Distributed Learning in the Presence of Disturbances

Chithrupa Ramesh, Marius Schmitt and John Lygeros*†

April 18, 2022

Abstract

We consider a problem where multiple agents must learn an action profile that maximises the sum of their utilities in a distributed manner. The agents are assumed to have no knowledge of either the utility functions or the actions and payoffs of other agents. These assumptions arise when modelling the interactions in a complex system and communicating between various components of the system are both difficult. In [1], a distributed algorithm was proposed, which learnt Pareto-efficient solutions in this problem setting. However, the approach assumes that all agents can choose their actions, which precludes disturbances. In this paper, we show that a modified version of this distributed learning algorithm can learn Pareto-efficient solutions, even in the presence of disturbances from a finite set. We apply our approach to the problem of ramp coordination in traffic control for different demand profiles.

1 Introduction

In complex systems, modelling the interactions between various components and their relationship to the system performance is not an easy task. This poses a challenge while designing controllers for such systems, as most design methods require a model of the system. Even when considerable effort has been expended in identifying suitable models for such systems, utilising these models to design online controllers is not always easy. This is because collecting measurements of a complex system, computing control signals using complex algorithms and applying these controls to actuators across the system is communication intensive and computationally demanding. The resulting delays are not well suited to the control of real-time complex systems.

An example is the real-time control of freeway traffic, where often traffic models are highly nonlinear and methods to design controllers using these models do not scale well [2]. Furthermore, to use these models, the traffic flow from every segment of the freeway must be measured and collected, and the control signals must be delivered to the ramps on the freeway. To reduce the communication and computation burden, distributed controllers that act on mostly local information are required.

One approach is to use a distributed randomised algorithm to explore the policy space and learn the optimal actions. Recently, a distributed learning algorithm has been proposed in [1] where agents learn action profiles that maximise the system welfare. This algorithm is payoff-based, and the agents require no prior knowledge of either the utility functions or the actions and payoffs of other agents. An implicit assumption in this approach is that every agent that influences the utility can choose its actions. In reality, there might always be disturbances which cannot be chosen in a desired manner. In this paper, we extend this approach to include the effects of disturbances.

Our main contribution is a modification to the algorithm in [1] to deal with disturbances. We show that agents learn Pareto-efficient solutions in a distributed manner using our algorithm, even in the presence of disturbances from a finite set. We verify the theoretical results on a small example. In this case, all

*This work was supported by the EU project SPEEDD (FP7-ICT 619435).

†C. Ramesh, M. Schmitt and J. Lygeros are with the Automatic Control Lab, Electrical Engineering, ETH, Zurich, Switzerland. {rameshc,schmittm,lygeros}@control.ee.ethz.ch

assumptions can be verified and strong convergence guarantees can be given. To demonstrate versatility of the approach, we also apply the results to a realistic coordination problem motivated by freeway traffic control. We use our newly developed algorithm to learn a high-level coordination strategy for a ramp metering problem with promising results, using simulation parameters and traffic demand data from a real-world use case.

The learning rule used in this paper is related to the trial and error learning procedure from [3] and its cognates [1,4]. These papers proposed algorithms that learnt Nash equilibria [3], Pareto efficient equilibria [4] and Pareto-efficient action profiles [1], respectively. Convergence guarantees for the latter were presented in [5]. Restrictions on the payoff structure, which are required for the result in [1] to hold, were eliminated through the use of explicit communication in [6]. We also draw on the analysis of deliberate experimentation using the theory of regular perturbed Markov processes from [7].

This paper is organised as follows: We describe the algorithm in Section 2, and present known results in Section 3. Our main result is presented in Section 4 and illustrated on a few examples in Section 5. The conclusion is in Section 6.

2 Problem Formulation

We consider a set of agents $N := \{1, \ldots, n\}$, each with a finite action set $\mathcal{A}_i$ for $i \in N$. The disturbance is modelled as an independent and identically distributed (i.i.d.) process w_k, which takes values from a finite set $\mathbb{W}$ according to a probability distribution $\mathbf{P}_w$ that is fully supported on $\mathbb{W}$. Given an action profile $a \in \mathcal{A}$, where $\mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$, and a disturbance $w \in \mathbb{W}$, the payoff for each agent is $u_i(a, w)$. The payoffs are generated by utility functions $\mathcal{U}_i : \mathcal{A} \times \mathbb{W} \to [0, 1)$ whose functional forms are unknown to the agents. The welfare of the network of agents is $\mathcal{W}(a, w) = \sum_{i \in N} \mathcal{U}_i(a, w)$.

The agents play a repeated game; in the k^{th} iteration, each agent chooses its action $a_{i,k}$ with probability $p_{i,k} \in \Delta(\mathcal{A}_i)$, where $\Delta(\mathcal{A}_i)$ is the simplex of distributions over $\mathcal{A}_i$. The strategy $p_{i,k}$ is completely uncoupled or pay-off based, i.e., $p_{i,k} = \psi_i(\{a_{i,\tau}, u_{i,\tau}(a_\tau, w_\tau)\}_{\tau=0}^{k-1})$. In other words, an agent does not know the actions or payoffs of any other agent in the network.

Each agent maintains an internal state $z_{i,k} := [\bar{a}_{i,k}, \bar{u}_{i,k}, m_{i,k}]$ in the k^{th} iteration, where $\bar{a}_{i,k} \in \mathcal{A}_i$ is the baseline action, $\bar{u}_{i,k}$ is the corresponding baseline utility that lies in the range of $\mathcal{U}_i$ and $m_{i,k} \in \{\mathcal{C}, \mathcal{D}\}$ is the mood variable that connotes whether the agent is content or discontent. The state $z_k := \{z_{1,k}, \ldots, z_{n,k}\}$ lies in the finite state space $\mathbb{Z}$.

The algorithm is initialised with all agents setting their moods to discontent, i.e., $m_{i,0} = \mathcal{D}$ for $i \in N$. An experimentation rate $0 < \varepsilon < 1$ is fixed and a constant $c > n$ is selected. Then, each agent selects an action $a_{i,k}$ according to its mood and the corresponding probabilistic rule:

$$\begin{aligned} m_{i,k} = \mathcal{C} :\ p_i(a_{i,k}) \quad &= \begin{cases} \frac{\varepsilon^c}{|\mathcal{A}_i|-1} & a_{i,k} \neq \bar{a}_{i,k} \\ 1 - \varepsilon^c & a_{i,k} = \bar{a}_{i,k} \end{cases} \\ m_{i,k} = \mathcal{D} :\ p_i(a_{i,k}) \quad &= \frac{1}{|\mathcal{A}_i|} \quad \forall \quad a_{i,k} \in \mathcal{A}_i \end{aligned} \tag{1}$$

The agents choose their strategies based on their moods. A content agent selects its baseline action with high probability and experiments by choosing other actions with low probability. A discontent agent selects an action with uniform probability.

