Highlights

Reinforcement Learning in a Prisoner's Dilemma

Arthur Dolgopolov

- Reinforcement learners cooperate in a prisoner's dilemma even without memory.
- Behavior in the limit is fully characterised by the learning rate and payoffs.
- Algorithms collude frequently under high learning rate and equal experimentation.
- Any dynamic process with a central state has a simple solution.
- Reinforcement learners without noise converge in games solvable by IESDS.

Reinforcement Learning in a Prisoner's Dilemma

Arthur Dolgopolov[a,1]

[a]*Bielefeld University, Center for Mathematical Economics, Germany*

Abstract

I fully characterize the outcomes of a wide class of model-free reinforcement learning algorithms, such as Q-learning, in a prisoner's dilemma. The behavior is studied in the limit as players explore their options sufficiently and eventually stop experimenting. Whether the players learn to cooperate or defect can be determined in a closed form from the relationship between the learning rate and the payoffs of the game. The results generalize to asymmetric learners and many experimentation rules with implications for the issue of algorithmic collusion.

Keywords: Q-learning, stochastic stability, evolutionary game theory, collusion, pricing-algorithms

JEL: C72, C73, D43, D83, L41

1. Introduction

Reinforcement learning, that is, adapting behavior through exploration and exploitation, is a natural counterpart to game-theoretic models of bounded rationality. However, reinforcement learning is rarely considered in game theory outside of simulations because of its large state complexity. In this paper, I offer a complete closed-form characterization of value-based reinforcement learning behavior in a prisoner's dilemma that removes the necessity for simulations.

Broad classes of learning rules have been studied in relation to the prisoner's dilemma: the overwhelming evidence in the literature is that the Nash equilibrium is very robust to the specification of bounded rationality. Learning rules that converge to the Nash equilibrium include adaptive dynamics (Milgrom and Roberts, 1990), which covers best-response dynamics, fictitious play, etc., as well as many others that rely on the weak acyclicity of the game, for example, Marden et al. (2009). This stands in stark contrast with the simulated behavior of reinforcement learning algorithms in a prisoner's dilemma, which has been shown to consistently reach cooperation (Calvano et al., 2020).

Email address: artur.dolgopolov@uni-bielefeld.de (Arthur Dolgopolov)

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

The present paper studies all algorithms in a class that generalizes the so-called Q-learning algorithm. A Q-learning player 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 a predetermined rule.

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 neither the reinforcement learning algorithm nor its designers observe anything beyond their own payoffs (Calvano et al., 2020; Klein, 2021). In this context, the goal is to find the conditions that make the cooperative or, in this case, collusive outcome in a prisoner's dilemma unsustainable.

Unlike some of the similar studies of learning in a prisoner's dilemma (Mengel, 2014; Calvano et al., 2020), I do not take into account 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 guarantee convergence to non-Nash outcomes even without relying on conditional strategies.

The proof uses techniques introduced 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 apply directly. However, the same idea can be used to construct a path to a certain "central" state, which narrows down the possible minimal spanning trees and ultimately leads to a characterization. This generalization of Newton and Sawa (2015) is of some independent interest. Specifically, I give a simple characterization of behavior 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 (Newton and Sawa, 2015; Nax, 2019) and games solvable by iterative elimination of strictly dominated strategies.

I, therefore, start with this general result and solve the prisoner's dilemma case as an application.

2. Literature

Computer science studies a similar topic under the name multi-agent reinforcement learning (MARL). Some MARL studies also consider competitive scenarios, but the existing theoretical results are conceptually different. The goal is usually to propose tailored algorithms for independent reinforcement learners agents to provably jointly converge (Hu and Wellman, 2003). Instead, I focus on a strategic interaction in an economic setting and therefore take the algorithms as given – set by individual firms in the hope of maximizing profits. As this is a vast literature, I refer the reader to the recent survey by Zhang et al. (2021). To my knowledge, only limited convergence results discussed below are available for standard value-based reinforcement learning algorithms, especially beyond zero-sum games.

The most closely related studies (Waltman and Kaymak, 2007, 2008) pursue the same goal and have partially characterised 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. In many applications however (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. Neither is it 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, the Nash equilibrium in the prisoner's dilemma as well. A 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. This algorithm is most similar to Bilancini and Boncinelli (2020) in the extreme case when the weight of the recent evidence in the learning process is the highest, but differs in that it keeps a separate record of the last payoff for each action.

The potential of reinforcement learning as a model for human behavior 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. However, I use reinforcement learning as a long-term equilibrium selection concept, instead of as a description of the medium-term dynamics.

The rest of the paper is organised as follows. I begin by introducing the game and learning rules in the next section, then I characterize the recurrent (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 reinforcement learning rules. I conclude by discussing possible extensions and comparing the results to other learning models.

3. Preliminaries

Consider a two-player strategic game. The possible actions of each player comprise the set A. Any pair of actions of the two players from the set A^2 is called a profile. Let π_{ab} denote the payoff of playing a when the opponent plays b. For a prisoner's dilemma $A = \{C, N\}$, and the four possible payoff values 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 or Nash action (defection).

The play of the game will be captured by a Markov process. Every state g of this process 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 Q-value of N for i in the state g will be denoted $Q_i^N(g)$, similarly for other Q-values. The set of all possible states, i.e. pairs of valid Q-vectors will be written as $\mathfrak{G}$.

In order to stay true to practical implementations of reinforcement learning and to avoid unnecessary continuity arguments while staying formal, 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 $\mathfrak{D}$ denotes some compact subset of $\{z\epsilon, z \in \mathbb{Z}\}$. Naturally, $\pi_{ab} \in \mathfrak{D}$ for any $a, b \in A$. Whenever the Q-value does not conform to this grid, it is rounded to the closest grid point (this will be formalised 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. Within this finite space, showing that a sequence of states is acyclic is enough for it to be finite. I will rely on this fact in the proofs to show that the process converges.

