REINFORCEMENT LEARNING IN A PRISONER'S DILEMMA

ARTHUR DOLGOPOLOV

European University Institute

Abstract. We fully characterize the outcomes of a wide-class of Q-value based model-free reinforcement learning algorithms, such as Q-learning, in a prisoner's dilemma. Learning is shown to always converge to one of the two states. Whether the players learn to cooperate or defect can be determined in closed-form from the relationship between the learning rate and the payoffs of the game. The results generalize to asymmetric learners and many experimentation dynamics.

1. Introduction

While a wide classes of learning rules have been studied with relation to the prisoner's dilemma, reinforcement learning algorithms, such as Q-learning, are rarely considered outside of simulation studies due to their large state complexity. In this paper we fill this gap by offering a complete closed-form characterization that removes the necessity for simulations.

This paper studies all algorithms in a class that is a generalization of the, so-called, Q-learning algorithm. A Q-learning agent maintains a vector of Q-values that encode her 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 some predetermined rule.

E-mail address: arthur.dolgopolov@eui.eu.

Date: September 2021.

Part of the attractiveness of reinforcement learning as a model for behavior lies with the minimal assumptions imposed by such algorithms 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 themselves only use their prior experience to play, having no knowledge of the game structure. It is therefore not surprising that the predictions derived from our model are in stark contrast with the predictions of other adaptive dynamics, such as best-response[1], or conventional analysis of repeated games through subgame-perfect equilibria and folk theorems. We shall further outline these differences in the conclusion, having the expression for the relevant payoff information at hand.

Unlike some of the similar studies of learning in a prisoner's dilemma (Mengel, 2014; Calvano et al., 2020), we also do not consider "memory", i.e. actions can not be conditioned on past 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 guarantee convergence to non-Nash outcomes even without relying on conditional strategies. In fact, this exact property has gained attention for the algorithm as a proof-of-concept technique for showing the possibility of algorithmic collusion in deceptively benign environments where neither the algorithm nor its designers observe anything beyond their own payoffs (Calvano et al., 2020; Klein, 2021).

The most closely related studies (Waltman and Kaymak, 2007, 2008) pursue the same goal as us and have partially characterized the convergence in the prisoner's dilemma game for high learning rates. In particular, when one experimentation step

[1]Best-responses will converge to Nash equilibrium as well as many other learning dynamics that rely on weak acyclicity of the game, see for example Marden et al. (2009).

is enough for a switch from a noncooperative state to a cooperative state and vice-versa, our analysis can be simplified by only considering the minimum cost paths. In many applications however (Calvano et al., 2020, e.g.), the learning rate may be expected to be low to ensure enough experimentation over a short period and full traversal of the state space. Neither is it clear if the high learning rate assumption would be restrictive for human subjects.

Unlike Waltman and Kaymak (2007, 2008), we follow the evolutionary game theory approach of characterizing stochastically stable sets through spanning trees (Young, 1993). This approach has been applied to prisoner's dilemma in particular to characterize learning rules based on sampling from past history Mengel (2014). The Q-learning algorithm instead maintains an "expectation" of the payoff from choosing actions in the state of the learning process.

Another set of closely related learning rules is adaptive dynamics (Milgrom and Roberts, 1990), which can be shown to always converge to Nash equilibria in supermodular games and thus also in a prisoner's dilemma. However, Q-learning is not in this class and will be shown to converge to non-equilibrium actions for some combinations of parameters and payoffs.

Finally, the proof uses techniques introduced by Newton and Sawa (2015) for learning in matching games. They show that in the class of games that they consider a minimum cost path always exists from any state to a state that is most robust to one-shot deviations. In the prisoner's dilemma this is not always the case, so the results do not apply directly. However we use the same idea to construct a path to a certain "central" state, which narrows down the possible minimal spanning trees and ultimately leads to a characterization.

The rest of the paper is organized as follows. We begin by introducing the game and learning rules in the next section, then we characterize the stable (absorbing) sets of states of the unperturbed process without experimentation, refine them to

stochastically stable states of the process with experimentation, and finally apply the results to the prisoner's dilemma game under two common learning rules. We conclude by discussing possible extensions and comparing the results to other learning rules.

2. Preliminaries

Let $\pi(a, x)$ denote the payoff of playing a when the opponent plays x. The two possible actions of each player comprise the set $A = \{C, N\}$. The four possible payoff values are $\pi_{CC}, \pi_{CN}, \pi_{NC}, \pi_{NN}$ with $\pi_{NC} > \pi_{CC} > \pi_{NN} > \pi_{CN}$. Under the last condition, the game is a Prisoner dilemma, C stands for cooperative action and N – for non-cooperative or Nash action.

Every state g can be identified with a pair of Q-vectors, $g = (Q_1, Q_2)$, each Q-vector in turn being a pair of two Q-values, i.e. $Q_i = (Q_i^N, Q_i^C)$ for both $i \in \{1, 2\}$. The set of all possible states, i.e. pairs of valid Q-vectors will be denoted $\mathfrak{G}$.

In order to stay true to practical implementations of reinforcement learning and to avoid unnecessary continuity arguments while staying formal, we assume $\mathfrak{G}$ to be a fine grid with $\epsilon > 0$ between consecutive Q-values, i.e. $\mathfrak{G} \subseteq \{(Q_1, Q_2) : Q_i \in \mathfrak{D}\}$, where $\mathfrak{D} = \{q = z\epsilon \text{ for some } z \in Z\}$, Z denoting some compact subset of $\mathbb{Z}$. Naturally $\pi_{CC}, \pi_{CN}, \pi_{NC}, \pi_{NN} \in \mathfrak{D}$. Whenever the Q-value does not conform to this grid, we assume that it is rounded to the closest grid point (this will be formalized below). This specification represents machine precision – a computer running a reinforcement learning algorithm would eventually reach the limit for the machine representation of a decimal number. All paths through the state space in our proofs are therefore finite, but this formulation will require additional steps to capture behavior at the boundary.

2.1. Unperturbed dynamics.

The unperturbed dynamics is denoted P_0. It corresponds to some reinforcement learning rule (e.g. Q-learning) *without* experimentation.

Cast in terms of a stochastic Markov process, it means that players always choose the action with the higher Q, obtain $\pi_{t,i}$, and then each update the Q-vector as follows:

$$\begin{array}{l} \text{for } a = a_t : Q^a_{t+1,i} \in (Q^a_{t,i}, \pi_{t,i}] \cap \mathfrak{D} \text{ if } Q^a_{t,i} \neq \pi_{t,i} \text{ and } Q^a_{t+1,i} = \pi_{t,i} \text{ otherwise,} \\ \text{for } a \neq a_t : Q^a_{t+1,i} = Q^a_{t,i}. \end{array} \tag{1}$$

In other words, it does not matter how the Q-values are updated, 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. If the player continues to obtain the same payoff π, we assume that she approaches this value in the limit of some convergent sequence, i.e. $\lim_t Q^a_t = \pi$.

The process that is usually called Q-learning is a particular kind of such updating rule when the speed of updating is captured by a single parameter α that is independent of the current Q-value:

$$Q^a_{t+1,i} = \begin{cases} \max \arg\min_{z \in Z} |z\epsilon - \left((1-\alpha_i)Q^a_{t,i} + \alpha_i \pi_{t,i}\right)| & \text{if } a = a_t \\ Q^a_{t,i} & \text{otherwise,} \end{cases} \tag{2}$$

where $1 \geq \alpha_i > 0$ is the learning parameter for player i. Note that this parameter can be different between players, and that we map the process to finite grid by taking the closest value in $\mathfrak{D}$ to $(1-\alpha_i)Q^a_{t,i} + \alpha_i \pi_{t,i}$. The main results are for the general setup in (1), and we will later discuss (2) as an illustration.

