Games and Economic Behavior 144 (2024) 84–103

Contents lists available at ScienceDirect

Games and Economic Behavior

journal homepage: www.elsevier.com/locate/geb

Reinforcement learning in a prisoner's dilemma

Arthur Dolgopolov

Center for Mathematical Economics (IMW), Postfach 100131, Bielefeld University, 33501 Bielefeld, Germany

ARTICLE INFO

JEL classification:
C72
C73
D43
D83
L41

Keywords:
Q-learning
Stochastic stability
Evolutionary game theory
Collusion
Pricing-algorithms

ABSTRACT

I characterize the outcomes of a class of model-free reinforcement learning algorithms, such as stateless Q-learning, in a prisoner's dilemma. The behavior is studied in the limit as players stop experimenting after sufficiently exploring their options. A closed form relationship between the learning rate and game payoffs reveals whether the players will learn to cooperate or defect. The findings have implications for algorithmic collusion and also apply to asymmetric learners with different experimentation rules.

1. Introduction

Reinforcement learning, that is, the adaptation of behavior through exploration and exploitation, is a natural counterpart to game-theoretic models of bounded rationality. In this paper, I offer a characterization of value-based reinforcement learning behavior in a prisoner's dilemma for algorithms that do not rely on conditional strategies.

Broad classes of learning rules have been studied in relation to the prisoner's dilemma. All individualistic (Newton, 2018) learning rules of best-/better-response type converge to the Nash equilibrium in the prisoner's dilemma, including adaptive dynamics (Milgrom and Roberts, 1990), fictitious play, and other rules that rely on the weak acyclicity of the game (Marden et al., 2009). This stands in contrast to the simulated behavior of reinforcement learning algorithms with memory, which consistently achieve cooperation (Calvano et al., 2020). To resolve this discrepancy, the present paper fills the gap between these two approaches by studying Q-learning algorithms without memory, with a singleton state. A Q-learning player maintains a vector of Q-values that encode her subjective expected payoff from taking the corresponding action. She then usually takes an action with the highest Q-value, but sometimes experiments with other actions according to a predetermined rule. Cooperation proves possible under this model, but may require a higher rate of learning.

The proof uses techniques developed by Newton and Sawa (2015) for learning in matching games. They show that in a certain class of matching games a minimum-cost path always exists from any state to the group of states that is most robust to one-shot deviations. In the prisoner's dilemma with reinforcement learners, this is not always the case, thus the results do not directly apply. However, this idea can be used to construct such minimum-cost paths to a certain "central" state, not necessarily the most robust to one-shot deviations, thus allowing one to characterize stochastically stable states with reference to this central state. This generalization of Newton and Sawa (2015) is of some independent interest. Specifically, I give a simple characterization of behavior

E-mail address: artur.dolgopolov@uni-bielefeld.de.

https://doi.org/10.1016/j.geb.2024.01.004
Received 29 September 2022

Available online 17 January 2024

in the small noise limit for any scenario with at least one central state. This includes other dynamics beyond reinforcement learning and other games beyond the prisoner's dilemma: central states appear ubiquitous in matching scenarios (Nax and Pradelski, 2015; Newton and Sawa, 2015) and games solvable by iterated elimination of strictly dominated strategies. I, therefore, start with this general result and solve the prisoner's dilemma case as an application.

Part of the appeal of this model, and reinforcement learning in general, lies with the minimal assumptions imposed on the players' understanding of the game. In the economics literature, the dynamics of such learning processes are often called "completely uncoupled" (Hart and Mas-Colell, 2003; Foster and Young, 2006; Nax, 2019) or "asynchronous" (Asker et al., 2021) as the players (or algorithms) themselves use only their prior experience to play, having no knowledge of the game structure. This exact property has brought attention to the issue of algorithmic collusion in deceptively benign environments, where both the reinforcement learning algorithm and its designers observe little more than their own payoffs (Calvano et al., 2020, 2021; Klein, 2021). Adoption of pricing algorithms in practice had also been shown to increase the margins in the gasoline market (Assad et al., 2022). In this context, the characterization offers the conditions that make the cooperative or collusive outcome in a prisoner's dilemma unsustainable.

2. Literature

Unlike some of the similar studies of learning in a prisoner's dilemma (Mengel, 2014; Calvano et al., 2020), I do not allow for "memory", i.e., actions cannot be conditioned on past play. The same goes for foresight—learning is only based on the immediate payoff and is not based on any forecasts of future play. This is intentional: the Q-learning algorithm, while being a very simple technique, proves capable of maintaining enough information in the Q-values to converge to non-Nash outcomes even without relying on conditional strategies. On the other hand, without such strategies, this algorithm cannot converge to cooperation if its learning rate is low. The prospect of collusion by independent algorithms is a controversial topic at least for two reasons. First, small changes to learning environment may allow, preclude, or change the degree of collusion (Dorner, 2021; Asker et al., 2021; den Boer et al., 2022; Asker et al., 2022; Abada and Lambin, 2023; Dawid et al., 2023; Banchio and Mantegazza, 2022). Second, the legal consequences of algorithmic collusion can be ambiguous, with some rules explicitly requiring strategies to punish deviations from the supra-competitive price to be classified as collusive (Harrington, 2018). Both of these concerns highlight the importance of establishing a baseline case with a simple "stateless" learning rule, where collusion is still possible even though agents do not and cannot hold punishment contingency plans that ensure cooperation.

Several of the above studies of algorithmic collusion cite imperfect exploration as its potential source or show that changes to the experimentation rule may mitigate it, see for example Asker et al. (2021); den Boer et al. (2022), as well as Calvano et al. (2023) for Q-learning without memory. Abada and Lambin (2023) show that artificially increasing the tendency to explore the competitive outcome brings the players closer to the Nash equilibrium. The stochastic stability approach similarly reveals the effects of the relative frequency of experimentation in the cooperative and non-cooperative states.

Computer science studies a similar topic under the name multi-agent reinforcement learning (MARL). Some MARL studies (Zhang et al., 2021) also consider competitive scenarios, but the existing theoretical results are conceptually different. Their goal usually is to propose tailored algorithms for independent reinforcement learners to jointly converge or adapt to each others' behavior. The agent-aware and agent-tracking (Buşoniu et al., 2010) learning algorithms directly maintain a model of the opponents to react to their policy changes. For example, the algorithm in Suematsu and Hayashi (2002) estimates the opponent's policy and acts in part to minimize the other agent's incentive to change its current policy. The agent-independent learning algorithms in Hu et al. (1998) and Hu and Wellman (2003) maintain an estimate of all other players' Q-values to be able to compute the future Nash equilibrium payoffs and use these estimates to update their own Q-values. This significantly departs from a regular, single-agent, Q-learning. In particular, the algorithms have to observe the payoffs of all other players to update these estimates. They also need to recompute Nash equilibria every period, which is a more demanding task then a simple Q-learning update. By incorporating best-response behavior into the updating rule, these papers establish conditional convergence to Nash equilibria. In contrast, I focus on a strategic interaction of algorithms that are exogenously fixed. For example, they may be set by individual firms in the hope of maximizing profits without access to other agents' payoffs. I specifically consider classical single-agent learning algorithms with no built-in best-response or equilibrium behavior. That is, learning agents do not maintain a model of their opponent but independently make decisions based on their own payoff signals. While single-agent reinforcement learning can be and is applied in multi-agent settings (Sen et al., 1994; Tuyls et al., 2006), there are almost no convergence guarantees, especially beyond zero-sum games.

The most closely related studies (Waltman and Kaymak, 2007, 2008) pursue the same goal and have partially characterized the convergence in the prisoner's dilemma game for high learning rates. In particular, when one experimentation step is enough for a switch from a non-cooperative state to a cooperative state and vice-versa, the analysis can be simplified by considering only the minimum-cost paths. Yet in many applications (e.g. Calvano et al., 2020), the learning rate may be expected to be low to ensure enough experimentation and full traversal of the state space. Moreover, it is not clear whether the high learning rate assumption would be restrictive for human subjects.

Unlike Waltman and Kaymak (2007, 2008), I follow the evolutionary game theory approach and characterize the convergence of learning through a "tree surgery" argument, using costs (resistances) and stochastically stable sets (Freidlin and Wentzell, 1984; Young, 1993). This allows me to solve the problem in full generality as well as provide the estimated time until players converge. There is a growing literature on the relationship between learning behavior and stochastically stable sets (Newton, 2018, contains an overview of recent results). Many rules select the risk-dominant equilibrium and, since it is the unique Nash equilibrium in dominant strategies, they select the Nash equilibrium in the prisoner's dilemma as well. The subset of this literature showing the possibility of cooperation usually incorporates a notion of memory. Mengel (2014) studies the learning rules based on sampling from past history in a prisoner's dilemma. Bilancini and Boncinelli (2020) consider the condition-dependent mistake model in a stag hunt game where experimentation probability depends on the payoff in the previous period. The Q-learning algorithm instead maintains an "expectation" of the payoff, an overall statistic of past experimentation. The two models are most similar in the extreme case when the weight of the recent evidence in the Q-learning process is the highest, but differ in that Q-learning keeps a separate record of the last payoff for each action.

Reinforcement learning also has potential as a model for human behavior. This is supported by previous studies such as Roth and Erev (1995) and Erev and Roth (1998), which combine simulations with experiments to show that reinforcement learning models have better predictive and descriptive power than standard equilibrium analysis. More recently, Mäs and Nax (2016) have shown that reinforcement learning models often fit the behavior of experimental subjects. Subjects choose specific actions more frequently after having received a high payoff from playing this action, which supports the main assumptions underlying the reinforcement learning models (section 3.2 of the Appendix). They also show that the deviations from myopic best responses by subjects in laboratory experiments do not appear uniform and vary with the realized payoffs, which highlights the importance of different, non-uniform, experimentation rules that are also considered here. Differently from these studies, I use reinforcement learning as a long-term equilibrium selection concept, abstracting from the medium-term dynamics. Bilancini et al. (2021) show that if the experimentation rule is chosen based on individual decisions in laboratory experiments, the final outcomes of these experiments are consistent with the long-term predictions of stochastic stability. In other words, models of long-term behavior are empirically grounded in experimental data.

The rest of the paper is organized as follows. I begin by introducing the game and learning rules (section 3), then I characterize the recurrent (absorbing) sets of states of the unperturbed process without experimentation (section 4.1), refine them to stochastically stable states of the process with experimentation (section 4.2), and finally apply the results to the prisoner's dilemma game under two reinforcement learning rules (section 5). I conclude by discussing possible extensions and comparing the results to other learning models (section 6).

3. Preliminaries

Consider a two-player symmetric strategic game with a player set $I = \{1, 2\}$ and a set of actions A for each player so that the set of action profiles is A^2. Let π_{ab} for $a, b \in A$ denote the payoff of playing a when the opponent plays b. For example, in a prisoner's dilemma, $A = \{C, N\}$ and payoffs are $\pi_{CC}, \pi_{CN}, \pi_{NC}, \pi_{NN}$ with $\pi_{NC} > \pi_{CC} > \pi_{NN} > \pi_{CN}$. In this case, C stands for the cooperative action and N for the non-cooperative action or defection.

The play of the game will be described by a Markov process. A state g of this process is a pair of Q-vectors, $g = (Q_1, Q_2)$, each Q-vector containing a Q-value for every element in A. For the prisoner's dilemma $Q_i = (Q_i^N, Q_i^C)$, $i \in I$. The Q-value of $a \in A$ for i at state g is denoted $Q_i^a(g)$. The set of all possible states is denoted $\mathfrak{G}$.

Assuming that Q-values are real numbers would require repeatedly proving convergence for paths over the infinite state-space without gaining additional intuition. In order to stay true to practical implementations of reinforcement learning and to avoid convergence issues, I assume that all Q-values belong to a fine grid with $\epsilon > 0$ between consecutive Q-values. In other words, $\mathfrak{G} \subseteq \{(Q_1, Q_2) : Q_i \in \mathfrak{D}^{\#A}\}$, where # is the cardinality of a set and $\mathfrak{D}$ denotes some compact subset of $\{\zeta\epsilon, \zeta \in \mathbb{Z}\}$. Let $\pi_{ab} \in \mathfrak{D}$ for any $a, b \in A$. Whenever the Q-value does not conform to this grid, it is rounded to the nearest grid point (this will be formalized below). This representation reflects machine precision. A computer running a reinforcement learning algorithm carries a built-in limit for the machine representation of decimal numbers.

3.1. Unperturbed dynamics

The unperturbed dynamic, denoted P_0, is defined through transition probabilities $P_0(g, g')$ for states $g, g' \in \mathfrak{G}$. It corresponds to some reinforcement learning rule *without* experimentation, wherein players always choose the action from the set $\arg\max_{a \in A} Q_i^a$, one of the actions with the highest Q-value. Suppose this action is a_i for each player i. Players then obtain the corresponding payoffs $\pi_{a_1 a_2}$, $\pi_{a_2 a_1}$, and each update the Q-vector. Player i's update is as follows:

$$
\begin{aligned}
&Q_i^{a_i}(g_{t+1}) = \mathcal{F}_i^{a_i a_{-i}}(g_t), \\
&Q_i^b(g_{t+1}) = Q_i^b(g_t), \qquad \text{for any } b \neq a_i,
\end{aligned}
\tag{1}
$$

where $\mathcal{F}_i^{a_i a_{-i}} : \mathfrak{G} \to \mathfrak{D}$ is some function of the state g_t, additionally parameterized by the action profile, s.t. $|\mathcal{F}_i^{a_i a_{-i}}(g_t) - \pi_{a_i a_{-i}}| < |Q_i^{a_i}(g_t) - \pi_{a_i a_{-i}}|$ if $Q_i^{a_i}(g_t) \neq \pi_{a_i a_{-i}}$ and $\mathcal{F}_i^{a_i a_{-i}}(g_t) = \pi_{a_i, a_{-i}}$ otherwise. Only the Q-value of the action that was taken is updated, which is the main idea behind reinforcement learning. The $\mathcal{F}_i^{a_i a_{-i}}$ will be called the *learning rule*. The Q-values can be updated in any way, as long as they get strictly closer to the obtained payoff, i.e., the player updates her expectation towards the realized payoff in full or in part. Notice that the rule $\mathcal{F}_i^{a_i a_{-i}}$ is nonetheless assumed to be deterministic. I will refer to the actions with the higher Q as the actions "played on path", i.e., the actions in $\arg\max_a Q_i^a(g)$ for each player i. If there is more than one such action, I further assume that the player randomizes over the full support of this set, so that any action with the maximum Q-value is taken with positive probability. I will write $\mathrm{path}_i(g) = \arg\max_{a \in A} Q_i^a(g)$ for the set of actions on path at state g for player i, and $\mathrm{path}(g) = \{(a_1, a_2) \in \mathrm{path}_1(g) \times \mathrm{path}_2(g)\}$ for the set of possible pairs of "on-path" actions of the two players.