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. It means that players always choose the action with the higher Q. 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:

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

where $\mathcal{F}_i^{a_i a_{-i}}$ is some function of the state g_t, additionally parametrised by the action profile, s.t. $\mathcal{F}_i^{a_i a_{-i}}(g_t) \in (Q_i^a(g_t), \pi_{a_i a_{-i}}] \cap \mathfrak{D}$ if $Q_i^a(g_t) \neq \pi_{a_i a_{-i}}$ and $\mathcal{F}_i^{a_i a_{-i}}(g_t) = \pi_{a_i, a_{-i}}$ otherwise. In other words, only the Q-values of the action that was taken is ever updated, which is the main idea behind reinforcement learning. The $\mathcal{F}_i^{a_i a_{-i}}$ will be called the *learning rule.* We do not specify 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 realised 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 $\{a : Q_i^a(g) = \max_{a'} Q_i^{a'}(g)\}$ 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, i.e. all actions with the maximum Q-value have some positive probability to be taken. I will write $\text{path}_i(g)$ for the set of actions on path in state g for player i, and $\text{path}(g) = \{(a_1, a_2) \in \text{path}_1(g) \times \text{path}_2(g)\}$ 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 convenience I will introduce the functions $\mathcal{F}^{a_1, a_2}(\cdot)$, which for any state g return the new state g' that results from updating the previous values $(Q_1(g), Q_2(g))$ for the two players after playing the actions (a_1, a_2) once. Formally, $\mathcal{F}^{a_1, a_2}(g) = g' = (\mathcal{F}_1^{a_1 a_2}(g), \mathcal{F}_2^{a_2 a_1}(g))$. The $\mathcal{F}^{a_1, a_2}(\cdot)$ functions are deterministic as well – 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 the state g' does not constitute a valid update, and $P_0(g, g') > 0$ if it does:

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

In terms of the unperturbed dynamic I will be interested in the set of recurrent states, the states that are 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 probability of transition 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) $P_\eta \xrightarrow{\eta \to 0} P_0$*, 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 $g \neq g', P_0(g, g') = 0, P_{\hat{\eta}}(g, g') > 0$ *for some* $\hat{\eta} > 0$*, then* $\lim_{\eta \to 0} -\eta \log P_\eta(g, g') = c$ *for some* $c > 0$*.*

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

(vi) For any $\eta > 0$ *and state* g*, such that* $\{(a_1, a_2)\} = \text{path}(g)$*,* $P_\eta\left(g, \mathcal{F}^{b_1, b_2}(g)\right) = P_\eta\left(g, \mathcal{F}^{b_1, a_2}(g)\right) \times P_\eta\left(g, \mathcal{F}^{a_1, b_2}(g)\right)$*, where* $b_1 \neq a_1, b_2 \neq a_2$*.*

(vii) For any $\eta > 0$ *and states* g, g'*, such that* $(a_1, a_2) \in \text{path}(g) \cap \text{path}(g')$*, if* $Q_i^{a_i}(g) \geq Q_i^{a_i}(g')$ *and* $Q_i^{b_i}(g) < Q_i^{b_i}(g')$*, where* $b_i \neq a_i$*, then, if* $i = 1$*,* $P_\eta(g, \mathcal{F}^{b_1, a_2}(g)) \leq P_\eta(g', \mathcal{F}^{b_1, a_2}(g'))$ *and similarly if* $i = 2$*,* $P_\eta(g, \mathcal{F}^{a_1, b_2}(g)) \leq P_\eta(g', \mathcal{F}^{a_1, b_2}(g'))$*.*

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). 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.

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 parametrised 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 probability of both players experimenting is the product of the probability that each of them experiments alone. This implies that one player experimenting is always more likely than two players experimenting simultaneously for any given state. Importantly however, 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 is the same or lower and the action off-path is higher in g' compared to g, then she is at least as likely to experiment in g' as in g. In other words, players are less likely to experiment if the other action seems to give low payoffs.

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 (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.

It follows from the regularity conditions that the set of possible states $\mathfrak{G}$ cannot contain a state with Q-values that are beyond the lowest or highest attainable payoffs in the game, i.e. $\mathfrak{G} \subseteq \{(Q_1, Q_2) : Q_i^{a_i} \in [\min_{a_{-i}} \pi_{a_i a_{-i}}, \max_{a_{-i}} \pi_{a_i a_{-i}}]$ for any $a_i \in A, i \in \{1, 2\}\}$, and I assume this everywhere below. Clearly no other state can be reached from within such set $\mathfrak{G}$, because any transition has to be a valid update. That is, from (1) it follows that there are no states $g_1 \in \mathfrak{G}$ and

$g_2 \notin \mathfrak{G}$ such that $\mathcal{F}^{a_1,a_2}(g_1) = g_2$ for some $a_1, a_2 \in A$. In the actual computer implementation of the algorithm, the restriction is irrelevant as the learning process on a finite grid will eventually reach $\mathfrak{G}$. Thus, the designer does not need to have prior knowledge of the payoffs or other information about the game to set up the initial conditions, provided there is sufficient experimentation.

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

First of all, I am concerned with the behavior of the learning process in the limit as $\eta \to 0$, i.e. after the players have experimented with actions sufficiently and converged to some behavior. 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 $\lim_{\eta \to 0} \mu_\eta > 0$. Therefore, the goal of the paper is the characterization of these states. I will refer to the set of all stochastically stable states as SS.

The 1-step cost of the 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 $c\left(g, g'\right)$ is the exponential decay rate of the probability of transition 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 $P_\eta\left(g, g'\right)$ does not approach zero as $\eta \to 0$. The cost can also be defined for any finite sequence of distinct states (called a path) $g_1, g_2, ..., g_r$ as $c(g_1, g_2, ..., g_r) = \sum_1^{r-1} c(g_l, g_{l+1})$. Denote the set of all paths between g and g' by $S(g, g')$. Then the overall minimum cost of transitioning from g to g' in any number of steps can be defined as:

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

A spanning tree rooted at $\hat{g} \in \mathfrak{G}$ is a directed graph over the set $\mathfrak{C}$ 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, such that for some fixed vertex, called the root, there is exactly one directed path from any other vertex to the root. 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[2] there exists a minimum-cost spanning tree rooted at $\hat{g}$ (Freidlin and Wentzell, 1984; 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 that minimize $cost(\cdot)$.

It is common to restrict attention to transitions within $\mathfrak{C}$ instead of $\mathfrak{G}$. This is also without loss of geneality, see Corollary 12.A.4 in Sandholm (2010). I use the whole set of states $\mathfrak{G}$ for the construction of trees to avoid the need to compute a minimum spanning tree altogether. Using the theoretical results in this paper, limiting behavior is much easier to describe in this way than by solving the minimum arborescence problem on the set of recurrent classes $\mathfrak{C}$.

I 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 actions are played or an update after the most likely experimentation.