These updates thus move the process to the new state $g' = (Q_{t+1,1}, Q_{t+1,2})$. For convenience we will introduce the functions $\mathcal{F}^{a_1,a_2}(\cdot)$, which for any state $g = (Q_{t,1}, Q_{t,2})$ return the new state g' that results from updating the previous values $Q_{t,i}$ for the two players after playing some action profile (a_1, a_2) once.

We will refer to the actions with the higher Q as the action profile "played on path", i.e. the actions in $\{a : Q^a_{t,i} = \max_b Q^b_{t,i}\}$ for each player i. If this set is not a singleton, we further assume that that player randomizes over its full support, i.e. all actions with the maximum Q-value have some positive probability to be taken.

In terms of the unperturbed dynamics we will be interested in the set of stable states. The set of such states is denoted $\mathfrak{C}$, and it is a set of all states that are reentered with probability 1 in the unperturbed process, i.e. $\mathfrak{C} = \{g \in \mathfrak{G} : P^t_0(g, g) = 1\}$ for some t.

2.2. **Perturbed dynamics.** Let $\{P_\beta\}_{\beta\in(0,\bar{\beta})}$ be the family of perturbed dynamics indexed by the experimentation parameter β. In particular, $P_\beta(g, g')$ denotes the transition probability from state g to state g'. It is assumed to satisfy the following conditions expanded from the list in Newton and Sawa (2015):

Assumption 1. *(Conditions on the perturbed dynamic).*

(i) $P_\beta \overset{\beta\to 0}{\longrightarrow} P_0$, *where* P_0 *are the transition probabilities for some unperturbed dynamic as described above.*

(ii) *For* $\beta > 0$, *the chain induced by* P_β *is irreducible.*

(iii) P_β *vary continuously in* β.

(iv) *If, for* $g \neq g', P_0(g, g') = 0, P_{\hat{\beta}}(g, g') > 0$ *for some* $\hat{\beta} > 0$, *then* $\lim_{\beta\to 0} -\beta \log P_\beta(g, g') = c$ *for some* $c > 0$.

(v) *For any* $\beta \geq 0, P_\beta(g, g') > 0$ *implies* $g' = \mathcal{F}^{a_1,a_2}(g)$ *for some* $a_1, a_2 \in A$.

(vi) *For any* $\beta > 0$, $P_\beta(g, g') > P_\beta(g, \hat{g})$, *for any* g *with* (a_1, a_2) *played on path and any* $g' = \mathcal{F}^{a_1,b_2}(g)$ *or* $g' = \mathcal{F}^{b_1,a_2}(g)$, *and* $\hat{g} = \mathcal{F}^{b_1,b_2}(g)$, *where* $b_1 \neq a_1$, *and* $b_2 \neq a_2$.

(vii) *For any* $\beta > 0$ *and states* $g = (Q_1, Q_2)$, $g' = (Q'_1, Q'_2)$, *such that* (a_1, a_2) *is played on path in both* g *and* g', *if* $Q^{a_i}_i \geq Q'^{a_i}_i$ *and* $Q^{b_i}_i < Q'^{b_i}_i$ *for some* $i \in \{1, 2\}$,

where $b_1 \neq a_1$, $b_2 \neq a_2$, *then* $P_\beta(g, \hat{g}) \leq P_\beta(g', \hat{g}')$ *for* $\hat{g} = \mathcal{F}^{b_i, a_{-i}}(g)$ *and* $\hat{g}' = \mathcal{F}^{b_i, a_{-i}}(g')$.

The first four conditions are borrowed directly from Newton and Sawa (2015). They connect perturbed and unperturbed processes and restrict the perturbed process to be "weakly regular" (Sandholm, 2010).

Condition (v) states that every transition is a valid Q-learning update, possibly an update on the profile that resulted from experimentation. Note that while the dynamics are parametrized by a single variable β, this condition admits different experimentation rules for players with different probabilities of experimentation or different processes altogether as long as the probability of experimentation decreases in β for all players and the other conditions are satisfied.

The remaining two conditions impose mild restrictions stemming from the interpretation of the Q-vector as an imperfect estimate of the value function. In general, if the two players are experimenting independently of each other and the state with probability that is lower for the actions with Q-values that are further from the on-path payoff, then both of the remaining conditions are satisfied. Condition (vi) requires that only one player experimenting be more likely than two players experimenting simultaneously, keeping their actions fixed. Importantly however, this does not imply that a two-player experimentation for some state can not be less costly than a single-player experimentation from another state. In particular, this is not true if the former state has no single-player experimentation that eventually leads to another stable state. Condition (vii) states that if for some player the Q-value for the action on path is the same or lower and the action off-path is higher in one of the two states, then she is at least as likely to experiment in this state as in the other one. In particular, (vii) is true if the probability of experimentation is increasing in the Q-value of the corresponding action.

Overall, these conditions are quite permissive, and the logit choice 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, etc. The results do not depend on the choice of perturbations as long as they satisfy these regularity assumptions.

Let $\mathfrak{G}^{PD}$ be the set of states, s.t. $Q_i^a \in [\min_X \pi(a_1, a_2), \max_{a_2}(\pi(a_1, a_2))]$ for any $a \in A$ and both $i \in \{1, 2\}$.[2] The initial seed for the process has to be chosen from $\mathfrak{G}^{PD}$ since the other states are not reachable from within the set. This follows by regularity conditions in Assumption 1 because any state g_2 reachable from g_1 has to be a valid update, i.e. $g_2 = \mathcal{F}^{a_1,a_2}(g_1)$ for some profile (a_1, a_2). However, from the definition of $\mathcal{F}$ in (1) it follows that there are no states $g_1 \in \mathfrak{G}^{PD}$ and $g_2 \notin \mathfrak{G}^{PD}$ such that $\mathcal{F}^{a_1,a_2}(g_2) = g_1$ for any $a_1, a_2 \in A$. Therefore there can be no path from any state in the set $\mathfrak{G}^{PD}$ to any state not in the set $\mathfrak{G}^{PD}$ and $P_\beta(g, g') = 0$ for any β.

While a common restriction for Q-learning simulations, it is not always vacuous. Consider a stable state where $Q_i^C = \pi_{CC}$ and $Q_i^N < \pi_{NC}$ for both $i \in \{1, 2\}$. Any experimentation by either player i from this state will bring Q_i^N closer to π_{NC}, but the process will never reach $\mathfrak{G}^{PD}$ unless either the state space is a finite grid up to machine precision or the learning parameter is one. And, on the other hand, if either of these conditions is true, including the actual computer implementation of the algorithm, the restriction is irrelevant as the finite learning process will eventually reach $\mathfrak{G}^C$. Thus the designer does not need to have prior knowledge of the game (payoffs).

As we mentioned in the introduction, the proof relies on some machinery of the "one-shot deviation principle" introduced in Newton and Sawa (2015) for matching games and uses the spanning trees approach from Young (1993). The definitions below are taken from these papers.

[2]Note that this is the set where *all* actions are within bounds unlike the region in Lemma 1 below that is only referring to the actions on path.

Definition 1. *The 1-step cost of the process moving from g to g' is defined as:*

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

adopting the convention that $-\log 0=\infty$.