These updates move the process to the new state g'. For every pair of actions $a_1, a_2 \in A$, I introduce the function $\mathcal{F}^{a_1, a_2}(\cdot)$, $\mathcal{F}^{a_1, a_2}(g) = (\mathcal{F}_1^{a_1 a_2}(g), \mathcal{F}_2^{a_2 a_1}(g))$, which for state g returns the new state g' resulting from updating the previous values $(Q_1(g), Q_2(g))$

after (a_1, a_2) is played once. While the choice of actions may be random, once the actions are fixed, the updated Q-values are a deterministic function of the state.

Cast in terms of a stochastic process, the unperturbed dynamic P_0 is then defined so that $P_0(g, g') > 0$ if and only if the state g' constitutes a valid update:

$$P_0(g, g') > 0 \text{ if and only if } g' = \mathcal{F}^{a_1 a_2}(g) \text{ with } (a_1, a_2) \in \text{path}(g).$$

In terms of the unperturbed dynamic I will be interested in the set of recurrent states. A recurrent state is a state that is visited infinitely often with probability one. The set of all recurrent states is denoted $\mathfrak{C}$.

3.2. Perturbed dynamics

Let $\{P_\eta\}_{\eta \in (0, \bar{\eta})}$ be the family of perturbed dynamics indexed by the experimentation parameter η. In particular, $P_\eta(g, g')$ denotes the transition probability from g to g'. It is assumed to satisfy seven conditions. The first four are regularity conditions that are borrowed directly from Newton and Sawa (2015).

The following notion of cost will be useful to motivate these regularity conditions. The 1-step cost of the perturbed process moving from g to g' is defined as:

$$c\left(g, g'\right) := \lim_{\eta \to 0} -\eta \log P_\eta\left(g, g'\right),$$

adopting the convention that $-\log 0 = \infty$. The 1-step cost $c(g, g')$ represents the exponential decay rate of the transition probability from state g to state g'. A higher cost is associated with rarer transitions, while impossible transitions are assigned an infinite cost. For any state g outside the set of recurrent states $\mathfrak{C}$, there exists a zero-cost transition from g. This is because there exists some state $g' \neq g$ for which the transition probability $P_\eta(g, g')$ does not approach zero as η tends to zero.

Assumption 1. *(Regularity conditions on the perturbed dynamic).*

(i) $P_\eta \xrightarrow{\eta \to 0} P_0$*, where* P_0 *are the transition probabilities for some unperturbed dynamic as described above.*
(ii) For $\eta > 0$*, the chain induced by* P_η *is irreducible.*
(iii) P_η *vary continuously in* η*.*
(iv) If, for $g \neq g', P_0\left(g, g'\right) = 0, P_{\hat{\eta}}\left(g, g'\right) > 0$ *for some* $\hat{\eta} > 0$*, then* $\lim_{\eta \to 0} -\eta \log P_\eta\left(g, g'\right) = c$ *for some* $c > 0$*.*

The conditions in Assumption 1 connect perturbed and unperturbed processes and restrict the perturbed process to be "weakly regular" (Sandholm, 2010). Weak regularity ensures that the limiting distribution of the process and the costs, which we use below to describe this limiting behavior, are well-defined.

The paper builds on the machinery of the "one-shot deviation principle" introduced in Newton and Sawa (2015) for matching games, uses spanning trees from Young (1993) and radius/coradius concepts from Ellison (2000). The definitions below are taken from these papers.

Consider the learning process in the limit as $\eta \to 0$. Irreducibility implies that every P_η has a unique invariant probability distribution μ_η. Following the literature, I will say that a state g is *stochastically stable* if it has strictly positive mass in the limiting distribution $\lim_{\eta \to 0} \mu_\eta > 0$.[1] The goal of the paper is the characterization of these states. I will refer to the set of all stochastically stable states as SS.

To consider the overall cost of moving between states g and g', even when multiple steps are involved, I define the cost for a finite sequence of distinct states (referred to as a path) $g_1, g_2, ..., g_r$ as $c(g_1, g_2, ..., g_r) = \sum_{l=1}^{r-1} c(g_l, g_{l+1})$. Let $S(g, g')$ denote the set of all paths between states g and g'. Then the minimum overall cost of transitioning from g to g', regardless of the number of steps required, can be defined as follows:

$$C(g, g') = \min_{g, ..., g' \in S(g, g')} c(g, ..., g')$$

A spanning tree rooted at $\hat{g} \in \mathfrak{G}$ is a directed graph over the set $\mathfrak{G}$ such that every $g \in \mathfrak{G}$ other than $\hat{g}$ has exactly one exiting edge, and the graph has no cycles. Alternatively, a spanning tree is a directed graph, where there exists a fixed vertex, known as the root, to which every other vertex is connected by a unique path.[2] The cost of a spanning tree is the sum of the costs of its edges given by $c(\cdot, \cdot)$. A minimum-cost spanning tree is a spanning tree whose cost is lower than or equal to the cost of any other spanning tree. A state $\hat{g} \in \mathfrak{G}$ is stochastically stable only if[3] there exists a minimum-cost spanning tree rooted at $\hat{g}$ (Freidlin and Wentzell, 1984;

[1] There is a subtle difference between weak stochastic stability, defined as $\lim_{\eta \to 0} \eta \log \mu_\eta(g) = 0$, and stochastic stability in a sense of $\lim_{\eta \to 0} \mu_\eta(g) > 0$. The latter asserts the presence of a non-vanishing probability mass at g in the limit, while the former only requires that it does not vanish at an exponential rate.

[2] Although conventionally called the root, $\hat{g}$ is the "sink" of the directed graph.

[3] The stronger definition of stochastic stability in a sense of $\lim_{\eta \to 0} \mu_\eta(g) > 0$ is the reason for the "only if" here instead of "if and only if". In practical terms, the "only if" is enough for applied results as the action profile in weakly stochastically stable states is typically unique except for borderline cases. Since all stochastically stable states are contained within the larger set of weakly stochastically stable states, this approach effectively determines whether cooperation or defection prevails

Young, 1993). I will use $cost(\hat{g})$ to denote the cost of a minimal spanning tree among all trees rooted in $\hat{g}$. This expression is also called the stochastic potential of $\hat{g}$. Thus, the stochastically stable states are among the states with the lowest stochastic potential.

A transition $g \to g'$ from $g \in \mathfrak{G}$ is a *least cost transition* from g if its cost is the lowest of all possible 1-step transitions from g.

Denote the set of least cost transitions from $g \in \mathfrak{G}$ by:

$$L(g) := \arg\min_{g' \neq g} c\left(g, g'\right)$$

The cost of these transitions is denoted $c_L(g)$:

$$c_L(g) := \min_{g' \neq g} c\left(g, g'\right).$$

For non-recurrent states, the least cost transition is an update of Q-values after the on-path actions are played (in which case $c_L = 0$). For recurrent states, the least cost transition is an update after the least costly experimentation ($c_L > 0$).

The costs of transitions driven by each of the two players would generally be different. For a recurrent state $g \in \mathfrak{C}$ with a unique action profile on path, $\text{path}(g) = \{(a_1, a_2)\} \in A^2$, one can define the least cost of transitions for each player. The least cost of the transition driven by the first player's experimentation is $\min_{b_1 \in A \setminus a_1} c(g, \mathcal{F}^{b_1, a_2}(g))$, and the least cost of the transition driven by the second player is $\min_{b_2 \in A \setminus a_2} c(g, \mathcal{F}^{a_1, b_2}(g))$. The smaller of these two numbers is the least cost $c_L(g)$. It will also be useful to introduce the counterpart $c_M(g)$, the cost of the least costly transition following experimentation by the player for whom the cost is *higher*. For a state $g \in \mathfrak{C}$ with $\text{path}(g) = \{(a_1, a_2)\} \in A^2$, this value is

$$c_M(g) = \max\{\min_{b_1 \in A \setminus a_1} c\left(g, \mathcal{F}^{b_1, a_2}(g)\right), \min_{b_2 \in A \setminus a_2} c\left(g, \mathcal{F}^{a_1, b_2}(g)\right)\}.$$

Define OS, the set of states which are most robust to one-shot deviation (Newton and Sawa, 2015) as

$$OS = \left\{ g \in \mathfrak{G} : c_L(g) = \max_{g' \in \mathfrak{G}} c_L\left(g'\right) \right\}.$$

As $c_L(g)$ is only strictly positive for the states in $\mathfrak{C}$, it must be that $OS \subseteq \mathfrak{C}$.

The *radius* $R(g) = \min_{g' \in \mathfrak{C} \setminus \{g\}} C\left(g, g'\right)$ of state g is the minimum cost necessary to leave the basin of attraction of g. The minimum cost does not generally equal the radius because the minimum cost transition may not leave the basin of attraction.

Finally, readers familiar with the radius-coradius theorems will find the following definition similar to the modified cost from Ellison (2000), with the exception that it uses minimum cost instead of a radius:

$$\bar{c}(g_1, g_2, ... g_r) = c(g_1, g_2, ... g_r) - \sum_{l=2}^{r-1} c_L(g_l).$$

$$\bar{C}(g_1, g_r) = \min_{g_1, ..., g_r \in S(g_1, g_r)} \left(c(g_1, g_2, ... g_r) - \sum_{l=2}^{r-1} c_L(g_l) \right).$$

The expression adjusts the total cost of transitions by subtracting the minimum costs along the path, which captures the cost of transitions conditional on the fact that a transition out of state occurs. If $c_L(g_l)$ is replaced with $R(g_l)$, one obtains the modified cost of Ellison (2000). I will call the latter *R-adjusted cost* to distinguish it from $\bar{C}(g_1, g_r)$, which I call *c_L-adjusted cost.*[4] Note that the first and the last states are not included in the sum. In many cases, including the present paper, adjusted cost naturally emerges in the expressions for the cost of the spanning trees.

It is now possible to state the remaining three conditions on the perturbed dynamic. These conditions are formulated in full generality in terms of transition costs.

Assumption 2. *(Additional conditions on the perturbed dynamic).*

(v) *If $P_\eta\left(g, g'\right) > 0$ for some $\eta \geq 0$, then $g' = \mathcal{F}^{a_1, a_2}(g)$ for some $a_1, a_2 \in A$.*

(vi) *For any state g, such that $\{(a_1, a_2)\} = \text{path}(g)$, and any $b_1 \neq a_1, b_2 \neq a_2$, it holds that $c\left(g, \mathcal{F}^{b_1, b_2}(g)\right) = c\left(g, \mathcal{F}^{b_1, a_2}(g)\right) + c\left(g, \mathcal{F}^{a_1, b_2}(g)\right)$.*

in the long run. Adopting a weaker definition of cost lets me cover a wider range of experimentation rules, for instance the probit rule, which would otherwise be excluded. In many cases the minimum tree characterization holds in both directions. Further details and related discussion can be found in Chapter 12 and Proposition 12.1.2 of Sandholm (2010) and in Newton and Sawa (2015). A drawback of this approach is the inability to assert that *both* cooperation and defection happen with a positive probability when the parameters lie exactly on the border between regions in the Fig. 2. Rather, it can only be claimed that either of these actions may occur. For many experimentation rules this can be refined to strictly positive probability of both actions as long as the transition probabilities satisfy the original definition in Young (1993): if $P_\eta(g, g^i) > 0$ for some η, then there is $c \geq 0$, s.t. $0 < \frac{P_\eta(g,g^i)}{\eta^c} < \infty$.

[4] See Ellison (2000) for more examples of this construction. Remark 2 below shows that for the class of problems in this paper, $R(g) = c_L(g)$ for almost any state g. However, as noted in Newton and Sawa (2015), this fact does not follow from the definitions.

(vii) For any states g, g', such that $(a_1, a_2) \in \mathrm{path}(g) \cap \mathrm{path}(g')$, it holds that:
if $Q_1^{a_1}(g) \geq Q_1^{a_1}(g')$ and $Q_1^{b_1}(g) < Q_1^{b_1}(g')$ for some $b_1 \neq a_1$, then $c(g, \mathcal{F}^{b_1,a_2}(g)) \geq c\left(g', \mathcal{F}^{b_1,a_2}\left(g'\right)\right)$;
if $Q_2^{a_2}(g) \geq Q_2^{a_2}(g')$ and $Q_2^{b_2}(g) < Q_2^{b_2}(g')$ for some $b_2 \neq a_2$, then $c(g, \mathcal{F}^{a_1,b_2}(g)) \geq c\left(g', \mathcal{F}^{a_1,b_2}\left(g'\right)\right)$.

Condition (v) states that every transition is a valid Q-learning update, possibly an update on the profile that resulted from experimentation. While the dynamics are parameterized by a single variable η, this condition admits different experimentation rules for the players—different probabilities of experimentation or different processes altogether—as long as the probability of experimentation decreases in η for all players.

The remaining two conditions impose mild restrictions stemming from the interpretation of the Q-vector in reinforcement learning as an imperfect estimate of the value function. In general, if the two players are experimenting independently of each other with probability that is increasing in the Q-values off path, then both of the remaining conditions are satisfied. Condition (vi) requires that players experiment independently, i.e., the cost of both players experimenting is the sum of costs of each of them experimenting alone. Importantly, this does not imply that a two-player experimentation for some state cannot be less costly than a single-player experimentation in another state. Condition (vii) states that for any pair of states g, g' if for some player the Q-value for the action on path in g is the same as or lower than in g' and at the same time the Q-value for the action off-path is higher in g' than in g, then the cost of her experimenting is at least as high at g as at g'. This captures the intuition that the player experiments more (or, more precisely, the experimentation decays slower) if the other action seems to give high payoffs than when it gives low payoffs.