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:

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

It will also be useful to introduce the counterpart, the cost of the transition following experimentation by the player who was *less* likely to experiment. For a state g with $\text{path}(g) = \{(a_1, a_2)\} \in A^2$,

[2] 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 $\lim_{\eta\to 0} \mu_\eta(g) > 0$. That is, instead of saying that there is a positive mass on g in the limit, the latter claim is that it does not vanish at an exponential rate. This difference is the reason for the "only if" here instead of "if and only if". Practically, for my results the "only if" part is enough, because in all except borderline cases the action profile in all weakly stable states is unique. Since all stochastically stable states are contained within this larger set, this approach already answers whether cooperation or defection happens in the limit. The weaker definition of cost lets me cover a wider range of experimentation rules, for instance the probit rule, which would otherwise fall out of scope. In many cases the minimum tree characterization holds in both directions. See (Sandholm, 2010) for the details and the related discussion in (Newton and Sawa, 2015). The downside is that I cannot claim that *both* cooperation and defection happen with a positive probability when the parameters fall exactly on the border between regions in the Figure 2. I only claim that either may happen. 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): $P_\eta(g, g^i) > 0 \implies 0 < \frac{P_\eta(g,g^i)}{\eta^c} < \infty$ for some cost $c \geq 0$.

this value can be defined as:

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

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

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

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

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

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:

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

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

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.[3] I will call $\bar{C}(g_1, g_r)$ a c_L-adjusted cost to distinguish it from the R-adjusted cost in Ellison (2000). Note that the first and the last states on the path are not included in the sum. In many cases, including the present paper, the expression for the adjusted cost appears naturally in the expressions for the cost of the spanning trees.

[3] See (Ellison, 2000) for more examples of this construction. The definition here is simplified since I do not have to consider non-singleton recurrent classes. This definition also has an important difference with the one in Ellison (2000): for every state g_l I subtract $c_L(g_l)$ instead of $R(g_l)$. I show below in Remark 2 that for the class of problems in this paper, $R(g) = c_L(g)$ for almost any state g, and therefore the definitions coincide. However, as already noted in Newton and Sawa (2015), this fact does not follow from the definitions. I thank the anonymous reviewer for noticing the issue.

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 $b \in A \setminus a$.

It will be convenient to separately show the first two steps as 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 second lemma states that all actions that are not currently played but would be played eventually in some recurrent class always have the same Q-value, possibly unique for each player. Informally, whenever the player changed her action, the past action could not have become more attractive – its Q-value had to decrease, or be equal to the Q-value of the new action. But, unless the Q-values are equal, it is impossible to cycle back to the original Q-value, because the action needs to be played for its Q-value to change. From this, it also follows that when the player changes her action without leaving a recurrent class, she randomizes fully over all actions played by her in this class, including the previously played action.

Lemma 2. *In any recurrent class G:*

(i) In any state $g \in G$ all Q-values of $\mathcal{A}_i(G) \setminus \mathrm{path}_i(g)$ equal some constant for every player i;

(ii) Any transition in G where player i switches between two actions in $\mathcal{A}_i(G)$, while the other player keeps playing the same action, has a strictly positive probability.

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

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

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

Moreover, if the game can be solved by IESDS then these 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 absorbing states can now be refined to stochastically stable states.

The following two lemmas will allow me 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 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 every minimal spanning tree can only have non-minimum-cost edges from the states on the path from g^c to the root of the tree.

Lemma 3. *If $g^c \in \mathfrak{C}$ is central then in any minimal spanning tree, for 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 path 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 allows for the case when $g^c \notin OS$, which is used in the next proposition.

Proposition 2. *If there is a central state $g^c \in \mathfrak{C}$, then the minimal trees are rooted in states that minimize*

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

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

Proposition 2 offers a convenient way to solve for stochastically stable states when at least one central state is known to exist. Since $cost(g^c)$ enters the expressions for the cost of each tree, it suffices to compare $\bar{C}(g^c, \hat{g})$ and $c_L(\hat{g})$, which are much easier to calculate. 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 (shown in black). Only the edges between g^c and the root $\hat{g}$ (shown in red) could have a cost higher than minimal, because otherwise they 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 ones shown in blue. The only other possibility is that there is a state $\hat{g}$, such that the cost of reaching it from g^c (dashed/red edges) is smaller than the total cost of all minimum cost edges along the way (dotted/blue edges).

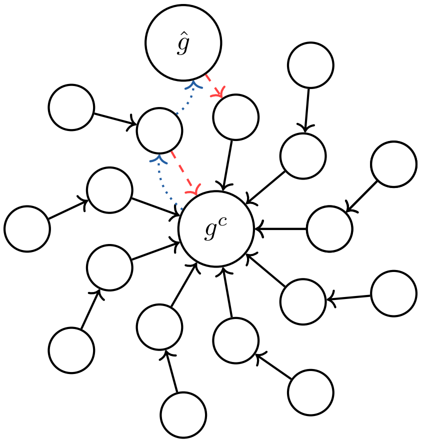


Figure 1: A central state and two spanning trees.

Conveniently, if one central state is known, there is no need to prove it is unique or search for all such states.

Remark 1. *It does not matter if there are multiple "central" states, and, if so, which is used as g^c to calculate the cost in Proposition 2.*

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 $g = g'_1, ..., g'_r = g'^c$, s.t. $C(g_l, g_{l+1}) = c_L(g_l)$ 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 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 will coincide with the R-modified cost as defined in Ellison (2000), which in turn 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 $g \in \mathfrak{G} \setminus g^c$*

Take first a state $g \in \mathfrak{C}$. 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, i.e. passes 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. Notice that $C(g, g_2) = C(g, g_1) + c(g_1, g_2) = C(g, g_1) = c_L(g)$. 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 $R(g) = c_L(g)$. The remark is automatically true for any unstable non-recurrent state $g \in \mathfrak{G} \setminus \mathfrak{C}$, because $R(g) = c_L(g) = 0$.

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

I will introduce two recurrent states, g^* and g^{**}, in $\mathfrak{C}$. The identity of the stochastically stable state depends on the choice rule, and I will show that we need only consider these two states. The first state g^* has defection on path with $Q_i^N(g^*) = \pi_{NN}, Q_i^C(g^*) = \pi_{CN}$ for both $i \in \{1, 2\}$. The other state g^{**} has cooperation on path with $Q_i^N(g^{**}) = \pi_{NN}, Q_i^C(g^{**}) = \pi_{CC}$ for both $i \in \{1, 2\}$. For the most part, I will only need to be concerned with these two states and the transitions between them. These states are the states with least likely experimentation among the states with N and C on path respectively by condition (vii) in Assumption 1.

To be able to use Proposition 2, I will show that indeed a variant of Newton and Sawa (2015) "getting closer" lemma holds for Q-learning in a prisoner's dilemma, but instead of approaching the OS set, the process will be approaching 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 will mean by "closer" to g^*. The distance to g^* will take a form of an ordering on the set $\mathfrak{C}$. It will have a different structure for states where players are cooperating and the states where they are already defecting, separating the adjustment process in two phases. For this purpose, I introduce $0 \geq m(g) \geq 2$ as the number of players who play N on path, $m(g) = |i : N \in \text{path}_i(g)|$.

for the case when the players are defecting on path, I define the quantity $D(g)$ as follows:

$$D(g) = \sum_{i \in \{1,2\}} |Q_i^C(g) - \pi_{CN}|,$$

Since the Q-value of N is π_{NN} in any of the states with (N, N) on path, $D(g)$ captures how far the Q-values of cooperation are from the lowest possible, which correspond to g^*.

