising commute time, firm profit maximisation) or a sufficiently accurate proxy can be constructed from data, designing incentives remains a challenge. This is due to the fact that changes to the agents’ joint behaviour (and the resulting system outcomes) after modifications to their rewards is generally difficult to predict.

In general, it is known that in many real-world MASs, human strategic interaction approximates Nash equilibrium strategies. Multi-agent reinforcement learning (MARL) is a powerful tool that enables computerised agents to learn strategic behaviour after repeated interactions in unknown systems - this enables MARL to serve as a useful tool to generate a proxy of outcomes in systems with human participants and simulate the behaviour of other computerised agents [9]. As with algorithmic methods in game theory, MARL does not offer a method of promoting efficient outcomes that maximise social welfare (e.g. minimise travel time in traffic networks) or optimise external objectives (e.g. maximise taxi firm efficiency) and typically converge to poor system outcomes [25].

We propose a new technique to tackle the issue of undesirable outcomes in MASs. In our framework, an incentive designer (ID)
We prove convergence to the reward modifier that induces efficient outcomes. This modification, as shown in one of our experiments, can represent a toll charge on a traffic network that induces even traffic flow leading to reduced congestion.

Using the known agents’ intrinsic goals, our framework firstly uses MARL to learn the NE of simulated MAS and thus generate a proxy for real-world outcomes. This then allows us to model the induced changes in agents’ behaviour given modifications to their rewards through incentives. The ID uses Bayesian optimisation in the simulated environment to determine the optimal modifications to the agents’ rewards to be implemented in the real-world settings.

The ID is not required to have a priori knowledge of the system performance metric but requires only the goal of the agents (e.g. arriving at work in the quickest time possible). We concern ourselves with Markov potential games (MPGs) — a class of MGs that model settings in which agents compete for a common resource such as selfish routing games (transportation networks) [22], spectrum sharing (wireless communications) [32], oligopoly [26], electric power grids [12] and cloud computing [3].

We prove theoretical results that demonstrate that within MPGs, the equilibrium set is continuous on the reward modifications. As we show, this allows us to prove existence of an optimal reward modifier.

We illustrate the framework in a set of experiments and its dynamic and learning variants [27]. These incomplete information models analyse the problem of constructing a mechanism — a system of rewards and transfers, among self-interested strategic agents that have private information about their reward functions. The problem is to incentivise truth-revealing announcements from the agents. A well-known result in MD rules out (strategy-proof) mechanisms that induce the desired agent behaviour for general agent reward functions [23]. Therefore, in MD, agents’ reward functions are (typically) limited to quasi-linear functions that are known up front [19]. Our framework permits a general rewards beyond quasi-linear functions.

This work relates to leader-follower games - sequential games in which a leader moves in advance of other agent(s) or followers, who each select a best response strategy [28]. However, in leader-follower games, the leader cannot induce efficient outcomes i.e. maximise its own objective (e.g. ex. 98.1 in [20]) since the leader’s reward is a function over a fixed joint action set. Our work also relates to reward shaping through which a reward is added with the aim of inducing convergence to a more desirable equilibrium [2]. The majority of the reward shaping literature is concerned with potential based reward shaping. Potential based reward shaping leaves the NE set unaltered and does not guarantee convergence to more efficient equilibria [6]. A number of papers handle non-potential based rewards shaping e.g. [21], however, such papers are limited in scope since they consider only specific normal form games settings e.g. the stag hunt game. We tackle the MG case which adds considerable complexity since it requires a method of incentivising sequences of state-action pairs (trajectories) in a stochastic environment.

2 PRELIMINARIES

Let \( \mathcal{N} = \{1, \ldots, N\} \) denote the (possibly infinite) set of agents where \( N \in \mathbb{N} \times \{\infty\} \). An MG is a tuple:

\[
\mathcal{G} = (\mathcal{N}, (\gamma_i)_{i \in \mathcal{N}}, \delta, (\mathcal{U}_i)_{i \in \mathcal{N}}, P, (\mathcal{R}_i)_{i \in \mathcal{N}})
\]

which can be described as follows: at each time step \( t = 1, 2, \ldots, T \in \mathbb{N} \), the state of the system is given by \( s \in \mathcal{S} \subseteq \mathbb{R}^p \) for some \( p \in \mathbb{N} \). The game is equipped with an action set \( \mathcal{U}_i = \mathcal{X}_{i \in \mathcal{N}} \mathcal{U}^j - \) a Cartesian product of each agent’s action set \( \mathcal{U}^j \). Each set \( \mathcal{U}^j \) is a compact, non-empty action set for each agent \( i \in \mathcal{N} \). We define by \( \mathcal{U}^{-i} = \mathcal{X}_{j \in \mathcal{N} \setminus \{i\}} \mathcal{U}^j \) - the Cartesian product of all agents’ action sets except agent \( i \). At each time step, the next state of the game is determined by a probability distribution \( P : \mathcal{S} \times \mathcal{U} \times \mathcal{S} \) so that \( \mathcal{P}(s, u, s) \) gives the probability distribution over next states given a current state \( s \) when the agents take a joint action \( u \in \mathcal{U} \). When the environment is at state \( s \) and the agents take action \( u \), each agent \( i \) receives a reward computed by a Lipschitz function \( R_i : \mathcal{S} \times \mathcal{U} \times \mathcal{S} \rightarrow \mathbb{R} \). The term \( \gamma_i \in [0, 1] \) is each agent \( i \)'s discount factor. Each agent has a stochastic policy \( \pi_i : \mathcal{S} \times \mathcal{U}^j \rightarrow \mathbb{R}^+ \) - a conditional distribution over the action set given the current state. Let \( \Pi^j \) be a non-empty set of stochastic policies over \( \mathcal{S} \times \mathcal{U}^j \) such that \( \pi^j \in \Pi^j \).