Together with (v), irreducibility (ii) implies that the set of possible states $\mathfrak{G}$ does not contain states with Q-values that are beyond the lowest or highest attainable payoffs in the game. Mathematically, $\mathfrak{G} \subseteq \{(Q_1, Q_2) : Q_i^{a_i} \in [\min_{a_{-i}} \pi_{a_i a_{-i}}, \max_{a_{-i}} \pi_{a_i a_{-i}}]$ for any $a_i \in A, i \in I\}$. No state $g \notin \mathfrak{G}$ can be reached from within $\mathfrak{G}$, because all transitions have to be valid updates (v), and for any such state $g \notin \mathfrak{G}$ there can be no state $g' \in \mathfrak{G}$ such that $\mathcal{F}^{a_1,a_2}(g') = g$ for $a_1, a_2 \in A$. This restriction is purely technical and irrelevant in the actual computer implementation of the algorithm. The learning process on a finite grid will always reach $\mathfrak{G}$. Therefore, the designer does not need to have prior knowledge of the payoffs or other information about the game to set up the initial conditions, and can set them arbitrarily as long as there will be time for sufficient experimentation. Overall, these conditions are quite permissive, and the logit rule (also called the Boltzmann softmax function) described in Waltman and Kaymak (2007, 2008) can be shown to satisfy these conditions as well as experimenting uniformly, probit (Sandholm, 2010; Newton and Sawa, 2015), etc. The general results do not depend on the choice of perturbations as long as they satisfy these regularity assumptions.

4. General results

4.1. Recurrent classes (unperturbed process)

Stochastically stable states belong to the recurrent classes (recurrent states or groups of states) of the unperturbed process (Young, 1993). Therefore, as common for the studies of stochastic dynamics, I begin with the characterization for the unperturbed process without experimentation. I will also show that the unperturbed process is always absorbed by a single state and cannot get "stuck" in a cycle.

Let $\mathcal{A}_i(G) \subseteq A$ be the set of actions that are played by i on path in a recurrent class G. That is, any action $a \in A$ is in $\mathcal{A}_i(G)$ if and only if there is $g \in G$, s.t. $Q_i^a(g) \geq Q_i^b(g)$ for any $b \in A \setminus a$.

The characterization of the unperturbed process is broken down into two lemmas. The intuition is kept in the text, but all formal proofs of the lemmas and propositions are collected in the Appendix A. The first step is to show that the Q-values are bounded by the lowest and highest payoffs that can happen on path.

Lemma 1. *For any recurrent class G, any state $g \in G$, and any action a_i that is played by i with a positive probability in G, $\min_{a_{-i} \in \mathcal{A}_{-i}(G)} \pi_{a_i, a_{-i}} \leq Q_i^{a_i}(g) \leq \max_{a_{-i} \in \mathcal{A}_{-i}(G)} \pi_{a_i a_{-i}}$.*

The actions played at least once in a recurrent class are also played infinitely often in this class with probability 1. The second lemma states that such actions always have the same Q-value when they are not played, possibly distinct for each player. The proof relies on the reinforcement nature of the learning algorithm: Q-value of any action a may only change when a is played. This implies that a player can stop playing an action only if its Q-value decreases. Conversely, starting from any state g, if the Q-value of a decreases, then to return to this state g, the Q-value of a has to eventually increase, which is impossible when a is not already played. Therefore, whenever the player changes the actions played on path, the Q-value can never be too high, or too low. In fact, it has to be a constant number for every player, so that all actions can recur with positive probability. In turn, since all Q-values are equal when the player changes actions, she randomizes fully over all actions played by her in the recurrent class.

Lemma 2. *Take any recurrent class G. Let $g \in G$, $i \in I$, and $a_i \in \mathcal{A}_i(G) \setminus \mathrm{path}_i(g)$. Then $Q_i^{a_i}(g) = q_i$ for some constant q_i.*

The two lemmas can now be used to characterize the recurrent (absorbing) states in any game solvable by iterated elimination of strictly dominated strategies (IESDS).

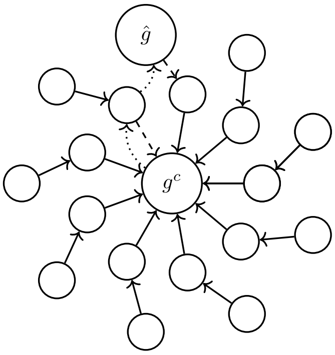

Fig. 1. A central state and two spanning trees.

Proposition 1. *A state g is absorbing in the unperturbed process, i.e., $g \in \mathfrak{C}$, if and only if for some $a_1, a_2 \in A$:*

$$\pi_{a_i a_{-i}} = Q_i^{a_i}(g) \text{ and } Q_i^{a_i}(g) > Q_i^{b_i}(g) \text{ for all } b_i \in A \setminus a_i.$$

If the game can be solved by IESDS, then such states are the only recurrent classes, i.e., there are no recurrent classes that are not singletons.

Proposition 1 says that in all recurrent states the actions on path have the Q-values that equal the payoffs, while all other actions have lower Q-values and are not played. This result shows that the Q-learning process (without noise) will always converge for games solvable by IESDS, although not necessarily to a Nash equilibrium.

4.2. Stochastically stable states (perturbed process)

Using the characterization for the unperturbed process, the recurrent states can now be refined to stochastically stable states.

The following two lemmas will help generalize the approach from Newton and Sawa (2015). Instead of showing that all states have a minimum-cost path to the OS set, which is no longer true for Q-learning, I will use the fact that all states have such paths to some "central" state, not necessarily in the OS. If there are minimum-cost paths to some state g^c that is not in OS, it is still possible to say that the minimal trees are of a particular form.

Definition 1 *(Central state). State $g^c \in \mathfrak{C}$ is said to be* central *if for any $g \in \mathfrak{C} \setminus g^c$ there is a path $g = g_1, ..., g^L = g^c$, s.t. $C(g_l, g_{l+1}) = c_L(g_l)$ for any $l \in \{1, ...r - 1\}$*

The next lemma says that if such state g^c exists, then in every minimal spanning tree all edges are minimum-cost edges, except, possibly, for the path from g^c to the root of the tree. Therefore the comparison of spanning trees is reduced to studying this path.

Lemma 3. *If $g^c \in \mathfrak{C}$ is central then for any minimal spanning tree and any $g' \in \mathfrak{C}$, either the outgoing edge from g' has the cost $c_L(g')$ or there is a path from g^c to g'.*

The construction of the minimum-cost tree is the same as the one in Newton and Sawa (2015) and if $g^c \in OS$ then, by their result, $SS = OS$. However Lemma 3 also captures $g^c \notin OS$, which is used in the next proposition.

Proposition 2. *Let $g^c \in \mathfrak{C}$ be a central state.*

(i) A minimum-cost tree is rooted in state $\hat{g} \neq g^c$ if and only if $\bar{C}(g^c, \hat{g}) - c_L(\hat{g}) \leq 0$, and $\hat{g}$ minimizes the difference $\bar{C}(g^c, \hat{g}) - c_L(\hat{g})$ among all $\hat{g} \in \mathfrak{C} \setminus g^c$.
(ii) A minimum-cost tree is rooted in central state g^c if and only if $\bar{C}(g^c, \hat{g}) - c_L(\hat{g}) \geq 0$ for all $\hat{g} \in \mathfrak{C} \setminus g^c$.

Proposition 2 offers a convenient way to solve for stochastically stable states when at least one central state is known to exist. It suffices to compare $\bar{C}(g^c, \hat{g})$ and $c_L(\hat{g})$, which are much easier to calculate than the cost of the whole tree. The intuition for the proof is gained from the illustration in Fig. 1. The presence of a central state implies that any minimal spanning tree would be of a similar form: the outgoing edges from nearly all states are already minimal (solid arrows). Only the edges between g^c and the root $\hat{g}$ (dotted arrows) could have a cost higher than minimal, because other edges could be replaced by the minimum cost edges leading to g^c (Lemma 3). If a minimum cost tree is rooted in g^c, then it can be constructed from the minimum cost edges by including the dashed ones. The only other possibility is that there is a state $\hat{g}$, such that the cost of reaching it from g^c (dotted arrows) is lower than the total cost of all minimum cost edges along the way (dashed arrows). The difference between these costs is precisely $\bar{C}(g^c, \hat{g}) - c_L(\hat{g})$.

Conveniently, if one central state is known, there is no need to prove it is unique or search for all such states to use Proposition 2. It does not matter if there are multiple central states, and, if so, which is used as g^c to calculate the cost $\bar{C}(g^c, \hat{g})$ in Proposition 2.

Remark 1. *The set of roots of minimum-cost trees obtained from Proposition 2 does not change if g^c is replaced by a different central state g'^c.*

Proof. Indeed, suppose there are two states g^c and g'^c, s.t. for any $g \in \mathfrak{C}$ there are paths $g = g_1, ..., g^L = g^c$ and $g = g'_1, ..., g'_r = g'^c$ with $C(g_l, g_{l+1}) = c_L(g_l)$ for any $l \in \{1, ...r-1\}$ and $C(g_{l'}, g_{l'+1}) = c_L(g_{l'})$ for any $l' \in \{1, ...r'-1\}$. It suffices to cover the two cases in Proposition 2. First, $cost(g^c) = cost(g'^c) + c_L(g'^c) - c_L(g^c) = cost(g'^c) + \bar{C}(g'^c, g^c) - c_L(g^c)$. The last equality follows from $\bar{C}(g'^c, g^c) = c_L(g'^c)$ because there is a path of minimum cost deviations between the central states and each difference in the definition of the c_L-adjusted cost is therefore zero except for the cost of the first transition. Second, by a similar argument, $cost(g^c) + \bar{C}(g^c, \hat{g}) - c_L(\hat{g}) = cost(g'^c) + c_L(g'^c) - c_L(g^c) + \bar{C}(g^c, \hat{g}) - c_L(\hat{g}) = cost(g'^c) + \bar{C}(g'^c, \hat{g}) - c_L(\hat{g})$ for any $\hat{g}$ using the fact that $\bar{C}(g^c, \hat{g}) = \bar{C}(g'^c, \hat{g}) + c_L(g'^c) - c_L(g^c)$. □

As a final remark of the section, let us show that for the problems with a central state the radius of any state $g \in \mathfrak{G}$ equals $c_L(g)$. Therefore the c_L-adjusted cost coincides with the R-modified cost as defined in Ellison (2000), which can be used to find bounds on convergence time.

Remark 2. *If there is a central state $g^c \in \mathfrak{C}$ then $R(g) = c_L(g)$ for any $g \in \mathfrak{G} \setminus g^c$.*

Proof. The remark is automatically true for any non-recurrent state $g \in \mathfrak{G} \setminus \mathfrak{C}$, because $R(g) = c_L(g) = 0$. Suppose that $g \in \mathfrak{C}$. There has to be a path from g to g^c that goes only through minimum cost edges. Suppose the shortest path from g to some g^c consists of a single edge. Then $R(g) = c_L(g)$ since $g^c \in \mathfrak{C} \setminus g$. Now suppose the path between g and g^c consists of multiple edges, passing through additional states: $g, g_1, g_2, ..., g^c$. Consider each state g_l on this path, starting with g_1. There are two possibilities. If $c(g_1, g) > 0$, then g_1 is not in the basin of attraction of g and $R(g) = c(g, g_1) = c_L(g)$. If $c(g_1, g) = 0$, then $c_L(g_1) = 0$ and therefore $c(g_1, g_2) = c_L(g_1) = 0$ because the path consists of minimum cost transitions. In this case, consider the next state g_2, where $C(g, g_2) = C(g, g_1) + c(g_1, g_2) = C(g, g_1) = c_L(g)$. Applying the same argument by induction for a state g_l, if for some g_l on this path $C(g_l, g) > 0$, then $C(g, g_l) = c_L(g) = R(g)$. If there is no such state and $C(g_l, g) = 0$ for all $l \geq 1$, then $C(g, g^c) = c(g, g_1) = c_L(g)$. In both cases $R(g) = c_L(g)$. □

There is one important case when the Remark does not apply. When there is a unique central state, this state itself may have $R(g^c) > c_L(g^c)$.

5. Prisoner's dilemma

Two specific recurrent states, g^* and g^{**} will be useful below for the prisoner's dilemma case. The first state g^* has defection on path with $Q_i^N(g^*) = \pi_{NN}, Q_i^C(g^*) = \pi_{CN}$ for both $i \in I$. The other state g^{**} has cooperation on path with $Q_i^N(g^{**}) = \pi_{NN}, Q_i^C(g^{**}) = \pi_{CC}$ for both $i \in I$. Stochastically stable states can be determined by only considering these two states and the transitions between them. These two states have the highest cost of experimentation among the states with N and C on path respectively. This follows by Assumption 2(vii).

To be able to use Proposition 2, I will show that indeed a variant of the "getting closer"-lemma by Newton and Sawa (2015) holds for Q-learning in a prisoner's dilemma, but instead of approaching the OS set, the process approaches the state g^* with Nash equilibrium actions on path. In other words, $g^c = g^*$ in terms of the Lemma 3.

Before stating the lemma let me formalize what I mean by "closer" to g^*. The states in $\mathfrak{C}$ can be ordered by their position on a path of minimum-cost transitions towards g^*. This path can be broken down into four phases, resulting in a lexicographic order $\prec$. In order to define this order, two measures will be useful: $m(g)$ and $D(g)$.

The value $0 \geq m(g) \geq 2$ is defined as the number of players who play N on path, $m(g) = \#\{i : N \in \text{path}_i(g)\}$. Neither (N, C) nor (C, N) can be played on path at $g \in \mathfrak{C}$ because then $\pi_{CN} < \pi_{NN} \leq Q_i^N(g) < Q_i^C(g) = \pi_{CN}$ by Proposition 1. Therefore $\text{path}(g) = \{(C, C)\}$ or $\text{path}(g) = \{(N, N)\}$, and $m(g) \in \{0, 2\}$ for any $g \in \mathfrak{C}$.

The value $D(g)$ measures the difference in Q-value of C between g and g^*:

$$D(g) = \sum_{i \in I} |Q_i^C(g) - \pi_{CN}|,$$

I now define the order $\prec$. I say that a state g_2 is "closer" to g^* than g_1, written $g_2 \prec g_1$, if:

$$\begin{cases} m(g_2) > m(g_1) \\ m(g_2) = m(g_1) = 2 \text{ and } D(g_2) < D(g_1) \\ m(g_2) = m(g_1) = 0 \text{ and } c_M(g_2) > c_M(g_1) \\ m(g_2) = m(g_1) = 0 \text{ and } c_M(g_2) = c_M(g_1) \text{ and } c_L(g_2) < c_L(g_1). \end{cases}$$

That is, $m(\cdot)$ is lexicographically more important than $D(\cdot)$ when players defect on path, and increasing c_M is more important than decreasing c_L when both players cooperate.

To gain intuition for $\prec$, consider each of the components of $\prec$ in the order of lexicographic importance. The biggest step, is the change of the actions on path from cooperation to defection, captured by $m(\cdot)$.

