444 Lecture 13

Prisoners Dilemma

Brian Weatherson

February 21, 2023

Day Plan

Prisoners' Dilemma

Basic Challenge:

- Each player is better off defecting;
- The players are collectively better off if both cooperate.

Tragedy of the Commons

- In a two-player setting, we normally call this Prisoners' Dilemma, or PD.
- In a multi-player setting it's sometimes called the Tragedy of the Commons.

Tragedy of the Commons

- The story (which is probably wildly ahistorical) is that everyone grazed their herds on the commons - which was a good thing to do or else the herd would die - but collectively this made the commons unusable.
- And in the standard story, private property was the solution to the tragedy.

Social Challenge

- How do we get to cooperation?
- First question is whether in this case we should want to get to cooperation.
- Second question is whether this really is PD.
- Let's assume that the answer in each case is *yes*, what do we do.

Change the Payouts

One possible social response is to change the payouts.

- *Snitches get stiches* is kind of a version of this response.

Change the Options

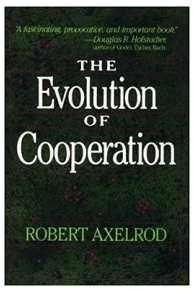
Another is to make it just impossible for everyone to do the defecting move.

- Enclosures are sort of like this.
- Just like with signaling games, the difference between making something expensive and making it impossible is a little vague, but it's useful conceptually to think of them as separate options.

Iterate the Game

- But the simplest way to handle this kind of problem is to iterate the game.
- Arguably it is in everyone's interests to be cooperative if they will have to interact with the other players repeatedly.

Robert Axelrod



Axelrod's Famous 1984 Book

The One Shot Game

Axelrod worked with this version of Prisoners' Dilemma (PD).

	c	d
C	3, 3	0, 5
D	5, 0	1, 1

Indefinite Iteration

In the fancier version of the game, he didn't tell people how long the game would go.

- Instead he just said there was a probability of it ending after each round; if I recall 0.005.
- This was used to avoid backwards induction reasoning.
- It turned out not to really mater a ton; no one uses backward induction reasoning in practice. But it's theoretically useful.

The Tournament

- There are *n* strategies submitted.
- Strategies are not quite full strategies in our sense; they just say what to do given what the other player did. (They don't account for possible errors in their own performance.)
- Each will play *k* rounds of PD with each of the other *n*-1 strategies.
- Their payouts will add up over the $k(n-1)$ rounds and the one with the highest total will win.

Cooperative and Competitive

- This is not entirely a cooperative game; ultimately if I'm a strategy I want to win, and that means I want the other strategy I'm interaction with to lose.
- But in the short run there is much to be gained by improving our mutual position vs the other $n-2$ strategies.
- So in the short run there is a benefit to cooperation, even if we're ultimately rivals.

Iterated Axelrod Game

- Axelrod famously ran a tournament just like the one described here.
- But we can iterate the whole tournament in an interesting way.
- Of course the Axelrod tournament involves iterating PD within each 'round'; the idea now is to play multiple rounds.

Iterated Axelrod Game

- Imagine at the start each strategy is $1/n$ of the overall 'population'.
- After playing all these games, where each strategy plays $k(n-1)$ versions of PD, each strategy gets a score.
- In the next round, it's share of the population is a function of (a) its initial population, and (b) its score in this round.
- And in future rounds, one's score is a weighted average of how well one does in games against the other strategies, where the weights are given by their populations.

Evolution of Cooperation

- This is a useful model for thinking about the phenomena in the title of Axelrod's book: The Evolution of Cooperation.
- We want strategies that do well not just when the world consists of random strategies, but when the world consists of strategies that themselves could have survived at least a little bit of evolution.

Evolution of Cooperation

- Theoretically this could make a difference.
- Strategies that exploit dumb strategies could do well initially, but then fade away.
- Alternatively, some strategies could do badly against bad strategies, but if they survive initial rounds, do well when there are sophisticated strategies around.

Spatial Evolution

- To be even more realistic, you could imagine that each strategy lives 'somewhere' in a large grid.
- And at each round, each strategy plays with a weighted average of strategies that live nearby.
- This really does make a difference; some strategies that aren't great against the world in general are fairly immune to invasion, and can even expand their territory under a range of conditions.