The 1-step cost $c\left(g,g'\right)$ is the exponential decay rate of the transition probability from g to g'. The rarer a transition, the higher its cost. Impossible transitions have infinite cost. Note that for $g\notin\mathfrak{C}$, there is a zero cost transition from g. This is because there is some $g'\neq g$, such that $P_{\beta}\left(g,g'\right)$ does not approach zero as $\beta\to 0$. We are also interested in the overall cost of moving between g and g', even if many steps are required. Let the t-step transition probabilities be given by $P_{\beta}^{t}\left(g,g'\right)\equiv P\left(g^{t}=g'\mid g^{0}=g,P_{\beta}(.,.)\right)$

Definition 2. *The overall cost of the process moving from g to g' is defined as:*

$$C\left(g,g'\right):=\min_{t\in\mathbb{N}}\lim_{\beta\to 0}-\beta\log P_{\beta}^{t}\left(g,g'\right)$$

A spanning tree rooted at $\hat{g}\in\mathfrak{C}$ is a directed graph over the set $\mathfrak{C}$ such that every $g\in\mathfrak{C}$ other than $\hat{g}$ has exactly one exiting edge, and the graph has no cycles (implying that $\hat{g}$ has no exiting edges). The cost of a spanning tree is the sum of the costs of its edges given by $C(.,.)$. 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 $\hat{g}\in\mathfrak{C}$ is stochastically stable only if there exists a minimum cost spanning tree rooted at $\hat{g}$. We will use $cost(\hat{g})$ to denote the cost of a minimal spanning tree among all trees rooted in $\hat{g}$.

We call a transition $g\to g'$ from $g\in\mathfrak{G}$ the *least cost transition* from g if it has the lowest cost of all possible 1-step transitions from g. This is either the regular update of Q-values after the on-path action profile is played or an update after the most likely experimentation.

Definition 3. *Denote the set of possible least cost transitions from* $g \in \mathfrak{G}$ *by:*

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

$c_{\mathrm{L}}(g)$ *will be used to denote the cost of the least cost transition from* g*.*

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

Define OS, the set of states which are most robust to one-shot deviation:

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

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

3. General results

3.1. **Recurrent classes (unperturbed process).** As usual for the study of stochastic stability, the first step is based on the fact that the stochastically stable states belong to the recurrent classes (absorbing states or repeating sequences of states) of the unperturbed process (Young, 1993).

We will show that the process is always absorbed by a single state and can not get "stuck" in a recurring 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 $g = (Q_1, Q_2) \in G$, s.t. $Q_i^a \geq Q_i^b$, for any $b \in A \setminus a$.

We first show that the Q-values are bounded by the lowest and highest payoffs that can happen on path. We can also get the same results by considering the limiting distribution.

Lemma 1. *For any recurring class G, any state $\hat{g} \in G$, $\hat{g} = (\hat{Q}_1, Q_2)$, and any action a_i that is played by i in this or any other state in G, $\max_{a_{-i} \in \mathcal{A}_{-i}(G)}(\pi(a_i, a_{-i})) \geq \hat{Q}_i^{a_i} \geq \min_{a_{-i} \in \mathcal{A}_{-i}(G)}(\pi(a_i, a_{-i}))$.*

Proof. By construction in (1),

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

for both players and any action a_i played on path.

From any state $g = (Q_1, Q_2) \in G$, the process transitions to a new state $g' = \mathcal{F}^{a_1,a_2}(g) = (Q'_{t,1}, Q'_{t,2})$, s.t.:

(i) if a_i is played and

- $\max_{a_{-i} \in \mathcal{A}_{-i}(G)}(\pi(a_i, a_{-i})) < Q_i^{a_i}$ then
  $Q_i'^{a} - \max_{a_{-i} \in \mathcal{A}_{-i}(G)}(\pi(a_i, a_{-i})) < Q_i^{a_i} - \max_{a_{-i} \in \mathcal{A}_{-i}(G)}(\pi(a_i, a_{-i}))$,
- $\min_{a_{-i} \in \mathcal{A}_{-i}(G)}(\pi(a_i, a_{-i})) > Q_i^{a_i}$ then
  $\min_{a_{-i} \in \mathcal{A}_{-i}(G)}(\pi(a_i, a_{-i})) - Q_i'^{a_i} < \min_{a_{-i} \in \mathcal{A}_{-i}(G)}(\pi(a_i, a_{-i})) - Q_i^{a_i}$,
- otherwise $Q_i'^{a_i} \in [\min_{a_{-i} \in \mathcal{A}_{-i}(G)}(\pi(a_i, a_{-i})), \max_{a_{-i} \in \mathcal{A}_{-i}(G)}(\pi(a_i, a_{-i}))]$,

(ii) if a_i is not played then $Q_i'^{a_i} = Q_i^{a_i}$.

From this we know that any state $\hat{g}$ for which $\hat{Q}_i^{a_i} > \max_{a_{-i} \in \mathcal{A}_{-i}(G)}(\pi(a_i, a_{-i}))$ or $Q_i^{a_i} < \min_{a_{-i} \in \mathcal{A}_{-i}(G)}(\pi(a_i, a_{-i}))$ is transient and therefore $\hat{g} \notin G$. □

We can now characterize the absorbing states.

Lemma 2. *A state $g = (Q_1, Q_2)$ is absorbing in the unperturbed process, i.e. $g \in \mathfrak{C}$ if and only if:*

(1) $Q_1^{a_1} = \pi_{a_1 a_2} > Q_1^{b_1}$, and

(2) $Q_2^{a_2} = \pi_{a_2 a_1} > Q_2^{b_2}$

for some $a_1, a_2, b_1, b_2 \in A, a_1 \neq b_1, a_2 \neq b_2$. Moreover, these states are the only recurrent classes, i.e. there are no recurrent classes that are not singletons.

Proof. We start with the "if" part. Each of the players 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 $\pi_{a_1a_2}$ and $\pi_{a_2a_1}$, the new Q-vectors are unchanged, $\mathcal{F}^{a_1,a_2}(g) = g$. Thus the process stays in g.

For the "only if" part suppose there is a recurring class with possibly more than one state 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 $\max_{x\in\mathcal{A}(G)_{-i}}(\pi(a_i, x)) = \pi(a_i, a_{-i}) \geq Q_i^{a_i} \geq \min_{x\in\mathcal{A}(G)}(\pi(a_i, x)) = \pi(a_i, a_{-i})$ and thus $Q_i^{a_i} = \pi(a_i, a_{-i})$ and similarly for the other player. Moreover, if $Q_i^{a_i} \leq Q_i^{b_i}$ 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, and in particular there are no other singleton absorbing states.

It remains to show that there are no non-singleton recurrent classes. Note first that from any state where $\pi(a_i, a_{-i}) > Q_i^{b_i}$ for each $i \in \{1, 2\}$, $b_i \in A \setminus a_i$, and any (a_1, a_2) played on path, i.e. $Q_i^{b_i} < Q_i^{a_i}$ for each $i \in \{1, 2\}$, and $b_i \in A \setminus a_i$, the process is eventually absorbed into a singleton absorbing state. This is because in any state that follows (a_i, a_{-i}) is played again and $|Q_i^{a_i} - \pi(a_i, a_{-i})| > |\hat{Q}_i^{a_i} - \pi(a_i, a_{-i})|$ with strict inequality (unless already $Q_i^{a_i} = \pi(a_i, a_{-i})$) by construction in (1). Thus either the new state is absorbing or, again, $\pi(a_i, a_{-i}) > Q_i^{a_i}$ and $Q_i^{a_i} > Q_i^{b_i}$ for both $i \in \{1, 2\}$ and any $b_i \in A \setminus a_i$. Since $|Q_i^{a_i} - \pi(a_i, a_{-i})|$ is bounded by 0 the process either eventually ends.

Suppose now that (C, C) is played in some state $g \in G$. Then $Q_i^C \geq Q_i^N \geq \pi_{CC}$ for at least one of $i \in \{1, 2\}$, the first inequality is because (C, C) is played, and the second because otherwise by the above remark the process is absorbed into a singleton state. But by Lemma 1 $Q_i^C \leq \pi_{CC}$. Moreover the inequality is strict – the

equality only holds if the same maximual payoff was obtained by i in the previous period, which, since all payoffs are distinct, is only possible if the same profile was played in the previous period by construction in (1). Therefore since at least some other profile is also recurring, $Q_i^C < \pi_{CC}$. This is a contradiction and therefore (C, C) is not played in G.

