The Darwin Game

The Darwin Game

The Darwin Game

Epistemic Status: True story

The plan is that this post will begin the sequence Zbybpu'f Nezl.

In college I once took a class called Rational Choice. Because obviously.

Each week we got the rules for, played and discussed a game. It was awesome.

For the grand finale, and to determine the winner of the prestigious Golden Shark Award for best overall performance, we submitted computer program architectures (we'd tell the professor what our program did, within reason, and he'd code it for us) to play The Darwin Game.

The Darwin Game is a variation slash extension of the iterated prisoner's dilemma. It works like this:

For the first round, each player gets 100 copies of their program in the pool, and the pool pairs those programs at random. You can and often will play against yourself.

Each pair now plays an iterated prisoner's dilemma variation, as follows. Each turn, each player simultaneously submits a number from 0 to 5. If the two numbers add up to 5 or less, both players earn points equal to their number. If the two numbers add up to 6 or more, neither player gets points. This game then lasts for a large but unknown number of turns, so no one knows when the game is about to end; for us this turned out to be 102 turns.

Each pairing is independent of every other pairing. You do not know what round of the game it is, whether you are facing a copy of yourself, or any history of the game to this point. Your decision algorithm does the same thing each pairing.

At the end of the round, all of the points scored by all of your copies are combined. Your percentage of all the points scored by all programs becomes the percentage of the pool your program gets in the next round. So if you score 10% more points, you get 10% more copies next round, and over time successful programs will displace less successful programs. Hence the name, The Darwin Game.

Your goal is to have as many copies in the pool at the end of the 200th round as possible, or failing that, to survive as many rounds as possible with at least one copy.

If both players coordinate to split the pot, they will score 2.5 per round.

To create some common terminology for discussions, 'attack' or means to submit 3 (or a higher number) more than half the time against an opponent willing to not do that, and to 'fold' or 'surrender' is to submit 2 (or a lower number) more than half the time, with 'full surrender' being to always submit 2. To 'cooperate' is to alternate 2 and 3 such that you each score 2.5 per round.

In this particular example we expected and got about 30 entries, and I was in second place in the points standings, so to win The Golden Shark, I had to beat David by a substantial amount and not lose horribly to the students in third or fourth.

What program do you submit?

(I recommend actually taking some time to think about this before you proceed.)

Some basic considerations I thought about:

1. The late game can come down to very small advantages that compound over time.

2. You need to survive the early game *and* win the late game. This means you need to succeed in a pool of mostly-not-smart programs, and then win in a pool of smart programs, and then in a pool of smart programs that outlasted other smart programs.

3. Scoring the maximum now regardless of what your opponent scores helps you early, but kills you late. In the late game, not letting your opponent score more points than you is very important, especially once you are down to two or three programs.

4. In the late game, how efficiently you cooperate with yourself is very important.

5. Your reaction to different programs in the mid game will help determine your opponents in the end game. If an opponent that outscores you in a pairing survives into the late game, and co-operates with itself, you lose.

6. It is all right to surrender, even fully surrender, to an opponent if and only if they will be wiped out by other programs before you become too big a portion of the pool, provided you can’t do better.

7. It is much more important to get to a good steady state than to get there quickly, although see point one. Getting people to surrender to you would be big game.

8. Some of the people entering care *way* more than others. Some programs will be complex and consider many cases and be trying hard to win, others will be very simple and not trying to be optimal.

9. It is hard to tell what others will interpret as cooperation and defection, and it might be easy to accidentally make them think you’re attacking them.

10. There will be some deeply silly programs out there at the start. One cannot assume early programs are doing remotely sensible things.

That leaves out many other considerations, including at least one central one. Next time, I’ll go over what happened on game day.

The Darwin Pregame

Epistemic Status: True story

This is intended as post two of the sequence Zbybpu'f Nezl.

Previously (required): The Darwin Game

I

This is my reconstruction of my thoughts at the time.

The Darwin Game requires surviving the early, middle and late games.

In the opening, you need to maximize scoring against whatever randomness people submit. Survival probably isn't enough. The more copies of yourself you bring to the middle game, the more you face yourself, which snowballs. Get as many points as you can.

In the middle game, you face whatever succeeded in the opening. Strategies that survived the opening in bad shape can make a comeback here, if they are better against this new pool. What strategies do well *against you* matters.

In the end game, you'll need to beat the successful middle game strategies, all of which have substantial percentages of the pool. Eventually you'll be heads up against one opponent. Not letting opponents outscore you in a pairing becomes vital.

How would the game play out? What types of strategies would thrive?

I divided the types as follows:

There were attackers, who would attempt to get the opponent to accept a 3/2 or 4/1 split. They might or might not give up on that if you refused, and presumably most would use a signal to self-cooperate, but not all. One person did submit "return 3."