We denote by \( \mathcal{P} \) the set of policies for all agents i.e. \( \mathcal{P} = \mathcal{X}_{i \in \mathcal{N}} \Pi^j \), where each \( \pi^j \), and by \( \Pi_i \equiv \mathcal{X}_{j \in \mathcal{N} \setminus \{i\}} \Pi^j \). For simplicity, we assume \( \Pi^j \equiv \Pi_i \), \( i \neq j \). The joint policy of all agents is denoted by \( \pi = (\pi^j)_{i \in \mathcal{N}} \in \mathcal{P} \), while the joint policy of all but the \( i \)-th agent is denoted \( \pi^{-i} = (\pi^j)_{j \in \mathcal{N} \setminus \{i\}} \).

We will sometimes write \( \pi = (\pi^1, \pi^2) \) for any \( i \in \mathcal{N} \).

Each agent \( i \in \mathcal{N} \) uses a value function, \( v_i^T : \mathcal{S} \times \mathcal{U} \rightarrow \mathbb{R} \), as its objective function:

\[
v_i^T(s) = \mathbb{E}[\sum_{t=0}^{T} \gamma_t^t R_i(s_t, u_{i,t}, u_{-i,t})|u_t \sim \pi^{-i}(\cdot|s_t), \nonumber
\]

\[
\mathcal{S}_{t+1} \sim \mathcal{P}(|s_t, u_t), s_0 = s \}
\]

\[
(1)
\]

where \( u_t = (u_{i,t}, u_{-i,t}) \) is the joint action at time \( t \). We now give some essential definitions:

Definition 2.1. The policy \( \pi_i \in \Pi^j \) is a best-response policy against \( \pi^{-i} \in \Pi^{-i} \) if: \( \pi^j \in \arg\max_{\pi^j \in \Pi^j} v_i^{\pi^j, \pi^{-i}} \).

\[\text{1}^{\text{In [21] some experiments on repeated games are performed but no theoretical analysis is provided.}}\]
We now describe how the ID modifies the MG played by the agents.

A Markov-Nash equilibrium (M-NE) is the solution concept for MGs in which every agent plays a best-response against other agents. A M-NE is defined by the following:

**Definition 2.2.** A strategy \( \pi = (\pi_i)_{i \in N} \in \Pi \) is an M-NE if
\[
\forall i \in N, \forall \pi_i \in \Pi_i, \forall \pi_i \in \Pi_i, \forall \pi_i \in \Pi_i, s \in S, \forall i \in N.
\]

The M-NE condition ensures no agent can improve their rewards by deviating unilaterally from their current strategy. We define \(NE(\mathcal{G})\) as the set of M-NE for the game \( \mathcal{G} \).

**Definition 2.3.** An MG is called an exact MPG or an MPG for short, if there exists a function \( \Phi : \delta \times \Pi \rightarrow \mathbb{R} \) such that:
\[
v_i^{(\pi_i, \pi_i)}(s) - v_i^{(\pi_i, \pi_i)}(s) = \Phi(\pi_i, \pi_i)(s) - \Phi(\pi_i, \pi_i)(s)
\]
\[
\forall \pi_i \in \Pi_i, \forall \pi_i \in \Pi_i, \forall s \in S, \forall i \in N.
\]

Note that \( \Phi \pi(s) \) gives the value function for all agents. We use \( \mathcal{G}(\mathcal{w}) \) to denote an MPG. In this paper, we focus exclusively on MPGs.

## 3 THE FRAMEWORK

We now describe how the ID modifies the MG played by the agents. The problem is arranged into a hierarchy in which the ID chooses the most natural alteration to an agent’s reward function is for it to be modified by the ID. This function describes the agents’ goals e.g. revenue management (e.g., ticket pricing), congestion management, and network design problems (e.g. tolling) [5]. The function \( \Psi \) can be interpreted as a system of wealth transfers for example, in the case of freelance taxis, \( \Psi \) represents rewards given to drivers for taking jobs at specific times and locations or surcharges to customers, and similarly for smart grid users at peak times. The following condition constrains the transfer of wealth to the set of agents:

**Definition 3.1.** We say that the ID’s choice of \( \mathcal{w} \in \mathcal{W} \) is weakly budget balanced, if there is no net transfer of wealth from the ID to the agents: \( \Psi(\mathcal{w}, \pi) \leq 0 \).

We consider two main types of reward function for the ID, depending on the ID’s goal:

1. **Trajectory targeted:** The ID’s payoff is a function of the state trajectories produced by the agents’ policies in the MG; i.e. is,
   \[
   J(\mathcal{w}, \pi) \triangleq \mathbb{E}[R_{ID}(\mathcal{w}, \pi, \zeta)],
   \]
   where \( X^\pi \) is Markov chain induced by the policy profile \( \pi \in \Pi \) in \( \mathcal{G}(\mathcal{w}) \) and \( \zeta \) is an i.i.d. random variable which captures outcome noise. An example is taxi firm seeking to match the location of a set of freelance taxi drivers with (predicted) customer locations in some region. Here, the ID’s objective could be given by a KL divergence between the distribution of taxis at every timestep, \( D_{\text{KL}}^T(\mathcal{w}, \pi) \), and the target distribution of demand, \( D_{\text{ID}}^T(\mathcal{w}, \pi) = \sum_{t=0}^T \mathbb{E} \left[ \sum_{i \in N} KL(D_t^\pi(\mathcal{w}, \pi)) | D_t^\pi \right] \). Other applications of trajectory targeted objectives are firms seeking to smoothen electricity consumption in smart grids through dynamic pricing [5] and modification of firm activity through taxation [17].