Each agent plays the action it has selected and receives a payoff $u_{i,k}(a_k, w_k)$, which it uses to update its state as

$$z_{i,k+1} = \begin{cases} z_{i,k} & \begin{array}{l} m_{i,k} = \mathcal{C}, a_{i,k} = \bar{a}_{i,k}, \\ |u_{i,k} - \bar{u}_{i,k}| \leq \rho \end{array} \\ \left.\begin{array}{l} z_{\mathcal{C}} \text{ w.p. } p_{\mathcal{C}} \\ z_{\mathcal{D}} \text{ w.p. } 1 - p_{\mathcal{C}} \end{array}\right\} & \text{otherwise} \end{cases} \tag{2}$$

where $z_{\mathcal{C}} = [a_{i,k}, u_{i,k}, \mathcal{C}]$, $z_{\mathcal{D}} = [a_{i,k}, u_{i,k}, \mathcal{D}]$, $p_{\mathcal{C}} = \varepsilon^{1-u_{i,k}}$ and ρ is the maximum deviation in the payoffs due to the disturbance process w, as defined in (3). The state update also depends on the mood of the agent. A content agent that chose to play its baseline action and received a payoff within the interval $u_{i,k} \in [\bar{u}_{i,k} - \rho, \bar{u}_{i,k} + \rho]$ retains its state. In all other cases, the state is updated to the played action and received payoff, and the mood is set to content or discontent with high probabilities for high or low payoffs, respectively. Thus, a content agent must receive a payoff outside an interval $\pm\rho$ of its baseline payoff to reevaluate its mood or change its state. This interval rule renders an agent insensitive to small changes in the payoff.

The variable ρ is defined as the maximum deviation in the payoffs received by any agent $i \in N$ for every action profile $a \in \mathcal{A}$ and every pair of disturbances $w_1, w_2 \in \mathbb{W}$, i.e.,

$$\rho := \arg\min_{r \in \mathbb{R}} \{|u_i(a, w_1) - u_i(a, w_2)| \leq r\} \quad \forall\, i \in N,\ \forall\, a \in \mathcal{A} \ \text{ and } \ \forall\, w_1, w_2 \in \mathbb{W}\,. \tag{3}$$

We are interested in identifying the set of states the above algorithm converges to. A necessary condition for this algorithm to function as desired is the interdependence property, stated below.

Definition 2.1 *An n-person game is* interdependent *if for every action profile $a \in \mathcal{A}$, every disturbance $w \in \mathbb{W}$ and every proper subset of agents $J \subset N$, there exists an agent $i \subset N \setminus J$, a choice of actions $a'_J \in \prod_{j \in J} \mathcal{A}_j$ and a disturbance $w' \in \mathbb{W}$, such that*

$$|\mathcal{U}_i(a'_J, a_{-J}, w') - \mathcal{U}_i(a_J, a_{-J}, w)| > \rho \tag{4}$$

This property ensures that the set of agents cannot be divided into two mutually non-interacting groups, and that a discontent agent always has recourse to actions that influence the utilities of other content agents despite the algorithm's insensitivity to the interval $[\bar{u}_i - \rho, \bar{u}_i + \rho]$ in (2).

Remark 2.1 (A remark on the state space $\mathbb{Z}$:) *The state z_k is an aggregation of the states $z_{i,k} := [\bar{a}_{i,k}, \bar{u}_{i,k}, m_{i,k}]$ of each of the agents. Thus, one would expect the cardinality of the state space to be $|\mathbb{Z}| = 2^N|\mathcal{A}||\mathbb{W}|$, because the payoffs obtained are completely determined by the choice of the actions and the disturbance. However, the interval rule in* (2) *results in more states becoming reachable and $|\mathbb{Z}| \leq 2^N|\mathcal{A}|^2|\mathbb{W}|$. The exact number of states depends on the payoffs. For the proofs presented in this paper, we define the state space in terms of the states reachable from the initial point of our algorithm, i.e., $\mathbb{Z} := \{z : \exists \tau > 0 \ s.t.\ \mathbf{P}(z_\tau = z|z_0) > 0\}$, where z_0 is any state where all agents are discontent.*

3 Preliminaries

We briefly outline Young's result on regular perturbed Markov processes [7]. Consider the Markov processes on a state space $\mathbb{X}$ with transition matrices $\mathbf{P}^0$ and $\mathbf{P}^\epsilon$, where a finite-valued $\epsilon > 0$ measures the noise level. The Markov chain induced by $\mathbf{P}^0$ describes some basic evolutionary process such as best response dynamics, while the chain induced by $\mathbf{P}^\epsilon$ represents the perturbed process obtained by introducing mistakes or experiments. This notion is formalised as follows.

Definition 3.1 *A family of Markov processes $\mathbf{P}^\epsilon$ is called a regular perturbation of a Markov chain with transition matrix $\mathbf{P}^0$ if it satisfies the following conditions:*

i. $\mathbf{P}^\epsilon$ *is aperiodic and irreducible for all finite $\epsilon > 0$.*

ii. $\lim_{\epsilon \to 0} \mathbf{P}^\epsilon_{xy} = \mathbf{P}^0_{xy}$, $\forall x, y \in \mathbb{X}$.

iii. If $\mathbf{P}^\epsilon_{xy} > 0$ for some ϵ, then $\exists\ r(x, y) \geq 0$, called the resistance *of the transition $x \to y$, such that*

$$0 < \lim_{\epsilon \to 0} \epsilon^{-r(x,y)}\, \mathbf{P}^\epsilon_{xy} < \infty \tag{5}$$

Property i ensures that there is a unique stationary distribution for all finite $\epsilon > 0$. Property ii ensures that the perturbed process converges to the unperturbed process in the limit as $\epsilon \to 0$. Property iii states that a transition $x \to y$ is either impossible under $\mathbf{P}^\epsilon$ or it occurs with a probability $\mathbf{P}^\epsilon_{xy}$ of order $\epsilon^{r(x,y)}$ for some unique, real $r(x,y)$ in the limit as $\epsilon \to 0$. Note that $r(x,y) = 0$ if and only if $\mathbf{P}^0_{xy} > 0$. Thus, the transitions of resistance zero are the same as the transitions that are feasible under $\mathbf{P}^0$.

Definition 3.2 *A state $x \in \mathbb{X}$ is said to be stochastically stable if $\mu^0_x > 0$, where μ^0 is a stationary distribution of $\mathbf{P}^0$.*

We are interested in characterizing the limiting distribution μ^0 of $\mathbf{P}^0$ through its support, or the set of stochastically stable states. To do this, we define two directed graphs. The first graph $G := (\mathbb{X}, \mathbb{E}_G)$ has as vertex set the set of states $\mathbb{X}$, and as directed edge set $\mathbb{E}_G := \{x \to y \mid \mathbf{P}^\epsilon_{xy} > 0,\ x, y \in \mathbb{X}\}$. Thus, a directed edge exists in G only if a single transition under $\mathbf{P}^\epsilon$ gets us from state x to y, for all values of $\epsilon \geq 0$. Finally, $r(x,y)$ in (5) defines the weight or resistance of this directed edge in G.