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

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

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.

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^*. Since it may take multiple steps, I will write the probabilities of these t-step transitions as $P_0^t(g, g') \equiv Prob\left(g^t = g' \mid g^0 = g, P_0(\cdot,\cdot)\right)$.

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

Since states with (C, N), (N, C) or mixed actions on path are unstable in the sense that they cannot be in $\mathfrak{C} \subseteq \mathfrak{G}$, one only needs to 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 cannot be minimal. This can be stated as a 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 smaller cost, which is a contradiction. □

In any state $g \in \mathfrak{C} \setminus g^*$ with $(N, N) \in \text{path}(g)$, it must be that $Q_1^N(g) = Q_2^N(g) = \pi_{NN} = Q_1^N(g^*) = Q_2^N(g^*)$, while $Q_i^C(g) \geq Q_i^C(g^*)$ for both players $i \in \{1, 2\}$ and with a strict inequality for at least one of them. Then by regularity conditions in Assumption 1 (vii) the probability of a single-player experimentation from g^* is less or equal than the probability of a single-player experimentation from g. Formally, $P_\eta(g^*, \mathcal{F}^{C,N}(g^*)) \leq P_\eta(g, \mathcal{F}^{C,N}(g))$ and $P_\eta(g^*, \mathcal{F}^{N,C}(g^*)) \leq P_\eta(g, \mathcal{F}^{N,C}(g))$ for any $eta > 0$. At the same time, a minimum cost transition from g^* requires a two-player experimentation.

Putting it in terms of the costs instead of probabilities, Assumption 1 (vi) implies that the cost of a two-player experimentation is the sum of the costs of each player experimenting individually, i.e. $c(g^*, \mathcal{F}^{C,C}(g^*)) = c(g^*, \mathcal{F}^{C,N}(g^*)) + c(g^*, \mathcal{F}^{N,C}(g^*))$. 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 single-player experimentation in g^* and therefore also than the cost of leaving any such $\hat{g}$ 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 noting 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 consists of two steps. First, lowering the Q-value of C would generally increase the cost of a (C, C) update. 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 the non-cooperative profile with (C, C) makes the remaining path less costly and possibly shorter. The path will also end in g^{**}, since the plays of (C, C) do not change the Q-value of N, which remains equal to π_{NN} in 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. The complete proof is given in Appendix A.

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 regularity conditions in Assumption 1 (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 $i \in \{1, 2\}$. 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.* [4]

[4] The $\bar{C}(g^*, g^{**}) < c_L(g^{**})$ is a stronger requirement than $c(g^*, g^{**}) < c_L(g^{**})$ used in 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 edges need to be taken.

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

(iii) If $\bar{C}(g^*, g^{**}) = c_L(g^{**})$ for one or more states, both defection and cooperation are possible.

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. First, I will now explicitly parameterize the learning rule. 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 that is independent of the current Q-value:

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

where α is a rational number called the learning parameter for player i and $1 \geq \alpha_i > 0$. Note that this parameter can be different between players. The arg min term only maps the process 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 avoid specifying 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 (with ties resolved uniformly), and with probability $e^{-\frac{k_i}{\eta}}$ the action is chosen by randomizing uniformly. Formally in my definitions,

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

if a is played by i on path in g and

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

otherwise[5]. The denominator in the former case only divides the amount among all actions played

[5] 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 $Pr_i^{greedy}(a|g) = e^{-\frac{k_i}{\eta}} = \beta^{k_i}$ and consider β variable instead of η in all limits. Both setups correspond to the same model of uniform

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 condition (i) in Assumption 1. If $k_1 = k_2 = 1$, I obtain the simpler case with symmetric experimentation probabilities.

Under the logit choice rule instead

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

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, $\prod_{i\in\{1,2\}} Pr_i^{greedy}(a_i|g)$ or $\prod_{i\in\{1,2\}} Pr_i^{logit}(a_i|g)$ for $g' = \mathcal{F}^{a_i,a_{-i}}(g)$.

For the 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. The characterization will rely on a commonly used property of the logit choice 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 $i \in \{1, 2\}$. Let z_i be the largest integer strictly 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

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.

play of (C, C). The expression can be obtained by rewriting the recursive equations $Q_i^C(g_{t+1}) = (1-\alpha_i)Q_i^C(g_t) + \alpha_i \pi_{CC} \geq \pi_{NN}$, $i \in \{1,2\}$, $Q_i^C(g_0) = \pi_{CN}$ for the first g_{t+1} where $Q_i^C(g_{t+1}) \geq \pi_{NN}$. 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 $q_{l,i}^C = \pi_{CC} + (1-\alpha_i)^{l-1}(\pi_{CN} - \pi_{CC})$. For sufficiently small ϵ, these 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 choice rule:*

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

(ii) Under the logit choice rule the set SS is determined by sign of the expression

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

- $SS = \{g^*\}$ *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 if* $\Delta = 0$, *i.e. SS may include states with defection, cooperation, or both.*

Here $\mathbb{1}_{z>1}$ equals 1 if $z > 0$ 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, $q_l^C = \pi_{CC} + (1-\alpha)^{l-1}(\pi_{CN} - \pi_{CC})$. Substituting, the corollary follows immediately:

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

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

(ii) Under the logit choice rule the set SS is determined by sign of the expression

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

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

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}. Therefore I fix $\pi_{CC} = 1$ and $\pi_{CN} = 0$ without loss of generality – the shapes of the regions are preserved for other values of the three payoffs π_{CN}, π_{NC}, and π_{CC}. Instead I 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 Figure 2. The boundary of the regions consists of the only values where both cooperation and defection are possible in the limit. The area up and left of the red dashed line is the region covered by the theoretical part of (Waltman and Kaymak, 2008).

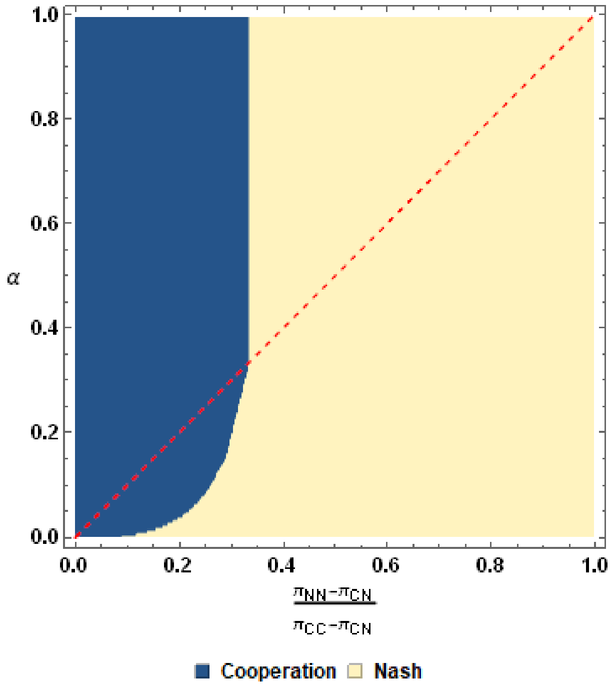


Figure 2: Trade-off between learning rate α and relative punishment payoff $\frac{\pi_{NN}-\pi_{CN}}{\pi_{CC}-\pi_{CN}}$ for symmetric learners.