2. **Welfare targeted:** The ID’s payoff is a function of the agents’ joint rewards, that is,
   \[
   J(\mathcal{w}, \pi) \triangleq \mathbb{E}[R_{ID}(\mathcal{w}, \pi, \zeta)],
   \]
   for some uniformly continuous function \( h \) and \( v_\pi^\mathcal{w} \triangleq (v_i^\mathcal{w})_{i \in N} \). One example a traffic network manager that seeks to minimise travel time of all agents. In this cases, the ID is the sum of agents’ negative costs (travel times) i.e.: \( k_{ID}^{\text{loc}} = \sum_{i \in N} v_i^\mathcal{w} \), which results in the
ID maximising social welfare. Similar examples are resource extraction and oligopoly intervention e.g. fishery problems using optimal taxation [26] in which the ID seeks to maximise firm welfare whilst seeking to sustain a minimum amount of the resource and worst-case optimisation (maxmin) problems (i.e. by setting $h = -1$).

The ID problem (5) is a bilevel optimisation problem (mathematical program with equilibrium constraints). Such problems are generally highly non-convex with unconnected feasible regions. For this reason, the problem is generally highly intractable using analytic methods but for simple cases (e.g. linear rewards) [4].

In the next section, we overcome these issues by expressing the NE constraint in terms of the potential function, and show that MARL methods can be applied to compute the set of NE for the MAS model, so that we can ensure feasibility for the ID problem without requiring closed analytic solutions. Crucially, this, as we show, allows us to compute the agents equilibrium policies to an MG the reward function of which, is chosen by the ID. We prove continuity properties of the MPG with respect to the ID’s changes to the reward function which allows the ID to produce an iterative sequence of reward functions. We then give a constructive formulation that allows to prove convergence to such an optimal solution for the ID. Finally, we provide an approximation bound when the optimal reward modifier is approximated with a truncated power series. We proceed to explain the details.

4 THEORETICAL ANALYSIS

We now show that $\mathcal{G}(w)$ is an MPG, which enables $NE(\mathcal{G}(w))$ to be described in terms of local maxima of function (not fixed points).

**Proposition 4.1.** There exists a function $\Phi : \mathcal{S} \times \mathcal{X} \rightarrow \mathbb{R}$ such that each agent’s best-response strategy in $\mathcal{G}(w)$ maximises $\Phi$.

Prop. 4.1 reduces the problem of finding the M-NE for $\mathcal{G}(w)$ to a single optimal control problem as opposed to finding a fixed point solution which is considerably more difficult. However, it is necessary to show that the game produced after the ID alters the agents’ rewards is still potential. The following lemma establishes that fact:

**Lemma 4.2.** The game $\mathcal{G}(w)$ is an MPG.

**Proof.** To prove the assertion we need to show that the transformation $R_i \rightarrow R_i^w$ preserves the potential game property.

For any function $\Xi : \mathcal{S} \times \mathcal{X} \rightarrow \mathcal{S}$ define $\Delta_\Xi \triangleq \Xi_i^w(s_t, u_i, u_{-i}, t) - \Xi_i^w(s_t, u_i, u_{-i}, t) \Phi_i^w(s)$. We claim that there exists a function $\Phi^\pi_i, \Phi^w(s)$ s.th. $\Delta_{R_i^w} = \Phi^\pi_i, \Phi^w(s)$. This follows directly from the additive form of the reward function modification. Indeed, consider the function $\Phi^\pi_i, \Phi^w(s) \triangleq \Phi^\pi_i(s) + \Theta(s, u^i, u^{-i}, w)(s)$. Since $\Xi_0$ is potential, by (3) and (4) we have that:

$$\Delta_{R_i} = \Delta_{\Theta} + \Delta_{\Pi} = \Delta_{\Phi^\pi} = \Delta_{\Phi^\pi_i, \Phi^w}(s)$$

which completes the proof. $\Box$

**Proposition 4.3.** $R_{ID}$ is uniformly continuous in $w$.

The proof of the proposition is deferred to the appendix.

**Corollary 4.4.** The following expression holds

$$\arg\max_{\pi \in \Pi} \Phi^\pi_i, \Phi^w(s), \forall s \in \delta \subseteq NE(\mathcal{G}(w)).$$

Cor. 4.4 expresses that in playing their best-response strategies $\mathcal{G}(w)$, each agent inadvertently maximises $\Phi^\pi_i, \Phi^w$, so the function $\Phi^\pi_i, \Phi^w$ is a potential of $\mathcal{G}(w)$.