When both players are already defecting on path, the Q-value of N equals π_{NN}, the same value as in g^*. Either player can still be too "optimistic" about the C action in the sense that its Q-value can be strictly between π_{CN} and π_{NN}, instead of exactly π_{CN}, which is the Q-value of C in g^* for both players. For the process to reach g^* it remains for each such player to update the Q-value of C down. The distance from any state $g \in \mathfrak{C}$ to g^* in this case is captured by $D(g)$, the total difference in Q-values of the cooperative action from the lowest value π_{CN} attained at g^*. Since $Q_i^N(g) = Q_i^N(g^*)$ for both $i \in I$, $D(g)$ measures the Euclidean distance to g^* in the space of Q-values.

When the players are cooperating on path, the process may need several steps to reach mutual defection. A minimum-cost deviation from a state with (C, C) on path would be for one player (player i) to experiment and play N, and for the other (player $-i$) to continue to play C. The costs c_M and c_L conveniently describe the adjustment by the non-experimenting and experimenting players respectively.

The c_L part is easy to interpret. The low c_L for a state with (C, C) on path is equivalent to a high Q-value of N for i and low cost of experimentation with N. After experimenting with N player i would update the Q-value of N up, and her opponent will update her Q-value of C down. If the players continue to cooperate after that, the other player recovers the Q-value of cooperation to π_{CC}, but the Q-value of N for player i has increased. This makes the experimentation with N less and less costly by Assumption 2(vii), until eventually, the player switches to N. Therefore part of the path to g^* is captured simply by the increasing Q-value of N for one of the players or, equivalently, the decreasing value of c_L.

The high c_M is equivalent to a low Q-value of N for $-i$. This low Q-value of N ensures that the non-experimenting player $-i$ persists in playing C even after getting the lowest possible payoff π_{CN} from i playing N. Then, when $-i$ eventually switches to N, the Q-value of C is very low, in particular it has to be lower than π_{NN} for the players to converge to a state with (N, N) on path and for the $m(\cdot)$ to increase from 0 to 2. Therefore the increasing value of c_M ensures that the negative update for the Q-value of C of player $-i$ is eventually strong enough for her to make a permanent switch to non-cooperative action N. Informally, player $-i$'s original aversion to non-cooperation (low Q_{-i}^N) ensures that the defection by i has a more profound effect (low Q_{-i}^C), resulting in a switch to mutual non-cooperation.

With $\prec$ defined, we can state the "getting closer"-lemma. It uses the fact that experimentation by two players is less likely than experimentation by one player (Assumption 2(vi)) to show that a single-player experimentation from any state, possibly followed by zero-cost deviations, moves the process closer to the state g^*.

Lemma 4 *(Getting closer to g^*). Suppose $g \in \mathfrak{C} \setminus g^*$. Let $g_1 \in L(g)$. Then there is $g' \in \mathfrak{C}$ and $t \in \mathbb{N}_+$, s.t. $g' \prec g$ and $C(g_1, g') = 0$.*

The above lemma helps establishes that g^* is the central state. It is a candidate for stochastic stability, but there may be many other states that could also be stochastically stable if they meet the conditions of Proposition 2. It is possible to narrow these possibilities to states with (C, C) on path. To begin with, recall that neither (N, C) nor (C, N) can be played on path at $g \in \mathfrak{C}$. Therefore one only needs to consider trees rooted in states with only (N, N) or only (C, C) on path.

A tree rooted in some state $\hat{g} \neq g^*$ with (N, N) on path cannot be minimal. This is implied by the following corollary of Proposition 2:

Corollary 1. *A minimum cost spanning tree has to be rooted in a state $\hat{g}$ such that $c_L(\hat{g}) \geq c_L(g^*)$.*

Proof. By definition, $\bar{C}(g^*, \hat{g}) \geq c_L(g^*)$ and therefore $c_L(\hat{g}) < c_L(g^*) \leq \bar{C}(g^*, \hat{g})$ would imply that $cost(g^*) - c_L(\hat{g}) + \bar{C}(g^*, \hat{g}) > cost(g^*)$. Then, by Proposition 2, the minimal tree rooted in g^* has lower cost, which is a contradiction. □

Let us show that Corollary 1 in particular implies that any state $g \in \mathfrak{C} \setminus g^*$ with $(N, N) \in \text{path}(g)$ cannot be a root of a minimum-cost spanning tree. Indeed, in any such state g, $Q_1^N(g) = Q_2^N(g) = \pi_{NN} = Q_1^N(g^*) = Q_2^N(g^*)$ by Proposition 1, while $Q_i^C(g) \geq Q_i^C(g^*)$ for both players $i \in I$ and with a strict inequality for at least one of them. The Q-values of C cannot both be equal to $Q_i^C(g^*)$ since otherwise $g = g^*$. Then by Assumption 2(vii) the cost of a single-player experimentation from g^* is higher or equal to the cost of a single-player experimentation from g. Formally, $c(g^*, \mathcal{F}^{C,N}(g^*)) \geq c(g, \mathcal{F}^{C,N}(g))$ and $c(g^*, \mathcal{F}^{N,C}(g^*)) \geq c(g, \mathcal{F}^{N,C}(g))$. A minimum cost transition from g^* requires a two-player experimentation. Assumption 2 (vi) implies that the cost of a two-player experimentation is the sum of the costs of each player experimenting individually, $c(g^*, \mathcal{F}^{C,C}(g^*)) = c(g^*, \mathcal{F}^{C,N}(g^*)) + c(g^*, \mathcal{F}^{N,C}(g^*))$. Then, since the cost of any experimentation is strictly positive, the minimum cost of leaving g^* through the two-player experimentation is strictly higher than the cost of a single-player experimentation in g^* and therefore also strictly higher than the cost of leaving any such g with (N, N) on path. Thus, the cost of a minimal tree rooted at g^* is strictly lower by Corollary 1.

This leaves only the state g^* and states in $\mathfrak{C}$ with (C, C) on path as candidates for stochastic stability. One can further refine the possibilities by showing that the least-cost path from g^* to a state with (C, C) on path has to consist only of plays of (C, C). This is proven as the following lemma:

Lemma 5. *The path $g^* = g_1, ..., g_r = g^{**}$, where (C, C) is played in every state, has the lowest c_L-adjusted cost among all paths between g^* and any state with (C, C) on path.*

The proof relies on two observations. First, lowering the Q-value of C would generally increase the cost of playing (C, C). Second, any other profile, $(C, N), (N, C)$, or (N, N) would decrease the Q-value of C. These two facts together imply that replacing any such non-cooperative profile with (C, C) makes the remaining path less costly and possibly shorter. The path ends at g^{**}, since the plays of (C, C) do not change the Q-value of N, which remains equal to π_{NN} at both g^* and g^{**}. Therefore going to a state with a different Q-value of N would require N to be played at least once.

It follows, that out of all the states with cooperation on path, I can limit the analysis to g^{**}.

Corollary 2. *If a minimum cost spanning tree is rooted in a state $\hat{g} \neq g^{**}$ and (C, C) is played on path in $\hat{g}$ then there is also a minimum cost spanning tree rooted in g^{**}.*

Proof. By the previous Lemma 5, $\bar{C}(g^c, \hat{g}) \geq \bar{C}(g^c, g^{**})$. At the same time, $c_L(\hat{g}) \leq c_L(g^{**})$ for any state $\hat{g}$ with (C, C) by Assumption 2(vii) because $Q_i^N(\hat{g}) \geq \pi_{NN} = Q_i^N(g^{**})$ and $Q_i^C(\hat{g}) = Q_i^C(g^{**}) = \pi_{CC}$ for $i \in I$. Then $cost(g^c) - c_L(\hat{g}) + \bar{C}(g^c, \hat{g}) \geq cost(g^c) - c_L(g^{**}) + \bar{C}(g^c, g^{**})$ and by Proposition 2 the result follows. □

Thus, to argue about the action profiles on path in the limit, one only needs to consider g^* and g^{**}. This leads to a characterization:

Corollary 3.

(i) If $\bar{C}(g^*, g^{**}) < c_L(g^{**})$ then $g^* \notin SS$ and players always converge to cooperation in any state in SS.[5]
(ii) If $\bar{C}(g^*, g^{**}) > c_L(g^{**})$ then $SS = \{g^*\}$ and players always converge to defection.
(iii) If $\bar{C}(g^*, g^{**}) = c_L(g^{**})$, both defection and cooperation are possible. SS may include states with defection, cooperation, or both.

Proof. Follows directly from Proposition 2 and from Corollaries 1, 2 by the remark above. □

I now illustrate these concepts with a particular learning process and two experimentation rules. The learning rule is explicitly parameterized. This rule, which is usually called Q-learning is a particular kind of the learning rule in (1) when the magnitude of updates is captured by a single parameter α_i (per player) that is independent of the current Q-value:

$$\mathcal{F}_i^{a,a_{-i}}(g_t) \in \arg\min_{z \in \mathfrak{D}} |z - \left((1-\alpha_i)Q_i^a(g_t) + \alpha_i \pi_{a,a_{-i}}\right)| \tag{2}$$

where α_i is a rational number called the learning parameter for player i and $1 \geq \alpha_i > 0$. The players may therefore have different learning rates. The $\arg\min$ term only maps the states to a finite grid by taking the closest value in $\mathfrak{D}$ to $(1-\alpha_i)Q_i^a(g_t) + \alpha_i \pi_{a,a_{-i}}$. I do not specify the tie-breaking rule for selecting a particular element from the $\arg\min$, assuming some element is chosen deterministically. I will discuss two standard experimentation rules: greedy (uniform experimentation probabilities) and logit (also called softmax or Boltzmann) rules.

Under the greedy rule with probability $(1 - e^{-\frac{k_i}{\eta}})$ player i chooses the actions with highest Q-values, and with probability $e^{-\frac{k_i}{\eta}}$ the action is chosen by randomizing uniformly. Here k_i is a constant that describes the player's tendency to experiment over time. Ties are also resolved uniformly. Formally in my definitions,

$$Pr_i^{\text{greedy}}(a|g) = \frac{1}{\#\{a : Q_i^a(g) = \max_{a'} Q_i^{a'}(g)\}}(1 - e^{-\frac{k_i}{\eta}}) + \frac{1}{2}e^{-\frac{k_i}{\eta}}$$

if a is played by i on path at g and

$$Pr_i^{\text{greedy}}(a|g) = \frac{1}{2}e^{-\frac{k_i}{\eta}}$$

otherwise.[6] The denominator in the former case only divides the amount among all actions played on path if there is more than one. The chosen experimentation function $e^{-\frac{k_i}{\eta}}$ ensures that the probability of experimentation is increasing in η for any choice of the constants $k_i > 0$ as required by Assumption 1(i). If $k_1 = k_2 = 1$, I obtain the simpler case with symmetric experimentation probabilities.

[5] The requirement that $\bar{C}(g^*, g^{**}) < c_L(g^{**})$ is stronger than $c(g^*, g^{**}) < c_L(g^{**})$ used in Waltman and Kaymak (2008). The weaker inequality is sufficient in Waltman and Kaymak (2008) because the lower bound on the learning rate α ensures that the cooperative state g^{**} is reachable by a single least cost transition from g^*. In the general case presented here, the learning parameter can take any value in $(0, 1]$, and then the path may have to go through more states. The least cost transitions are not enough to reach g^{**} and costlier edges need to be taken. The $\bar{C}(g^*, g^{**})$ term accounts for the corresponding costs.

[6] This verbose exponential setup ensures that the cost adheres to the definition $\lim_{\eta \to 0} -\eta \log P_\eta(g, g')$. This is the result of a standard substitution of the noise parameter $\eta \equiv -(\log \beta)^{-1}$. One can equivalently write $Pr_i^{\text{greedy}}(a|g) = \frac{1}{2}e^{-\frac{k_i}{\eta}} = \frac{1}{2}\beta^{k_i}$ and consider β variable instead of η in all limits. Both setups correspond to the same model of uniform mistakes, but the β specification relies on a different, stricter set of regularity conditions. The difference is thus purely technical. Please see more about lenient costs in Sandholm (2010), section 12.A.5.

Under the logit rule instead

$$Pr_i^{\text{logit}}(a|g) = \frac{e^{Q_i^a(g)/(k_i\eta)}}{\sum_{a'\in\{C,N\}} e^{Q_i^{a'}(g)/(k_i\eta)}},$$

with no restriction on k_i and η as long as they are positive. The $(k_i\eta)$ component is sometimes called the temperature: with higher values of η, experimentation becomes more likely and less dependent on Q-values. When η approaches zero, the actions with the highest Q-value are always chosen and the process approaches the unperturbed dynamic P_0. For the symmetric setup in the limit with η approaching infinity, the actions are chosen with equal probability.

In both cases then $P_\eta(g,g')$ is the product of these probabilities, $\prod_{i\in I} Pr_i^{\text{greedy}}(a_i|g)$ or $\prod_{i\in I} Pr_i^{\text{logit}}(a_i|g)$ for $g' = \mathcal{F}^{a_i,a_{-i}}(g)$.

Under the greedy rule, experimentation by 2 players in any state has a higher cost than experimentation by 1 player in any state, which, as shown below, always leads players to defection.

The logit rule, on the other hand, can lead to cooperation depending on the values of the parameters. The characterization will rely on a commonly used property of the logit rule: the costs of transitions in the limit $\eta \to 0$ are determined by the absolute difference in Q-values between the states.

For the remainder of this section I am going to assume that ϵ is negligibly small. More precisely, the ϵ is small enough, so that all the Q-values in the proposition below fall directly on the grid $\mathfrak{D}$, which is possible because α is assumed to be rational. It is straightforward to obtain similar results from Corollary 3 for any ϵ, s.t. $\alpha_i > \frac{\epsilon}{2}$ for both $i \in I$. Let $z_i = \lfloor \log_{(1-\alpha_i)} \frac{\pi_{NN}-\pi_{CC}}{\pi_{CN}-\pi_{CC}} \rfloor$ be the largest integer less than $\log_{(1-\alpha_i)} \frac{\pi_{NN}-\pi_{CC}}{\pi_{CN}-\pi_{CC}}$, which is the necessary number of intermediate updates on the profile (C,C) for player i to get from g^* to a state where i cooperates on path under the logit rule when ϵ is negligible. If $\alpha_i = 1$, then let $z_i = 1$, i.e., the update is immediate after just one play of (C,C). The expression can be obtained by rewriting the recursive equations $Q_i^C(g_{t+1}) = (1-\alpha_i)Q_i^C(g_t) + \alpha_i\pi_{CC}$, $i \in I$, $Q_i^C(g_0) = \pi_{CN}$ until $Q_i^C(g_{t+1}) \geq \pi_{NN}$. Let me also introduce $z = \max\{z_1, z_2\}$ as the necessary number of updates for both players to reach cooperation. Lemma 5 says that these equations describe the minimum-cost path from g^* to g^{**}.