To define the second graph, we first enumerate the recurrence classes of $\mathbf{P}^0$ as $X_1, \ldots, X_L$. Then, we can define the resistance between two classes as the minimum resistance between any two states belonging to these classes, i.e.,

$$r_{\ell_1 \ell_2} := \min_{x \in X_{\ell_1}, y \in X_{\ell_2}} r(x,y), \qquad \text{for} \quad \ell_1, \ell_2 \in \{1, \ldots, L\}\ . \tag{6}$$

Note that there is at least one path from every class to every other because $\mathbf{P}^\epsilon$ is irreducible. We now define the second graph as $\mathcal{G} := (\{1, \ldots, L\}, \mathbb{E}_\mathcal{G})$. This graph has as vertex set the set of indices of the recurrence classes of $\mathbf{P}^0$, and as edge set the set of directed edges between members of the recurrence classes. Also, $r_{\ell_1 \ell_2}$ defines the resistance or weight of this directed edge.

Definition 3.3 *Let an ℓ-tree in $\mathcal{G}$ be a spanning sub-tree of $\mathcal{G}$, such that for every vertex $\ell' \neq \ell$, there exists exactly one directed path from ℓ' to ℓ. Then, the stochastic potential γ_ℓ of the recurrence class X_ℓ is defined as*

$$\gamma_\ell := \min_{T \in \mathcal{T}_\ell} \sum_{(a,b) \in T} r_{ab} \tag{7}$$

where $\mathcal{T}_\ell$ is the set of all ℓ-trees in $\mathcal{G}$.

We can now state Young's result for perturbed Markov processes [7].

Theorem 3.1 (Theorem 4 from [7]) *Let $\mathbf{P}^0$ be a time-homogenous Markov process on the finite state space $\mathbb{X}$ with recurrence classes $X_1, \ldots, X_L$. Let $\mathbf{P}^\epsilon$ be a regular perturbation of $\mathbf{P}^0$, and let μ^ϵ be its unique stationary distribution for every small positive ϵ. Then,*

i. as $\epsilon \to 0$, μ^ϵ converges to a stationary distribution μ^0 of $\mathbf{P}^0$,

ii. the recurrence class X_{ℓ^*}, with stochastic potential $\gamma_{\ell^*} := \min_{\ell \in \{1,\ldots,L\}} \gamma_\ell$, contains the stochastically stable states $\{x \in \mathbb{X} : \mu^0_x > 0\}$.

4 Learning Pareto-efficient solutions

We begin by establishing that Young's result applies to our system, resulting in a distributed algorithm for Pareto-efficient learning in the presence of disturbances. To prove this result, we enumerate the recurrence classes of $\mathbf{P}^0$ and the resistances between the classes. We use these values to identify the structure of the tree with minimum stochastic potential.

4.1 Main Result

For $\varepsilon = 0$, the transition matrix $\mathbf{P}^0$ corresponds to an unperturbed Markov process, and we begin by showing that $\mathbf{P}^\varepsilon$ is a regular perturbation on $\mathbf{P}^0$.

Lemma 4.1 *The Markov process with transition matrix $\mathbf{P}^\varepsilon$ is a regular perturbation on $\mathbf{P}^0$.*

The proof is presented in Appendix A. Next, we use Young's result from Theorem 3.1 to obtain the distributed learning outcome stated below.

Theorem 4.2 *Let G be an interdependent n-person game on a finite joint action space $\mathcal{A}$, subject to i.i.d. disturbances from a finite set $\mathbb{W}$. Under the dynamics defined in (1)–(2), a state $z = [\bar{a}, \bar{u}, m] \in \mathbb{Z}$ is stochastically stable if and only if the following conditions are satisfied:*

i. The action profile $\bar{a}$ maximises the network welfare, i.e.,

$$(\bar{a}, \bar{w}) \in \underset{a\in\mathcal{A}, w\in\mathbb{W}}{\arg\max}\, \mathcal{W} = \underset{a\in\mathcal{A}, w\in\mathbb{W}}{\arg\max} \sum_{i\in N} \mathcal{U}_i(a, w) \tag{8}$$

ii. The benchmark actions and payoffs are aligned for the maximising disturbance, i.e., $\bar{u}_i = \mathcal{U}_i(\bar{a}, \bar{w})$.

iii. The mood of each agent is content.

We present the proof for this theorem in the next section.

4.2 Recurrence Classes

The states $z \in \mathbb{Z}$ can be classified into three categories: states where all agents are content or discontent and states where some agents are content and others discontent. By inspecting the algorithm in (1)–(2), it is easy to see that as $\varepsilon \to 0$ the former states can be recurrent, but not the latter. We formalise this notion below, by defining the recurrence classes $\mathbb{D}$ and $\mathbb{C}^m$ for $0 \le m < n$ and showing that there are no other recurrence classes.

Discontent Class $\mathbb{D}$: The states in this recurrence class correspond to those where all agents are discontent.

$$\mathbb{D} := \Big\{ z \in \mathbb{Z} \quad \Big| \quad m_i(z) = \mathcal{D},\ \forall i \in N \Big\} \tag{9}$$

Note that the payoffs and action profiles are aligned, i.e., $\bar{u}_i(z) = \mathcal{U}_i(\bar{a}(z), w)$, $\forall z \in \mathbb{D}$, $\forall i \in N$ and for some $w \in \mathbb{W}$. Also, corresponding to each action profile and disturbance pair $(a, w) \in \mathcal{A} \times \mathbb{W}$, there is a discontent state in this recurrence class.

States containing only content agents can be categorised further into n classes, $\mathbb{C}^m$ for $0 \le m < n$, as follows.

0^{th}-Content Class $\mathbb{C}^0$: This recurrence class contains singleton states where all agents are content, and where the payoffs of all agents are aligned with the action profile for some value of the disturbance $w \in \mathbb{W}$, while satisfying the interval rule in (2) for all other values of the disturbance. Let B_i denote the set of states that satisfy these conditions on the payoffs of the i^{th} agent:

$$\begin{aligned} B_i := \Big\{ z \in \mathbb{Z} \big| \bar{u}_i(z) = \mathcal{U}_i(\bar{a}(z), w), &\text{for some } w \in \mathbb{W}, \\ &|\bar{u}_i(z) - \mathcal{U}_i(\bar{a}(z), \tilde{w})| \le \rho,\ \ \forall\ \tilde{w} \in \mathbb{W} \Big\}. \end{aligned} \tag{10}$$

Then, the recurrence class $\mathbb{C}^0$ is defined as

$$\mathbb{C}^0 := \Big\{ z \in \mathbb{Z} \quad \Big| \quad m_i(z) = \mathcal{C}, z \in B_i, \forall i \in N \Big\}. \tag{11}$$

From the definition of ρ in (3), we know that corresponding to each action profile and disturbance pair $(a, w) \in \mathcal{A} \times \mathbb{W}$, there is a state in this recurrence class with payoffs satisfying (10).

There might also be states where the payoffs of all agents are not aligned with the action profile for any single value of the disturbance $w \in \mathbb{W}$. Some of these states can be recurrent, and belong to the classes defined below.