Having reduced the problem of finding $NE(\mathcal{G}(w))$ to an optimal control problem, we now establish that the ID’s problem is a constrained optimisation problem:

**Theorem 4.5.** ID’s problem is equivalent to:

$$\max_{w \in W, \pi \in \Pi} f(w, \pi) \ s.t. \ \nabla_x \Phi^\pi_i, \Phi^w = 0, \ \nabla^2 \Phi^\pi_i, \Phi^w \leq 0.$$  (8)

**Sketch.** The proof of the theorem consists of the following components: proving that $\Phi \in C^1$ and that ID’s problem can be rewritten as a constrained optimisation problem and the set of constraints of the problem are expressed by (8). By Rademacher’s lemma we have that $\Phi$ is Lipschitz continuous on some open subset of its domain then $\Phi$ is differentiable almost everywhere (in that set). Since the $\Phi$ is defined over $\delta \subseteq \mathbb{R}^n$, we can construct an open subset for which Rademacher’s lemma holds. To deduce the remainder of the theorem, we note that by Corollary 4.4, $NE(\mathcal{G}(w))$ coincide with the set of the local maxima of $\Phi$. The result then follows by noting that conditions (8) are first and second order conditions for local maxima of the function $\Phi$. $\Box$

Theorem 4.5 establishes that the ID’s problem reduces to a constrained optimisation problem where the feasibility set is given by the set of points that are local maxima of $\Phi$. In the next section, we show that we can apply MARL to constrain the set of points in $W$ to lie within the feasibility set.

We now prove that $NE(\mathcal{G}(w))$ is continuous on $w$ — this enables the ID to generate an iterative sequence of games and permits use of black-box optimisation to solve the ID’s problem. We firstly study the effect of modifying $w$ on $NE(\mathcal{G}(w))$. To establish a formal notion of continuity of $NE(\mathcal{G}(w))$ w.r.t $w$, we introduce essentiality:

**Definition 4.6.** Given metric space $X$, let $B_{\delta}(x) \triangleq \{y \in X : \|x - y\| < \alpha\}$ denote the open ball with radius $\alpha > 0$ around $x \in X$. Then $x \in NE(\mathcal{G}(w))$ is **essential** in $w$ if for any $\epsilon > 0, \exists \delta > 0 : w' \in B_\delta(w) \implies x' \in B_\delta(x)$, for any $x' \in NE(\mathcal{G}(w'))$.

The following results establish the continuity in ID’s reward under changes in $w$ which underpin the existence of a solution for ID’s problem and a method for computing the solution. We begin by demonstrating that small changes in ID’s action lead to small changes in the game, that is, the game itself is continuous in $w$.

**Proposition 4.7.** $NE(\mathcal{G}(w))$ is an essential set in $w$.

**Proof.** We begin the proof by proving that the value function for each agent $i \in \mathcal{N}$ is Lipschitz continuous w.r.t. $w$:

$$\|v_i^\pi - v_i^w(s) - u_i^\pi(s)\| = \|\max_{\pi \in \Pi} R_i(s, u_i, u_{-i}, t, w) - R_i(s, u_i, u_{-i}, t, w')\| \leq \max_{\pi \in \Pi} \|\max_{s' \in S} R_i(s, u_i, u_{-i}, t, w')\| \leq \max_{\pi \in \Pi} \|\max_{s' \in S} R_i(s, u_i, u_{-i}, t, w')\| \leq \max_{\pi \in \Pi} \|\max_{s' \in S} R_i(s, u_i, u_{-i}, t, w')\|$$
We can modify the framework to tackle the case in which the ID \( \Theta \) w.r.t. The solution constant, it is straightforward to deduce that the NE set is preserved. The equilibrium selection convergence to the highest welfare equilibria within a fixed M-NE 5 PRESERVING THE NASH EQUILIBRIA. The method uses MARL to generate a model of the strategic (equilibrium) behaviour among the agents for a given value of \( \pi \) is then updated. The specifics are as follows: the function \( F \) is fixed and may lead to convergence to less desirable equilibria [6], now the function \( F(\cdot, \pi) \) is determined as a solution to the ID’s problem in \( \pi \).

By similar reasoning as the method of the previous section, we deduce that the following: maximise \( J(w, \pi) \) s.t. \( \nabla_{\pi} \Phi(\pi, \mu_{0}(W)) = 0 \), \( \nabla_{\pi}^{2} \Phi(\pi, \mu_{0}(W)) \leq 0 \). (10)

In this case, the M-NE constraint is defined over the M-NE set before the ID alters the game. The formulation of the problem ensures that the agents’ rewards are modified in a way that the agents play efficient policies, the constraint ensures that the policy remains within the original NE set. In order to formally describe the notion of efficiency, we need the following concepts:

Definition 5.1. The strategy profile \( \pi \in \Pi \) is said to be a welfare optimal strategy profile of \( \mathcal{G}(w) \) if: \( \sum_{i \in N} v_{i}^{\pi_{i}, \pi_{-i}, w} \geq \sum_{i \in N} v_{i}^{\pi_{i}', \pi_{-i}', w} \) for an \( i \in N \).

Definition 5.2. For a given \( w \in W \), \( \pi \in \Pi \) is said to be a Pareto efficient (PE) strategy profile of \( \mathcal{G}(w) \) if: i) \( v_{i}^{\pi_{i}, \pi_{-i}, w} = v_{i}^{\pi_{i}', \pi_{-i}', w} \) for all \( i \in N \), ii) \( v_{i}^{\pi_{i}', \pi_{-i}', w} > v_{i}^{\pi_{i}', \pi_{-i}, w} \) for an \( i \in N \).

PE implies that no agent increases their reward whenever some other strategy profile \( \pi' \in \Pi \) is played and, at least one agent is strictly best off under \( \pi' \) so that all agents prefer the PE outcome. PE is a criterion for a welfare maximising ID. We say that strategy profile \( \pi \) is payoff dominant if \( \pi \in NE(\mathcal{G}(w)) \) and \( \pi \) is PE.