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 other words, one of the learners can always preclude cooperation if her learning parameter is low enough.

Corollary 5. *There is always a low-enough $\alpha > 0$ for any π_{NN} such that $\{g^*\} = SS$ under either rule.*

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

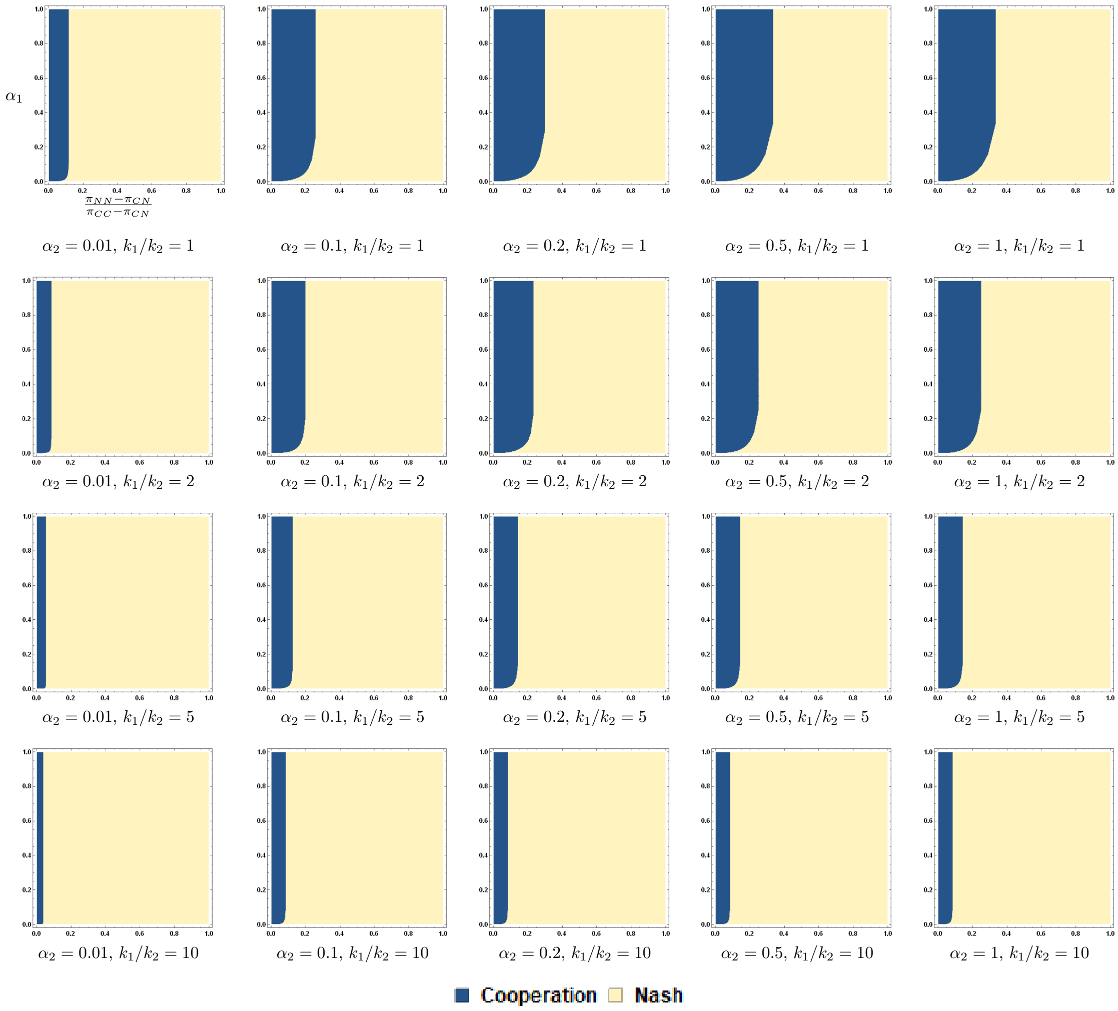
It is illustrative to also consider a supergame of choosing a learning algorithm against an opponent. In a supergame of choosing the parameters α_i and k_i, since the algorithms can only converge to (N, N) or (C, C) on path, low values of α_i are dominated. This fact follows mechanically from the Δ expression in Proposition 3, which is decreasing in α_i through the decreasing values of the $q^C_{l,i}$ variables:

$$\Delta = (\frac{1}{k_1}+\frac{1}{k_2})(\pi_{NN}-\pi_{CN}) + \mathbb{1}_{z>1}\sum_{l=2}^{z}\max_{i\in\{1,2\}}(\max(0,\frac{1}{k_i}(\pi_{NN}-q^C_{l,i})) - \min_{i\in\{1,2\}}\frac{1}{k_i}(\pi_{CC}-\pi_{NN});$$

Slower learning makes the cooperative state g^{**} harder to reach, but does not affect the minimum cost deviations that could lead the process 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. In other words, setting $\alpha_i = 1$ is never a bad strategy.

Assuming the players do not play the dominated strategies and 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 can be seen again by inspecting the Δ expression. Suppose $k_i > k_{-i}$. Then player i can decrease the value of Δ (and weakly increase the chance of cooperative outcome) by lowering k_i: the first two terms would decrease and the last term is independent of the higher k_i. And, vice versa, the other player $-i$ can decrease the value of Δ by increasing k_{-i}. The second term is zero because learning is immediate under $\alpha_1 = \alpha_2 = 1$ with $q^C_{2,i} = \pi_{CC}$ and $z_1 = z_2 = 0$. If $(\pi_{NN} - \pi_{CN}) > (\pi_{CC} - \pi_{NN})$ then $\Delta > 0$ for any value of k_{-i}. 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 increasing k_{-i} increases also the chance of converging to cooperation. The other player i, the one with the higher cost of experimentation, can safely increase the experimentation rate in the sense of lowering the cost by decreasing k_i. 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 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 3.



