PRISONER'S DILEMMA

ARVID LUNNEMARK

Definition 1. The *prisoner's dilemma* is a symmetric two-player game with two actions, cooperate (C) and defect (D), where, if player 1 plays a and player 2 plays b, player 1 gets payoff

$$p(a,b) = \begin{bmatrix} R & S \\ T & P \end{bmatrix}$$

if player 1 is the row player, and cooperate is the first action. We have $T > R > P > S$, and typically, we have the concrete values $T = 5$, $R = 3$, $P = 1$ and $S = 0$.

phrase this in a nicer way

Definition 2. A *strategy* is a Moore machine (finite automaton with outputs) over the input and output alphabet $\{C, D\}$, with probability $1 - p$ of following the correct transition and probability p of following the incorrect transition. The strategy can be viewed as a function s from a history string h of the opponents moves to an action, i.e., $s : \{C, D\}^* \to \{C, D\}$.

(could also think about including the $2R > T + S$ condition here)

(could also define it as probability p of outputting the wrong character, but should be largely equivalent)

Definition 3. The *expected time-average payoff* of strategy s_1 when played against strategy s_2 is

$$v(s_1, s_2) = \lim_{T \to \infty} E\left[\frac{1}{T}\sum_{t=1}^{T} p(s_1(h_{t-1,2}), s_2(h_{t-1,1}))\right]$$

where the history is defined as

$$h_{t,1} = h_{t-1,1} \circ s_1(h_{t-1,2})$$
$$h_{t,2} = h_{t-1,2} \circ s_2(h_{t-1,1})$$

Definition 4. A *population* of strategies $P = (S, f)$ is a set S of strategies and a function $f : S \to (0, 1]$ such that $\sum_{s \in S} f(s) = 1$, representing the frequency of each strategy in the population.

Definition 5. The *fitness* of a strategy s in a population $P = (S, f)$ is

$$F(s) = \sum_{s' \in S} f(s) v(s, s').$$

(TODO redo this. the actual definition i'm using is the distribution in the stationary state of the markov chain as explained below)

Definition 6. A strategy s_2 can *invade* a strategy s_1, if there exists a real number $\alpha \in (0, 1]$, such that for all populations P with $S = \{s_1, s_2\}$ and $f(s_2) < \alpha$, we have

$$F(s_2) > F(s_1).$$

a question is what to do with evolutionary drift, that is, if $F(s_2) = F(s_1)$, becuase it could potentially pave the way for

Remark. Thus, s_2 can invade s_1 if and only if it can start as an infinitesimally small part of the population and grow to become a constant fraction α of it.

Definition 7. A strategy s_1 is *evolutionary stable* if it cannot be invaded by any strategy s_2.

Corollary 1. A strategy s_1 is evolutionary stable if and only if

$$v(s_1, s_1) > v(s_2, s_1)$$

or

$$v(s_1, s_1) = v(s_2, s_1) \text{ and } v(s_1, s_2) \geq v(s_2, s_2)$$

for all s_2.

apparently the last inequality is strict in the traditional definition. given what I've seen with Pavlov it may even be possible to have that stricter definition be in place

Proof. Yeah this is true, the proof is easy by examination of cases. I just don't want to write it now. □

Theorem 1. Suppose a strategy s is evolutionary stable. Then $v(s, s) = R$. In other words, s is efficient with itself.

This looks like a very similar result to the theorem presented in the original paper. In fact, it looks like merely a simplication of it. To the contrary, however, it is very different, which stems from the definition of $v(\cdot, \cdot)$ it captures the entire behavior, the actual expected value, instead of some weird artificial value of pretending that no mistakes occur at all which doesn't really make sense. Therefore, I think that this definition of evolutionary

OK we need a different notion of what it means to be efficient with oneself now. *Probably.* In fact, maybe we don't. Because AllC will get R. But neither TFT nor Pavlov will. And, in fact, any strategy that is not stupid, will have opposite behavior when $p = 0$ and when $p = 1$ which suggests that it cannnot possibly get R in both situations. COULD MAKE IT ONLY WORK IN THE LIMIT AS $p \to 0$.

wait..... maybe allC can invade TFT now. seems to depend on the specific values of R, S, T, P which is the worst possible situation one can end up in. I think AllC can invade Pavlov too, which is not very promising

hmm okay so Pavlov is better against itself than TFT against itself, which, however, we have seen doesn't matter much.

I think the proof strategy is like this: suppose $v(s_1, s_1) < R$. then create a strategy s_2 that is more or less exactly like s_1, so $v(s_2, s_1) = v(s_1, s_2) = v(s_1, s_1)$, but s_2 identifies itself and has $v(s_2, s_2) = R$.

I think that captures the idea of an outside species pretending to be the same as everyone else but secretly getting a lot of value from cooperating with other members of the infiltrators, which I think makes a lot of sense from the anthropomorphic viewpoint.

A reasonable condition is

$$\operatorname*{argmax}_v \left[(R, T, S, P) \cdot v\right] = (1, 0, 0, 0).$$

I think this is needed and crucial. Maybe one should even require that it's the unique maximum but I'm not exactly sure. This also seems like a very reasonable thing to assume, and is a generalization of the commonly used $R \geq (S + T)/2$. I like this. omg wait this condition makes no sense at all.

If we require unique maximum, then TFT can be invaded by allC, which kinda makes sense.

NICEEEE: As p tends to 0, Pavlov tends to characteristic vector $(1, 0, 0, 0)$ (i.e. payoff R) so I actually think that this theorem might be true (especially considering that AllC fares poorly against Pavlov). Since TFT have characteristic vector $(1/4, 1/4, 1/4, 1/4)$, and would thus not be evolutionary stable here, this theorem indicates a stronger result than the theorem that it is based on.

omg wait the condition I proposed makes no sense at all hmmmm

oh wow read https://search.proquest.com/docview/235762955?accountid=12492 they came up with exactly the same Markov chain as I did and seems to have done more analysis on it. they seem to not do much with their results at all, however.