1st-Content Class $\mathbb{C}^1$: Suppose that a proper subset of agents $J_1 \subset N$ from a state $z' \in \mathbb{C}^0$ experiment with different actions despite being content, and become content with their new utilities. If the rest of the agents $j_0 \in J_0 = N \setminus J_1$ do not notice this change, because their new utilities lie within the interval $[\bar{u}_{j_0}(z') - \rho, \bar{u}_{j_0}(z') + \rho]$ for all values of the disturbance, then the agents find themselves in a state z in a recurrence class $\mathbb{C}^1$.

$$\begin{aligned}\mathbb{C}^1 := \Big\{ z \in \mathbb{Z} \;\Big|\; & m_i(z) = \mathcal{C}, \forall i \in N, \\ & \exists (J_0, J_1) \text{ s.t. } J_0 \cup J_1 = N, z \in B_{j_1}, \quad \forall\, j_1 \in J_1, \\ & \exists z' \in \mathbb{C}^0 \text{ s.t. } z_{j_0} = z'_{j_0}, |\bar{u}_{j_0}(z') - \mathcal{U}_{j_0}(\bar{a}(z), \tilde{w})| \le \rho, \quad \forall\, \tilde{w} \in \mathbb{W}, \forall\, j_0 \in J_0 \Big\},\end{aligned} \tag{12}$$

where the symbol $\cup$ denotes a disjoint union of the subsets.

A subset of agents from a state in $\mathbb{C}^1$ could experiment and find themselves in a state in a recurrence class $\mathbb{C}^2$. In general, states in the recurrence class $\mathbb{C}^m$ can be reached from a state in $\mathbb{C}^{m-1}$, following a similar procedure. The recurrence class $\mathbb{C}^m$ is defined below.

mth-Content Class $\mathbb{C}^m$: These recurrence classes contain singleton states where all agents are content, and where the agents can be divided into $m+1$ mutually disjoint subsets $J_0, \ldots, J_m$, such that the utilities of the agents within each subset are aligned with an action profile for some value of the disturbance.

$$\begin{aligned}\mathbb{C}^m := \Big\{ z \in \mathbb{Z} \;\Big|\; & m_i(z) = \mathcal{C}, \forall i \in N, \\ & \exists (J_0, \ldots, J_m) \text{ s.t. } \cup_{l=0}^{m} J_l = N, z \in B_{j_m}, \quad \forall\, j_m \in J_m, \\ & \exists z' \in \mathbb{C}^{m-1} \text{ s.t. } z_{j_\ell} = z'_{j_\ell}, |\bar{u}_{j_\ell}(z') - \mathcal{U}_{j_\ell}(\bar{a}(z), \tilde{w})| \le \rho, \quad \forall\, \tilde{w} \in \mathbb{W}, \forall\, j_\ell \in N \setminus J_m \Big\}.\end{aligned} \tag{13}$$

There can be at most n disjoint subsets from a set of n agents, and hence $m < n$. Clearly, there might be many states in $\mathbb{Z}$, where the baseline payoffs and actions satisfy some, but not all, of the above conditions for classes $\mathbb{C}^m$, $1 \le m < n$. These states are not recurrent, as we show below.

Lemma 4.3 *The recurrence classes corresponding to the n-person interdependent game described by* (1)–(2) *are* $\mathbb{D}$*, and the singletons in* $\mathbb{C}^0$ *and* $\mathbb{C}^m$*, for* $0 < m < n$*, as defined in* (9)–(13)*, respectively.*

The proof of this result is presented in Appendix A.

4.3 Resistances and Trees

Transitions can occur between all three recurrence class types, namely $\mathbb{D} \to \mathbb{C}^0$ and vice versa, $\mathbb{D} \to \mathbb{C}^m$ and vice versa, and $\mathbb{C}^l \to \mathbb{C}^m$ for $0 \le l, m < n$ and $l \ne m$. In addition, the singleton states in $\mathbb{C}^0$ and $\mathbb{C}^m$ can transition to other singleton states within the same classes. All these transitions are enumerated along with the corresponding resistances in Table 1. In this table, we use $d \in \mathbb{D}$, $z^0_\cdot \in \mathbb{C}^0$ and $z^m_\cdot \in \mathbb{C}^m$ to denote states in the respective recurrence classes. Some of the entries contain the term $\bar{r}_m$, which is given by

$$\bar{r}_m = mc^2 + \frac{4 + m - m^2}{2} c - \frac{m(m+1)}{2}, \; 0 < m < n\,. \tag{14}$$

The calculations for the entries in Table 1 are presented in Appendix B. We can now compute the stochastic potential of a state in $\mathbb{C}^0$ and show that a minimum potential tree is rooted at a singleton in $\mathbb{C}^0$.

Table 1: Resistances Between Recurrence Classes

No.	Path	Resistance Relationship
1	$\mathbb{D} \to \mathbb{C}^0$	$r_{dz^0} = \sum_{i \in N} 1 - \bar{u}_i(z^0)$
2	$\mathbb{D} \to \mathbb{C}^m$	$r_{dz^m} = \min_{z^0 \in \mathbb{C}^0} r_{dz^0} + r_{z^0 z^m}$
3	$\mathbb{C}^0 \to \mathbb{D}$	$r_{z^0 d} = c$
4	$\mathbb{C}^m \to \mathbb{D}$	$r_{z^m d} = c$
5	$\mathbb{C}^0 \to \mathbb{C}^0$	$c \leq r_{z_1^0 z_2^0} \leq 2c$
6	$\mathbb{C}^m \to \mathbb{C}^m$	$c \leq r_{z_1^m z_2^m} \leq \bar{r}_m$
7	$\mathbb{C}^l \to \mathbb{C}^m$, $0 \leq l, m < n,\ m \neq l$	$\lvert m - l \rvert c \leq r_{z_1^m z_2^m} \leq \bar{r}_m$

Lemma 4.4 *The stochastic potential of a state $z^0 \in \mathbb{C}^0$ is*

$$\gamma(z^0) = c\left(\sum_{m=0}^{n-1} |\mathbb{C}^m| - 1\right) + \sum_{i \in N} \left(1 - \bar{u}_i(z^0)\right) \ . \tag{15}$$

Lemma 4.5 *The states in the recurrence class $\mathbb{D}$ and the singletons $\mathbb{C}^m$, for $0 < m < n$, are not stochastically stable.*

The proofs for both Lemmas are presented in Appendix A.

4.4 Proof of the Main Result

We now present the proof of Theorem 4.2.

Proof The stochastically stable states are contained in the recurrence class of $\mathbf{P}^0$ with minimum stochastic potential (from Theorem 3.1). From Lemma 4.5, we also know that the recurrence class with minimum stochastic potential is rooted at a singleton in $\mathbb{C}^0$.

Lemma 4.4 gives us the minimum stochastic potential as

$$\gamma(z^{0,*}) = \min_{z_0 \in \mathbb{C}^0} c(\sum_{m=0}^{n-1} |\mathbb{C}^m| - 1) + \sum_{i \in N} (1 - \bar{u}_i(z^0))$$

Thus, the action profile corresponding to the state $z^{0,*}$ must satisfy $\bar{a}(z^{0,*}) \in \arg\max_{a \in \mathcal{A}, w \in \mathbb{W}} \sum_{i \in N} \mathcal{U}_i(a, w)$.