Day Plan

The Basic Game

Axelrod

Overview

This lecture covers some of the lessons from the Iterated Prisoners' Dilemma tournaments that Michigan professor Robert Axelrod ran in the early 1980s.

Four Papers

- Effective Choice in the Prisoner's Dilemma, *Journal of Conflict Resolution* 24 (1980): 3-25.
- More Effective Choice in the Prisoner's Dilemma, *Journal of Conflict Resolution* 24 (1980): 379-403.
- The Emergence of Cooperation among Egoists, *The American Political Science Review* 75 (1981): 306-318.
- The Evolution of Cooperation with William Hamilton, *Science* 211 (1981): 1390-1396.

The First Tournament

- Axelrod advertised the first round of his tournament, and called for submissions.
- This was far from trivial in pre-internet days, and he only got 13 submissions.
- In the first tournament he said that k would be 100, but no one actually exploited that fact.

The Winner

Tit-for-Tat

Tit-for-Tat

Two rules.

1. Play C at round 1.
2. In all subsequent rounds, do whatever the other player just did.

The Second Tournament

- So Axelrod wrote this up, including saying who won.
- He called for more submissions, and now got 66.
- Some of these were typed, some came to Ann Arbor on the huge magnetic disks that were used way back then.
- He ran the tournament again, this time with a random number of rounds.
- And Tit-for-Tat won again.

Logic and Victory

- This doesn't mean Tit-for-Tat is the best strategy.
- Indeed, in each tournament it was easy in retrospect to describe strategies that would have beaten everyone, including TFT, if they had been entered.
- But still, it's pretty impressive.

Four Features

Tit-for-Tat has five striking characteristics, each of which was positively correlated with success in the tournaments.

- Nice
- Provocable
- Forgiving
- Not envious
- Simple

Nice

The clearest distinction in the tournament was between strategies that were Nice and those that were Nasty.

- By definition, a strategy is Nice iff it is never the first to defect.
- You don't have to be very nice in the intuitive sense to count as Nice.

Grim Trigger

Here is one nice strategy, one Axelrod calls Grim Trigger.

1. Cooperate on move 1.
2. If the other player ever defects, defect on every subsequent move.

This strategy did really badly; it was the worst Nice strategy in round 2. But still many Nasty strategies did worse.

Nice Strategies

- In the evolutionary versions of the game, there can be a tendency for strategies to tend towards being Nice.
- Then evolution stops, because when two Nice strategies meet, the payout is inevitably 3k to each.
- Although the best strategies are all Nice, it is how they interact with Nasty strategies that determines who wins.

Provocable

- It's bad to get pushed around.
- Nasty strategies are always looking for how much they can get away with.
- So you want to send a clear message that defections will not be tolerated.
- Obviously TFT does that.

Forgiving

- But you don't want to be Grim Trigger.
- It's bad to be pushed around, but it's not much better to end up in all defect land.
- You need a way back to all cooperate land.
- TFT has that, though notably it isn't perfect at this.
- TFT can get into CD-DC-CD-etc cycles with a bunch of strategies.

Not Envious

- In any interaction, TFT never does better than who it is playing with.
- Yet it comes out first overall.
- This is kind of amazing.
- It just does not care at all about winning against who it is facing off with.

Not Envious To a Fault

- Note that TFT doesn't always do that well in evolutionary games.
- This is because it might take this a bit too far.
- It doesn't look to exploit weaknesses in opponents.

Simple

- Other strategies try to figure out what their rivals are doing.
- They normally get this wrong.
- Or they try and send complex signals.
- These are usually misinterpreted.
- TFT keeps things simple, and doesn't lose points messing around looking for any edges.

Variant Games

- The most interesting variant to me is the one where a strategy only gets implemented with probability 0.99 on each move.
- Sometimes there are performance errors.
- TFT does terribly in this; it can't get out of randomly generated defection cycles.
- In this kind of game you need to be a bit more forgiving.
- But also you can try to get away with a bit more; if the other person will treat a defection as random, you can plan a few.

Rest of Day

- I'm not going to do slides about Oyun.
- But the plan for the rest of the day is to go over the assignment, and talk about how the tournament software works.