A Case for Simulated Self-Play in Decision Models with Learning

Christopher Zosh

Economics, Binghamton University

March 3, 2025

Motivation and Contribution

The Problem: Models of learning can have trouble explaining early play in lab data. This can be attributed in part to assumptions about initial beliefs, which themselves can be unrealistic in certain contexts.

The Contribution: I discuss the notion of *Simulated Self-Play*, where agents play simulated rounds of the game against themselves to develop intuition about the nature of the game before the first round of play, and discuss both its theoretical coherence and provide evidence of its value through empirical evidence.

The Agenda

1. Discuss common structural assumptions about initial beliefs in decision-theoretic models and their issues/implications.
   - No priors / uniform priors
   - Priors as free parameters
   - Burn-in
2. Discuss how *Simulated Self-Play* works and how it addresses many of the aforementioned issues.
3. Provide some empirical evidence for the value of Simulated Self-Play as a modeling assumption for decision-theoretic models with learning.
   - Evaluation Criteria: Out of sample fit using
   - Data: Variations of the repeated Beauty Contest Game (BCG) lab data

1 - No Priors / Uniform Priors

Interpretation(s):

- Agents have no initial understanding of the game
- Agents do understand the game but have no preferences over action initially.

Some Problem(s):

- Often poorly explains early play in lab data.
- Particularly unrealistic when...
  - Agents are told how the game works (e.g. in the lab).
  - There exist dominant or dominated strategies.

1 - Priors as Free Parameters

Interpretation(s):

- Agents do have priors about the game, but no structural claim is made about their source.

Some Problem(s):

- Less feasible for large action sets.[1]
- Not a compatible solution for some learning models (e.g. Case-Based Reasoning).
- It is unclear to what degree these parameters tell us about behavior more generally.
- It is unclear to what degree these parameters can tell us about other similar but distinct games.

[1] One option is to coarse-grain the action when assigning priors, similar to what is done by Chen and Du (2016)

1 - Priors via a Burn-In Period

Interpretation(s):

- All agents have experiences consistent with the same events which they use to form the basis of their priors.
- Agents have played this game with each other before a number of times.

Some Problem(s):

- Agents can only enter with experiences which are consistent with the events all other players experienced.
  - For example, in a 2 player repeated PD, it is unlikely that one agent enters with the belief that repeated C can be sustained while the other enters with the belief that it cannot.

2 - Simulated Self-Play

How it works:
Agents start with no priors / uniform priors. For each of the N agents, repeat the process below b times to form priors before the first round of game play.

1. The agent makes N choices (one for each player that would be playing the game).
2. The result of the game is tabulated using those N choices.
3. The agent updates their memory / attractions / priors with N experiences. Each experience is stored as if they chose one of the N actions, and got its corresponding payoff when faced with the other N-1 actions.

Intuition: It's a modified burn-in where, instead of agents playing each other, they play themselves. They store experiences from the perspective of each of their choices.

2 - Simulated Self-Play

Interpretation(s):

- Agents form an intuition of how the game works by playing a number of imaginary rounds of the game against themselves and observing the outcomes.

Desirable Properties:

- Allows agents to enter with non-trivial beliefs
- Allows those beliefs to be derived from different events.
- Adds only 1 free parameter to existing models, regardless of action set size. [Parsimony]
- Agent beliefs are derived from features of the games themselves.
- Can be applied to other games or variants in a well-defined way. [Has testable External Validity]

3 - Empirical Exercise Summary

I fit an ABM of learning agents playing the repeated BCG to some training data using each of the four methods for establishing agent priors:

- Uniform Priors
- Fitted Priors (Using 5 Bins)
- Burned-In Priors
- Priors generated using Simulated Self-Play

Then I then use those best fitting parameters to evaluate how well play matches what's observed in the evaluation dataset under each of the four and compare their performance.

