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

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

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.

A Major Problem

The definitions in the preceding section were all taken from the existing literature. In particular, the corollary is the traditional definition of evolutionary stability (and the underlying definitions were reversed engineered from that). However, these definitions lead to the following big problem: *AllD can no longer be invaded*. The reason is simple: Suppose strategy S can invade AllD. In the stationary distribution, there can be no state of S that cooperates. Thus, one can contract all independent connected components into single nodes of D, reachable by different branches. One can then easily see that running S against itself also always defects, so it must effectively just be an AllD copy, and will thus not be able to invade.

Thus, **I need to change the definitions** to get something meaningful out of this. Actually, what this also means is that it further goes to show that the results presented using the weird model of noise are weird as well, as the result they proved using that model is provably not reproducible in this model.

I can think of two ways of fixing this problem by changing the definitions:

(1) Make $p \to 0$ earlier. i.e. instead of fixing p and then letting $\alpha \to 0$, fix α and let $p \to 0$ or something similar.
(2) Define invasion differently: a strategy is ES iff given an invasion fraction α (for sufficiently small α) there exists no strategy that can achieve a higher fitness.

There might be other ways of solving this; if so, add to this list later. For now, I should analyze the two methods in separation. Later, I can determine which of the two methods is more reasonable to adopt.

Make $p \to 0$ happen earlier. The simplest way: define $v(s_1, s_2)$ as the limit $\lim_{p\to 0}$ of the stationary distribution payoff. In this situation, theorem 1 is true, and can be proved by using a very simple strategy that just self-identifies on the first turn and then copies. **Pros:** This gives us exactly the result we want. **Cons:** I don't think that result means much, because it is impossible to relate it to a real situation since there is no actual value of p

for which the result would hold (not even $p = 0$) which just means that it's some sort of fictional result.

Can we take a middle ground? Could say something like: A strategy s_2 can invade a strategy s_1 if there exists a real

Definition 8. 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 $f(s_2) < \alpha$, there exists a value of $p > 0$ such that

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

This means that the same strategy s_2 can invade s_1 but that the p can possibly change from time to time. This would also solve the AllD problem, and it would also prove theorem 1, while making more sense (albeit not much sense either).

Redefine invasion. This is sort of similar to our second approach above actually. But here we take a different approach: instead of different p for every α, we take different s_2 for different α.

The following is one way of defining invasion in this way.

Definition 9. A strategy s_1 can be invaded if there exists a real number $\alpha \in (0, 1]$ such that for all $\epsilon < \alpha$, there exists a strategy s_2 such that in the population P with $S = \{s_1, s_2\}$ and $f(s_2) = \epsilon$, we have

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

Does this alleviate the AllD problem? Yes, and it also proves theorem 1, using the same (boring) construction (except that the self-identification now may take k steps instead of only 1 depending on how small the population is). Is this reasonable? Not really, because it doesn't say anything about what happens to that strategy next.

Other approaches. So we can see that all three possible ways to solve the problem are somewhat unsatisfactory. New ideas:

(1) Include evolutionary drift into the picture: a strategy is not evolutionary stable if it is vulnerable to a 2-stage attack whereby 1 startegy first enters and survives because of evolutionary drift, and then another comes in and colludes with the second one to take over.
(2) Move away from infinitely repeated games (no no seems very messy and sad)
(3) Think of another model that better captures the finiteness of real population (which also helps with the $\alpha \to 0$ problem).
(4) A different model of computation. Note that the Moore machines I'm using have a very clear start. This, one could say, does not really model reality that well, because there will not necessarily be a clear starting point.

(5) Another way of evaluating fitness: think that the individuals are spread out geographically which means that even in a very large population, the initial strategy mutation injection will first face a small fraction of the population, then take over that part, then face a larger proportion but now be larger itself, and so on and so forth. Thinking this way, as long as the span of higher fitness grows by something like r^2 in a 2D model, and there is some α where the invader is better, then it can eventually take over most of the population which is very interesting.

