Learning to Cooperate and Compete via Self Play

Noam Brown
The scientist named the population, after their distinctive horn, Ovid’s Unicorn.
A popular strategy game from the 50s

- 7 players trying to conquer Europe in WW1
- JFK and Kissinger's favorite game

- Each turn involves **private natural language negotiation**
- Moves are done simultaneously
  - e.g. F CLY - NWG, A DEN H, F SKA S A SWE – NWY, ...

- Alliances and trust-building are key!

- Long considered a **challenge problem for AI** [1]
  - Research going back to the 80’s
  - Research picked up in 2019 with work from MILA, DeepMind, ourselves, others

---

[1] Dafoe et al. "Cooperative AI: machines must learn to find common ground". Nature comment, 0/2021
If you’ve ever heard of Diplomacy, chances are you know it as “the game that ruins friendships.” It’s also likely you’ve never finished an entire game. The game is an all-too-real metaphor for political Realpolitik, and its brutality is the reason why few players ever finish it. Yet its influence is as far-reaching as its brutality. Diplomacy: The Map That Ruined a Thousand Friendships

HENRY GRABAR  MARCH 7, 2013

Diplomacy: The Most Evil Board Game Ever Made

Haoran Un | Nov 10, 2017 10:30am · Filed to: board games ·
“Diplomacy is ultimately about building trust in an environment that encourages you to not trust anyone.”

-Andrew Goff
3-Time Diplomacy World Champion
Self-Play in 2p 0-Sum Games
Who is the better poker player?

Option 1: Someone who, over a large enough sample size, wins head-to-head vs. any other player

Option 2: Someone who makes more money playing poker than anyone else
Who is the better poker player?

**Minimax Equilibrium**
Option 1: Someone who, over a large enough sample size, wins head-to-head vs. any other player

**Population Best Response**
Option 2: Someone who makes more money playing poker than anyone else
Minimax Equilibrium

Minimax Equilibrium in 2p0sum: each player’s strategy is optimal given the other player’s policy

In balanced games, playing minimax ensures you will not lose on average

**Exploitability**: How much we’d lose to a best response

<table>
<thead>
<tr>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Us</td>
<td>![Rock]</td>
<td>![Rock]</td>
</tr>
<tr>
<td>Best Response</td>
<td>![Paper]</td>
<td>![Paper]</td>
</tr>
</tbody>
</table>

Our Exploitability = 1
Minimax Equilibrium

Minimax Equilibrium in 2p0sum: each player’s strategy is optimal given the other player’s policy

In balanced games, playing minimax ensures you will not lose on average

**Exploitability**: How much we’d lose to a best response

<table>
<thead>
<tr>
<th></th>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Us</td>
<td><img src="image1" alt="Us" /></td>
<td><img src="image2" alt="Us" /></td>
<td><img src="image3" alt="Us" /></td>
</tr>
<tr>
<td>Best Response</td>
<td><img src="image4" alt="Best Response" /></td>
<td><img src="image5" alt="Best Response" /></td>
<td><img src="image6" alt="Best Response" /></td>
</tr>
</tbody>
</table>

Our Exploitability = 1
Minimax Equilibrium

Minimax Equilibrium in 2p0sum: each player’s strategy is optimal given the other player’s policy

In balanced games, playing minimax ensures you will not lose on average

Exploitability: How much we’d lose to a best response
Minimax Equilibrium

“Poker is simple, as your opponents make mistakes, you profit.”
-Ryan Fee’s Poker Strategy Guide

<table>
<thead>
<tr>
<th></th>
<th>Round 1</th>
<th>Round 2</th>
<th>Round 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Us</strong></td>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
</tr>
<tr>
<td><strong>Best Response</strong></td>
<td><img src="image4" alt="Image" /></td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
</tr>
</tbody>
</table>

Our Exploitability = 0
Self-play in two-player zero-sum games

- In **self-play**, an agent gradually improves by playing against copies of itself.

- Initial strategy can be completely random.

- In balanced **two-player zero-sum** games, **sound self-play** provably converges to a **minimax equilibrium**.