From the definition of the recurrence class $\mathbb{C}^0$ in (11), we obtain statements ii and iii of the theorem.

5 Examples

We present a simple example of a two-agent interdependent game to illustrate the results of Theorem 4.2, and then apply this method to the ramp coordination problem.

Example 1 *Consider a simple game G_2 with $n = 2$ agents. The action sets, disturbance set and payoffs are given in Table 2. The disturbance process is uniformly distributed on $\{0, 1\}$. It is easy to verify that $\rho = 0.1$ (from (3)), and that the interdependence property (from Definition 2.1) is satisfied, for G_2.*

Table 2: Payoffs in Example 1

$\{a_1, a_2, w\}$	u_1	u_2	$\{a_1, a_2, w\}$	u_1	u_2
$\{0, 0, 0\}$	0.30	0.40	$\{1, 1, 0\}$	0.80	0.90
$\{0, 0, 1\}$	0.40	0.30	$\{1, 1, 1\}$	0.90	0.80
$\{0, 1, 0\}$	0.20	0.10	$\{2, 0, 0\}$	0.65	0.55
$\{0, 1, 1\}$	0.10	0.20	$\{2, 0, 1\}$	0.55	0.65
$\{1, 0, 0\}$	0.60	0.50	$\{2, 1, 0\}$	0.75	0.85
$\{1, 0, 1\}$	0.50	0.60	$\{2, 1, 1\}$	0.85	0.75

Table 3: Fraction of occurrence of states in Example 1

State z	Normalised Number of Instances	State z	Normalised Number of Instances
$\{[0, 0.10, \mathcal{C}], [1, 0.20, \mathcal{C}]\}$	0.0009	$\{[1, 0.60, \mathcal{C}], [0, 0.50, \mathcal{C}]\}$	0.0013
$\{[1, 0.50, \mathcal{C}], [0, 0.60, \mathcal{C}]\}$	0.0016	$\{[1, 0.50, \mathcal{C}], [0, 0.50, \mathcal{C}]\}$	0.0004
$\{[\mathbf{1}, \mathbf{0.80}, \mathcal{C}], [\mathbf{1}, \mathbf{0.90}, \mathcal{C}]\}$	**0.9532**	$\{[1, 0.90, \mathcal{C}], [1, 0.90, \mathcal{C}]\}$	0.0123
$\{[1, 0.90, \mathcal{C}], [1, 0.80, \mathcal{C}]\}$	0.0032	$\{[1, 0.90, \mathcal{C}], [1, 0.85, \mathcal{C}]\}$	0.0034
$\{[1, 0.8, \mathcal{C}], [1, 0.85, \mathcal{C}]\}$	0.0002	$\{[2, 0.65, \mathcal{C}], [0, 0.65, \mathcal{C}]\}$	0.0014
$\{[2, 0.55, \mathcal{C}], [0, 0.65, \mathcal{C}]\}$	0.0007	$\{[2, 0.65, \mathcal{C}], [0, 0.55, \mathcal{C}]\}$	0.0004
$\{[2, 0.55, \mathcal{C}], [0, 0.55, \mathcal{C}]\}$	0.0002	$\{[2, 0.65, \mathcal{C}], [0, 0.60, \mathcal{C}]\}$	0.0007
$\{[2, 0.85, \mathcal{C}], [1, 0.85, \mathcal{C}]\}$	0.0022	$\{[2, 0.75, \mathcal{C}], [1, 0.85, \mathcal{C}]\}$	0.0015
$\{[2, 0.85, \mathcal{C}], [1, 0.75, \mathcal{C}]\}$	0.0059	$\{[2, 0.75, \mathcal{C}], [1, 0.80, \mathcal{C}]\}$	0.0088
$\{[2, 0.85, \mathcal{C}], [1, 0.80, \mathcal{C}]\}$	0.0006		

We simulated 10^6 *iterations of the algorithm* (1)–(2) *in Matlab, with a time-varying* ε*-sequence, and* $c = 2$*. The experimentation rate was modified by setting* $\varepsilon_{k+1} = 0.99995\varepsilon_k$*, with an initial value of* $\varepsilon_1 = 0.1$*. The results of a typical sample run of our simulation are presented in Table 3, validating the results of Theorem 4.2. The average welfare over all the iterations was* 1.6937*.*

In Table 3, we have displayed a list of states and the normalised number of occurrences of these states, only when this figure was larger than 0.0001*. This is because a total of* 101 *states were explored by this simulation. Note that some of the states, such as* $\{[2, 0.85, \mathcal{C}], [1, 0.80, \mathcal{C}]\}$ *are examples of states in* $\mathbb{C}^1$*.*

The average welfare obtained from playing the optimal action profile(s) will, in general, be different from $\mathcal{W}^*$, the maximum welfare in (8). This is because the optimal action profile maximises the welfare for the most favourable value of the disturbance as per Theorem 4.2. When averaged over all possible values of the disturbance, the welfare will lie in the interval $[\mathcal{W}^* - n\rho, \mathcal{W}^* + n\rho]$, depending on $\mathbf{P}_w$. For the above example, the average welfare equals $\mathcal{W}^*$, which may not always be the case as we see in the next example.

Example 2 *In freeway traffic control, one seeks to estimate the occupancy of a freeway, usually via loop detectors [9], and subsequently adjust speed limits [10] or traffic lights on the onramps [11], a technique known as ramp metering, to improve traffic flow. However, popular freeway traffic models such as the cell transmission model [12, 13] or the Metanet model [14], see also [15] for an overview of traffic models, are highly nonlinear and methods to design controllers for traffic networks often do not scale well. To reduce the communication and computation burden, distributed controllers that act on mostly local information are required. Model-based approaches for decomposition exist [16, 17], but in fact, local feedback [18] and the combination of local feedback and heuristic, high-level coordination [19] are among the most popular and*

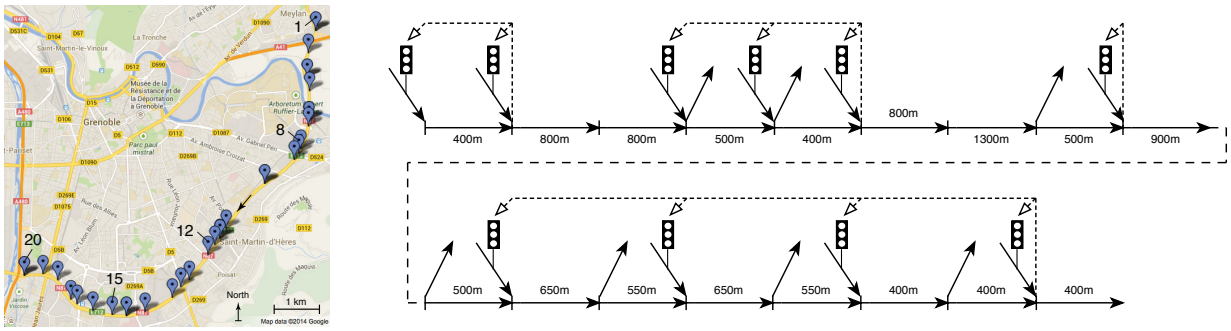

Figure 1: Map of the Grenoble South Link as depicted in [8] and the corresponding freeway topology. Also shown is a particular coordination pattern between onramps, corresponding to the action profile $[\text{COR}, \text{LOC}, \text{COR}, \text{COR}, \text{LOC}, \text{LOC}, \text{COR}, \text{COR}, \text{COR}, \text{LOC}]^\top$.

practically successful strategies. To demonstrate its versatility, we will evaluate the efficacy of our method in learning a ramp coordination pattern, for given local controllers.