Then there were cooperators, who attempt to split the pot evenly. I assumed that meant alternating 3/2 splits. This then divided into those who would fold if attacked, allowing you to score above 2.5 per turn, those that would let themselves be outscored but would make sure you scored less than 2.5 per turn, and those that would not allow themselves to be outscored. The last group might or might not forgive an early attempt to attack them.

There would also be bad programs. People do dumb things. Someone might play all 2s, or pick numbers fully at random, or who knows what else.

As a list (attackers from here on means both AttackBot and BullyBot):

AttackBot. Attackers who don't give up.

BullyBot. Attackers who give up.

CarefulBot. Cooperators who harshly punish attackers.

DefenseBot. Cooperators who don't let you outscore them but don't otherwise punish.

EquityBot. Cooperators who let you outscore them, but make sure you don’t benefit.

FoldBot. Cooperators who accept full unfavorable 3/2 splits.

GoofBot. Weird stuff.

My prior was we’d see all seven, with most looking to cooperate.

Was attacking a good strategy?

Attacking only works against FoldBots. When attacking fails, even DefenseBots might take a while to re-establish cooperation. CarefulBots could wipe you out. It was also impossible to know how long to keep attacking before concluding opponents weren’t going to fold.

With a pool of bots chosen by humans, attacking strategies (AttackBot or BullyBot) likely would fail hard in the opening.

The endgame was a different story. All GoofBots would be dead. Unless FoldBots fold too quickly to a BullyBot, in a given round they strictly outscore CarefulBots, DefenseBots and EquityBots. Each round, provided they exist, FoldBots would become a bigger portion of the cooperative pool. If you were an AttackBot or BullyBot, and survived long enough, you would kill off the CarefulBots, then the DefenseBots and finally the EquityBots as the FoldBots out-competed them, leaving a world of AttackBots, BullyBots and FoldBots. If all but one attacker was gone, the last attacker to survive would win if it cooperated efficiently against itself, since it would score above average each round. In theory a steady state could exist with multiple attackers keeping each other in check, but that isn’t stable since advantages in size snowball.

CarefulBots are strictly worse than DefenseBots, so those were out. GoofBots are terrible.

This meant there were five choices:

I could submit an AttackBot that cooperates with itself, and hope to survive into the endgame. I quickly dismissed this as unlikely to work.

I could submit a BullyBot that cooperates with itself, attacks but accepts an even split against stubborn opponents. But this rewards stubborn opponents while wiping out non-stubborn opponents in the mid-game, which means your endgame trump card stops working. I dismissed this as well.

DefenseBots don’t lose heads-up by non-tiny amounts, and punish anyone who tries to outscore them, wiping them out in the mid-game. But you score nothing against AttackBots in the opening, before you can shape the pool much. At best you take a smaller pool into the mid-game, where efficient cooperation with your own copies starts to snowball.

I saw the *emotional* appeal of DefenseBots, but using one didn’t made sense. Its defenses were too robust and expensive, and you still lose to a smart AttackBot heads-up if you’re outnumbered. I’d need to take more risk.

That was the problem with being a FoldBot. FoldBots feed attackers. You are free riding on the rest of the cooperative pool. You hope they kill attackers despite that. The problem is that if even one copy of an attacker survives, as you and other

FoldBots grow strong, attacking becomes a better and better strategy. I decided this wasn't worth that risk.

I would submit an EquityBot. I wouldn't protect against them *outscoring me.* I would protect against them *outscoring what cooperation would have gotten them.* If at any point they wanted to split the remaining pie, I would accept. Even if they refused, I'd give them *some* points on a 3/2 split, so long as they were punished for it, and I wasn't growing their portion of the pool.

This raised the threshold percentage of the pool I needed to win heads-up against an attacker, but with a size disadvantage I'd lose no matter what, and I'd still win if I had a sufficiently large size edge, which was more likely if I did better early on.

Too much folding and you strengthen someone who beats you. Too little and you fall behind letting others snowball.

I decided to alternate 3/2 even if my opponent was going 3/3. This said both 'I'm not going to give up' and 'you are welcome to cooperate at any time,' and still punished the opponent reasonably hard. After long enough I even risked throwing in a few more 2s.

I considered sending a signal to recognize myself, but realized there was no point. Better to start coordinating right away. I'd randomize my first turn to 2 or 3, and once my opponent didn't match me I would alternate. I figured opponents would start 2 more often than 3, so I decided to do a 50/50 split to take advantage of that, coordinating faster and with a slight edge, at the expense of doing slightly worse against myself, but this was probably just a mistake and I should have done an uneven split (but not quite the fully maximizing-for-self-play ratio). However, in an endgame against a similar program, you can definitely get an edge by being slightly more willing to play 3s early than your opponent.

Opponents that wanted to cooperate would have a very easy time recognizing my offer and cooperating. That left special case logic.

