Response to reviews

October 31, 2017

1 Comments from editor

We would like to start our responses by stating that we feel that the reviews we have received are fair and detailed. It is clear that the reviewers are aiming to help us improve the manuscript and we are grateful for their time.

We were particularly happy to note that the reviewers not only felt the work was of interest but made a point to note that our open scientific approach was noteworthy (publishing of source code and data).

> `Remember that it is possible to put material into an on-line appendix which is not page-limited if there are things you feel require more space to display.`

This is probably the most important change to the manuscript. We have moved most of the figures to a supplementary materials document which not only allows for them to be presented in a clearer way but also allows for all cases (variety of population sizes) to be included.

We have referred to this throughout the text simply as 'supplementary material', if there is a better way to do this we would be happy to change.

2 Comments from first referee

> `It's not easy to solve, so I can only point out the problems, and ask you to do what is possible to fix them. One possibility would be to put all figures in a downloadable Appendix, or some other online resource; and maybe remove some of the figures from the paper, in order to make more room for the others.`

Thank you for acknowledging the difficulty we're presented with. We feel that having opted to put most of the figures in supplementary materials we arrive at a more concise paper with much clearer plots (as they have more space).

We have also endeavoured to clear up the plots that have remained in the paper (for example by removing cluttering labelling).

font in the Appendix could be reduced (for example to the same size of the font in the references)

This has been done.

When you speak about evolutionary computation applied to the IPD, you could add some words about reference [21], that is otherwise only cited for GTFT. The basic idea of the publication was to create an EA able to adapt its behavior by building a model of the opponent 'online', during the game, and then exploiting the opponent's weaknesses. It's unpractical to use in your experiments (reported computational time is considerable, the technique is tailored for 1vs1 games as it can only model one opponent), but it could be interesting for the reader.

Thank you for pointing this out, we have included these details in our discussion.

- capitalization of words in the title is not consistent; why is "Emergence" with a capital

We have fixed this.

- page 1
-- Recently, some strategies...can manipulate some... -> remove 'some'

- page 2
-- The strongest resistors specifically evolve or have -> possess
-- Some work has looked -> A few works looked...
-- http link should be put in a footnote, and using \url{} (if you are writing in Latex)

- page 4
-- ...to analytical results in some cases -> a few selected cases

- page 5
-- The difference is that CS will
defect after the handshake if the opponent defects while
handshake will not. -> last Handshake should be capitalized, to separate the player from the
-- Tables II -> Table II

- page 8
-- This will be explored
further in the next section, looking not only at x 1 and x N 1
but also consider x N/2 . -> considering...

- page 10
-- procceses -> processes
-- TABLE VI: replace 'some' with 'a few selected'

All of these have been addressed.

- page 3
-- why did you present TF1, TF2, TF3 in reverse ordering?

We were not too sure what happened here, looking at our version the ordering was correct. We will specifically inspect the proofs upon submission.

-- Using this it is a known results -> I could not understand what you mean here

-- These
are no longer a good match which highlights the weakness of
assuming a given interaction between two IPD strategies can be
summarised with a set of utilities as shown in (1). -> this phrase is also not clear

We have addressed both of these comment by rephrasing.

3 Comments from second referee

> 1. Title -- I don't think your title represents your paper well. It isn't really about cooperation and it uses the Moran process rather than studying it. I suggest:
>
> A Comprehensive Empirical Study Of Iterated Prisoner's Dilemma Strategies Using The Moran Process
>
> Or, if you prefer to focus on results:
>
> Iterated Prisoner's Dilemma Strategies With Handshakes Are More Evolutionarily Stable
>
> Whatever you decide on, make sure all words in the title are capitalized.

Many thanks for the suggestion, having reflected on this and as this wasn't raised by any of the other reviewers we have decided to keep with the original title. We feel that the self recognition mechanisms serve as a way to reinforce cooperation which we would prefer to keep in the title. If you feel particularly strongly about this we would re consider.