Consider a freeway with a number of onramps. The idea of ramp metering is to control the traffic inflow from the ramps via traffic lights so as to avoid congestion on the mainline. Both theoretical [20] and practical [21] studies have demonstrated that this approach can potentially avoid traffic breakdown in congestion and reduce the sum of travel times of all drivers (TTS, Total Time Spent). An effective metering strategy is to control the inflow such that the local traffic density does not exceed the threshold to congestion, the so-called critical density [18]. However, there are limits to this strategy. Multiple ramps, each controlling the local traffic densities, are coupled by the mainline flow as traffic travels downstream and congestion queues can spill back upstream. If no control action of a single ramp is sufficient to prevent congestion of an adjacent bottleneck, then coordination between ramps may hold the answer [19].

In this example, we aim to learn a coordination pattern, while the low-level metering policy remains fixed. We consider ten ramps on a freeway located in Grenoble, as presented in [8] and depicted in Figure 1. We allow for ramps either to control only the local traffic density (LOC) or to coordinate with downstream ramps (COR) and control the ramp occupancy, i.e. the queue length divided by the ramp length, according to the occupancy of the next downstream ramp. Therefore, the action set for every agent, i.e., every ramp i is $\mathcal{A}_i = \{LOC, COR\}$. The utility is computed by simulations of the freeway using the modified cell-transmission model as described in [22], which uses a non-monotonic demand function to model the capacity drop empirically observed in a congested freeway. The local utility for agent i is computed as the sum of the total travel time of all cars in the adjacent section of the freeway and the total waiting time in the onramp queue, which is then mapped to the interval $[0, 1]$ via a linear transformation. The utilities do not only depend on the action profile but also on the traffic demand, which acts as an external disturbance. We consider real traffic demands during peak hours of the weekdays May 11^{th} - May 15^{th}, and hence, the disturbance set is $\mathbb{W} = \{Mon, Tue, Wed, Thu, Fri\}$.

We do not try to identify ρ as per (3) or verify the interdependence property in this example. Instead, we simply choose ρ to be sufficiently large to ensure convergence of the algorithm within a reasonable number of iterations. We ensure that interdependence holds by complimenting the interaction graph with a communication graph, as suggested in [6]. Each agent broadcasts the mood it computes in (2). It then receives all the other agents' moods and performs the following update step to finalize its own mood, as per

$$m_{i,k+1} = \begin{cases} \mathcal{D} & \tilde{m}_{i,k+1} = \mathcal{D} \\ \mathcal{C} & \tilde{m}_{j,k+1} = \mathcal{C},\ \forall j \in N \\ \left.\begin{array}{l} \mathcal{C} \text{ w.p. } \varepsilon^\beta \\ \mathcal{D} \text{ w.p. } 1-\varepsilon^\beta \end{array}\right\} & \text{otherwise} \end{cases} \tag{16}$$

where $\tilde{m}_{i,k+1}$ is the mood of the i^{th} agent updated locally as per (2). The above update compliments the interaction between the agents by coupling the moods, and is controlled by the parameter β. If each agent broadcasts its mood to all other agents, this update alone will suffice to ensure the interdependence property, irrespective of the utility functions. Thus, all the results in this paper, including Theorem 4.2, can be shown to hold for this modified algorithm, for ρ chosen as per (3). However, in a real-world setting it might be difficult to compute a suitable bound on ρ beforehand. Instead, we chose ρ empirically to facilitate quick convergence, sacrificing the guarantees that come with Theorem 4.2. The performance is then checked a posteriori.

We simulated 1000 iterations of the algorithm (1)–(2), (16), in Matlab, with $\varepsilon = 0.0001$, $c = 10$, $\beta = 0.00005$ and $\rho = 0.6$. The algorithm explored 36 different action profiles before settling on the ramp coordination schedule $[COR, LOC, LOC, COR, COR, COR, LOC, LOC, COR, LOC]^\top$. The corresponding baseline utility was 9.6 and the algorithm spent 890 out of 1000 iterations in the above state. The average utility obtained over the entire simulation run was 8.72, in comparison to an average utility of 6.30 for the uncontrolled case. In terms of travel times, this corresponds to savings of 31% over the uncoordinated case. Note that we compute the savings just for the rush-hour period and therefore this value might exceed the savings typically reported for ramp metering field trials, which are usually computed for the entire day [11].

6 Conclusions

We presented a distributed learning algorithm, based on the algorithm in [1], that can be used to learn Pareto-efficient solutions in the presence of disturbances. Our algorithm learns efficient action profiles corresponding to the most favourable disturbance, and specifies a range for the average welfare. In general, the approach outlined in this paper is particularly well suited to problems where the disturbances can be modelled as a finite set of small perturbations from a nominal model. Our examples validated the main result in our paper, and also illustrated the potential of this randomised approach. In many applications, the average welfare is an important performance metric. In future work, we wish to explore randomized approaches that optimize the average welfare obtained.

References

[1] J. R. Marden, H. P. Young, and L. Y. Pao, "Achieving pareto optimality through distributed learning," *SIAM Journal on Control and Optimization*, vol. 52, no. 5, pp. 2753–2770, 2014.

[2] R. E. Allsop, "Transport networks and their use: how real can modelling get?," *Phil. Trans. R. Soc. A*, vol. 366, pp. 1879 – 1892, 2008.

[3] H. P. Young, "Learning by trial and error," *Games and Economic Behavior*, vol. 65, pp. 626–643, March 2009.

[4] B. S. Pradelski and H. P. Young, "Learning efficient nash equilibria in distributed systems," *Games and Economic Behavior*, vol. 75, no. 2, pp. 882 – 897, 2012.

[5] A. Menon and J. S. Baras, "Convergence guarantees for a decentralized algorithm achieving pareto optimality," in *Proceedings of the 2013 American Control Conference*, pp. 1935–1940, 2013.

[6] A. Menon and J. S. Baras, "A distributed learning algorithm with bit-valued communications for multi-agent welfare optimization," in *Proceedings of the 52nd IEEE Conference on Decision and Control (CDC)*, pp. 2406–2411, 2013.

[7] H. P. Young, "The evolution of conventions," *Econometrica*, vol. 61, no. 1, pp. pp. 57–84, 1993.

[8] C. C. de Wit, F. Morbidi, L. L. Ojeda, A. Y. Kibangou, I. Bellicot, and P. Bellemain, "Grenoble traffic lab: An experimental platform for advanced traffic monitoring and forecasting," *Control Systems, IEEE*, vol. 35, no. 3, pp. 23–39, 2015.

[9] R. Gibbens and Y. Saatci, "Data, modelling and inference in road traffic networks," *Phil. Trans. R. Soc. A*, vol. 366, pp. 1907–1919, 2008.