If my opponent was alternating on the same schedule as me (somehow we started 2/3, but then we'd 2/2 then 3/3 then 2/2), then I'd play 2 twice in a row to break that up. Ideally, if the opponent was offering a different cycle that was fair, I'd match that (so if they went 1/4/1/4, I'd submit 4 next time, and if they did 1 I'd start alternating), but I didn't expect such cases so I didn't make that logic robust, as the professor had already thrown out part of a previous submission for being too complex, and I wanted to preserve the more important parts.

If my opponent was playing all 2s even after I started alternating, I put in logic to play all 3s. If they played even one 3, I'd back down permanently. I also put in logic against a few other bizarre simple bots (like all 1s, all 4s, seems to be completely random, etc) but didn't worry about it too much since they'd be wiped out very quickly and complexity is bad.

If my opponent was playing all 3s without a starting signal, and kept it up long enough, that meant he'd defect against himself, which meant he couldn't win an endgame, and also meant that he was highly unlikely to ever give up, so I'd eventually fold. If they were going to lose in the long run, better to get what I could. Letting them survive longer would only help me.

II

David took a different approach.

David knew about the class mailing list.

David assembled a large group. They agreed to submit 2-0-2 as their first three moves. If both sides sent the signal, they'd cooperate using a reasonable randomization system. If they didn't get the signal back, they'd play all 3s. They'd be pure CliqueBots, cooperating with each other and defecting against everyone else. With a large enough group, they'd wipe out the other players and share the victory. David would win The Golden Shark and his guaranteed A+.

I would find out about the coalition after round one.

III

We were all set for game night. We had each chosen the logical output of our decision functions. The professor set up a website where we could see the game played out in real time over the course of several hours (due to a combination of that's more fun and the game was slow to run), with a discussion board for him to offer observations and us to comment.

Next time I'll reveal what happened on game night. Predictions are encouraged.

The Darwin Results

Epistemic Status: True story (numbers are best recollections)

This is post three in the sequence Zbybpu’f Nezl.

Previously (required): The Darwin Game, The Darwin Pregame.

I

It was Friday night and time to play The Darwin Game. Excited players gathered around their computers to view the scoreboard and message board.

In the first round, my score went up slightly, to something like 109 from the starting 100. One other player had a similar score. A large group scored around 98. Others did poorly to varying degrees, with one doing especially poorly. That one played all 3s.

Three, including David, shot up to around 130.

If it isn't obvious what happened, take a minute to think about it before proceeding.

II

The CliqueBots had scores of 98 or so. They quickly figured out what happened.

David lied. He sent the 2-0-2 signal, and cooperated with CliqueBots, but instead of playing all 3s against others, he and two others cooperated with others too.

Whoops.

CliqueBots had been betrayed by MimicBots. The three defectors prospered, and the CliqueBots would lose.

Without those three members, the CliqueBots lacked critical mass. Members would die slowly, then increasingly quickly. If the three defectors had submitted CliqueBots, the CliqueBots would have grown in the first round, reaching critical mass. The rest of us would have been wiped out.

Instead, the three defectors would take a huge early lead, and the remaining members would constitute, as our professor put it, their 'packed lunch.'

The opening consisted of CliqueBots being wiped out, along with G-type weirdos, A-type attackers and D-style cooperators that got zero points from the CliqueBots.

Meanwhile, on the message board, the coalition members were *pissed.*

III

Everyone who survived into the middle game cooperated with everyone else. Victory would come down to efficiency, and size boosted efficiency. Four players soon owned the entire pool: Me and the three defectors.

I thought I had won. The coalition members wasted three turns on 2-0-2. Nothing could make up for that. My self-cooperation was far stronger, and I would outscore

them over the first two rounds when we met due to the 0. It wouldn't go fast, but I would grind them out.

It did not work out that way. David had the least efficient algorithm and finished fourth, but I was slowly dying off as the game ended after round 200. Maybe there was a bug or mistake somewhere. Maybe I was being exploited a tiny bit in the early turns, in ways that seem hard to reconcile with the other program being efficient. I never saw their exact programs, so I'm not sure. I'd taken this risk, being willing to be slightly outscored in early turns to better signal and get cooperation, so that's probably what cost me in the end. Either way, I didn't win The Darwin Game, but did survive long enough to win the Golden Shark. If I hadn't done as well as I did in the opening I might not have, so I was pretty happy.

IV

Many of us went to a class party at the professor's apartment. I was presented with my prize, a wooden block with a stick glued on, at the top of which was a little plastic shark, with plaque on the front saying Golden Shark 2001.

Everyone wanted to talk about was how awful David was and how glad they were I had won while not being him. They loved that my core strategy was so simple and elegant.