Further, let

$$q_{l,i}^C = \pi_{CC} + (1-\alpha_i)^{l-1}(\pi_{CN} - \pi_{CC}). \tag{3}$$

For sufficiently small ϵ, these $q_{l,i}^C$ are the Q-values of cooperation for each player on the minimum-cost path (of repeatedly playing (C,C)) from g^* to g^{**} under the logit rule. For a rational α, $(q_{l+1,i}^C - q_{l,i}^C)$ is a rational multiple of $(\pi_{CN} - \pi_{CC})$, and the values can therefore be placed on a grid for sufficiently small ϵ.

Then the characterization for the two rules is as follows:

Proposition 3. *For the asymmetric Q-learners the SS set depends on the experimentation rule:*

(i) $SS = g^*$ under the greedy rule.
(ii) Under the logit rule the set SS is determined by sign of the expression

$$\Delta = \left(\frac{1}{k_1} + \frac{1}{k_2}\right)\left(\pi_{NN} - \pi_{CN}\right) + \mathbb{1}_{z>1}\sum_{l=2}^{z}\max_{i\in I}\max\{0, \frac{1}{k_i}\left(\pi_{NN} - q_{l,i}^C\right)\} - \min_{i\in I}\frac{1}{k_i}\left(\pi_{CC} - \pi_{NN}\right); \tag{4}$$

- *if $\Delta > 0$, $SS = \{g^*\}$ and N is played,*
- *if $\Delta < 0$, $g^* \notin SS$ and C is played in all states in SS,*
- *if $\Delta = 0$, both defection and cooperation are possible. SS may include states with defection, cooperation, or both.*

Here $\mathbb{1}_{z>1}$ equals 1 if $z > 1$ and 0 otherwise.

The characterization is simpler for the symmetric case where $\alpha_1 = \alpha_2 = \alpha$, $k_1 = k_2 = 1$, and therefore $\bar{C}(g^*, g^{**}) = 2(\pi_{NN} - \pi_{CN}) + \mathbb{1}_{z>1}\sum_{l=2}^{z}(\pi_{NN} - q_l^C)$ and $c_L(g^{**}) = \pi_{CC} - \pi_{NN}$ with $z = z_1 = z_2$. Here, as before, $q_l^C = \pi_{CC} + (1-\alpha)^{l-1}(\pi_{CN} - \pi_{CC})$. After substitution, the corollary follows immediately:

Corollary 4. *If $\alpha_1 = \alpha_2$ and $k_1 = k_2 = 1$ then:*

(i) $SS = g^*$ under the greedy rule.
(ii) Under the logit rule the set SS is determined by sign of the expression

$$\Delta = 2\left(\pi_{NN} - \pi_{CN}\right) + \mathbb{1}_{z>1}\sum_{l=2}^{z}\left(\pi_{NN} - q_l^C\right) - \left(\pi_{CC} - \pi_{NN}\right);$$

- *if $\Delta > 0$, $SS = \{g^*\}$,*
- *if $\Delta < 0$, $g^* \notin SS$ and C is played in all states in SS,*
- *if $\Delta = 0$, both defection and cooperation are possible i.e., SS may include states with defection, cooperation, or both.*

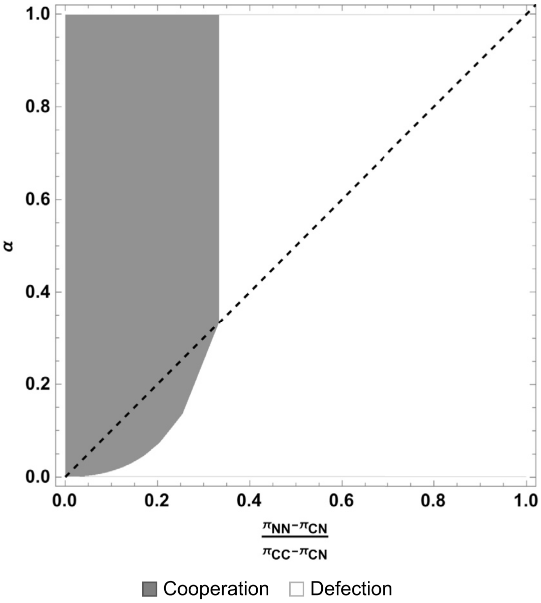

Fig. 2. Trade-off between learning rate α and relative punishment payoff $\frac{\pi_{NN}-\pi_{CN}}{\pi_{CC}-\pi_{CN}}$ for symmetric learners.

I illustrate the regions with cooperation and defection for symmetric learners with a two-dimensional graph because the only relevant factors are the learning rate α and the position of the π_{NN} payoff between π_{CN} and π_{CC}, captured by the ratio $\frac{\pi_{NN}-\pi_{CN}}{\pi_{CC}-\pi_{CN}}$. The regions are shown in Fig. 2. The boundary of the regions consists of the only values where both cooperation and defection are possible in the limit.

Corollary 4 is similar to Theorem 1 of Waltman and Kaymak (2008), and directly extends these results. The use of the stochastic stability toolset however allows me to drop their assumption of a high enough learning rate, $\alpha \geq \frac{\pi_{NN}-\pi_{CN}}{\pi_{CC}-\pi_{CN}}$. The more general result in Proposition 3 can be used with asymmetric players and other experimentation rules. The area up and left of the dashed line in Fig. 2 is the region covered by Theorem 1 of (Waltman and Kaymak, 2008).

Proposition 3 suggests several practical results. It implies that there is always a low-enough $\alpha = \min\{\alpha_1, \alpha_2\}$ for any π_{NN} such that cooperation occurs with probability zero in the limit and the g^* state persists. In the asymmetric case, one of the learners can always preclude cooperation if her learning parameter is low enough.

Corollary 5. *For any π_{NN} there is a $\alpha^* > 0$, such that $\{g^*\} = SS$ for all $\alpha \leq \alpha^*$ under both logit and greedy rules.*

Proof. z_i increases without bound as α_i approaches 0. Then $\mathbb{1}_{z>1} \sum_{l=2}^{z} \max_{i\in I} \max\{0, \frac{1}{k_i}\left(\pi_{NN} - q^{C}_{l,i}\right)\}$ also increases without bound and by Proposition 3 $\{g^*\} = SS$. □

The analysis easily extends to other learning and experimentation rules, so long as the regularity conditions are satisfied. The difference in learning parameters will only affect the regions through the changing costs $C(g^*, g^{**})$ and $c_L(g^*)$, so Corollary 3 can again be used to obtain the characterization.

6. Discussion

6.1. Supergame

The asymmetric scenario under the logit experimentation rule is conveniently illustrated by a supergame of choosing a learning algorithm against an opponent. Suppose two parties play a game, where they simultaneously choose the parameters α_i and k_i. The payoffs of the game are then calculated according to the behavior of the algorithms in the limit.

According to Proposition 3, three parameters determine the outcomes for the asymmetric case: the ratio of experimentation parameters $\frac{k_1}{k_2}$, the learning rates α_1, α_2, and the payoffs. Most of the effect of the payoffs is captured by the ratio $\frac{\pi_{NN}-\pi_{CN}}{\pi_{CC}-\pi_{CN}}$. The threshold value of the latter is shown in Fig. 3. For values above the curve, the two players would converge to defection, for values below the curve, they would converge to cooperation. For example, for $\alpha \geq \frac{1}{3}$ and $k_1 = 2k_2$, the players will converge to cooperation if $\frac{\pi_{NN}-\pi_{CN}}{\pi_{CC}-\pi_{CN}} < \frac{1}{4}$ and to defection if $\frac{\pi_{NN}-\pi_{CN}}{\pi_{CC}-\pi_{CN}} > \frac{1}{4}$. For any game with $\frac{\pi_{NN}-\pi_{CN}}{\pi_{CC}-\pi_{CN}} \geq \frac{1}{3}$, the Q-learning players will always converge to defection no matter the parameters of the algorithms.

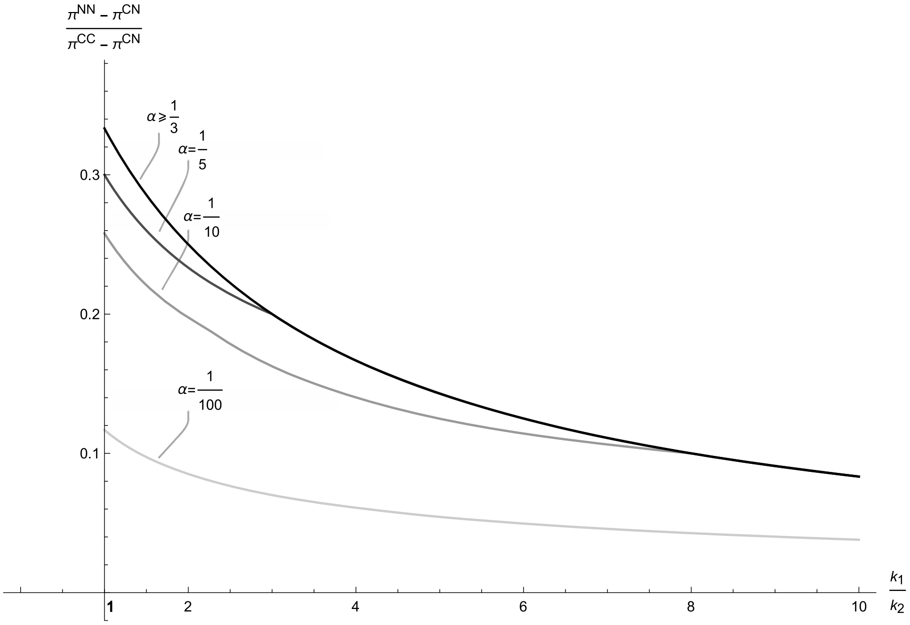

Fig. 3. Threshold payoff values that ensure cooperation in the asymmetric case.

In the supergame, since the algorithms can only converge to (N, N) or (C, C) on path, low values of α_i are dominated by setting $\alpha_i = 1$. This fact follows mechanically from the Δ expression (4) in Proposition 3, which is decreasing in α_i through the decreasing values of the $q^C_{l,i}$ (3). The intuition behind this relationship is that learning cooperation requires multiple costly steps, while learning defection requires only one. Slower learning makes the cooperative state g^{**} harder to reach, but does not affect the minimum cost deviations back to g^*. Even though more steps may be necessary to return to g^* under lower α, the cost of each step is always the minimum cost. Due to the max term in the expression, the slower learning player determines the overall cost, and therefore increasing α either has no effect or raises the probability of cooperation.

It remains to select the experimentation parameter k_i. Assuming that the players always set $\alpha_i = 1$, a good strategy for either player i is to try to match the opponent's value of k_{-i}. This also follows from the Δ expression (4). Suppose $k_i > k_{-i}$. Then player i can decrease Δ and weakly increase the chance of cooperative outcome by lowering k_i. Indeed, the first two terms would decrease and the last term is independent of k_i. And, vice versa, the other player $-i$ can decrease the value of Δ by increasing k_{-i}. If $(\pi_{NN} - \pi_{CN}) > (\pi_{CC} - \pi_{NN})$ then $\Delta > 0$ for any value of k_{-i}, and there is no choice of action that avoids the Nash outcome. Otherwise, if $(\pi_{NN} - \pi_{CN}) \leq (\pi_{CC} - \pi_{NN})$, $k_{-i} = k_i$ minimizes the value of Δ. The interpretation is that the player with the lower cost of experimentation, i in this case, will usually be the one to randomly play N in the cooperative state g^{**}. Therefore raising this cost by decreasing k_i increases the chance of converging to cooperation. The other player $-i$, the one with the higher cost of experimentation, can safely increase the experimentation rate k_{-i}, lowering the cost of experimentation. It would not affect the cost of leaving the cooperative state g^{**}, which is determined by the other player's k_i, but would increase the chance of both players experimenting simultaneously in the non-cooperative state g^*.

In sum, it is always best for the algorithm to remember only the immediately previous payoff, disregarding prior history of play, while trying to experiment about as often as the opponent. This exercise assumes that the players view the limiting behavior as a good approximation of their payoffs over the long time horizon. These payoffs are determined by stochastic stability, because after sufficient experimentation, the process spends most of the time in the stochastically stable outcomes. In a more realistic scenario, a company may want to exploit the opponent's learning phase by occasionally defecting in the short term. However, doing so would require a strategy that conditions the actions on previous observations or on time period. The supergame describes a situation where such conditional policies are impossible, for example, due to legal constraints on collusive behavior.

6.2. Convergence time

It is possible to use the modified coradius of Ellison (2000) to get a bound on convergence time. The modified coradius $C^*(G)$ of a state or group of states G is the expression

$$\min_{g_r \in G} \max_{g_1 \notin G} C^*(g_1, g_r),$$

where $C^*(g_1, g_r)$ is the R-adjusted cost of transition from g_1 to g_r:

$$C^*(g_1, g_r) = \min_{g_1,\dots,g_r \in S(g_1,g_r)} \left(c(g_1, g_2, \dots g_r) - \sum_{l=2}^{r-1} R(g_l) \right).$$

The modified coradius is the highest R-adjusted cost C^* to reach the set G from any other state. In the present paper I adjusted $\bar{C}$ by the minimum cost instead. The values $c_L(g^*)$ and $R(g^*)$ do not generally coincide for the problems with a unique central state g^* like Q-learning in the prisoner's dilemma.

Since experimentation only disappears in the limit, the value of interest is the hitting time: the expected time until the stochastically stable state is first observed. From Lemma 6 in Ellison (2000), the hitting time of state g is $O(\beta^{-C^*(g)}) = O(e^{C^*(g)/\eta})$. By Remark 2, the expected hitting time of any central state g^c, including g^* is no more than $O(e^{\frac{1}{\eta}(\max_{g \neq g^c} c_L(g))}) = O(e^{\frac{1}{\eta}(c_L(g^{**}))})$. On the other hand, the expected hitting time of g^{**} is no more than $O(e^{\frac{1}{\eta}(\bar{C}(g^c, g^{**}) - R(g^c) + \max_{g \neq g^{**}} c_L(g))})$. Here the substitution $\eta \equiv -(\log \beta)^{-1}$ or $\beta \equiv e^{-\frac{1}{\eta}}$ is used once more to obtain the same tight bound as in Remark 3.6 in Newton and Sawa (2015).