Proposition 5.3. Let \( w \in W \) be a solution to ID’s ES problem, then \( \exists \pi \in NE(\mathcal{G}_{0}) \) which is a payoff dominant policy profile of \( \mathcal{G}_{0} \). The issue of how to compute \( w^{\star} \) remains; we now describe its computation using black-box optimisation and MARL.

5 PRESERVING THE NASH EQUILIBRIA.

We can modify the framework to tackle the case in which the ID modifies the rewards to maximise some efficiency criterion subject to the condition that M-NE set of the game is preserved. Inducing convergence to the highest welfare equilibrium within a fixed M-NE set is known as equilibrium selection (ES) and represents a major challenge in GT and MARL [10]. The ID framework can be used to address ES within the context of MPGs.

Let \( \mu_{k}(W) = \{ w \in W : \Psi(w) = k \in \mathbb{R} \} \). Since \( \mathcal{G}(\mu_{k}(W)) \) is just the MG in which the agents’ rewards are modified by at most a constant, it is straightforward to deduce that the NE set is preserved.
for some \( w_k \in W \), the belief is used to form a posterior distribution which is used to construct an acquisition function (e.g., expected improvement) that indicates which parameter \( w_{k+1} \) should be evaluated next, guiding exploration over \( W \). We use MARL to solve the game \( \mathcal{G}(w) \) allowing the agents (of the simulated game) to observe only their individual (modified) rewards after their joint policy \( \pi \) is played. The agents sample trajectories of experience tuples \( (s_t, u_t, (R_{ik}(w_k, s_{t+1}), i \in N, s_{t+1}), (c.f. Prop. 6.1) ) \) which are used to estimate the joint value function, \( v^\pi_{w_k} \). Then, they update their policies by performing stochastic gradient ascent.

The optimisation objective in (8) is nested; the ID chooses \( w \) of \( \mathcal{G}(w) \) and the agents select a joint policy which generates a reward signal for the ID. Simultaneous updates of both the ID parameters and the agents’ policies, in general, lack converge guarantees due to non-stationarity. Therefore, in order to compute the solution iteratively, after an initial choice by the ID, we let the MARL algorithm run until convergence which fulfils the M-NE constraint for the ID’s problem (c.f. Prop. 6.1); the ID receives feedback from the outcome of the game \( \mathcal{G}(w) \), then updates its choice of \( w \). This results in an inner-outer loop method. Each step performed by the ID is computationally costly. As such, gradient-based algorithms require a substantial number of iterations to converge to a solution. Therefore, we use a sample-efficient optimisation algorithm, namely Bayesian optimisation which also allows scaling of the framework. BO also has strong theoretical guarantees for non-convex problems [24] and can handle large dimensional problems [8, 30]. Inner-outer loop methods are widely used in single agent problems to tune hyperparameters of learning algorithms [14].

**Inputs**: Maximum number of BO evaluations \( K \), and maximum number of MARL iterations \( M \).

1. Initialise ID’s dataset \( \mathcal{D}_0 = \{ \} \) and reward modifier parameter \( w_0 \).
2. for \( k = 0, \ldots, K \) do
3.   Initialise agents’ strategy profile \( \pi_0 \).
4.   for \( m = 0, \ldots, M \) do
5.     Agents sample data from the environment following strategy profile \( \pi_m \).
6.     Estimate joint value function (critic) \( v^\pi_{m,w_k} \).
7.     Update joint policy (actor) \( \pi_{m+1} \).
8.   end for
9.   Estimate ID’s payoff function \( J(w_k, \pi_M) \).
10. Select new \( w_{k+1} \) guided by current data \( \mathcal{D}_k \) using BO with expected improvement criterion.
11. Augment dataset \( \mathcal{D}_{k+1} = \{ \mathcal{D}_k, (w_k, J(w_k, \pi_M)) \} \).
12. end for
13. Return \( w_T \).

Algorithm 1: The ID framework

### 6.1 Discussion on the method

**Convergence.** In order to ensure the algorithm converges to an optimal solution for the ID both the inner and outer loop are required to converge. Theorem 4.8 guarantees the existence of a solution for \( w^* \). Convergence of the inner loop is required to obtain the equilibria of the simulated MPG. Consequently, the method is subject to conditions under which MARL methods converge. Hence, the method is subject to conditions under which MARL methods converge. MARL methods have been shown in general, to have strong convergence guarantees to M-NE solutions for MPGs [11, 13, 29]. The following proposition provides this guarantee:

**Proposition 6.1 (Convergence).** Algorithm 1 converges to a stable point, moreover the set of stable points of algorithm 1 correspond to M-NE for the MPG.

Another consideration is the growth in decision complexity of the ID’s problem with the number of parameters over which the BO is performed. This depends on the size of the state space of the MAS model. Theorem 3, however, proves that approximate solutions are computable with fewer parameters for a given error bound.

### 7 EXPERIMENTS

#### 7.1 Experiment 1: Optimising a Traffic Network

The following experiment illustrates the application of the method to a traffic network problem. We consider road traffic network examples, one of which is a subsection of the city of London. In this setting, each agent seeks to traverse the graph from a source node (labelled 1) to a goal node (labelled 8) - this, for example can represent agents performing a commute. The agents incur costs which represent the travel time. When traversing an edge, each agent incurs a unit cost plus an additional cost which is a convex function of the number of agents traversing the edge at that time - the latter cost represents additional time delays due to traffic congestion.

The goal of each agent is to minimise its own costs. It is well-known that in such systems (e.g. Braess’ Paradox, Pigou example), the agents’ selfish behaviour of leads to congestion on ‘more desirable’ paths leading poor system efficiency [22].