I tried gently pointing out David's actions were utterly predictable. I didn't know about the CliqueBot agreement, but I was deeply confused how they didn't see this 'betrayal' coming a mile away. Yes, the fact that they were only one or two CliqueBots short of critical mass had to sting, but was David really going to leave all that value on the table? Even if betraying them hadn't been the plan all along?

They were having none of it. I didn't press. Why spoil the party?

V

Several tipping points could have led to very different outcomes.

If there had been roughly two more loyal CliqueBots, the CliqueBots would have snowballed. Everyone not sending 2-0-2 would have been wiped out in order of how much they gave in to the coalition (which in turn accelerates their victory). Betrayers would have bigger pools, but from there all would cooperate with all and victory would come down to if anyone tweaked their cooperation algorithms to be slightly more efficient. David's betrayal may have cost him the Golden Shark.

If someone had said out loud "I notice that anyone who cares about winning is unlikely to submit the CliqueBot program, but instead will start 2-0-2 and then cooperate with others anyway" perhaps the CliqueBots reconsider.

If enough other players had played more 2s against the CliqueBots, as each of us was individually rewarded for doing, the CliqueBots would have won. If the signal had been 2-5-2 instead of 2-0-2, preventing rivals from scoring points on turn two, that might have been enough.

If I had arrived in the late game with a slightly larger pool, I would have snowballed and won. If another player submits my program, we each end up with half the pool.

Playing more 2s against attackers might have won me the entire game. It also might have handed victory to the CliqueBots.

If I had played a better split of 2s and 3s at the start, the result would have still depended on the exact response of other programs to starting 2s and 3s, but that too might have been enough.

Thus these paths were all possible:

The game ended mostly in MimicBots winning from the momentum they got from the CliqueBots.

It could have ended in an EquityBot (or even a DefenseBot) riding its efficiency edge in the first few turns to victory after the CliqueBots died out. Scenarios with far fewer CliqueBots end this way; without the large initial size boost, those first three signaling turns are a killer handicap.

It could have ended in MimicBots and CliqueBots winning together and dividing the pool. This could happen even if their numbers declined slightly early on, if they survived long enough while creating sufficient growth of FoldBot.

CliqueBots could have died early but sufficiently rewarded FoldBots to create a world where a BullyBot could succeed, and any BullyBot that survived could turn around and win.

It could have had MimicBots and CliqueBots wipe out everyone else, then ended in victory for *very subtle* MimicBots, perhaps that in fact played 3s against outsiders, that exploited the early setup turns to get a tiny edge. Choosing an algorithm that can't be gamed this way would mean choosing a less efficient one.

In various worlds with variously sized initial groups of CliqueBots and associated MimicBots, and various other programs, the correct program to submit might be a CliqueBot, a MimicBot that attacks everyone else but cheats on the coordination algorithm, a MimicBot that also cooperates with others, a BullyBot with various tactics, an EquityBot with various levels of folding, or an FoldBot. There are even scenarios where *all marginal submissions lose,* because the program that would win without you is poisoning the pool for its early advantage, so adding another similar program kills you both.

This is in addition to various tactical settings and methods of coordination that depend on exactly what else is out there.

Everyone's short term interest in points directly conflicts with their long term goal of having a favorable pool. The more you poison the pool, the better you do now, but if bots like you poison the pool too much, you'll all lose.

There is no 'right' answer, and no equilibrium.

What would have happened if the group had played again?

If we consider it only as a game, my guess is that this group would have been unable to trust each other enough to form a coalition, so cooperative bots in the second game would send no signal. Since cooperative bots won the first game, most entries would be cooperative bots. Victory would likely come down to who could get a slight edge during the coordination phase, and players would be tempted to enter true

FoldBots and otherwise work with attackers, since they would expect attackers to die quickly. So there's some chance a well-built BullyBot could survive long enough to win, and I'd have been tempted to try it.

If we include the broader picture, I would expect an attempt to use out-of-game incentives to enforce the rules of a coalition. The rise of a true CliqueBot.

VI

I spent so long on the Darwin Game story and my thinking process about it for several reasons.

One, it's a fun true story.

Two, it's an interesting game for its own sake.

Three, because it's a framework we can extend and work with, that has a lot of nice properties. There's lots to maximize and balance at different levels, no 'right' answer and no equilibrium. It isn't obvious what to reward and what to punish.

Four, it naturally ties your current decisions to your future and past decisions, and to what the world looks like and what situations you will find yourself in.

Five, it was encountered 'in the wild' and doesn't involve superhuman-level predictors. A natural objection to models is 'you engineered that to give the answer you want'. Another is 'let's figure out how to fool the predictor.' Hopefully minimizing such issues will help people take these ideas seriously.

There are many worthwhile paths forward. I have begun work on several. I am curious which ones seem most valuable and interesting, or where people think I will go next, and encourage such discussion and speculation.