However, the process may spend a lot of time in various states with cooperation on path, even if the Q-values are not exactly the ones of g^{**}. Instead, I can compare the expected hitting times of switching between N and C regardless of the specific Q-values. These hitting times can also be obtained from the modified coradius; in this case, for the groups of states with a specific action on path. The states g^* and g^{**} are the states with the highest minimum cost among the states with (N, N) and (C, C) on path respectively by Assumption 2(vii). Then by Lemma 5 the modified coradius of the states with (C, C) on path is $\bar{C}(g^*, g^{**})$ and the modified coradius of the states with (N, N) on path is $c_L(g^{**})$. When both players are playing N on path, the expected hitting time of learning to play C is then $O(e^{\frac{1}{\eta}\bar{C}(g^*, g^{**})})$. When both players are playing C on path, the expected hitting time of learning to play N is $O(e^{\frac{1}{\eta}c_L(g^{**})})$. Thus, the Δ expression, which is the difference between these two exponents, captures the effects of the parameters on the convergence rates as well. As the Δ gets further away from 0, the time to converge generally gets shorter, and the convergent actions persist for a longer time. It is also possible to derive practical effects of the parameters from the definitions of $c_L(g^{**})$ and $\bar{C}(g^*, g^{**})$:

Remark 3.

(i) For any learning and experimentation rules, the expected hitting time of states with (N, N) is independent of α.
(ii) For Q-learning and any experimentation rule, the expected hitting time of states with (C, C) is lowest at $\alpha = 1$.
(iii) For Q-learning and the logit experimentation rule: the expected hitting time of states with (C, C) on path is non-increasing in α and non-decreasing in $(\pi_{NN} - \pi_{CN})$; the expected hitting time of states with (N, N) on path is non-decreasing in $(\pi_{CC} - \pi_{NN})$.

Proof. The first part follows from the fact that g^* is central, established in Lemma 4. Thus the adjusted cost of reaching g^* from any state is determined by a single update and the speed of learning is irrelevant. The second part follows from the term $\mathbb{1}_{z>1} \sum_{l=2}^{z} \max_{i \in I} \max\{0, \frac{1}{k_i} \left(\pi_{NN} - q_{l,i}^C \right)\}$ in (4). This term may be positive for $\alpha < 1$ but it is always 0 for $\alpha = 1$. In other words, the cost cannot be higher when learning is immediate than when there are additional steps. The last part is by inspection of the signs on the payoff differences in the $c_L(g^{**})$ and $\bar{C}(g^*, g^{**})$ expressions in Proposition 3 and because z is non-decreasing in $(\pi_{NN} - \pi_{CN})$. □

In short, the time spent in cooperation is generally higher with faster learning. This holds true for many other learning/experimentation rules because learning to defect is always immediate. To support these theoretical predictions I conducted simulations for a symmetric model under different learning rates and payoffs. I simulated 1000 runs for every parameter pair and random starting Q-values.[7] The decreasing experimentation probability is modelled as $0.015^{\frac{2}{100000}t}$, similarly to Hettich (2021). Since there is always a non-zero probability of players learning to play a different action, a simulated run is deemed to have converged if the chosen actions do not change for 10000 consecutive periods. Fig. 4 reports the mean convergence times. The pattern from Fig. 2 emerges in this plot—the shortest convergence times are observed for the parameter values that are further away from the region with $\Delta = 0$, which is represented by the dashed curve. Closer to the border between the cooperation and defection regions, the convergence is expectantly more difficult and takes longer. As suggested by the previous analysis, the convergence is almost immediate when defection is stochastically stable and α is low (the dark region in the lower-right corner). For this region the time to convergence is little more than the lowest possible number of 10000 steps.

6.3. Condition-dependent mistakes, Bilancini and Boncinelli (2020)

When $\alpha = 1$, only the most recent payoffs for the two actions determine the transitions (2). Similarly, the condition-dependent mistake model describes transitions in terms of the previous period's payoff but regardless of the action. The behavior in the limit is then driven by the difference in probabilities of experimentation after each action profile. The same logic is at the root of the difference between the logit and uniform cases in the model: under logit, players experiment more frequently when they play the Nash equilibrium than when they cooperate. The last case in Proposition 3 in Bilancini and Boncinelli (2020) describes the

[7] The Python code for simulations is on GitHub: https://github.com/artdol/Qlearning_sim.

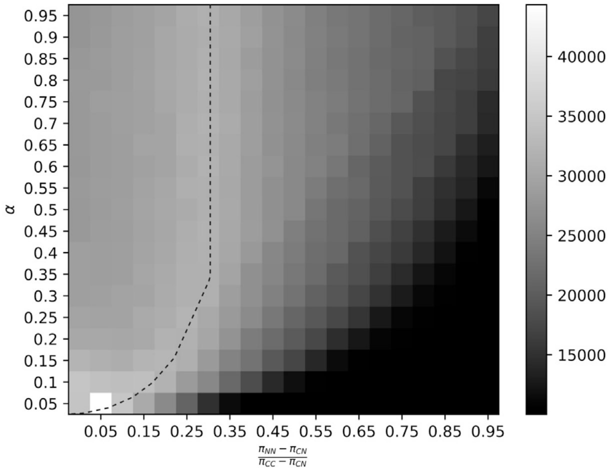

Fig. 4. Mean convergence time for simulated players.

stochastically stable states for a Stag Hunt. A similar argument can be used for a prisoner's dilemma where the selection between the payoff-dominant (C, C) and the risk-dominant (N, N) depends on the (decay of) probability of experimentation in the respective states. The difference with reinforcement learning is that the model does not account for the agent's experience with the other action.

6.4. *Subgame-perfection*

The textbook approach to repeated games focuses on the existence of cooperative equilibria in terms of the discount rate δ and three of the four payoffs of the game. According to the folk theorem, a Nash equilibrium that gives both players cooperative payoffs will exist if $\delta \geq \frac{\pi_{NC}-\pi_{CC}}{\pi_{NC}-\pi_{NN}}$. This expression differs from the payoff information relevant for cooperation of reinforcement learners, $\frac{\pi_{NN}-\pi_{CN}}{\pi_{CC}-\pi_{CN}}$. In fact, the payoff π_{CN} is completely irrelevant in the former, and the temptation payoff π_{NC} in the latter. Reinforcement learners learn from experience and cannot be "tempted" to defect by the prospect of a high payoff. Blonski et al. (2011) argue in favor of a third view, that all four payoffs should matter axiomatically with extremely low and extremely high π_{CN} corresponding to Nash and cooperative outcomes respectively, provided that the discount is high enough to support cooperation in the first place. All three approaches differ in what is assumed about the decision-making process, and their predictions can be clearly demarcated based on the payoffs and parameters, such as δ or α.

6.5. *Other reinforcement learning rules*

The states $\mathfrak{S}$ faithfully encode the state of the value-based reinforcement learning algorithm, but the state of the game itself is a singleton and never changes because the players are unable to condition actions on the history of play. Possnig (2023) shows that the structure of the state space of the reinforcement learning algorithms is yet another crucial factor determining the possibility of collusion. The setup is intentionally stripped from such features, which would help players converge to a cooperative outcome. Simulations by Calvano et al. (2020) and Kasberger et al. (2023) show that the algorithms with memory can learn to cooperate through what appears to be versions of tit-for-tat and other strategies of conditional cooperation. In contrast, Proposition 3 shows that the greedy experimentation policy will never lead to cooperation regardless of payoffs or the learning rate.

Yet surprisingly, Proposition 3 also shows that this memory feature is not always *necessary*, and very simple algorithms can learn to cooperate without memory at least for some experimentation rules. Convergence however requires a high learning rate that is rarely observed in applications. The result of Proposition 3 can therefore be taken as a negative: for low enough learning rates cooperation without memory is impossible. The learning rules discussed in the present paper are contained in parameter spaces of on-policy and off-policy learning algorithms such as Q-learning with forward-looking behavior, SARSA (state–action–reward–state–action, Rummery and Niranjan, 1994), and others. Moreover, the distinctions between these algorithms usually manifest as adjustments in the learning process based on the predicted future payoff—the maximum Q-value for Q-learning or the on-policy action draw for SARSA. The simplicity of the model without histories or forecasting erases the distinctions between these algorithms. It would be interesting to apply the approach to learning algorithms that have a more complex structure of the Q-matrix like deep Q-learning, which have been shown to collude faster than traditional counterparts (Hettich, 2021; Dawid et al., 2023). While my approach applies directly after appropriate transformation of the state space, the central state may not always exist for more complex learning rules.

7. Conclusion

Characterization of learning equilibria in this paper addresses two issues. The results differ from predictions of other learning processes, making Q-learning a testable theory given enough variation in payoffs. It is then a practical question whether subjects think in terms of adjusting their best responses, or instead keep a mental model of expected valuations of different actions, the Q-vector.

A more apparent setting for this research is the field of algorithmic pricing. Due to their simplicity, Q-learning algorithms provide a low-complexity baseline for algorithmic pricing systems. Even these simple algorithms have been shown empirically to support supra-competitive prices when allowed to condition actions on history. Depending on the acceptable values of the parameters and the interpretation of the learning rate, Proposition 3 either shows that algorithmic collusion is possible even without memory, or that it is impossible unless the parameter values are degenerate. Moreover, given the opportunity, a rational designer will choose the optimal set of parameters for the algorithm to maximize the chance of collusion, namely the highest learning rate and the best guess for the experimentation rate of the opponent.

Reinforcement learning has been entering the traditional areas of economics research such as taxation (Zheng et al., 2020) and oligopoly pricing (Calvano et al., 2020). This paper is an attempt to go in the other direction: use the tools of evolutionary game theory to study convergence of multiple reinforcement learning agents. This can be a useful approach to understanding the behavior of large groups of reinforcement learning agents in complex economic environments. A prisoner's dilemma is the simplest form of several motivating models, including a public goods game and a Bertrand oligopoly. A natural extension of this analysis is a differentiated Bertrand competition or a similar game. Unfortunately, not all results extend in a straightforward manner to games without a dominant strategy equilibrium as the minimum-cost path to a central state may no longer exist.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

I am indebted to Giacomo Calzolari for continuing support and advice. I am very grateful to Jonathan Newton, Heinrich Nax, Dominik Karos, Yves Breitmoser, Ennio Bilancini, Marina Isabel Sanchez Del Villar, and Philip Hanspach for the invaluable comments. I also thank Alyson Price for proof reading the earlier version. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Remaining proofs

Proof of Lemma 1. By construction in (1),

$$Q_i^{a_i}(g) \in [\min_{a_{-i}}(\pi_{a_i a_{-i}}), \max_{a_{-i}}(\pi_{a_i a_{-i}})]$$

for both players and any action $a_i \in \text{path}_i(g)$.

From the state g, the process transitions to some new state $g' = \mathcal{F}^{a_1,a_2}(g)$.

Suppose first that a_i is played in g'. Then

(i) if $\max_{a_{-i}\in\mathcal{A}_{-i}(G)} \pi_{a_i a_{-i}} < Q_i^{a_i}(g)$, then
$$Q_i^{a}(g') - \max_{a_{-i}\in\mathcal{A}_{-i}(G)} \pi_{a_i a_{-i}} < Q_i^{a_i}(g) - \max_{a_{-i}\in\mathcal{A}_{-i}(G)} \pi_{a_i a_{-i}};$$
(ii) if $\min_{a_{-i}\in\mathcal{A}_{-i}(G)} \pi_{a_i a_{-i}} > Q_i^{a_i}(g)$, then
$$\min_{a_{-i}\in\mathcal{A}_{-i}(G)} \pi_{a_i a_{-i}} - Q_i^{a_i}(g') < \min_{a_{-i}\in\mathcal{A}_{-i}(G)} \pi_{a_i a_{-i}} - Q_i^{a_i}(g);$$
(iii) otherwise $Q_i^{a_i}(g') \in [\min_{a_{-i}\in\mathcal{A}_{-i}(G)} \pi_{a_i a_{-i}}, \max_{a_{-i}\in\mathcal{A}_{-i}(G)} \pi_{a_i, a_{-i}}]$.

If instead a_i is not played in g' then $Q_i^{a_i}(g') = Q_i^{a_i}(g)$.

Hence any state g for which $Q_i^{a_i}(g) > \max_{a_{-i}\in\mathcal{A}_{-i}(G)} \pi_{a_i a_{-i}}$ or $Q_i^{a_i}(g) < \min_{a_{-i}\in\mathcal{A}_{-i}(G)} \pi_{a_i a_{-i}}$ is transient and therefore $g \notin G$. □

Proof of Lemma 2. The lemma is trivial if there is only one action played by i in all states of G, so let there be at least two such actions.

Suppose there is a sequence of positive probability transitions $g_0, g_1, g_2, ..., g_l$ through the states of the recurrent class, such that player i plays some action $a \in A$ in all of these states except the first g_0 and the last g_l, so that $a \in \text{path}_i(g_1) \cap \text{path}_i(g_2) \cap ... \cap \text{path}_i(g_{l-1})$.

Suppose also that in g_0 she plays some $\hat{a} \in \text{path}_i(g_0), \hat{a} \neq a$, and at g_l she plays $a' \in \text{path}_i(g_l), a' \neq a$. For the player to switch her action from a to a', the Q-value of the action a has to decrease to the level of a' or below it. Thus, the Q-value of a cannot be strictly lower at g_1 than at g_l, $Q_i^a(g_l) \leq Q_i^{a'}(g_l) = Q_i^{a'}(g_1) \leq Q_i^a(g_1)$. Otherwise, player i would have played a' instead of a already in g_1. This is true for any such sequence of states that ends in a switch of an action. By the recurrent class property, the process must return to g_1 with positive probability. If the inequality was strict so that $Q_i^a(g_l) < Q_i^a(g_1)$, then after the process leaves the state g_l it must eventually go through a sequence of states, such that i always plays a and by the end of the sequence updates its Q-value back to the level $Q_i^a(g_1)$. Let us similarly denote this sequence by $\hat{g}_0, \hat{g}_1, \hat{g}_2, ..., \hat{g}_l$, where a is played in $\hat{g}_1, ...\hat{g}_{l-1}$. For the sequence to reach the Q-value level of $Q_i^a(g_1)$, it must be that $Q_i^a(\hat{g}_l) > Q_i^a(\hat{g}_1)$ for some such sequence. This contradicts the fact just established above, that for any such sequence of states, the corresponding Q-value cannot increase: $Q_i^a(\hat{g}_l) \leq Q_i^a(\hat{g}_1)$. Therefore it must be that $Q_i^a(g_l) = Q_i^{a'}(g_l) = Q_i^{a'}(g_1) = Q_i^a(g_1)$. Denote this value q_i.