The problem is modelled as a selfish routing game (SRG) - a widely studied potential game [22] that models traffic networks. In this setting, agents pursuing their individual objectives produce outcomes that result in high travel times for all [31]. In this problem, a set of \( N \) self-interested agents direct its commodity flow through a network \( G = (V, E) \) where \( V \) is the set of nodes and \( E \subseteq V \times V \) is the set of edges of \( G \). Each agent seeks to direct a single commodity e.g., a taxi firm directing only its fleet. When traversing an edge their commodity produces congestion incurring a negative externality (cost) on all agents. Each agent’s commodity is infinitely divisible so that at each node the agents may split their commodity flow over each outgoing edge. Each agent’s goal is to direct its commodity through paths that minimise its own costs.

A central planner (CP) seeks to minimise delays due to congestion by devising a dynamic system of toll charges that induces an even commodity flow over a given subset of edges of the network \( \hat{E} \subseteq E \) at all times. The CP’s problem is to maximise \( R_{ID}(w) = -\sum_{t=1}^{T} \sum_{l \in \hat{E}} \left| f^*(t) - f_l(t) \right|^{1/2} \) where \( f_l(t, w) \) is the flow on edge \( l \in \hat{E} \) at time \( t \) and \( f^*(t) \) \( = \sum_{l \in \hat{E}} f_l(t) \). To induce changes in the agents’ commodity flows, the CP adds to \( R_{id} \) the function \( \Theta(f) \) which is a power series of order 5.

We consider two cases, we firstly provide an intuitive example known as Braess’ example, a widely studied problem that clearly demonstrates the inefficiencies of traffic networks [22]. We then apply the method to a subsection of the traffic network in the city of
7.1 Braess’ Example. Fig. 1 shows a diagrammatic illustration of the Nash equilibrium agent flow (the size of the flow of agents through an edge is represented by the edge width) through the network after convergence without ID. As is shown in Figs. 1a) and 1b), when an ID is included, it learns how to set tolls (costs) on the middle edge that induce equal flow over the graph which maximises social welfare.

7.1.1 Braess’ Example. Consider 2,000 agents each seeking to locate themselves at desirable points in space over some time horizon. The desirability of a region changes with time and decreases with the number of agents located within their neighbourhood. The resulting NE distribution is in general, highly inefficient (and may not conform to external objectives) due to agent clustering [15]. The problem is a dynamic generalisation of the El Faro bar problem and encapsulates spectrum sharing problems in wireless communications [1]. The problem also models spatio-economics problems such as firms locating their supply with dynamic demand e.g. freelance taxis. To handle large strategic populations, we use an mean field game framework [15].

A formal description is as follows: the game has a finite set of agents \( N \triangleq \{1, \ldots, N\} \), where \( N \in \mathbb{N} \). At time \( t < T \), the state of the system is \( x_t = (x_{i,t})_{i \in N} \in \mathcal{S} \) where \( x_{i,t} \) denotes the location of agent \( i \) at time \( t \) and \( \mathcal{S} \subseteq \mathbb{R}^2 \). Each agent \( i \) selects action \( u_{i,t} \in \mathbb{R}^2 \) which is a vector movement towards some location \( x_{i,t+1} \in \mathcal{S} \). London, UK. We show that our framework finds an optimal system of tolls that leads to maximal system efficiency.

7.1.2 Extended City Case. We test our method in a complex network consisting of 8 nodes and 13 edges which represents a subsection of the London road network. We show that our method produces socially optimal (M-NE) outcomes. The ID is able to isolate the 3 roads edges to apply tolls in only 150 outer loop iterations.

Our method shows that the ID was able to isolate three nodes to apply a toll which led to a reduction in congestion (as indicated in Fig. 2) through the network in only 150 iterations of BO (outer loop). Fig. 2 c) shows the social welfare function (which is the sum of all agents’ returns) after 6,000 iterations of the MARL algorithm (inner loop) without the ID (orange curve) and with the ID (blue curve), and demonstrates a significant increase in social welfare. This technique is a first example of reinforcement learning in an SRG that handles large networks and populations of users. This is in contrast to current methods in which agents choose paths resulting in exponential scaling in decision complexity with graph size [18].

7.2 Experiment 2: Supply & demand matching with thousands of agents

Consider 2,000 agents each seeking to locate themselves at desirable points in space over some time horizon. The desirability of a region changes with time and decreases with the number of agents located within their neighbourhood. The resulting NE distribution is in general, highly inefficient (and may not conform to external objectives) due to agent clustering [15]. The problem is a dynamic generalisation of the El Faro bar problem and encapsulates spectrum sharing problems in wireless communications [1]. The problem also models spatio-economics problems such as firms locating their supply with dynamic demand e.g. freelance taxis. To handle large strategic populations, we use an mean field game framework [15].