- Thus, given sufficient memory and compute, any **finite two-player zero-sum game** can be “solved” via self-play.
Self-play in two-player zero-sum games

- In **self-play**, an agent gradually improves by playing against copies of itself
- Initial strategy can be completely random
- In balanced **two-player zero-sum** games, **sound self-play** provably converges to a **minimax equilibrium**
- Thus, given sufficient memory and compute, **any finite two-player zero-sum game** can be “solved” via self-play
**Question:** Why is self play limited to two-player zero-sum games?

**Answer:** Because outside two-player zero-sum games, unlimited memory and compute isn’t enough. You may need human data as well!
Ultimatum Game

- Alice is given $100
- Alice must offer $0 - $100 to Bob
- Then, Bob must decide whether to accept or reject
  - If Bob accepts, then Alice and Bob keep their money
  - If Bob rejects, then Alice and Bob get nothing

\[ \text{Proposer} \]
\[
\begin{array}{c}
\text{stake}: x \\
\text{offer}: y \\
\end{array}
\]
\[ \text{Responder} \]
\[
\begin{array}{c}
\text{accept} \\
x-y \\
y \\
\text{reject} \\
0 \\
0 \\
\end{array}
\]

That was an incredibly counterintuitive result to nobody but economists. The humans aren’t doing what the math says. The humans must be broken.
DORA: No-press Diplomacy from Scratch [1]

- DORA learns no-press Diplomacy through self-play
  - Similar to AlphaZero

- Performance with humans in 2-player no-press Diplomacy:
  - **Win rate**: 86.5% +/- 6.1% vs human experts

- Performance with bots in 7-player no-press Diplomacy:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1x ↓ vs 6x →</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DipNet [24]</td>
<td></td>
<td>0.8%±0.4%</td>
<td>0.0%±0.0%</td>
<td>0.1%±0.0%</td>
</tr>
<tr>
<td>SearchBot [11]</td>
<td>49.4%±2.6%</td>
<td>-</td>
<td>1.1%±0.4%</td>
<td>0.5%±0.2%</td>
</tr>
<tr>
<td>DORA</td>
<td>22.8%±2.2%</td>
<td>11.0%±1.5%</td>
<td>-</td>
<td>2.2%±0.4%</td>
</tr>
<tr>
<td>HumanDNVI-NPU</td>
<td>45.6%±2.6%</td>
<td>36.3%±2.4%</td>
<td>3.2%±0.7%</td>
<td>-</td>
</tr>
</tbody>
</table>

piKL- Human-regularized RL and planning
(Jacob et al. 2022)

Idea: Given anchor policy $\tau$ from human imitation learning, when optimizing policy $\pi$, optimize the regularized utility:

$$u(\pi) = EV(\pi) - \lambda D_{KL}(\pi \| \tau)$$

$\lambda$ is the anchor strength:

- $\lambda = 0$: self-play from scratch
- $\lambda = \infty$: human behavioral cloning
- Choosing $\lambda$ in-between gains benefits of both.

Results: Significant policy improvement while maintaining high human compatibility.
We entered CICERO anonymously in an online Diplomacy league.

CICERO was not detected as an AI agent after 40 games with 82 unique players *, sending and receiving an average of 292 messages per game.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Avg Score</th>
<th># Games</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>35.0%</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>25.8%</td>
<td>40</td>
</tr>
<tr>
<td>3</td>
<td>24.5%</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>22.7%</td>
<td>8</td>
</tr>
<tr>
<td>5</td>
<td>21.0%</td>
<td>5</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>19</td>
<td>3.0%</td>
<td>6</td>
</tr>
<tr>
<td>20</td>
<td>2.6%</td>
<td>7</td>
</tr>
</tbody>
</table>

* One player mentioned in post-game Discord that they were suspicious that our account was a bot after a game, but didn’t follow up about it.
Recap

- Sound self play will compute a minimax equilibrium in any two-player zero-sum given sufficient memory and compute.
- Outside two-player zero-sum games, self play isn’t enough.
- Self-play with KL regularization toward a human imitation policy (i.e., piKL) works well in general-sum games!
- See our papers for details:
- Code and models (along with those of our work in full-press):
  Diplomacy with dialogue) available at: https://github.com/facebookresearch/diplomacy_cicero