$\alpha_2 = 0.01$, $k_1/k_2 = 1$ $\alpha_2 = 0.1$, $k_1/k_2 = 1$ $\alpha_2 = 0.2$, $k_1/k_2 = 1$ $\alpha_2 = 0.5$, $k_1/k_2 = 1$ $\alpha_2 = 1$, $k_1/k_2 = 1$

$\alpha_2 = 0.01$, $k_1/k_2 = 2$ $\alpha_2 = 0.1$, $k_1/k_2 = 2$ $\alpha_2 = 0.2$, $k_1/k_2 = 2$ $\alpha_2 = 0.5$, $k_1/k_2 = 2$ $\alpha_2 = 1$, $k_1/k_2 = 2$

$\alpha_2 = 0.01$, $k_1/k_2 = 5$ $\alpha_2 = 0.1$, $k_1/k_2 = 5$ $\alpha_2 = 0.2$, $k_1/k_2 = 5$ $\alpha_2 = 0.5$, $k_1/k_2 = 5$ $\alpha_2 = 1$, $k_1/k_2 = 5$

$\alpha_2 = 0.01$, $k_1/k_2 = 10$ $\alpha_2 = 0.1$, $k_1/k_2 = 10$ $\alpha_2 = 0.2$, $k_1/k_2 = 10$ $\alpha_2 = 0.5$, $k_1/k_2 = 10$ $\alpha_2 = 1$, $k_1/k_2 = 10$

Figure 3: 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.

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. 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

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

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

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

In other words, the modified coradius is the highest R-adjusted cost C^* to reach G from any other state. In the present paper I adjusted $\bar{C}$ by the minimum cost instead. The values $c_L(g^*)$ and $R(g^*)$ are not generally the same for the unique central state g^*. 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 $O(e^{\frac{1}{\eta}(\max_{g \neq g^c} c_L(g))}) = O(e^{\frac{1}{\eta}(c_L(g^{**}))})$. 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). On the other hand, the expected hitting time of g^{**} is no more than $O(e^{\frac{1}{\eta}(\bar{C}(g^c, g^{**}) - R(g^c) + \max_{g \neq g^{**}} c_L(g))})$.

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 condition (vii) in Assumption 1. 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 from 0, the time to converge generally gets shorter, or 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 $(\pi_{CC} - \pi_{NN})$.

The first part follows from the fact that g^* is central, and thus the adjusted cost or reaching g^* from any state is determined by a single update. The speed of learning is therefore irrelevant. The second part follows from the term $\mathbb{1}_{z>1} \sum_{l=2}^{z} \max_{i \in \{1,2\}}(\max(0, \frac{1}{k_i}(\pi_{NN} - q^C_{l,i}))$ in the expression for $\bar{C}(g^*, g^{**})$. This term may be positive for $\alpha < 1$ but it is always 0 for $\alpha = 1$. In other words, the cost can not 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 (higher α for Q-learning). This would also be true for many other learning/experimentation rules because learning to defect is always immediate.

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

When $\alpha = 1$, the most recent payoffs for the two actions determine the transitions. 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 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 probability of experimentation in the respective states. The difference with reinforcement learning is that this probability does not account for the agent's experience with the other action.

6.3. 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. 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 (parametrised 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. 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.

6.4. Other reinforcement learning rules

The states $\mathfrak{G}$ faithfully encode the state of the value-based reinforcement learning algorithm, but the state of the game itself never changes because the players are unable to condition actions on the history of play. The setup is intentionally stripped from such features, which would help players converge to a cooperative outcome. Indeed, allowing the algorithms to maintain a policy vector and to condition their actions on history would lead to a folk theorem and cooperation as has been documented empirically by Calvano et al. (2020). Perhaps surprisingly, the illustration in this paper shows that this feature is not *necessary*, and very simple algorithms can learn to cooperate. This in turn implies that the more sophisticated algorithms have the capacity to converge as well. 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), state–action–reward–state–action or SARSA (Rummery and Niranjan, 1994), and others. Moreover, the distinctions between these algorithms usually lie with the adjustment of the learning process by 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.

7. Conclusion

Characterization of learning equilibria in this paper 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.

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. The present paper confirms these simulation results theoretically and shows that algorithmic collusion is possible even without this memory assumption. Moreover, 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 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 like taxation (Zheng et al., 2020) and oligopoly pricing (Calvano et al., 2020). This paper is an attempt to do the opposite: study the difficult problem of convergence of multiple reinforcement learning agents by modelling their interaction faithfully as a Markov chain. 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 therefore to expand the scope from a two-action game to 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.

8. Acknowledgements

I am indebted to Giacomo Calzolari for continuing support and advice. I am very grateful to Jonathan Newton, 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. Declarations of interest: none.

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, doi: 10.3386/w28535.

Bilancini, E. and L. Boncinelli (2020): "The evolution of conventions under condition-dependent mistakes," *Economic Theory*, 69, 497–521, doi: 10.1007/s00199-019-01174-y .

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, doi: 10.1257/mic.3.3.164.

Calvano, E., G. Calzolari, V. Denicolo, and S. Pastorello (2020): "Artificial intelligence, algorithmic pricing, and collusion," *American Economic Review*, 110, 3267–97, doi: 10.4337/9781786439055.00038 .

Ellison, G. (2000): "Basins of attraction, long-run stochastic stability, and the speed of step-by-step evolution," *The Review of Economic Studies*, 67, 17–45, doi: 10.1111/1467-937x.00119.

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, doi: 10.1109/ieecon.2014.6925916.

Freidlin, M. I. and A. D. Wentzell (1984): "Random Perturbations of Dynamical Systems," Springer, New York, doi: 10.1007/978-1-4684-0176-9_2.

Hart, S. and A. Mas-Colell (2003): "Uncoupled dynamics do not lead to Nash equilibrium," *American Economic Review*, 93, 1830–1836, doi: 10.1257/000282803322655581.

Hu, J. and M. P. Wellman (2003): "Nash Q-learning for general-sum stochastic games," *Journal of machine learning research*, 4, 1039–1069.

Klein, T. (2021): "Autonomous algorithmic collusion: Q-learning under sequential pricing," *The RAND Journal of Economics*, doi: 10.1111/1756-2171.12383.

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, doi: 10.1137/070680199.

Mengel, F. (2014): "Learning by (limited) forward looking players," *Journal of Economic Behavior & Organization*, 108, 59–77, doi: 10.26481/umamet.2008053.

Milgrom, P. and J. Roberts (1990): "Rationalizability, learning, and equilibrium in games with strategic complementarities," *Econometrica: Journal of the Econometric Society*, 1255–1277, doi: 10.2307/2938316.

Nax, H. H. (2019): "Uncoupled aspiration adaptation dynamics into the core," *German Economic Review*, 20, 243–256, doi: 10.1111/geer.12160.

Newton, J. (2018): "Evolutionary game theory: A renaissance," *Games*, 9, 31, doi: 10.3390/g9020031 .

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, doi: 10.1016/s0899-8256(05)80020-x.

Rummery, G. A. and M. Niranjan (1994): *On-line Q-learning using connectionist systems*, vol. 37, Citeseer.