A formal description is as follows: the game has a finite set of agents \( N \triangleq \{1, \ldots, N\} \), where \( N \in \mathbb{N} \). At time \( t < T \), the state of the system is \( x_t = (x_{i,t})_{i \in N} \in \mathcal{S} \) where \( x_{i,t} \) denotes the location of agent \( i \) at time \( t \) and \( \mathcal{S} \subseteq \mathbb{R}^2 \). Each agent \( i \) selects action \( u_{i,t} \in \mathbb{R}^2 \) which is a vector movement towards some location \( x_{i,t+1} \in \mathcal{S} \).
The transition dynamics are given by $x_{i,t+1} = \alpha x_{i,t} + \beta u_{i,t} + \epsilon_{i,t}$, where $\alpha$, $\beta$ are scalars, and $\epsilon_{i,t} \sim \mathcal{N}(0, \Sigma)$, for some covariance matrix $\Sigma$. The agents’ joint action produces a distribution $M^*_{t+1}$ of agents over $\mathcal{S}$. Let $m^a_{x,t}$ be the density of agents at some location $x_t \in \mathcal{S}$ at time $t \in [0, T]$, where $\mathcal{P}(\mathcal{S})$ denotes the space of probability measures. Each point in $\mathcal{S}$ has some level of desirability $\Gamma : \mathcal{S} \times \mathcal{P}(\mathcal{S}) \rightarrow \mathbb{R}$, which is determined by the agent’s location and the density of agents at that point. Each agent’s reward, $R_i$, is given for any $\pi \in \Pi$ by: $R_i(x_t, m^a_{x,t+1}) = \mathbb{E} \left[ \sum_{t=0}^{T} \Gamma(x_t, m^a_{x,t}) - \frac{1}{2}u^T_{i,t}Ku_{i,t} \right]$, where $\Gamma(x_t, m^a_{x,t}) = (\pi - \hat{\pi})^2 - \alpha(m^a_{x,t})^2$, where the expectation is taken over the state-action trajectory induced by the system dynamics and joint policy $\pi$. The term, $\Psi$, rewards the agent for locating closer to the point $\hat{x}_t \in \mathcal{S}$ at time $t \leq T$ whilst penalising the agent for remaining in areas with a large concentrations of agents. The quadratic term levies a movement penalty control cost.

A principal aims to incentivise the self-interested agents to adopt a target distribution $M^*_{t}$ at each time step $t \leq T$. The principal’s objective $J$ is given by a KL divergence between $M^a_{t}$ and $M^*_{t}$ i.e. $J(w, \pi) = \mathbb{E} \left[ \sum_{t=0}^{T} KL(M^a_{t}(w, \pi)||M^*_{t}) \right]$. To incentivise the agents to adopt its preferred distribution, the principal adds a reward modifier $\Theta$ - a function parameterised by $w \in \mathcal{W}$. We test our method both one-shot dynamic scenarios.

In the one-shot game the ID seeks to induce an agent distribution (shown by the left heat map in Fig.3) - this differs from the distribution obtained when agents’ maximise only their intrinsic reward function (central heat map in Fig.3). When the modifier function $\Theta$ is added to the agents’ rewards, the average KL divergence converges almost to zero which demonstrating a close match of the agents’ distribution (right heat-map in Fig.3) with the desired one.²

In the dynamic game the ID’s desired distribution changes over time. In our experiment, $M^*_{t}$ for $t = 0, 1, 2$ are as shown by the heat maps in the top row of Fig. 4 (left), while the bottom row presents the agents’ distributions achieved with the ID framework.

8 CONCLUSION

In this paper, we introduce an incentive designer (ID) framework - a technique that enables self-interested adaptive learners to converge to efficient Nash equilibria in Markov games. By adding a modifier function to the agents’ rewards, our method learns to modify the rewards of self-interested agents to induce efficient, desirable equilibrium outcomes. We prove a continuity property in the ID’s modifications to the game which permits a broad range of black-box optimisation techniques to be applied.

9 APPENDIX

Lemma A.1. Let $A$ and $B$ be sets and let $f : A \times B \rightarrow \mathbb{R}$ and $h : A \times B \rightarrow \mathbb{R}$ be two real-valued maps s.t. the following expression holds $\forall a \in A, b \in B$ and for some constant $c$:

$$|f(a, b) - h(a, b)| < c, \quad (11)$$

It then follows that:

$$| \max_{a \in A, b \in B} f(a, b) - \max_{a \in A, b \in B} h(a, b) | < c$$

²The small discrepancy from 0 is due to the the Gaussian approximation of the agent density.

![Figure 4: Dynamic case. (Top) Heat maps represent (first row) the ID’s preferred distribution $M^*_t$, (second row) the induced agent distribution $M^a_t$ at time-steps $t = 0, 1, 2$. (Bottom) Average episodic cumulative KL divergences for each evaluation of the ID’s BO outer loop (averaged over 100 independent tests per evaluation for 4 independent runs). Without the influence of the ID, the agents behave similar to the default behaviour displayed in Fig. 3-Top middle.](image)

Proof of Proposition 4.3.

Proof. To prove the proposition, we consider the two cases (trajectory targeted and welfare targeted) of the MA’s goal separately.

Case I: Welfare Targeted

For the welfare targeted case, we firstly make the observation that the agents’ reward functions $R_i(w)$ are Lipschitz continuous in $w$. This follows from the fact that the composite function $g_1 \circ (g_2 \circ \ldots \circ (g_n(\ldots)))$ of $n < \infty$ Lipschitzian functions $g_1, g_2, \ldots, g_n$ is itself Lipschitzian (moreover we can then apply Rademacher’s lemma to ascertain differentiability almost everywhere).

Specifically, we have for the function $R_{ID}$ that

$$R_{ID}(w, h(v \cdot w)) - R_{ID}(w', h(v' \cdot w')) \leq L_{R_{ID}} \|w - w'\| + \left( h(v' \cdot w) - h(v \cdot w) \right) \leq L' \|w - w'\|$$

where $L' \triangleq L_{R_{ID}} + L_k$ and $L_{R_{ID}}$ and $L_k$ are the Lipschitz constants of $R_{ID}$ and $h$, respectively. Since $J(w, \pi) \triangleq \mathbb{E} \left[ R_{ID}(w, h(v_{\pi}^T w), \pi) \right]$ and the function $h$ is uniformly continuous, it follows $J$ is expressible as a composite function of uniformly continuous functions and hence is itself uniformly continuous (since it is in fact Lipschitz continuous).