Then by Lemma 1, $Q_i^C = \pi_{CN}$ because in all profiles where i plays C the opponent plays N. At the same time since $Q_i^N \geq \pi_{NN} > \pi_{CN} = Q_i^C$, C can not be played in any state in G. Thus only (N, N) can be played on path, which is again a contradiction. □

3.2. **Stochastically stable states (perturbed process).** Using this characterization, we can now refine the absorbing states to stochastically stable states.

The following two lemmas will allow us to 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 not true in our case, we 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, we can still say that the minimal trees are of a particular form. The next lemma says that if there is such state g^c, then every minimal spanning tree rooted in some state $\hat{g}$ can only have non-minimum-cost arcs from the states on the path from g^c to the root $\hat{g}$.

Lemma 3. *If for any $g \in \mathfrak{C}$ 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 $l \in \{1, ...L-1\}$ then in any minimal spanning tree rooted in $\hat{g}$ for any $g' \in \mathfrak{C}$, either the outgoing arc from g' has the cost $c_L(g')$ or there is a path from g^c to g'.*

Proof. Suppose to the contrary that there is a state $g' \in \mathfrak{C}$ with the cost of the outgoing arc greater than $c_L(g')$ and there is no path from g^c to g'. Since the graph is a spanning tree, there is then a path from g' to g^c and it is not minimal. Replacing this arc with a minimum cost arc then yields a tree with a lower cost. □

If $g^c \in OS$ then by the result in Newton and Sawa (2015), $SS = OS$. However the lemma also allows for the case when $g \notin OS$, which is what we will be needing for the next proposition.

Proposition 1. *The minimal trees are rooted in states that minimize*

$$cost(\hat{g}) = \begin{cases} cost(g^c) - c_L(\hat{g}) + C(g^c, \hat{g}) & \text{if } \hat{g} \neq g^c, \\ cost(g^c) & \text{if } \hat{g} = g^c \end{cases}$$

among all possible roots $\hat{g} \in \mathfrak{C}$.

Proof. By Lemma 3 any minimal spanning tree has all minimal cost outgoing arcs except for the path between g^c and $\hat{g}$. The difference with the minimal tree rooted in $\hat{g}$ is then the cost of this path and the least cost transition from g^c. The cost of the tree rooted in $\hat{g}$ is then $cost(g^c) - c_L(\hat{g}) + C(g^c, \hat{g})$. □

4. Prisoner's Dilemma

We will introduce two states $g^*, g^{**} \in \mathfrak{C}$. Depending on the choice rule, one of these two states will be shown to be the unique state in SS. The first state $g^* = (Q^*_{t,1}, Q^*_{t,2})$ has defection on path with $Q^{*N}_i = \pi_{NN}, Q^{*C}_i = \pi_{CN}$ for both $i \in \{1, 2\}$. The second state $g^{**} = (Q^{**C}_{t,1}, Q^{**C}_{t,2})$ has cooperation on path with $Q^{**N}_i = \pi_{NN}, Q^{**C}_i = \pi_{CC}$ for both $i \in \{1, 2\}$.

To be able to use Proposition 1, we will need to show that indeed a variant of a "getting closer" lemma holds, but instead of approaching the OS set, the process will be approaching the stable state g^* with Nash equilibrium actions on path. In other words, $g^c = g^*$ in the Lemma 3.

Before stating the lemma we formalize what "closer" will mean. We define the distance $D(g, g')$ as follows:

$$D(g, g') = \sum_{i\in\{1,2\},A\in\{C,N\}} |Q_i^a - \bar{Q}_i^a(g)|, \tag{3}$$

where the Q-vectors $\bar{Q}(g)$ are constructed based on what is played on path in g as follows. If a_i is played on path in g by each player $i \in \{1, 2\}$, then $\bar{Q}_i^{a_i}(g) = \pi(a_i, a_{-i})$ and $\bar{Q}_i^{b_i}(g) = \pi(b_i, a_{-i})$ for $b_i \neq a_i$ for $i \in \{1, 2\}$. That is, the $\bar{Q}$-values for the "on path" action profile equal the respective payoffs for the two players, i.e. $\pi(a_1, a_2)$ and $\pi(a_2, a_1)$, and the values for the "off-path" actions are the payoffs of the action profile resulting from a single-player experimentation, i.e. $\pi(b_1, a_2)$, and $\pi(b_2, a_1)$[3], and therefore single player experimentation does not change the Q-vector.

We also introduce $0 \geq m(g, g') \geq 2$ as the number of actions that differ on path in g and g'. We will use this for the situations when exactly one player cooperates.

Finally we introduce the values $\bar{d}(g)$ and $\underline{d}(g)$, the probabilities of experimentation for the player who is more likely to experiment and for the player who is less likely to experiment.

$$\bar{d}(g) = \max_{i\in\{1,2\}} P_{\hat{\beta}}(g, g^i) \qquad \underline{d}(g) = \min_{i\in\{1,2\}} P_{\hat{\beta}}(g, g^i), \tag{4}$$

where g^i is the state after i experiments from g and $\hat{\beta}$ is any strictly positive value. These values will be used for states where both players cooperate.

[3]We will only use the values of N for states in $\mathfrak{C}$, so we need only consider singletons played on path.

We will say that a state g_2 is "closer" to g^* than g_1, written $g_2 \prec g_1$, if:

$$\begin{cases} m(g^*, g_2) < m(g^*, g_1) \\ m(g^*, g_2) = m(g^*, g_1) = 1 \text{ and } D(g^*, g_2) < D(g^*, g_1) \\ m(g^*, g_2) = m(g^*, g_1) = 2 \text{ and } \underline{d}(g_2) < \underline{d}(g_1) \\ m(g^*, g_2) = m(g^*, g_1) = 2 \text{ and } \underline{d}(g_2) = \underline{d}(g_1) \text{ and } \bar{d}(g_2) > \bar{d}(g_1) \end{cases}$$

That is, $m(\cdot, \cdot)$ is lexicographically more important than $D(\cdot, \cdot)$ when at least one player defects on path, and decreasing $\underline{d}$ is more important than increasing $\bar{d}$ when both players cooperate.

The next lemma uses the fact that experimentation by two players is less likely than experimentation by one player (Assumption 1, vi) to show that a single-player experimentation from any state, possibly followed by zero-cost deviations, will get the process closer to the state g^* from which only a two-player experimentation can lead to a new state.

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 $P_0^t(g_1, g') > 0$.*

Proof. Let $g = (Q_1, Q_2)$, $g' = (Q_1', Q_2')$, $g_1 = (Q_{1,1}, Q_{1,2})$, $g_2 = (Q_{2,1}, Q_{2,2})$ and so on.

Take any state $g \in \mathfrak{C} \setminus g^*$ with (a_1, a_2) played on path and $b_i \neq a_i$ for $i \in \{1, 2\}$.

Neither (N, C) nor (C, N) can be played on path in $g \in \mathfrak{C}$ because then $Q_i^N < Q_i^C = \pi_{CN} = \min_x \pi(C, x)$ for one of the players and $g \notin \mathfrak{G}$.

We proceed by cases.