Evolutionary drift and multiple different strategies. hmm. This has the same AllD problem as the original setup, using the same argument. The thing with noise really is that you can't really hide a secret that is useful (and when AllD finds out it will end up punishing you for it)

[other approaches here]

2D model. Okay so this is cool. Think of 1 strategy as covering a large circle. then start another strategy as a very very tiny circle. who does it interact with? could also think of this in terms of the grid. Hmmmm ok at the very least this would motivate why it is moew reasonable to say that a strategy can invade another strategy as long as there exists some α for which it can invade: because if a strategy only engages with other strategies within a certain fixed radius, the same strategy could then invade a given strategy regardless of how big the population is, which isn't really principally important but still kinda gives the definition a tiny bit more legitimacy. The nice thing about the circle 2D model is that it also doesn't require that the population size stays constant throughout – it could actually fluctuate quite a bit.

Redefine invasion. Can we use definition 9 but prove that it implies that it works for all $\alpha \geq \varepsilon$ and $\alpha <$ maybe $\frac{1}{2}$? That would be cool because it would show that that model actually makes a lot of sense. OK so that is not true. But generally, if we find a strategy that can invade using definition 9 we can almost certainly find a strategy that can invade using the natural definition. Wait so I think that that means that they're equivalent. Let me state it

Definition 10. A strategy s_1 can be invaded if for all $\varepsilon \in [0, 1]$ (and for all $p \in (0, 1)$????) there exists a strategy s_2 such that in all populations P with $S = \{s_1, s_2\}$ and $f(s_2) \geq \epsilon$, we have

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

So we have the concept of invadable9 and invadable10. If something is invadable9 then I think we can also make it be invadable10. And if it is invadable10 then certainly it is invadable9. So yay then definition 9 and definition 10 are equivalent which is amazing.

(The proof showing invadable9 implies invadable10 is somewhat interesting: you take a strategy that can invade according to 9 and then you modify it so that it is more or less the same but also has $v(s_1, s_1) + v(s_2, s_2) \geq v(s_1, s_2) + v(s_2, s_1)$ which then would show according to ipad calculations that it the new s_2 would be able to invade for all bigger ε which completes the proof.)

So we can use whichever definition we like. Note that definition 10 has a very nice interpretation: ε and p can be thought of as two constants of nature (representing the number of interactions and probability of failure, respectively), and then the definition says that a strategy can be invaded if regardless of what those parameters are, there is some bad strategy that can pop up and take over. In other words, a strategy is evolutionary stable if there exists some constants of nature that can make it prevail. So in this sense the evolutionary stability becomes a pretty weak concept. Showing a strong result about a weak concept – e.g. the utilitarianism result – would thus be kind of remarkable. Is this notion of ESS weaker than the usual notion of ESS? No because it excludes AllD. But at the same time, the result will be stronger than any results proved earlier, because probably only Pavlov will be good (although I'm still worried about the $p \to 0$ necessities of this statement. But we'll see about that when we get there.)

Note also that definition 9 is nice for proving stuff, because the condition is only to find 1 epsilon and not infinitely many.

Implications of definitions 9 and 10.

Interesting Findings

Using definition 9/10, we can prove the following:

(1) For a given p, if a strategy s is ESS then $v(s, s) \geq \frac{S+T}{2}$.
(2) For small p, if a strategy s is ESS then $\lim_{p \to 0} v(s, s) \geq R$.

This is great. The only problematic thing is that not every strategy that tends to R will be a good one I think (although I'm not sure about that actually). And I also need to prove that Pavlov, for example, actually is ESS, because otherwise the definition is kinda stupid. But this is still very nice and promising.

Note that all of this can be proven using the obvious self-identifying strategy, which goes to show that the original idea of having someone try to mimic the behavior of the other group but at the same time try to exploit it behind their backs actually works in theory too which is really cool I think.

OK how can we prove that Pavlov is actually ESS? Which definition is easier to use? (probably 9a again).

OK I will use definition 9 I think. ahhhh so much to do here and I'm not even sure that my result is valid.

Lemma 1. Suppose s_1 is a strategy such that $v(s_1, s_1) < \frac{S+T}{2}$. Then s_1 is not ESS according to definition 9.

Proof. We will show this by exhibiting a strategy s_2 and a parameter ϵ. Let the strategy s_2 be such that it "tests" its opponent for k steps. (uhhh ok this gets a little more complicated I think because it is just not a simple exponential). □

For reference

Pavlov against itself gets the payoff:

$v(s_{\text{Pavlov}}, s_{\text{Pavlov}}) = (-4p^3+7p^2-4p+1, p(1-p), p(1-p), 4p^3-5p^2+2p)\cdot(R, S, T, P)$

which tends to 1 as $p \to 0$.