> On page 3 you give noise values for TF1, TF2, and TF3. Some explanation of this would be nice. Also, how did you choose TF1, TF2, and TF3? Were they chosen randomly from many runs or were they chosen based on some selection principle?

The paragraphs just before 'There are three further strategies trained specifically for this study' describe the training mechanism used to obtain TF1, TF2 and TF3. Given that the Moran processes used are not in a noisy environment

we did not feel the need to discuss this further (2/3 of the trained strategies were not trained in noise).

> 3. Figures 2, 3, 4 -- hard to see arrow heads, make them larger
>
> 4. Figures 5 and 6 -- These are too small. It would be better to show fewer examples at a readable size.
>
> 5. Figure 13 -- text labeling axes and scale is too small.
>
> 6. Figure 14 -- Heat map labels should be larger as well as the x-axis label "Rounds." Given space restrictions, it is not possible to make the strategy names larger, but on your other figures it is possible to enlarge the image to read the strategy names and on these figures it is not. Maybe change the file format or the resolution.
>
> 7. Typos
>
> Section IIIB: "given in Tables II" should be "Table II"
>
> Section IIIC: "There are are only two new strategies" -- cut repeated "are"
>
> Section IIID: "Figure 13 show" should be "shows"
>
> Section V: "allowing the each strategy" -- cut the word "the"

We have hopefully addressed all these points satisfactorily: making images larger and more readable as well as picking up the typos. In particular for the heatmap, we have enlarged the text and simply removed the name labels (they are not important). The main point being to see the initial play handshake, and to this end we have chosen to only display 15 rounds (as opposed to 200). Thank you for bringing this to our attention.

4 Comments from third reviewer

> The first overarching comment is on the density and unreadability of the figures. Displaying as much data as possible is laudable, but to my eyes the condition "as possible" no longer holds.

We agree and have (as detailed in previous comments) hopefully addressed this.

> Particular examples: Fig. 5 is unreadably small, and without extreme magnification it is impossible to read the titles/axes/any words in the subfigure graphics. Note that this *can* be remedied by repeating such

information outside of the graphics, and explicitly writing that "the red lines are x_n-1, x-axis is N from 2 to 14, y-axis is fixation probability from 0 to 1" in the caption would allow those graphs to be understood.

These pointers have been added to the caption but also the number of these specific plots have been reduced which also hopes with clarity.

For this, tables may well be more readable in conjunction with the figures. Fig. 2-4 are on the edge of being uninterpretable, and a transition table would also help with implementing the strategy.

We have improved the size of Fig 2-4. We have however chosen to stick with diagrams as this is often the way they are represented in the literature (for example in Ashlock's papers). Note that we've added a description of the interpretation of the figures. If you feel strongly about a table we will reconsider.

Fig. 13,14 have unreadable axes. Additionally, Fig.7,9,11 are not only unreadable, but nowhere is it described what the error bars are (I am presuming you treat "fixation probability against opponent X" as one data point and are displaying stats over X, but other possibilities abound). Same goes for Fig 14: what is the cooperation rate per round? Is it fraction of C played by that strategy (as opposed to payoff R) observed in round n (out of how many?), over 10k trials of one single head-to-head match?

Fig 13,14 have improved axes but also are now improved due to the fact that only 15 rounds are presented. As well as this, the requested explanations have all been included.

Another general observation is with the strategy listing. The "memory depth", in number of rounds of history necessary, is given. This is one particular way of considering strategies, but in my opinion a far more important measure is minimal number of states in a FSM representation. An example is given where GRIM(Grudger) would need infinite memory *as number of rounds input to a lookup table*, but clearly GRIM needs just one bit of memory: whether the opponent has ever defected, which obviously maps to its 2^1 states. Allowing the FSM to specify what information it stores is much more flexible. I strongly suggest including the state size as well.