[10] A. Hegyi, B. D. Schutter, and J. Hellendoorn, "Optimal coordination of variable speed limits to suppress shock waves," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 1, pp. 102 – 112, 2005.

[11] M. Papageorgiou and A. Kotsialos, "Freeway ramp metering: An overview," in *Intelligent Transportation Systems, 2000. Proceedings. 2000 IEEE*, pp. 228–239, IEEE, 2000.

[12] C. F. Daganzo, "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transportation Research Part B: Methodological*, vol. 28, no. 4, pp. 269–287, 1994.

[13] C. F. Daganzo, "The cell transmission model, part ii: network traffic," *Transportation Research Part B: Methodological*, vol. 29, no. 2, pp. 79–93, 1995.

[14] A. Messner and M. Papageorgiou, "Metanet: A macroscopic simulation program for motorway networks," *Traffic Engineering & Control*, vol. 31, no. 8-9, pp. 466–470, 1990.

[15] H. J. Payne, "Models of freeway traffic and control," *Simulation Council Proc.*, vol. 1, pp. 51 – 61, 1971.

[16] W. B. Dunbar and R. M. Murray, "Distributed receding horizon control for multi-vehicle formation stabilization," *Automatica*, vol. 42, no. 4, pp. 549 – 558, 2006.

[17] L. B. de Oliveira and E. Camponogara, "Multi-agent model predictive control of signaling split in urban traffic networks," *Transportation Research Part C: Emerging Technologies*, vol. 18, no. 1, pp. 120 – 139, 2010.

[18] M. Papageorgiou, H. Hadj-Salem, and J.-M. Blosseville, "Alinea: A local feedback control law for on-ramp metering," *Transportation Research Record*, no. 1320, pp. 58–64, 1991.

[19] I. Papamichail, M. Papageorgiou, V. Vong, and J. Gaffney, "Heuristic ramp-metering coordination strategy implemented at monash freeway, australia," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2178, no. 1, pp. 10–20, 2010.

[20] G. Gomes and R. Horowitz, "Optimal freeway ramp metering using the asymmetric cell transmission model," *Transportation Research Part C: Emerging Technologies*, vol. 14, no. 4, pp. 244–262, 2006.

[21] M. Papageorgiou, C. Diakaki, V. Dinopoulou, A. Kotsialos, and Y. Wang, "Review of road traffic control strategies," *Proceedings of the IEEE*, vol. 91, no. 12, pp. 2043–2067, 2003.

[22] I. Karafyllis, M. Kontorinaki, and M. Papageorgiou, "Global exponential stabilization of freeway models," *arXiv preprint arXiv:1408.5833*, 2014.

A Proof of Lemmas

Proof of Lemma 4.1 We first show that Property i holds under $\mathbf{P}^{\varepsilon}$ for $\varepsilon > 0$. Note that all states are accessible from any state $d \in \mathcal{D}$, from the definition of the state space $\mathbb{Z}$. In other words, there exists an integer $\tau > 0$ such that $\mathbf{P}(z_{k+\tau} = z' | z_k = d) > 0$ for all $z' \in \mathbb{Z}$. Next, note that both content and discontent agents chooses an action with a probability distribution that is fully supported on $\mathbb{A}_i$, as per (1). Due to

this, one or more agents can become discontent. Along with the interdependence property, this ensures that all agents can consequently become discontent, and thus, a state such as d is accessible from any other state. In other words, there exists an integer $\tau' > 0$ such that $\mathbf{P}(z_{k+\tau'} = d | z_k = z') > 0$ for all $z' \in \mathbb{Z}$. This proves that the Markov chain is irreducible. Furthermore, many of these states permit a return to the same state with some positive probability, i.e., there exist states $z \in \mathbb{Z}$ such that $\mathbf{P}(z_{k+1} = z | z_k = z) > 0$. These states are aperiodic, which in combination with the irreducibility property effectively renders the Markov chain aperiodic.

By inspection of (1)–(2), it is clear that Property ii is satisfied. We now show that Property iii holds. Note that the transition probabilities contain terms with exponents of ε or its complement. The resistance $r = 0$ for transition probabilities containing complements of ε, because these transitions occur under $\mathbf{P}^0$. All other transition probabilities contain negative exponents of ε resulting in positive resistances, as required by Property iii.

Proof of Lemma 4.3 Consider a state $z \in \mathbb{D}$. Under $\mathbf{P}^0$, each agent picks an action with uniform probability and no utility ever makes an agent content. Thus, any accessible state remains in $\mathbb{D}$.

Consider a singleton state in any of the classes $\mathbb{C}^0$ or $\mathbb{C}^m$, for $0 < m < n$. Under $\mathbf{P}^0$, each agent plays the same action again, and the utilities received by the agents satisfy the interval rule. Thus, the agents remain in the same state.

In the $\mathbb{C}^m$ states, for $0 \leq m < n-1$, when a subset of agents $J \subset N$ choose a new action and become content, there are two circumstances under which the new state is not recurrent. If a value of the disturbance results in a payoff that does not satisfy the interval rule, the corresponding agent(s) become discontent and a new state is reached, which has a mix of content and discontent agents. Under $\mathbf{P}^0$, a discontent agent remains discontent. Furthermore, due to the interdependence property, a subset of discontent agents will cause at least one content agent to violate the interval rule and become discontent. This repeats until all agents are discontent, thus reaching $\mathbb{D}$. A similar situation occurs if the new action causes the payoff received by any agent in $N \setminus J$ to fall outside the prescribed interval. Thus, in general, no state with a mix of content and discontent agents, is recurrent.

Proof of Lemma 4.4 First we show that $\gamma(z^0)$ is less than or equal to the right hand side in (15), and then we show the reverse, thus proving the equality relationship in the lemma.

To show the first path, construct the following tree T rooted at z^0: Add a directed link $z^{0'} \to d$ with resistance c between every $z^{0'} \in \mathbb{C}^0 \setminus z^0$ and some $d \in \mathbb{D}$. Then, add a directed link $z^m \to d$ with resistance c between every $z^m \in \mathbb{C}^m$, for $0 < m < n$, and some $d \in \mathbb{D}$. Finally add a directed link $d \to z^0$ from some $d \in \mathbb{D}$, with resistance $\sum_{i \in N}(1 - \bar{u}_i(z^0))$. The resistance of this tree $\gamma(T) = c(\sum_{m=0}^{n-1} |\mathbb{C}^m| - 1) + \sum_{i \in N}(1 - \bar{u}_i(z^0))$, thus establishing that $\gamma(z^0) \leq \gamma(T)$.

To show the reverse, consider a general tree T' rooted at z^0. It may differ from T in one or more of the following aspects:

(a) It may contain a path of length q between the discontent class and the singleton z^0, such as $d \to z_1^{m_1} \to \ldots z_q^{m_q} \to z^0$, where $d \in \mathbb{D}$, $z_1^{m_1} \in \mathbb{C}^{m_1}$, $\ldots$, $z_q^{m_q} \in \mathbb{C}^{m_q}$.