The same argument can now be applied to the next sequence of states $g_{l-1} = g'_0, g_l = g'_1, g'_2, ...g'_l$ where a' is played until the next switch to some action in $A \setminus a'$ at state g'_l. Thus, $Q_i^{a'}(g'_l) = Q_i^{a'}(g_l) = q_i$. Since a is not played, its value is unchanged, $Q_i^a(g'_1) = Q_i^a(g'_l) = q_i$. Then the lemma holds until g'_l. By the same argument, it also holds for the next sequence of states after g'_l until the action changes again. By induction it also holds for the whole remaining chain. Since the first states $g_0, g_1, g_2, ..., g_l$ are reached with positive probability, the lemma is true for the whole sequence of transitions through G. □

Proof of Proposition 1. I start with the "if" part. Each player is taking the action with the higher Q-value, a_1 and a_2 respectively in the unperturbed process. Since the payoffs from this profile are exactly $\pi_{a_1 a_2}$ and $\pi_{a_2 a_1}$, the new Q-vectors are unchanged, $\mathcal{F}^{a_1,a_2}(g) = g$. Thus, the process stays at g.

For the "only if" part, take a recurrent class G. Suppose first that only one profile is played in G, i.e., $\mathcal{A}_1(G) = a_1$, $\mathcal{A}_2(G) = a_2$ for some pair of actions $a_1, a_2 \in A$. For the action profile played on path, the Q values should equal the expected value of playing these actions by Lemma 1. That is for any state $g \in G$, $\min_{x \in \mathcal{A}(G)} \pi_{a_i x} = \pi_{a_i a_{-i}} \leq Q_i^{a_i}(g) \leq \max_{x \in \mathcal{A}(G)_{-i}} \pi_{a_i x} = \pi_{a_i a_{-i}}$. Then $Q_i^{a_i}(g) = \pi_{a_i a_{-i}}$ for $i \in I$. Moreover, if $Q_i^{a_i}(g) \leq Q_i^{b_i}(g)$ for some i and some action $b_i \in A, b_i \neq a_i$ then a different action profile is played, which is a contradiction. Therefore there are no other recurrent classes where only one action profile is played. It remains to show that there are no non-singleton recurrent classes where multiple action profiles occur. The game is known to be solvable by IESDS. Therefore, unless a unique profile is played in G there is a strictly dominated strategy for some player i that is played with a positive probability. That is, there is a pair of actions $a_i, b_i \in \mathcal{A}_i(G)$, s.t. $\pi_{a_i a_{-i}} > \pi_{b_i a_{-i}}$ for all $a_{-i} \in A$. Consider then the case when i switches from playing a_i in some state in G to playing b_i in the next state in G. Call the latter state g. Lemma 2 implies that this transition occurs with strictly positive probability within the recurrent class: when player i changes the played action, she can switch to any action in $\mathcal{A}_i(G)$, because all actions in $\mathcal{A}_i(G) \setminus \text{path}_i(g)$ have Q-value q_i. Suppose the other player plays some action a_{-i} at state g. Then $Q_i^{a_i}(g) \geq \pi_{a_i a_{-i}} > \pi_{b_i a_{-i}}$. The first inequality is because otherwise the Q-value of a_i would have strictly increased and i would keep playing a_i at g. The second, strict inequality is due to strict dominance. Then by Lemma 2, $Q_i^{b_i}(g) = Q_i^{a_i}(g) > \pi_{b_i a_{-i}}$. However, this implies that player i immediately updates the Q-value of b_i down to some value in the interval $[\pi_{b_i a_{-i}}, Q_i^{b_i}(g))$. Thus, she has to switch to a different action at the state following g, leaving the Q-value of b_i strictly below $Q_i^{a_i}(g) = Q_i^{b_i}(g)$ and contradicting Lemma 2. □

Proof of Lemma 3. Let $\hat{g}$ be the root of the tree. Suppose to the contrary that there is a state $g' \in \mathfrak{C}$ with the cost of the outgoing edge strictly greater than $c_L(g')$ and there is no path from g^c to g'. Then one can construct another spanning graph by progressively adding the minimum cost edges starting from g' until the process either reaches g^c or some state on the path from g^c to $\hat{g}$. Removing any previous outgoing edges along this path, including the one from g', yields a tree with a lower cost, because every edge on this path now has the minimum cost by construction. Moreover, since we stopped adding edges once we reached the path from g^c to $\hat{g}$, there is a path to $\hat{g}$ from any state and the graph is therefore a spanning tree. □

Proof of Proposition 2. Take $\hat{g} \neq g^c$. By Lemma 3 any minimal spanning tree has all minimum cost outgoing edges except for the path between g^c and $\hat{g}$. The difference in the cost between the minimal tree rooted in $\hat{g}$, $cost(\hat{g})$ and the minimal trees rooted in g^c, $cost(g^c)$, is then the cost of this path and the sum of the least cost transitions from every state on this path. The difference equals the c_L-adjusted cost $\bar{C}(g^c, \hat{g})$ minus the minimum cost of a transition from the root $c_L(\hat{g})$, which is not included in the definition of the c_L-adjusted cost. Then the cost of a tree rooted in a state $\hat{g} \neq g^c$ is $cost(\hat{g}) = cost(g^c) + \bar{C}(g^c, \hat{g}) - c_L(\hat{g})$. The cost of the tree rooted in the central state g^c is $cost(g^c)$. The $cost(g^c)$ enters the expressions for costs of all trees and can be subtracted out, yielding $\bar{C}(g^c, \hat{g}) - c_L(\hat{g})$. If it is non-positive for some $\hat{g} \neq g^c$ then $cost(\hat{g}) \leq cost(g^c)$ (case i). If it is non-negative for all $\hat{g} \neq g^c$ then $cost(g^c)$ is minimal (case ii). □

Proof of Lemma 4. Take any state $g \in \mathfrak{C} \setminus g^*$ with (a_1, a_2) played on path and $b_i \neq a_i$ for $i \in I$.

Neither (N, C) nor (C, N) can be played on path at $g \in \mathfrak{C}$ because then $Q_i^N(g) < Q_i^C(g) = \pi_{CN}$ by Proposition 1 for one of the players. This contradicts $g \in \mathfrak{G}$, because π_{NN} is the lowest payoff to N and therefore $\pi_{CN} < \pi_{NN} \leq Q_i^N(g)$.

The remaining proof is by cases.

1. Suppose $(N, N) \in \text{path}(g)$. Then $Q_i^C(g) \neq \pi_{CN}$ for one of the players $i \in I$ in order for $g \neq g^*$. If the non-equality holds for both players, suppose without loss of generality that the least cost transition is by player i. Then in all cases experimentation leads i to

play C. By Proposition 1 since $g \in \mathfrak{C}$, $Q_i^N(g) = Q_{-i}^N(g) = \pi_{NN}$ and $Q_i^N(g) > Q_i^C(g)$, $Q_{-i}^N(g) > Q_{-i}^C(g)$. Player i then obtains $\pi_{CN} < \pi_{NN}$ at g_1 and $Q_i^C(g_1) < Q_i^N(g_1) = \pi_{NN}$. The player $-i$ obtains $\pi_{NC} > \pi_{NN}$ at g_1 and since $Q_{-i}^N(g) = \pi_{NN}$, $Q_{-i}^N(g_1) > Q_{-i}^N(g)$. Therefore at g_1 again $Q_i^N(g_1) > Q_i^C(g_1)$, $Q_{-i}^N(g_1) > Q_{-i}^C(g_1)$ and $(N,N) \in \text{path}(g_1)$. Then, with a positive probability and zero cost at states $g_2, g_3, \ldots g'$ that follow g_1, $Q_i^N(g_2) = Q_i^N(g_3) \ldots = \pi_{NN}$ continues to hold and $|Q_{-i}^N(g_2) - \pi_{NN}|, |Q_{-i}^N(g_3) - \pi_{NN}|, \ldots$ decrease until $Q_{-i}^N(g') = \pi_{NN}$ again for some g'. Thus, in the new recurrent state $Q_i^N(g') = Q_{-i}^N(g') = \pi_{NN}$, $Q_{-i}^C(g') = Q_{-i}^C(g)$, but $|Q_i^C(g^*) - Q_i^C(g')| = |\pi_{CN} - Q_i^C(g')| < |\pi_{CN} - Q_i^C(g)|$. Thus, $D(g') < D(g)$, while $m(g') = m(g) = 2$, and so $g' \prec g$ as required.

2. Suppose instead $\text{path}(g) = \{(C,C)\}$. Then experimentation leads i to play N. By Proposition 1 since $g \in \mathfrak{C}$, $Q_i^C(g) = Q_{-i}^C(g) = \pi_{CC}$ and $Q_i^C(g) > Q_i^N(g)$, $Q_{-i}^C(g) > Q_{-i}^N(g)$. Player i then obtains $\pi_{NC} > \pi_{CC}$ at g_1, and thus $Q_i^N(g_1) > Q_i^N(g)$. Player $-i$ obtains $\pi_{CN} < \pi_{CC}$ at g_1, and thus $Q_{-i}^C(g_1) < Q_{-i}^C(g)$.

   Two subcases are possible.

   (a) Both $Q_i^C(g_1) \geq Q_i^N(g_1)$, $Q_{-i}^C(g_1) \geq Q_{-i}^N(g_1)$ at g_1 and (C,C) is played again with positive probability. Then, with a positive probability and zero cost at states $g_2, g_3, \ldots g'$ that follow $Q_i^C(g_2) = Q_i^C(g_3) = \ldots = \pi_{CC}$ continues to hold and $|Q_{-i}^C(g_2) - \pi_{CC}|, |Q_{-i}^C(g_3) - \pi_{CC}|, \ldots$ decrease until $Q_{-i}^C(g') = \pi_{CC}$ again for some g'. Thus, in the new recurrent state $Q_i^C(g') = Q_{-i}^C(g') = \pi_{CC}$, $Q_{-i}^N(g') = Q_{-i}^N(g)$, but $\pi_{CC} > Q_i^N(g') = Q_i^N(g_1) > Q_i^N(g)$. By Assumption 2(vii) the cost of player i experimentation is at least as high at g' as at g. Moreover, since i was the experimenting player at g, player i is also at least as likely to experiment at g' as the other player. Then $0 = m(g') = m(g)$, $D(g') = D(g) = 2(\pi_{CC} - \pi_{CN})$, $c_M(g') = c_M(g)$, and $c_L(g') < c_L(g)$, so $g' \prec g$ as required.

   (b) In all remaining cases with probability 1 the process P_0 leads to some state g'', $Q_{-i}^C(g'') < Q_{-i}^N(g)$. If $Q_{-i}^C(g_1) < Q_{-i}^N(g_1) = Q_{-i}^N(g)$, this is true immediately at $g_1 = g''$. If instead $Q_{-i}^C(g_1) \geq Q_{-i}^N(g_1)$, but $Q_i^C(g_1) < Q_i^N(g_1)$, then i plays N and $-i$ plays C with positive probability. The payoff of player i is the highest possible, and the payoff of the other player is the lowest possible. The Q-value of C for player $-i$ decreases until (N,N) is the only profile played in some g''. For that state $Q_{-i}^C(g'') < Q_{-i}^N(g'') = Q_{-i}^N(g)$ as required.

   If the process reaches a state $g' \in \mathfrak{C}$ where the players defect, then $2 = m(g') > m(g)$, and $g' \prec g$ as required. So suppose instead that a state in $\mathfrak{C}$ with only (C,C) on path is reached. Let $\hat{g}$ be the first state where $\text{path}(\hat{g}) = \{(C,C)\}$, not necessarily in $\mathfrak{C}$. One can assume that in every transition between g' and $\hat{g}$ at least one player always defected, so $Q_{-i}^C(g'') \geq Q_{-i}^C(\hat{g})$. Then $Q_{-i}^N(g) > Q_{-i}^C(g'') \geq Q_{-i}^C(\hat{g}) > Q_{-i}^N(\hat{g})$. If $g' \notin \mathfrak{C}$, the process continues with plays of (C,C) until a state in $\mathfrak{C}$ is reached. Denote this state by g'. Since only C was played by either player between $\hat{g}$ and g', in this state g', $Q_{-i}^N(\hat{g}) = Q_{-i}^N(g') < Q_{-i}^N(g)$ and, because $g' \in \mathfrak{C}$, $Q_i^C(g') = Q_{-i}^C(g') = \pi_{CC} = Q_{-i}^C(g)$. Then, by Assumption 2(vii), the cost of experimentation by $-i$ is higher at g' than at g. The latter also equals $c_M(g)$ since i originally had lower cost of experimentation at g by assumption. Regardless of which player is less likely to experiment at g', the cost of this experimentation $c_M(g')$ is no less than the cost of experimentation by player $-i$ at g'. Therefore $c_M(g') > c_M(g)$ and $g' \prec g$ as required. □

Proof of Lemma 5. I will show that the c_L-adjusted cost of a sequence of (C,C) updates $g^* = g_1, \ldots, g_r = g^{**}$ is always lower than the cost of a sequence where N is played at least by one player at least once. Suppose now, to contradiction, that there is a lower cost path $g^* = g_1, g_2' \ldots g_{r'}' = g^{**}$ where for some state g_i', $g_i' \neq \mathcal{F}^{C,C}(g_{i-1}')$.

The total c_L-adjusted cost of the path without defection is $\bar{c}(g_1, \ldots g_r) = c(g_1, g_2) + \sum_{l=2}^{r-1} \bar{c}(g_l, g_{l+1}) = c_L(g^*) + \sum_{l=1}^{r-1} \bar{c}(g_l, g_{l+1})$. Similarly, the total c_L-adjusted cost of the path with defection is $\bar{c}(g_1', \ldots g_r') = c_L(g^*) + \sum_{l=1}^{r-1} \bar{c}(g_l', g_{l+1}')$. The $c_L(g^*)$ component is the same on both paths. I will compare only the costs of transitions for the states where (C,C) is played. On the path without defection $g^* = g_1, \ldots, g_r = g^{**}$, these are all of the transitions. On the path with defection $g^* = g_1, g_2' \ldots g_{r'}' = g^{**}$, the cost is at least as high as the total cost of transitions from these states as the $(C,N), (N,N)$ or (N,C) transitions are also present. Let $\bar{g}_l'$ be the state on this path where (C,C) is played for the l-th time. There is at least one such state, g^{**}. It suffices to show that the cost of these transitions is already higher on the path with defection than on the path without defection, without accounting for the remaining states.

The c_L-adjusted cost of these (C,C) transitions for $1 \leq l < r'$ depends on the actions on path. There are two cases.