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

Waltman, L. and U. Kaymak (2007): "A theoretical analysis of cooperative behavior in multiagent Q-learning," in *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, IEEE, 84–91, doi: 10.1109/adprl.2007.368173.

——— (2008): "Q-learning agents in a Cournot oligopoly model," *Journal of Economic Dynamics and Control*, 32, 3275–3293, doi: 10.1016/j.jedc.2008.01.003.

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

Zhang, K., Z. Yang, and T. Başar (2021): "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control*, 321–384.

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

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

(i) if a_i is played and

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

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

From this I know that 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. Suppose there is a sequence of positive probability transitions $g_1, g_2, ..., g_l$ through the states of the recurrent class, such that player i plays some action $a \in A$ in all of these states except g_l, i.e. $a \in \text{path}_i(g_1) \cap \text{path}_i(g_2), ..., \text{path}_i(g_{l-1})$. Suppose also that in the last state g_l of this sequence she plays some other action $a' \in \text{path}_i(g_l), a' \neq a$. Then the Q-value of a cannot be strictly lower in g_1 than in g_l, because $Q^a(g_l) \leq Q^{a'}(g_l) = Q^{a'}(g_1) \leq Q^a(g_1)$. Otherwise, player i would have started playing a' sooner. At the same time, the Q-value of a only changes when a is played. Therefore g_1 cannot be reached again after g_l (contradicting the recurrent class property) unless $Q^a(g_l) = Q^a(g_1)$. Moreover, since a' has a positive probability of being played in g_l, $Q^{a'}(g_l) = Q^a(g_l) = Q^a(g_1)$. The same argument can now be applied to the next sequence of states $g_l = g'_1, g'_2, ...g'_l$ where a' is played until the next switch in g'_l, where the player switches to some action in $A \setminus a'$. Thus, $Q^{a'}(g'_l) = Q^{a'}(g_l) = Q^a(g_l) = Q^a(g_1)$. Call this constant value q. Continuing this argument, the Q-value of any action just before the player switches to a different action has to be q. This Q-value also remains unchanged until that action is played again. This proves that any action $a \notin \text{path}(g)$ also has Q-value q in any state g within a recurrent class. This is the part (i). For the part (ii), it is enough to notice that the Q-values of all actions played in the class are the same and equal q whenever the player switches her action.

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 in 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 $g \in G$, $\min_{x \in \mathcal{A}(G)}(\pi_{a_i x}) = \pi_{a_i a_{-i}} \leq Q_i^{a_i}(g) \leq \max_{x \in \mathcal{A}(G)_{-i}}(\pi_{a_i x}) = \pi_{a_i a_{-i}}$, i.e. $\pi_{a_i a_{-i}} \leq Q_i^{a_i}(g) \leq \pi_{a_i a_{-i}}$ and similarly for the other player. 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 a_i to b_i and let a_{-i} be the actions of the other player in the new state g, where b_i is played. Lemma 2 proves that this transition occurs with strictly positive probability. 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 has strictly increased and i would keep playing a_i in 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 immediately switch to a different action, leaving the Q-value of b_i strictly below $Q_i^{a_i}(g) = Q_i^{b_i}(g)$, 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. In other words 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.

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 $i \in \{1, 2\}$.

Neither (N, C) nor (C, N) can be played on path in $g \in \mathfrak{C}$ 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 $\pi_{CN} < \pi_{NN} < \pi_{NC}$.

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 $i \in \{1, 2\}$ 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 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 $\pi_{CN} < \pi_{NN}$ in g_1 and $Q_i^C(g_1) < Q_i^N(g_1) = \pi_{NN}$. The player $-i$ obtains $\pi_{NC} > \pi_{NN}$ in g_1 and since $Q_{-i}^N(g) = \pi_{NN}$, $Q_{-i}^N(g_1) > Q_{-i}^N(g)$. Therefore in 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 in the states $g_2, g_3, ...g'$ that follow g_1, $Q_i^N(g_2) = Q_i^N(g_3)... = \pi_{NN}$ continues to hold and $|Q_{-i}^N(g_2) - \pi_{NN}|, |Q_{-i}^N(g_3) - \pi_{NN}|, ...$ decreases until $Q_{-i}^N(g') = \pi_{NN}$ again for some g'. Thus, in the new absorbing 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 $\{(C, C)\} = \text{path}(g)$. 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 $\pi_{NC} > \pi_{CC}$ in g_1 and thus $Q_i^N(g_1) > Q_i^N(g)$. Player $-i$ obtains $\pi_{CN} < \pi_{CC}$ in g_1 and thus $Q_{-i}^C(g_1) < Q_{-i}^C(g)$. If the process converges to a state where at least one player defects, then $2 = m(g') > m(g)$, and $g' \prec g$ as required. So suppose instead that eventually a state in $\mathfrak{C}$ with (C, C) on path is reached. Two subcases are possible.

   (a) Both $Q_i^C(g_1) > Q_i^N(g_1)$, $Q_{-i}^C(g_1) > Q_{-i}^N(g_1)$ continue to hold in g_1 and (C, C) is played again. Then, with a positive probability and zero cost in the states $g_2, g_3, ...g'$ that follow $Q_i^C(g_2) = Q_i^C(g_3) = ... = \pi_{CC}$ continues to hold and $|Q_{-i}^C(g_2) - \pi_{CC}|, |Q_{-i}^C(g_3) - \pi_{CC}|, ...$ decrease until $Q_{-i}^C(g') = \pi_{CC}$ again for some g'. Thus, in the new absorbing state $Q_i^C(g') = Q_{-i}^C(g') = \pi_{CC}$, $Q_{-i}^N(g') = Q_{-i}^N(g)$, but $\pi_{CC} > Q_i^N(g') > Q_i^N(g)$. By condition

(vii) in Assumption 1 the cost of player i experimentation is at least as high in g' as in g. 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. Then $c_M(g') = c_M(g)$ and $c_L(g') < c_L(g)$, so $g' \prec g$ as required.

(b) In all remaining cases eventually, with probability 1 the process P_0 leads to some state g'', $Q^C_{-i}(g'') < Q^N_{-i}(g'')$ and this Q-value will not increase unless (C, C) is played again. If $Q^C_{-i}(g_1) < Q^N_{-i}(g_1)$ and (C, N), or (N, N) is played next, this is true immediately in $g_1 = g''$. If instead $Q^C_{-i}(g_1) > Q^N_{-i}(g_1)$ continues to hold, but $Q^C_i(g_1) < Q^N_i(g_1)$, 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 $Q^C_{-i}(g'') < Q^N_{-i}(g)$ in some g'' as required.