(1) Suppose (N, N) is played on path. Then $Q_i^C \neq \pi(C, N)$ for one of the players $i \in \{1, 2\}$ 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 Lemma 2 since

$g \in \mathfrak{C}$, we have $Q_i^N = Q_{-i}^N = \pi_{NN}$ and $Q_i^N > Q_i^C$, $Q_{-i}^N > Q_{-i}^C$. The player i then obtains $\pi_{CN} < \pi_{NN}$ in g_1 and $Q_{1,i}^C < Q_{1,i}^N = \pi_{NN}$. The player $-i$ obtains $\pi_{NC} > \pi_{NN}$ in g_1 and since $Q_{-i}^N = \pi_{NN}$, $Q_{1,-i}^N > Q_{-i}^N$. Therefore in g_1 again $Q_i^N > Q_i^C$, $Q_{-i}^N > Q_{-i}^C$ and (N, N) is played again. Then with positive probability and zero cost in the states $g_2, g_3, ...g'$ that follow g_1 then $Q_{2,i}^N = Q_{3,i}^N ... = \pi_{NN}$ continues to hold and $|Q_{2,-i}^N - \pi_{NN}|, |Q_{3,-i}^N - \pi_{NN}|, ...$ decreases until $Q_{-i}'^N = \pi_{NN}$ again for some g'. Thus in the new absorbing state $Q_i'^N = Q_{-i}'^N = \pi_{NN}$, $Q_{-i}'^C = Q_{-i}^C$, but $|Q_i^{*C} - Q_i'^C| = |\pi_{CN} - Q_i'^C| < |\pi_{CN} - Q_i^C|$, i.e. the distance has decreased and thus $D(g^*, g') < D(g^*, g)$, while $m(g^*, g') = m(g^*, g) = 2$ so $g' \prec g$ as required.

(2) Suppose (C, C) is played on path. Then experimentation leads i to play N. By Lemma 2 since $g \in \mathfrak{C}$, we have $Q_i^C = Q_{-i}^C = \pi_{CC}$ and $Q_i^C > Q_i^N$, $Q_{-i}^C > Q_{-i}^N$. The player i then obtains $\pi_{NC} > \pi_{CC}$ in g_1 and thus $Q_{1,i}^N > Q_i^N$. The player $-i$ then obtains $\pi_{CN} < \pi_{CC}$ in g_1 and thus $Q_{1,-i}^C < Q_{-i}^C$. If the process converges to a state where at least one player defects, then $2 = m(g^*, g') > m(g^*, g) = 1$, and we have $g' \prec g$ as required. So suppose instead that eventually a stable state with (C, C) on path is reached. Two subcases are possible.

(a) Both $Q_{1,i}^C > Q_{1,i}^N$, $Q_{1,-i}^C > Q_{1,-i}^N$ continue to hold in g_1 and (C, C) is played again. Then with positive probability and zero cost in the states $g_2, g_3, ...g'$ that follow $Q_{2i}^C = Q_{3i}^C = ... = \pi_{CC}$ continues to hold and $|Q_{2,-i}^C - \pi_{CC}|, |Q_{3,-i}^C - \pi_{CC}|, ...$, decreases until $Q_{-i}'^C = \pi_{CC}$ again for some g'. Thus in the new absorbing state $Q_i'^C = \pi_{CC}$, $Q_{-i}'^C = \pi_{CC}$, $Q_{-i}'^N = Q_{-i}^N$, but $Q_i'^N > Q_i^N$. Moreover, since i was the experimenting player in g, player i is also at least as likely to experiment in g' as the other player by condition (vii) in Assumption 1. Then $\underline{d}(g) = \underline{d}(g')$ and $\bar{d}(g) < \bar{d}(g')$ and $g' \prec g$ as required.

(b) In all remaining cases eventually $\hat{Q}^C_{-i} < Q^N_{-i}$ and will not increase unless (C,C) is played again. If $Q^C_{1,-i} < Q^N_{1,-i}$ and (C,N), or (N,N) is played next, this is true immediately in g^1. If instead $Q^C_{1,-i} > Q^N_{1,-i}$ continues to hold, but $Q^C_{1,i} < Q^N_{1,i}$, then (N,C) is played. The payoff of player i is the highest possible, and the payoff of the other player is the lowest possible, so the Q-values of their actions increase and decrease respectively. Then eventually (N,N) is played and again $\hat{Q}^C_{-i} < Q^N_{-i}$ in some $\hat{g}$ and will not increase unless (C,C) is played again. Once (C,C) is played again, the game continues with (C,C) until convergence to a stable state. We also know that eventually (C,C) is reached and thus at some future state $Q^N_{-i} > \hat{Q}^C_{-i} > Q'^N_{-i}$. Moreover since i experimented in g, one of the following must occur. Either $-i$ is also less likely to experiment than i in g' and then again by condition (vii) in Assumption 1 and because $Q^N_{-i} > Q'^N_{-i}$, $\underline{d}(g) > \underline{d}(g')$, or i is at least as likely to experiment as $-i$ in g' but $P_{-i}(g) > P_{-i}(g') \geq P_i(g')$ and again $\underline{d}(g) > \underline{d}(g')$. Thus $g' \prec g$ as required.

□

Since states with (C,N), (N,C) or mixed actions on path can not be stable in $\mathfrak{C} \subseteq \mathfrak{G}^{PD}$, we need only consider trees rooted in states with (N,N) and (C,C) on path.

At the same time, a tree rooted in some state $\hat{g} \neq g^*$ with (N,N) on path can not be minimal. We can formulate this as a corollary of Proposition 1:

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

Proof. By Proposition 1 for any minimal spanning tree $cost(g^*) - c_L(\hat{g}) + C(g^*, \hat{g})$, where $\hat{g}$ is the root. By definition, $C(g^*, \hat{g}) \geq c_L(g^*)$ and therefore $c_L(\hat{g}) < c_L(g^*) \leq$

$C(g^*, \hat{g})$ would imply that $cost(g^*) - c_L(\hat{g}) + C(g^*, \hat{g}) > cost(g^*)$. Then the minimal tree rooted in g^* has smaller cost, which is a contradiction. □

With (N, N) on path the cost of leaving any state $\hat{g} \neq g^*$ with (N, N) on path is at the most $c_L(\hat{g}) = \pi_{NN} - \pi_{CN}$, while the minimal cost of leaving g^* is strictly greater than $\pi_{CN} - \pi_{NN}$ by regularity conditions in Assumption 1 (vi) because both players have to experiment simultaneously. Therefore the cost of a minimal tree rooted at g^* is strictly lower.

This leaves only the state g^* and states in $\mathfrak{C}$ with (C, C) on path as candidates for stochastic stability. We can further refine the possibilities by the following corollary:

Corollary 2. *If a minimal 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 minimal cost spanning tree rooted in g^{**}.*

Proof. From any state with (N, C) or (C, N) on path there is a zero-cost path to a state in (N, N). Therefore in any minimum cost path to g^{**} any state with (C, C) has to follow another state with (C, C) or a state with (N, N). Take the first state with (C, C) then there is a state with (N, N) before it. But then in this state $Q_i^N = \pi_{NN}$ for both $i \in \{1, 2\}$ and this state is therefore g^{**}. So any minimum cost path from g^* to a state with (C, C) passes through g^{**} and has the cost no less than $C(g^*, g^{**})$. At the same time, $c_L(\hat{g}) \leq c_L(g^{**})$ for any state $\hat{g}$ with (C, C) by regularity conditions in Assumption 1 (vii) because $\hat{Q}_i^N \geq \pi_{NN} = Q_i^{**N}$ and $\hat{Q}_i^C = Q_i^{**C} = \pi_{CC}$ for $i \in \{1, 2\}$.

Then $cost(g^c) - c_L(\hat{g}) + C(g^c, \hat{g}) \geq cost(g^c) - c_L(g^{**}) + C(g^c, g^{**})$ and by Proposition 1 the result follows. □

Thus if we are interested in action profiles on path in the limit, we need only consider g^* and g^{**}. This directly leads to a characterization:

Corollary 3. *(i) If $C(g^*, g^{**}) < c_L(g^{**})$ then $g^{**} \in SS$ and players always converge to cooperation in any state in SS.* [4]

(ii) If $C(g^*, g^{**}) > c_L(g^{**})$ then $SS = \{g^*\}$ and players always converge to defection.