(b) It may contain paths of length s between a singleton from any of the m^{th}-content classes and the discontent class, such as $z^m \to z_1^{m_1} \to \ldots z_s^{m_s} \to d$, where $z^m \in \mathbb{C}^m$, $z_1^{m_1} \in \mathbb{C}^{m_1}$, $\ldots$, $z_s^{m_s} \in \mathbb{C}^{m_s}$ and $d \in \mathbb{D}$.

From Table 1, we note that the path of length q in case (a) has a resistance $r(d \to z_1^{m_1} \to \ldots z_q^{m_q} \to z^0) \geq qc + \sum_{i \in N}(1 - \bar{u}_i(z^0))$. Construct a tree $T_{(a)}$ by replacing this path in T' with a set of links $z_i^{m_i} \to d$, for $1 \leq i \leq q$, and $d \to z^0$. By making these changes, we obtain a total resistance of $qc + \sum_{i \in N}(1 - \bar{u}_i(z^0))$. Thus, we have constructed a tree with $\gamma(T_{(a)}) \leq \gamma(T')$.

Next, note that the path in case (b), $z^m \to z_1^{m_1} \to \ldots z_s^{m_s} \to d$ has a resistance $r \geq (s+1)c$. Construct a tree $T_{(b)}$ by replacing the links in the path with the links $z^m \to d$ and $z_i^{m_i} \to d$, for $1 \leq i < s$, each of resistance c. Then, $\gamma(T_{(b)}) \leq \gamma(T')$. Thus, we have shown that $\gamma(T^*) = \gamma(T) \leq \gamma(z^0)$.

Proof of Lemma 4.5 Consider a tree rooted at $\mathbb{D}$. Then, it must contain a link $z^0 \to d$, for some $d \in \mathbb{D}$, of resistance c. Replace it with the link $d \to z^0$ that incurs a lesser resistance $\sum_{i\in N}(1-\bar{u}_i(z^0)) \leq n < c$. Thus, we have constructed a tree rooted at z^0 with lower potential.

Similarly, consider a tree rooted at $z^m \in \mathbb{C}^m$, for $0 < m < n$. This tree must contain a path $d \to z^0 \to z^1 \to \cdots \to z^m$, with resistance $r \geq mc + \sum_{i\in N}(1-\bar{u}_i(z^0))$. Replace the links in the path $z^0 \to z^m$ with the links $z^l \to d'$, for $0 < l \leq m$ and $d' \in \mathbb{D}$, which results in a resistance of exactly $mc + \sum_{i\in N}(1-\bar{u}_i(z^0))$. Thus, we have constructed a tree rooted at z^0 with lower stochastic potential, and proved our result.

B Calculation of Resistances

In this section, we use d, z^0 and z^m to denote a state $d \in \mathbb{D}$ and singleton states $z^0 \in \mathbb{C}^0$ and $z^m \in \mathbb{C}^m$, respectively.

Let us begin with row 1 of Table 1. The transition $d \to z^0$ occurs only when all agents are content with the received payoffs, which happens with probability $\prod_{i\in N}\varepsilon^{1-u_i}$. This gives us the resistance r_{dz^0} in row 1. The resistance r_{dz^m} in row 2 follows from the definition of a resistance between recurrence classes in (6) and the fact that the transition $d \to z^m$ only occurs through a state z^0.

For the transition $z^0 \to d$ to occur, at least one content agent must experiment and become discontent, which happens with probability of order $O(\varepsilon^c)$. Then, all agents become discontent eventually. Thus, $r_{z^0 d} = c$ in row 3. The same holds for the transition $z^m \to d$ in row 4.

In row 5, the transition $z_1^0 \to z_2^0$ between the singleton states $z_1^0, z_2^0 \in \mathbb{C}^0$ can occur in multiple ways. The transition with the least resistance occurs when an agent experiments with a new action and becomes content with the payoff it receives, while not affecting the payoffs of any other agent. Thus, $r_{z_1^0 z_2^0} \geq c + 1 - \bar{u}_i(z_2^0) \geq c$. In general, however, this transition occurs through intermediate states in $\mathbb{D}$, resulting in $r_{z_1^0 z_2^0} \leq c + \sum_{i\in N} 1 - \bar{u}_i(z_2^0) \leq c + n \leq 2c$. Transitions through states in $\mathbb{C}^m$ are not considered because these incur resistances of greater than $2c$.

A similar argument can be applied to calculate the resistance $r_{z_1^m z_2^m}$ of a transition between two singleton states $z_1^m, z_2^m \in \mathbb{C}^m$ in row 6. When the transition only requires one agent to experiment and be content with a new action, $r_{z_1^m z_2^m} \geq c + 1 - \bar{u}_i(z_2^m) \geq c$. Other transitions occur through intermediate states in $\mathbb{D}$, resulting in

$$r_{z_1^m z_2^m} \leq \min_{z^0\in\mathbb{C}^0, z^1\in\mathbb{C}^1,\ldots,z^{m-1}\in\mathbb{C}^{m-1}} c + \sum_{i\in N} 1 - \bar{u}_i(z^0) + r_{z^0 z^1} + \cdots + r_{z^{m-1} z_2^m}$$

The worst case least resistance path $z^0 \to z_2^m$ occurs when $n-1$ agents experiment and become content to ensure $z^0 \to z^1$, $n-2$ agents experiment and become content to ensure $z^1 \to z^2$ and so on until $n-m$ agents experiment and become content to ensure $z^{m-1} \to z_2^m$. This can happen when intermediate states, which are required to ensure that $z^0 \to z_2^m$ occurs with fewer experimenting agents, are not recurrent. Thus, we get

$$\begin{aligned} r_{z_1^m z_2^m} \leq \min_{z^0\in\mathbb{C}^0, z^1\in\mathbb{C}^1,\ldots,z^{m-1}\in\mathbb{C}^{m-1}} \quad & c + n - \sum_{i\in N} \bar{u}_i(z^0) \\ & + \underbrace{(n-1)c + \sum_{j_1=1}^{n-1} 1 - \bar{u}_{j_1}(z^1)}_{r_{z^0 z^1}} + \cdots + \underbrace{(n-m)c + \sum_{j_m=1}^{n-m} 1 - \bar{u}_{j_m}(z_2^m)}_{r_{z^{m-1} z_2^m}} \\ \leq \quad & c + n + (n-1)c + n - 1 + \cdots + (n-m)c + n - m \end{aligned}$$

Using the fact that $c > n$ and summing over the series, we obtain the upperbound in (14). Again, transitions through states in $\mathbb{C}^l$, for $l \neq m$, are not considered because these may incur resistances of greater than $\bar{r}_m$.

Similar arguments can be used to calculate the resistance in row 7 of a transition between two singleton states $z^l \in \mathbb{C}^l$ and $z^m \in \mathbb{C}^m$, for $0 \leq l, m < n$ and $m \neq l$. The least resistant paths require $|m-l|$ agents to

experiment and become content, and other transitions occur through an intermediate state in $\mathbb{D}$. Transitions through other states in $\mathbb{C}^s$, for $1 \leq s < n$, are not considered as the resistances of such paths can be higher than $\bar{r}_m$.