Since other cases were covered in the beginning, (C, C) has to be played eventually. Once (C, C) is played again, the game continues with (C, C) until convergence to a recurrent state – the payoff π_{CC} is the highest payoff for action C so the Q-value can not decrease after (C, C). Then at some future state g', $Q^N_{-i}(g) > Q^C_{-i}(g'') > Q^N_{-i}(g')$. Moreover, since i experimented in g, one of the following cases must occur. By condition (vii) in Assumption 1 and because $Q^N_{-i}(g) > Q^N_{-i}(g')$, $Q^C_i(g') = Q^C_{-i}(g') = \pi_{CC}$, the cost of experimentation by $-i$ is higher in g'. Regardless of which player is less likely to experiment in g', the cost of this experimentation $c_M(g')$ is no less than the cost of experimentation by player $-i$. 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 $g^* = g_1, ..., g_r = g^{**}$ is always lower than the cost of a sequence with at least one N. The expression for the c_L-adjusted cost of a sequence of (C, C) consists of the costs of the first and last transitions, as well as the costs of updates adjusted by the minimum costs, $c(g, \mathcal{F}^{C,C}(g)) - c_L(g)$. Let me first rewrite these expressions. From the regularity condition vi in Assumption 1, $c(g, \mathcal{F}^{C,C}(g)) = c(g, \mathcal{F}^{C,N}(g)) + c(g, \mathcal{F}^{N,C}(g))$. If $\text{path}(g) = (N, N)$ then the lowest cost deviation is the probability of experimentation by the player who has the smallest cost of experimentation. If, on the other hand, $(C, N) \in \text{path}(g)$ or $(N, C) \in \text{path}(g)$ then $c_L(g) = 0$. Thus, in both cases $c(g, \mathcal{F}^{C,C}(g)) - c_L(g) = \max(c(g, \mathcal{F}^{C,N}(g)), c(g, \mathcal{F}^{N,C}(g))) = c_M(g)$, i.e. the cost can be written as the cost of a single-player experimentation by the player who has a *higher* cost of experimentation.

Suppose now to contradiction that there is a lower cost path $g^* = g_1, g'_2...g'_{r'} = g^{**}$ with N played at least once in some state. Let $\bar{g}'_l$ be the state (if any) on this path where (C, C) is played for the

l-th (not necessarily consecutive) time. The rest of the proof relies on two facts about individual updates along the paths $g_1, g_2 ... g_r$ and $g_1, g'_2 ... g'_{r'}$.

First, decreasing the Q-value of C cannot decrease the cost of a (C, C) update. By the regularity condition vii in Assumption 1, $c_M(g_l)$ is the lowest among all $c_M(g, \mathcal{F}^{CC}(g))$ for any state $g \in \mathfrak{C}$, such that $Q^C_i(g) \leq Q^C_i(g_l)$ for both $i \in \{1, 2\}$. Indeed the $Q^N_i(g)$ takes the lowest possible value for a state in $\mathfrak{C}$, and a lower Q^C_i would decrease the probability of playing C. Moreover, the Q-value of C does not decrease from g to $\mathcal{F}^{CC}(g)$ for either player.

Second, a play of $(C, N), (N, C)$, or (N, N) cannot increase the Q-value of cooperation for either player, i.e. $Q^C_i(g'_l) \leq Q^C_i(g'_{l+1})$ if $(C, N), (N, C)$, or (N, N) is played in g'_l. Then for every pair of states g_l and $\bar{g}'_l$, $l \in 1..r$, on the paths without and with defection respectively, it must be that $Q^C_i(\bar{g}'_l) \leq Q^C_i(g_l)$ for both $i \in \{1, 2\}$. Moreover there are at least r states with (C, C) among $g'_2, ..., g'_{r'}$.

Then the cost $c_M(\bar{g}'_l)$ at every $\bar{g}'_l$ state for $l \in 1..r$ is at least as high as $c_M(g_l)$ on the other path. The overall c_L-adjusted cost of the path with defection is $\bar{c}(g_1, g'_2, ... g'_{r'}) = c(g_1, g'_2, ... g'_{r'}) - \sum_{l=2}^{r'-1} c_L(g'_l) \geq c(g_1, g'_2) + \sum_{l=1}^{r-1} c_M(\bar{g}_l)$. The cost of the path with no defection $g_1, ..., g_r$ is $\bar{c}(g_1, ... g_r) = c(g_1, g_2) + \sum_{l=2}^{r-1} c_M(g_l)$. This can be written as $c_L(g_1) + \sum_{l=1}^{r-1} c_M(g_l)$ because $c_M(g_1) = c(g_1, g_2) - c_L(g_1)$. Then $\bar{c}(g_1, ... g'_{r'}) \geq c(g_1, g'_2) + \sum_{l=1}^{r-1} c_M(g_l) = \bar{c}(g_1, ... g_r) - c_L(g_1) + c(g_1, g'_2) \geq \bar{c}(g_1, ... g_r)$. The last inequality follows by definition of minimum cost, which implies $c(g_1, g'_2) \geq c_L(g_1)$.

Finally, a path to any state with cooperation on path other than g^{**} requires a higher cost. Since the Q-value of N is the same in g^* and in g^{**}, $Q^C_i(g^*) = Q^C_i(g^{**}) = \pi_{NN}$ for both $i \in \{1, 2\}$, the sequence $g_1, g_2, ... g_r$ 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 $\frac{1}{k_1\eta}(\pi_{NN} - q^C_{l,1}) + \frac{1}{k_2\eta}(\pi_{NN} - q^C_{l,2})$. That is, $C(g^*, g^{**}) = \sum_{l=1}^{z_1} \frac{1}{k_1\eta}(\pi_{NN} - q^C_{l,1}) + \sum_{l=1}^{z_2} \frac{1}{k_2\eta}(\pi_{NN} - q^C_{l,2})$. 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 $2 \leq l \leq z$, which gives the second term, similarly to the argument in the proof of Lemma 5. In particular, $c(g_l, \mathcal{F}^{C,C}(g_l)) - c_L(g_l) =$

$\max(c(g_l, \mathcal{F}^{C,N}(g_l)), c(g_l, \mathcal{F}^{N,C}(g_l))) = \max_{i\in\{1,2\}}(\max(0, \frac{1}{k_i}(\pi_{NN} - q^C_{l,i}))$ for any $2 \leq l \leq z$ if $z > 1$. This leaves the $c(g_1, g_2)$ term for $l = 1$, which equals $c_l(g^*) = (\frac{1}{k_1} + \frac{1}{k_2})(\pi_{NN} - \pi_{CN})$. Finally, $c_L(g^{**}) = \min_{i\in\{1,2\}} \frac{1}{k_i\eta}(\pi_{CC} - \pi_{NN})$. Substituting these expressions into the characterization into $\bar{C}(g^*, g^{**}) - c_L(g^{**})$ yields the expression for Δ and then the result follows by Corollary 3.