(iii) If $C(g^*, g^{**}) = c_L(g^{**})$ for one or more states, both defection and cooperation may occur in the limit with positive probability.

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

We can then apply the concepts to particular experimentation rules: ϵ-greedy (which we call β-greedy in our notation) and logit (also called softmax or Boltzmann) rules.

Under the β-greedy rule with probability $(1-k_i\beta)$ player i chooses the actions with highest Q-values (with ties resolved uniformly), and with probability $k_i\beta$ the action is chosen by randomizing uniformly. Formally in our definitions, $Pr_i^{greedy}(a|g) = \frac{1}{|\{a:Q_i^a=\max_{a'} Q_i^{a'}\}|}(1-k_i\beta)+\frac{1}{2}k_i\beta$ if a is played by i on path in g and $Pr_i^{greedy}(a|g) = \frac{1}{2}k_i\beta$ otherwise. The denominator in the former case only divides the amount among all actions played on path if there are more than one. The chosen experimentation function $\beta_i(\beta) = (k_i\beta)$ ensures that the probability of experimentation is decreasing in β for any choice of the constants $k_i > 0$ as required by condition (i) in Assumption 1. For this rule, the starting value of β has to be taken so that $k_i\beta$ is less than 1 to ensure that the resulting probability of experimentation is well-defined. If $k_1 = k_2 = 1$, we obtain the simpler case with symmetric experimentation probabilities $\beta_1(\beta) = \beta_2(\beta) = \beta$.

[4]The $C(g^*, g^{**}) < c_L(g^{**})$ is a stronger requirement than $c(g^*, g^{**}) < c_L(g^{**})$ used in the Waltman and Kaymak (2008) due to high learning rate assumption. This is sufficient in Waltman and Kaymak (2008) because the cooperative state g^{**} is reachable by a single least cost transition from g^*, while in the general case the path has to go through other states, where the least cost transition does not approach g^{**} and costlier arcs need to be taken, which leads to $C(g, g^{**})$.

Under the logit choice rule instead

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

with no restriction on k_i and β as long as they are positive. The $\beta_i(\beta) = k_i\beta$ is sometimes called the temperature – with higher values of β experimentation becomes more likely and less dependent on Q-values. When β, and therefore also $\beta_i(\beta)$, approach zero, the actions with the highest Q-value are always chosen and the process approaches unperturbed dynamics P_0. For the symmetric setup in the limit with $\beta_1(\beta) = \beta_2(\beta) = \beta$ approaching infinity, the actions are chosen with equal probability.

In both cases then $P_\beta(g, g')$ is the product of these probabilities, $\prod_{i\in\{1,2\}} Pr_i^{greedy}(a_i|g)$ or $\prod_{i\in\{1,2\}} Pr_i^{logit}(a_i|g)$ for $g' = \mathcal{F}^{a_i,a_{-i}}(g)$.

For β-greedy choice rule experimentation by 2 players in any state is less likely than experimentation by 1 player in any state, which will be shown to always lead to players converging to defection.

The logit rule case on the other hand can lead to cooperation depending on the values of parameters. A commonly used property of the logit choice rule is that the transitions probabilities in the limit $\beta \to 0$ are determined by the absolute difference in payoffs between the states. Let z_i^{logit} be the smallest integer equal or greater than $\log_{(1-\alpha_i)} \frac{\pi_{NN}-\pi_{CC}}{\pi_{CN}-\pi_{CC}}$, which is the number of updates on the profile (C, C) that it takes player i to get from g^* to a state where i cooperates on path under the logit rule. The expression can be obtained by rewriting the recursive equations $Q_{t+1,i} = (1-\alpha_i)Q_{t,i} + \alpha_i\pi_{CC} \geq \pi_{NN}$, $i \in \{1, 2\}$, $Q_0 = \pi_{CN}$ for the first t where $Q_{t+1,i} \geq \pi_{NN}$. These equations describe the minimum cost path from g^* to g^{**} because the only possible updates that increase Q-values for C are the updates on the profile (C, C),

i.e. $\mathcal{F}^{a_i,a_{-i}}(g) = g'$ can only hold if $a_i = C$ for both $i \in \{1,2\}$ if g' has a higher Q-value for C than g. This is proven in the following lemma:

Lemma 5. *The path $g^* = g_1, ...g_L = g^{**}$, where (C,C) is played in every state g_l, is the lowest cost path between g^* and g^{**}.*

Proof. Let $g_l = (Q_{l,1}, Q_{l,2})$ and suppose there is a lower cost path $g^* = g'_1, ...g'_{L'} = g^{**}$, $g'_l = (Q'_{l,1}, Q'_{l,2})$ with at least one defection on path. Let $g'^{C,C}_l$ be the state (if any) on this path where (C,C) is played for the l-th time after $l-1$ (not necessarily consecutive) plays of (C,C) on this path. That is, the next state on the path after $g'^{C,C}_l$ is $\mathcal{F}^{C,C}(g'^{C,C}_l)$.

By the regularity condition vii in Assumption 1, $c(g_l, g_{l+1})$ is the lowest 1-step cost $c(g,\hat{g})$ among all pairs of states $g, \hat{g} \in \mathfrak{C}$, $g = (Q_1, Q_2)$ and $\hat{g} = \mathcal{F}^{CC}(g) = (\hat{Q}_1, \hat{Q}_2)$, such that $Q^C_i \leq Q^C_{l,i}$ for both $i \in \{1,2\}$ and (N,N) is played on path. Moreover, for any such pair of states, $\hat{Q}^C_i \leq Q^C_{l+1,i}$ by construction in (1).

At the same time, any profile where at least one player defects can not increase the Q-value of cooperation for either player, i.e. $Q'^C_{l+1,i} \leq Q'^C_{l,i}$ if $(C,N), (N,C)$, or (N,N) is played in g'_l. Then for every pair of states g_l and $g'^{C,C}_l$ on their respective paths for any $l \in 1..L$, $Q'^C_{l,i} \leq Q^C_{l,i}$ for both $i \in \{1,2\}$. Moreover, $L' > L$ as there is at least one state with defection that does not increase the Q-value of cooperation. But then the cost at every $(g'^{C,C}_l)$ is at least as high as the cost $c(g_l, g_{l+1})$ on the other path. Since there are no other states on the cooperative path $g^* = g_1, ...g_L = g^{**}$, its cost is the same or lower.

□

Further, let $q^C_{l,i} = \pi_{CC} + (1-\alpha_i)^{l-1}(\pi_{CN} - \pi_{CC})$. These are the Q-values of cooperation for each player on the minimum-cost path in Lemma 5 from g^* to g^{**} under the logit rule. Then the characterization for the two rules is as follows:

Proposition 2. *For the asymmetric Q-learners the SS set depends on the choice rule:*

(i) $SS = g^*$ *under* β*-greedy choice rule*

(ii) Under the logit choice rule:

- $SS = \{g^*\}$ *and* N *is played if*
  $\sum_{l=1}^{z_1^{logit}} \frac{1}{k_1}(\pi_{NN} - q_{l,1}^C) + \sum_{l=1}^{z_2^{logit}} \frac{1}{k_2}(\pi_{NN} - q_{l,2}^C) > \min_{i\in\{1,2\}} \frac{1}{k_i}(\pi_{CC} - \pi_{NN})$,
- $g^{**} \in SS$ *and* C *is played in all states in* SS *if*
  $\sum_{l=1}^{z_1^{logit}} \frac{1}{k_1}(\pi_{NN} - q_{l,1}^C) + \sum_{l=1}^{z_2^{logit}} \frac{1}{k_2}(\pi_{NN} - q_{l,2}^C) < \min_{i\in\{1,2\}} \frac{1}{k_i}(\pi_{CC} - \pi_{NN})$
- $g^{**}, g^* \in SS$ *if*
  $\sum_{l=1}^{z_1^{logit}} \frac{1}{k_1}(\pi_{NN} - q_{l,1}^C) + \sum_{l=1}^{z_2^{logit}} \frac{1}{k_2}(\pi_{NN} - q_{l,2}^C) = \min_{i\in\{1,2\}} \frac{1}{k_i}(\pi_{CC} - \pi_{NN})$