8
To prove the remaining part of the proposition we consider now the trajectory targeted case.

Case II: Trajectory Targeted

Let us now consider a sequence \( \{w_n\} \) s.th. \( w_n \to w \) as \( n \) tends to infinity, then there exists positive scalar values \( c \) and \( d \), s.th.: 

\[
\mathbb{E}[|f(w_n, X^w_n) - f(w_n, X^w(w_n))|] 
\leq c|w - w_n| + d|X^w_n - X^w(w_n)|, 
\]

where we have used the Lipschitzianity of \( f \) to deduce the inequality. Since \( X^w_n \to X^w(w) \) as \( n \to \infty \), then by (13) and the dominated convergence theorem we can deduce that \( \exists M \in \mathbb{N} \) s.th. for \( n \geq M \) such that:

\[
\mathbb{E}[|f(w_n, X^w(w_n)) - f(w_n, X^w(w_n))|] < c \delta
\]

for some constants \( c > 0 \) and \( \delta > 0 \) s.th \( \delta \to 0 \) as \( n \to \infty \). □

REFERENCES

[1] S Ahmad, C. Tekin, M. Liu, R. Southwell, and J. Huang. 2010. Spectrum sharing as spatial congestion games. Preprint arXiv:1011.5384 (2010).

[2] Monica Babes, Enrique Munoz De Cote, and Michael I. Littman. 2008. Social reward shaping in the prisoner’s dilemma. In Proc. 7th Int. joint Conf. on Auton. agents and multiagent systems-Volume 3. International Foundation for Autonomous Agents and Multiagent Systems, 1389–1392.

[3] Xu Chen, Lei Jiao, Wenzhong Li, and Xiaoming Fu. 2016. Efficient multi-user computation offloading for mobile-edge cloud computing. IEEE/ACM Transactions on Networking 5 (2016), 2795–2808.

[4] B. Colson, P. Marcotte, and G. Savard. 2007. An overview of bilevel optimization. Annals of operations research 153, 1 (2007), 235–256.

[5] André de Palma and Robin Lindsey. 2011. Traffic congestion pricing methodologies and technologies. Transportation Research Part C: Emerging Technologies 19, 6 (2011), 1377–1399.

[6] Sam Devlin and Daniel Kudenko. 2011. Theoretical considerations of potential-based reward shaping for multi-agent systems. In The 10th Int. Conf. on Auton. Agents and Multiagent Systems-Volume 1. International Foundation for Autonomous Agents and Multiagent Systems, 225–232.

[7] Pradeep Dubey. 1986. Inefficiency of Nash equilibria. Mathematics of Operations Research 11, 1 (1986), 1–8.

[8] Nicolas Durrande. 2001. Étude de classes de noyaux adaptées à la simplification et à l’interprétation des modèles d’approximation. Une approche fonctionnelle et probabiliste. Ph.D. Dissertation. Ecole Nationale Supérieure des Mines de Saint-Etienne.

[9] Jordi Grau-Moya, Felix Leibfried, and Haitham Bou-Ammar. 2018. Balancing Two-Player Stochastic Games with Soft Q-Learning. arXiv:1802.03216 (2018).

[10] John C. Harsanyi, Reinhard Selten, et al. 1988. A general theory of equilibrium selection in games. MIT Press Books 1 (1988).

[11] Johannes Heinrich, Marc Lanctot, and David Silver. 2015. Fictitious self-play in extensive-form games. In Int. Conf. on Machine Learning. 805–813.

[12] Christian Ibars, Monica Navarro, and Lorenzo Giupponi. 2010. Distributed demand management in smart grid with a congestion game. In Smart grid communications (SmartGridComm), 2010 1st IEEE Int. Conf. IEEE, 495–500.

[13] David S. Leslie and Edmund J. Collins. 2006. Generalised weakened fictitious play. Games and Economic Behavior 56 (2006), 285 – 298.

[14] Dougal MacLaury, David Dovvenaud, and Ryan Adams. 2015. Gradient-based hyperparameter optimization through reversible learning. In Int. Conf. on Mach. Learning. 2113–2122.

[15] D. Mgini, J. Jennings, and E. Munoz de Cote. 2018. Decentralised Learning in Systems with Many, Many Strategic Agents. In Proc. of 31st AAAI Conf. on AI (AAAI-18).

[16] Fei Miao, Shao Han, Shan Lin, John A Stankovic, Desheng Zhang, Sirajum Munir, Hua Huang, Tian He, and George J Pappas. 2016. Taxi dispatch with real-time sensing data in metropolitan areas: A receding horizon control approach. IEEE Transactions on Automation Science and Engineering 13, 2 (2016), 463–478.

[17] Amya A Moledina, Jay S Coggins, Stephen Polasky, and Christopher Costello. 2003. Dynamic environmental policy with strategic firms: prices versus quantities. Journal of Environmental Economics and Management 45, 2 (2003), 356–376.

[18] Luca Moscardelli. 2013. Convergence Issues in Congestion Games. Bulletin of EATCS 3, 111 (2013).

[19] Noam Nisan and Amir Ronen. 2001. Algorithmic mechanism design. Games and Economic Behavior 35, 1-2 (2001), 166–196.

[20] M. J. Osborne and A. Rubinstein. 1994. A Course in Game Theory. MIT Press.