We have added the number of states in all FSM strategies to this listing. Note that we have not attempted to quantify the minimal state representation of all strategies (this in itself would be an impressive piece of work outside of the scope of this paper).

Several base details of the experiments which I would need to understand, particularly in Algorithm 2:
* rounds per game in the actual IPD played
I cannot locate this fundamental number, which matters (see DOI

10.1109/CIG.2015.7317950)
* results <- cache(match)
There is a discussion that stochastic strategies exist, and the interpretation of Fig. 6 implies the result of a matchoff remains stochastic (as opposed to taking expected score). Does this mean the cache is a probabilistic function? This is unexpected and should be noted as well in the algorithm.
* parent <- selected randomly in proportion to total match payoffs
Each individual is selected randomly in proportion to *its* total match payoffs against every other individual? This could be clearer
* kill off <- random player from population
Is this uniformly at random? This should be specified

This have all been addressed, in particular we have added to the brief explanation of the sampling that occurs from the cached results but made this very explicit.

I must take particular issue with the last paragraph of Section II. This makes two significant claims with implications for future work:
1. Sampled fixation probabilities for deterministic strategies are a "perfect match" with the theoretical probabilities computed above (the latter is strongly appreciated)
2. Such sampled fixation probabilities for stochastic strategies are "no longer a good match" with the theoretical probabilities computed using (presumably) expected payoff between the pairs

I cannot accept either statement given the current manuscript, nor accept the manuscript without them. To claim 1., a cursory glance at Fig. 5(a,b,e) belie a perfect match. To claim 2., Fig. 6(a,b,f) look better than the three aforementioned graphs. The issue is, of course, that these probabilities are (each) binomially-distributed samples. Wherever there is sampling, and matching to theoretical is mentioned, I expect a statistical test. Here, I request confidence intervals computed using Bonferroni- (or equivalent) corrected binomial models be at least displayed in Fig. 5-6. Claim 2., which seems true, requires a corrected exact test against H0.

We feel that the reviewer has made an excellent point. To make things clearer we have chosen to only show one pair of strategies for both types (deterministic and stochastic) and have:

- Added simulated confidence bars obtained using an asymptotic normal approximation.
- Carried out a t test over all values in the plots which show the level of significance of the differences (for stochastic) and similarities (for deterministic) of the results.

We contemplated including similar plots/tables that show similar outcomes to the supplementary materials but felt that this was perhaps not necessary and would be an overflow of information.

Firstly, the extremely wide error bars (many are basically 0-1) imply, as we all know, that the relation "X beats Y at a particular type of IPD competition" is not even close to transitive. Computing the mean fixation probability *against a particular set of opponents* is highly sensitive to the composition of the competition, in the instant case "wide selection across the literature" with unknown weighting. There is no reason to accept the authors' choice, indeed any choice, as authoritative.

I strongly suggest that particular subgroups of strategies (ZD, simple FSM, handshakers, etc.) be evaluated in block fashion against other subgroups. This would allow statements of the form "ZD-type strategies fail (relative to neutral) to invade [group] at N=? pairwise", which is alluded to in the paper, with the bonus of readable data for each claim. Please ensure statistical analysis for these are multiple-testing corrected.

We acknowledge the point made here by the referee: it is a good one. We would suggest that as well as the large number of strategies they also represent quite a diverse selection. However the point made by the referee stands. We have included this point in our discussion. Note however that in terms of carrying out a subgroup analysis, we feel that this would not fall within the scope of this work. To do it properly would constitute a fundamental rewrite of all the results. This has been noted in the discussion, including the important point that by making our data available anyone could do this (and we hope to do so ourselves).

Secondly, with the results as computed globally, reporting the mean is insufficient for these wide ranges; for each of Fig. 7(b) and Tables II-III other generic descriptive stats are necessary.

A number of further statistics have been reported.

Minor detail: references [11] and [13] are the same work. Formatting of the references in general is also worth checking again.

This has been done.