Proof. (i) Under β-greedy rule, a two-player simultaneous experimentation has the probability $\beta_1(\beta) \times \beta_2(\beta) = k_1 k_2 \beta^2$, while a single-player experimentation has the probability at most $\max_i \beta_i(\beta) = \max\{k_1, k_2\}\beta$. By construction, leaving g^* requires a two-player simultaneous experimentation, while a least cost transition from any state $\hat{g} \in \mathfrak{C}$ with (C, C) on path requires only a single-player experimentation. Then

$$C(g^*, \hat{g}) \geq c_1(g^*, \hat{g}) \geq c_L(g^*) > c_L(\hat{g})$$

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 $\frac{1}{k_1\beta}(\pi_{NN} - q_{l,1}^C) + \frac{1}{k_2\beta}(\pi_{NN} - q_{l,2}^C)$ while both players have N on path (both players have to experiment), i.e. for $l = 1$ to $\min_i z_i^{logit}$. For $l = \min_i z_i^{logit}$ to $\max_i z_i^{logit}$ only the player who still has N on path has to experiment. Then $C(g^*, g^{**}) = \sum_{l=1}^{z_1^{logit}} \frac{1}{k_1\beta}(\pi_{NN} - q_{l,1}^C) + \sum_{l=1}^{z_2^{logit}} \frac{1}{k_2\beta}(\pi_{NN} - q_{l,2}^C)$ and $c_L(g^{**}) = \min_{i\in\{1,2\}} \frac{1}{k_i\beta}(\pi_{CC} - \pi_{NN})$. The result then follows by applying Proposition 1.

□

In particular, the previous proposition implies that there is always a low-enough $\alpha = \min\{\alpha_1, \alpha_2\}$ for any π_{NN} such that defect-defect (the g^* state played repeatedly) is the unique action profile in the limit. In other words, one of the learners can always preclude cooperation if her learning parameter is low enough.

Corollary 4. *There is always a low-enough $\alpha > 0$ for any π_{NN} such that $\{g^*\} = SS$ under either rule.*

Proof. z_i^{logit} increases without bound as α_i approaches 1. Then $\sum_{l=1}^{z_i^{logit}} 2(\pi_{NN} - q_l^C)$ also increases without bound and by Proposition 2 $\{g^*\} = SS$. □

We can also simplify the characterization for the symmetric case where $\alpha_1 = \alpha_2 = \alpha$, $\beta_1(\beta) = \beta_2(\beta) = \beta$, and therefore $C(g^*, g^{**}) = \sum_{l=1}^{z^{logit}} 2(\pi_{NN} - q_l^C)$ and $c_L(g^{**}) = \pi_{CC} - \pi_{NN}$ with $z^{logit} = z_1^{logit} = z_2^{logit}$. Substituting, we immediately obtain the corollary:

Corollary 5. *If $\alpha_1 = \alpha_2$ and $\beta_1(\beta) = \beta_2(\beta) = \beta$ then:*

(i) $SS = g^*$ under β-greedy choice rule

(ii) Under the logit choice rule:

- *$SS = \{g^*\}$ if $\sum_{l=1}^{z^{logit}} 2(\pi_{NN} - q_l^C) > \pi_{CC} - \pi_{NN}$,*
- *$g^{**} \in SS$ and C is played in all states in SS if $\sum_{l=1}^{z^{logit}} 2(\pi_{NN} - q_l^C) < \pi_{CC} - \pi_{NN}$*
- *$g^{**}, g^* \in SS$ if $\sum_{l=1}^{z^{logit}} 2(\pi_{NN} - q_l^C) = \pi_{CC} - \pi_{NN}$*

We can 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}. Therefore we could fix $\pi_{CC} = 1$ and $\pi_{CN} = 0$ without loss of generality – the shape of the regions is preserved for other values of the three payoffs π_{CN}, π_{NC}, and π_{CC}. Instead we will equivalently use the value $\frac{\pi_{NN} - \pi_{CN}}{\pi_{CC} - \pi_{CN}}$, which captures the relative position of π_{NN} between π_{CN} and

π_{CC}. The regions are shown in the Figure 1. The boundary of the regions consists of the only values where both cooperation and defection is possible in the limit. The area to the up and left of the red dashed line is the region covered by the theoretical part of (Waltman and Kaymak, 2008).

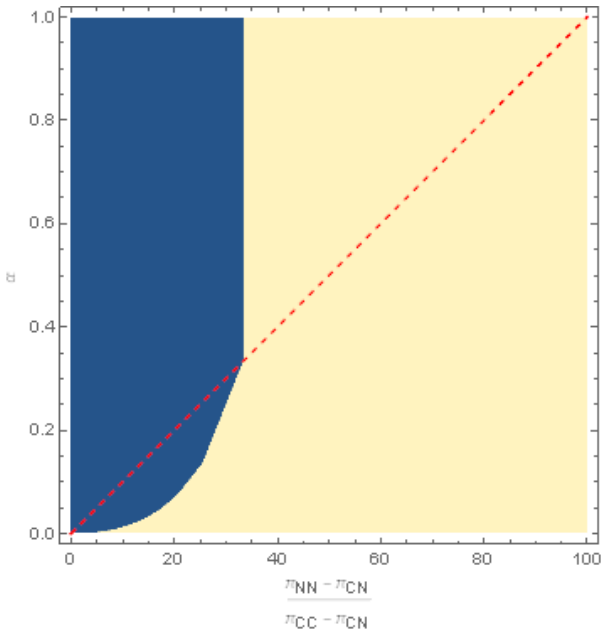

FIGURE 1. Trade-off between learning rate α and relative punishment payoff $\frac{\pi_{NN}-\pi_{CN}}{\pi_{CC}-\pi_{CN}}$ for symmetric learners. The blue region has (C, C) in the limit, the light region has (N, N) in the limit.

It is illustrative to also consider a supergame of choosing a learning algorithm against an opponent. In a supergame of choosing the parameters α_i and k_i, since the algorithms can only converge to (N, N) or (C, C) on path, low values of α_i are dominated. In other words, setting $\alpha_i = 1$ is never a bad strategy. For experimentation parameter k_i it is best for the player in the supergame to try to match the oppponent's value of k_{-i}. This can be seen by moving the cost of the player with the higher k_i to the right side in the expression in Proposition 2. In sum, it is always best to remember only the immediately previous payoff, disregarding prior history of

play, while trying to experiment about as often as the opponent. The shapes of the regions for the asymmetric case are similar and can be seen in Figure 2.

It is easy to extend the analysis 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.

5. Conclusion

The textbook approach to repeated games focuses on the existence of cooperative equilibria in terms of the discount rate δ and the three of the four payoffs of the game. Namely, the cooperative equilibrium will exist if $\delta \geq \frac{\pi_{NC}-\pi_{CC}}{\pi_{NC}-\pi_{NN}}$. One can perhaps think of this as an alternative learning concept (parametrized by δ) that relies on players understanding the repeated nature of the game. 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. 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.

Characterization of learning equilibria in this paper thus addresses two issues. With the results that differ from predictions of other learning processes, Q-learning becomes a testable theory given enough variation in payoffs – 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 – has implications for observable behavior. The potential of the reinforcement learning models is supported by previous studies such as Roth and Erev (1995) and Erev and Roth (1998), which combined simulations with experiments to show that reinforcement learning models have better predictive

and descriptive power than standard equilibrium analysis. At the same time, we only considered reinforcement learning as a long-term equilibrium selection concept, instead of focusing on the intermediate term dynamics studied in these papers, which may be more relevant for discerning learning algorithms in human behavior.