1. If $\text{path}(g_l')=\{(N,N)\}$ then, by Assumption 2(vi), $\bar{c}(g_l', g_{l+1}') = c\left(g_l', \mathcal{F}^{CN}\left(g_l'\right)\right) + c\left(g_l', \mathcal{F}^{NC}\left(g_l'\right)\right) - c_L(g_l') \geq \max\{\left(c\left(g_l', \mathcal{F}^{CN}\left(g_l'\right)\right), c\left(g_l', \mathcal{F}^{NC}\left(g_l'\right)\right)\right)\}$. If the state g_l' is recurrent, $g_l' \in \mathfrak{C}$, then c_L equals the cost of playing C for the player that is more likely to experiment, the previous expression holds with equality, and $\bar{c}(g_l', g_{l+1}') = c_M(g_l')$. If $g_l' \notin \mathfrak{C}$ then $c_L(g_l') = 0$ and the expression holds with strict inequality.
2. If instead $\text{path}(g_l') = \{(C,N)\}$ or $\text{path}(g_l') = \{(N,C)\}$ then $\bar{c}(g_l', g_{l+1}') = c\left(g_l', \mathcal{F}^{CC}\left(g_l'\right)\right) - c_L\left(g_l'\right) = c\left(g_l', \mathcal{F}^{CC}\left(g_l'\right)\right)$. These states are never recurrent by Proposition 1 and therefore $c_L(g_l') = 0$.

In both cases, the c_L-adjusted cost of (C,C) transition is at least as high as the cost of one player playing C for the player for whom it is the highest. On the path with only cooperation, the cost exactly equals this value. Indeed, in all states $Q_i^N(g^*) = Q_i^N(g_l) = Q_i^N(g^{**}) = \pi_{NN}$ for both $i \in I$ and all $1 \leq l \leq r$. So all states where $\text{path}(g_l) = \{(N,N)\}$ are recurrent and $c\left(g_l, \mathcal{F}^{CN}\left(g_l\right)\right) + c\left(g_l, \mathcal{F}^{NC}\left(g_l\right)\right) - c_L(g_l) = \max\{c\left(g_l, \mathcal{F}^{CN}\left(g_l\right)\right), c\left(g_l, \mathcal{F}^{NC}\left(g_l\right)\right)\}$. The states where $\text{path}(g_l) = \{(C,N)\}$ or $\text{path}(g_l) = \{(N,C)\}$ are never recurrent and $c\left(g_l, \mathcal{F}^{CC}\left(g_l\right)\right) - c_L(g_l) = c\left(g_l, \mathcal{F}^{CC}\left(g_l\right)\right)$.

A play of $(C,N),(N,C)$, or (N,N) cannot increase the Q-value of cooperation for either player. Formally, by Assumption 2(v), $Q_i^C(g_l') \le Q_i^C(g_{l+1}')$ if $(C,N),(N,C)$, or (N,N) is played in g_l', so that g_{l+1}' equals $\mathcal{F}^{CN}(g_l'),\mathcal{F}^{NC}(g_l')$ or $\mathcal{F}^{NN}(g_l')$. The Q-value of non-cooperation always takes the smallest possible value π_{NN} on the path without defection. These two facts together imply that the Q-values of cooperation (defection) are at most (at least) as high when (C,C) is played on the path with defection compared to the path where only (C,C) is played consistently without defection. Formally, $Q_i^C(\bar{g}_l') \le Q_i^C(g_l)$ and $Q_i^N(\bar{g}_l') \ge Q_i^N(g_l) = \pi_{NN}$. Then, by Assumption 2(vii), the cost of any player i with $\text{path}_i(g) = \{N\}$ experimenting and playing C instead of N is lower in any g_l then in g_l' for $1 \le l \le \min\{r, r'\}$. By the same argument, the total number of such states is no less on the path with defection than on the path without defection. Therefore, the c_L-adjusted cost of every transition is higher, and the overall c_L-adjusted cost is also higher with defection.

It remains to show that a path to any state with cooperation on path other than g^{**} also requires a higher cost. Since the Q-value of N is the same at g^* and at g^{**}, $Q_i^N(g^*) = Q_i^N(g^{**}) = \pi_{NN}$ for both $i \in I$, the sequence $g_1, g_2, \ldots g_r$ ends exactly in g^{**}. Any other state with (C,C) on path would require at least one play of $(C,N),(N,C)$, or (N,N) to change the Q-value of N. □

Proof of Proposition 3. (i) Under the greedy rule, a two-player simultaneous experimentation has the probability $e^{-\frac{k_1+k_2}{\eta}}$ and the cost k_1+k_2, while a single-player experimentation has the probability at least $\min\{e^{-\frac{k_1}{\eta}}, e^{-\frac{k_2}{\eta}}\}$ and the cost at most $\max\{k_1,k_2\}$. By construction, leaving g^* requires a two-player simultaneous experimentation, while a least cost transition from any other state $\hat{g} \in \mathfrak{C}$ requires only a single-player experimentation. Then $c_L(g^*) > c_L(\hat{g})$ and by Corollary 1 $SS = \{g^*\}$.

(ii) For $C(g^*,g^{**})$ on the minimum-cost path where all players cooperate (by Lemma 5), the cost of transitions is $\frac{1}{k_1\eta}(\pi_{NN} - q_{l,1}^C) + \frac{1}{k_2\eta}(\pi_{NN} - q_{l,2}^C)$. That is, $C(g^*,g^{**}) = \sum_{l=1}^{z_1} \frac{1}{k_1\eta}(\pi_{NN} - q_{l,1}^C) + \sum_{l=1}^{z_2} \frac{1}{k_2\eta}(\pi_{NN} - q_{l,2}^C)$. For the $\bar{C}(g^*,g^{**})$ two cases are possible. If $z=1$ then $\bar{C}(g^*,g^{**}) = C(g^*,g^{**})$. Otherwise, subtract $c_L(g_l)$ of every g_l for $2 \le l \le z$, which gives the second term, similarly to the argument in the proof of Lemma 5. In particular, $c\left(g_l, \mathcal{F}^{C,C}\left(g_l\right)\right) - c_L(g_l) = \max\{c\left(g_l, \mathcal{F}^{C,N}\left(g_l\right)\right), c\left(g_l, \mathcal{F}^{N,C}\left(g_l\right)\right)\} = \max_{i\in I} \max\{0, \frac{1}{k_i}\left(\pi_{NN} - q_{l,i}^C\right)\}$ for any $2 \le l \le z$ if $z > 1$. This leaves the $c(g_1,g_2)$ term for $l=1$, which equals $c_l(g^*) = \left(\frac{1}{k_1} + \frac{1}{k_2}\right)(\pi_{NN} - \pi_{CN})$. Finally, $c_L(g^{**}) = \min_{i\in I} \frac{1}{k_i\eta}(\pi_{CC} - \pi_{NN})$. By substituting these terms into $\bar{C}(g^*,g^{**}) - c_L(g^{**})$, one obtains the expression for Δ. The result then follows by Corollary 3. □

References

Abada, I., Lambin, X., 2023. Artificial intelligence: can seemingly collusive outcomes be avoided? Manag. Sci.

Asker, J., Fershtman, C., Pakes, A., 2021. Artificial Intelligence and Pricing: the Impact of Algorithm Design. Tech. Rep.. National Bureau of Economic Research. https://doi.org/10.3386/w28535.

Asker, J., Fershtman, C., Pakes, A., et al., 2022. Artificial Intelligence, Algorithm Design, and Pricing. AEA Papers and Proceedings, vol. 112. American Economic Association, pp. 452–456.

Assad, S., Clark, R., Ershov, D., Xu, L., 2022. Identifying Algorithmic Pricing Technology Adoption in Retail Gasoline Markets. AEA Papers and Proceedings, vol. 112, pp. 457–460.

Banchio, M., Mantegazza, G., 2022. Adaptive algorithms and collusion via coupling. arXiv preprint. arXiv:2202.05946.

Bilancini, E., Boncinelli, L., 2020. The evolution of conventions under condition-dependent mistakes. Econ. Theory 69, 497–521. https://doi.org/10.1007/s00199-019-01174-y.

Bilancini, E., Boncinelli, L., Nax, H.H., 2021. What noise matters? Experimental evidence for stochastic deviations in social norms. J. Behav. Exp. Econ. 90, 101626.

Blonski, M., Ockenfels, P., Spagnolo, G., 2011. Equilibrium selection in the repeated prisoner's dilemma: axiomatic approach and experimental evidence. Am. Econ. J. Microecon. 3, 164–192. https://doi.org/10.1257/mic.3.3.164.

Buşoniu, L., Babuška, R., De Schutter, B., 2010. Multi-agent reinforcement learning: an overview. In: Innovations in Multi-Agent Systems and Applications-1, pp. 183–221.

Calvano, E., Calzolari, G., Denicolo, V., Pastorello, S., 2020. Artificial intelligence, algorithmic pricing, and collusion. Am. Econ. Rev. 110, 3267–3297. https://doi.org/10.4337/9781786439055.00038.

Calvano, E., Calzolari, G., Denicoló, V., Pastorello, S., 2021. Algorithmic collusion with imperfect monitoring. Int. J. Ind. Organ. 79, 102712.

Calvano, E., Calzolari, G., Denicolò, V., Pastorello, S., 2023. Algorithmic collusion: genuine or spurious? Int. J. Ind. Organ. 90, 102973.

Dawid, H., Harting, P., Neugart, M., 2023. Implications of Algorithmic Wage Setting on Online Labor Platforms: a Simulation-Based Analysis. Bielefeld Working Papers in Economics and Management.

den Boer, A.V., Meylahn, J.M., Schinkel, M.P., 2022. Artificial collusion: Examining supracompetitive pricing by Q-learning algorithms. Amsterdam Law School Research Paper.

Dorner, F.E., 2021. Algorithmic collusion: a critical review. arXiv preprint. arXiv:2110.04740.

Ellison, G., 2000. Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution. Rev. Econ. Stud. 67, 17–45. https://doi.org/10.1111/1467-937x.00119.

Erev, I., Roth, A.E., 1998. Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. Am. Econ. Rev., 848–881.

Foster, D., Young, H.P., 2006. Regret testing: learning to play Nash equilibrium without knowing you have an opponent. Theor. Econ. 1, 341–367. https://doi.org/10.1109/ieecon.2014.6925916.

Freidlin, M.I., Wentzell, A.D., 1984. Random Perturbations. Springer US, New York, NY, pp. 15–43.

Harrington, J.E., 2018. Developing competition law for collusion by autonomous artificial agents. J. Compet. Law Econ. 14, 331–363.

Hart, S., Mas-Colell, A., 2003. Uncoupled dynamics do not lead to Nash equilibrium. Am. Econ. Rev. 93, 1830–1836. https://doi.org/10.1257/000282803322655581.

Hettich, M., 2021. Algorithmic collusion: Insights from deep learning. Available at SSRN 3785966.

Hu, J., Wellman, M.P., 2003. Nash Q-learning for general-sum stochastic games. J. Mach. Learn. Res. 4, 1039–1069.

Hu, J., Wellman, M.P., et al., 1998. Multiagent reinforcement learning: theoretical framework and an algorithm. In: ICML, vol. 98, pp. 242–250.

Kasberger, B., Martin, S., Normann, H.-T., Werner, T., 2023. Algorithmic Cooperation. Available at SSRN 4389647.

Klein, T., 2021. Autonomous algorithmic collusion: Q-learning under sequential pricing. Rand J. Econ. https://doi.org/10.1111/1756-2171.12383.

Marden, J.R., Young, H.P., Arslan, G., Shamma, J.S., 2009. Payoff-based dynamics for multiplayer weakly acyclic games. SIAM J. Control Optim. 48, 373–396. https://doi.org/10.1137/070680199.

Mäs, M., Nax, H.H., 2016. A behavioral study of "noise" in coordination games. J. Econ. Theory 162, 195–208.

Mengel, F., 2014. Learning by (limited) forward looking players. J. Econ. Behav. Organ. 108, 59–77. https://doi.org/10.26481/umamet.2008053.

Milgrom, P., Roberts, J., 1990. Rationalizability, learning, and equilibrium in games with strategic complementarities. Econometrica, 1255–1277. https://doi.org/10.2307/2938316.

Nax, H.H., 2019. Uncoupled aspiration adaptation dynamics into the core. Ger. Econ. Rev. 20, 243–256. https://doi.org/10.1111/geer.12160.

Nax, H.H., Pradelski, B.S., 2015. Evolutionary dynamics and equitable core selection in assignment games. Int. J. Game Theory 44, 903–932.

Newton, J., 2018. Evolutionary game theory: a renaissance. Games 9, 31. https://doi.org/10.3390/g9020031.

Newton, J., Sawa, R., 2015. A one-shot deviation principle for stability in matching problems. J. Econ. Theory 157, 1–27. https://doi.org/10.1016/j.jet.2014.11.015.

Possnig, C., 2023. Reinforcement learning and collusion. Preprint.

Roth, A.E., Erev, I., 1995. Learning in extensive-form games: experimental data and simple dynamic models in the intermediate term. Games Econ. Behav. 8, 164–212. https://doi.org/10.1016/s0899-8256(05)80020-x.

Rummery, G.A., Niranjan, M., 1994. On-line Q-learning using connectionist systems, vol. 37. Citeseer.

Sandholm, W.H., 2010. Population Games and Evolutionary Dynamics. MIT Press.

Sen, S., Sekaran, M., Hale, J., 1994. Learning to coordinate without sharing information. In: Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence, pp. 426–431.

Suematsu, N., Hayashi, A., 2002. A multiagent reinforcement learning algorithm using extended optimal response. In: Proceedings of the First International Joint Conference on Autonomous Agents and Multiagent Systems: Part 1, pp. 370–377.

Tuyls, K., Hoen, P.J.T., Vanschoenwinkel, B., 2006. An evolutionary dynamical analysis of multi-agent learning in iterated games. Auton. Agents Multi-Agent Syst. 12, 115–153.

Waltman, L., Kaymak, U., 2007. A theoretical analysis of cooperative behavior in multi-agent Q-learning. In: 2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning. IEEE, pp. 84–91.

Waltman, L., Kaymak, U., 2008. Q-learning agents in a Cournot oligopoly model. J. Econ. Dyn. Control 32, 3275–3293. https://doi.org/10.1016/j.jedc.2008.01.003.

Young, H.P., 1993. The evolution of conventions. Econometrica, 57–84. https://doi.org/10.2307/2951778.

Zhang, K., Yang, Z., Başar, T., 2021. Multi-agent reinforcement learning: a selective overview of theories and algorithms. In: Handbook of Reinforcement Learning and Control, pp. 321–384.

Zheng, S., Trott, A., Srinivasa, S., Naik, N., Gruesbeck, M., Parkes, D.C., Socher, R., 2020. The ai economist: improving equality and productivity with ai-driven tax policies. arXiv preprint. arXiv:2004.13332.