Another, more direct target of this research is the field of algorithmic pricing. Due to its simplicity, Q-learning algorithms are a natural candidate for building reinforcement learning into automated pricing systems. However even these simple algorithms have been shown empirically to be able to learn to support supracompetitive prices. We confirm these simulation results theoretically and, moreover, show that there is an optimal set of parameters that will always be chosen by a rational designer of the algorithm to maximize the chance of collusion, namely – a highest learning rate and a best attempt to match the experimentation rate of the opponent.

The most natural extension of this analysis is to expand the scope from a two-action game to a discretized Bertrand competition or a similar game. Unfortunately, not all results extend in a straightforward manner, most importantly, in a differentiated Bertrand competition a minimum cost path to a central state g^* no longer has to exist.

References

Asker, J., C. Fershtman, and A. Pakes (2021): "Artificial Intelligence and Pricing: The Impact of Algorithm Design," Tech. rep., National Bureau of Economic Research.

Blonski, M., P. Ockenfels, and G. Spagnolo (2011): "Equilibrium selection in the repeated prisoner's dilemma: Axiomatic approach and experimental evidence," *American Economic Journal: Microeconomics*, 3, 164–92.

Calvano, E., G. Calzolari, V. Denicolo, and S. Pastorello (2020): "Artificial intelligence, algorithmic pricing, and collusion," *American Economic Review*,

110, 3267–97.

Erev, I. and A. E. Roth (1998): "Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria," *American Economic Review*, 848–881.

Foster, D. and H. P. Young (2006): "Regret testing: Learning to play Nash equilibrium without knowing you have an opponent," *Theoretical Economics*, 1, 341–367.

Hart, S. and A. Mas-Colell (2003): "Uncoupled dynamics do not lead to Nash equilibrium," *American Economic Review*, 93, 1830–1836.

Klein, T. (2021): "Autonomous algorithmic collusion: Q-learning under sequential pricing," *The RAND Journal of Economics*.

Marden, J. R., H. P. Young, G. Arslan, and J. S. Shamma (2009): "Payoff-based dynamics for multiplayer weakly acyclic games," *SIAM Journal on Control and Optimization*, 48, 373–396.

Mengel, F. (2014): "Learning by (limited) forward looking players," *Journal of Economic Behavior & Organization*, 108, 59–77.

Milgrom, P. and J. Roberts (1990): "Rationalizability, learning, and equilibrium in games with strategic complementarities," *Econometrica: Journal of the Econometric Society*, 1255–1277.

Nax, H. H. (2019): "Uncoupled aspiration adaptation dynamics into the core," *German Economic Review*, 20, 243–256.

Newton, J. and R. Sawa (2015): "A one-shot deviation principle for stability in matching problems," *Journal of Economic Theory*, 157, 1–27, doi: 10.1016/j.jet.2014.11.015.

Roth, A. E. and I. Erev (1995): "Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term," *Games and Economic Behavior*, 8, 164–212.

Sandholm, W. H. (2010): *Population games and evolutionary dynamics*, MIT press.

Waltman, L. and U. Kaymak (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, 84–91.

——— (2008): "Q-learning agents in a Cournot oligopoly model," *Journal of Economic Dynamics and Control*, 32, 3275–3293.

Young, H. P. (1993): "The evolution of conventions," *Econometrica: Journal of the Econometric Society*, 57–84, doi: 10.2307/2951778.

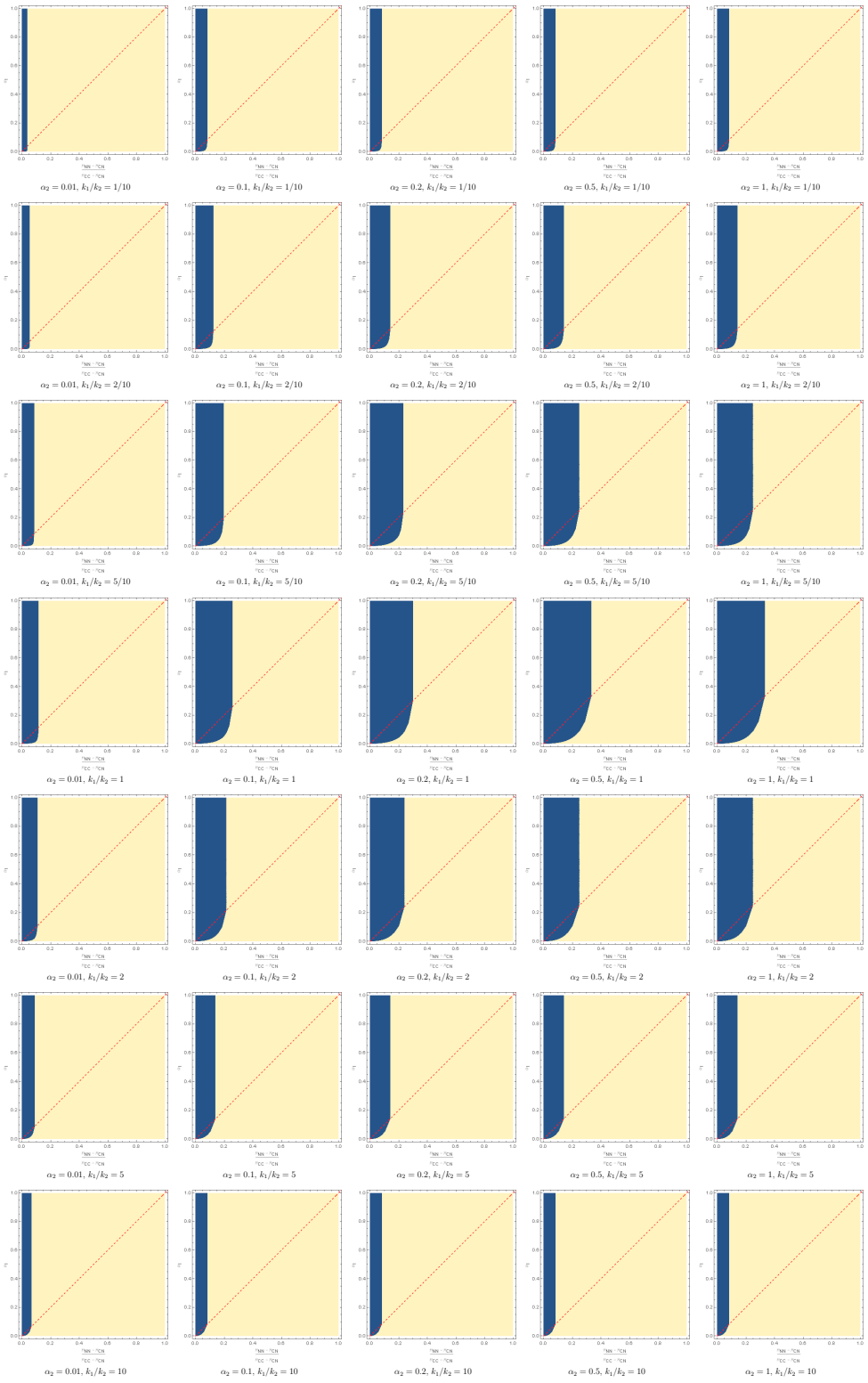

FIGURE 2. Trade-off between learning rates α_1, α_2, ratio of experimentation probabilities k_1/k_2 and relative punishment payoff $\frac{\pi_{NN}-\pi_{CN}}{\pi_{CC}-\pi_{CN}}$ for asymmetric learners.