3 - Empirical Exercise Summary

Model Details:

- Decision Models with Learning
  - Simple Reinforcement Learning (Erev and Roth 1998)
- Repeated Beauty Contest Game (BCG) lab data (Duffy and Nagel 1997)
- Fitted parameters via Behavior Search Algorithm
- Compute loss using weighted squared difference in mean and variance of actions chosen each round of play.

3 - The Learning Model(s)

From Erev and Roth 1998 (simplified):

Model Parameters:

- Strength of Priors $S \in (0, \text{inf})$.
- Recency Bias $R \in (0, 1)$.

How it works:

- Each agent i starts with a vector of attractions, where each action starts with attraction S.
- Each round, agents choose an action $a_{i,t}$ randomly, proportional to their attraction to each action.
- After payoffs are received, agents update their attractions using the following formula:

$$Attr_{i,t+1}(\hat{a}) = \begin{cases} (1-R) * Attr_{i,t}(\hat{a}) + \pi_{i,t} & \text{if } \hat{a} = a_{i,t} \\ (1-R) * Attr_{i,t}(\hat{a}) & \text{otherwise} \end{cases} \quad (1)$$

3 - The Game: The Repeated Beauty Contest Game (BCG)

How it works: Each round ...

- Every player simultaneously picks a number from [0,100].
- The target number is calculated by aggregating agents actions, then multiplying by p.
- The player who's choice is closest to the target wins some payment. If tied, the prize is split.

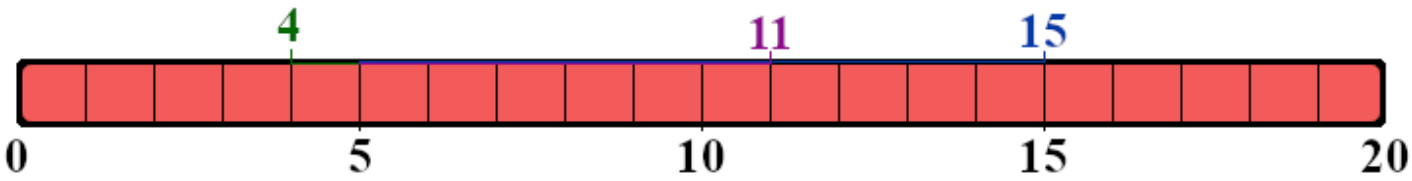
Game Features:

- *Agg(.)* - How the choices are aggregated. Either Mean, Median, or Max of choices.
- p - The multiplier to aggregated choices.
- N - Number of players.

3 - The Game: The Repeated Beauty Contest Game (BCG)

An Example:

Consider player count (N) = 3, multiplier (p) = $\frac{1}{2}$, and *Agg(.)* is mean(.)



1. Players choose 4, 11, and 15.
2. The target number = mean(4 + 11 + 15) * $\frac{1}{2}$ = 5
3. Green wins, as their choice 4 is closest to the target number 5.

Intuition: Players win by undercutting the aggregated choices by the right amount.

3 - The Data

From Duffy and Nagel (1997) - 868 data-points, BCG lab data.

Session	p	N	Agg(.)	Rounds
1	0.5	15	Median	4
2	0.5	15	Median	4
3	0.5	13	Median	4
4	0.5	13	Median	10
5	0.5	16	Mean	4
6	0.5	14	Mean	4
7	0.5	15	Mean	4
8	0.5	14	Mean	10
9	0.5	15	Max	4
10	0.5	15	Max	4
11	0.5	15	Max	4
12	0.5	15	Max	10

3 - Evaluation Criteria

$$Loss(ABM(\theta, r)) = \sum_{s \in D_j} \sum_{t=0}^{Rounds} \Big[(\overline{y_t} - \overline{\hat{y}_t(\theta, r)})^2 + \alpha(Var_t(y_t) - Var_t(\hat{y}_t(\theta, r)))^2 \Big] \quad (2)$$

where...

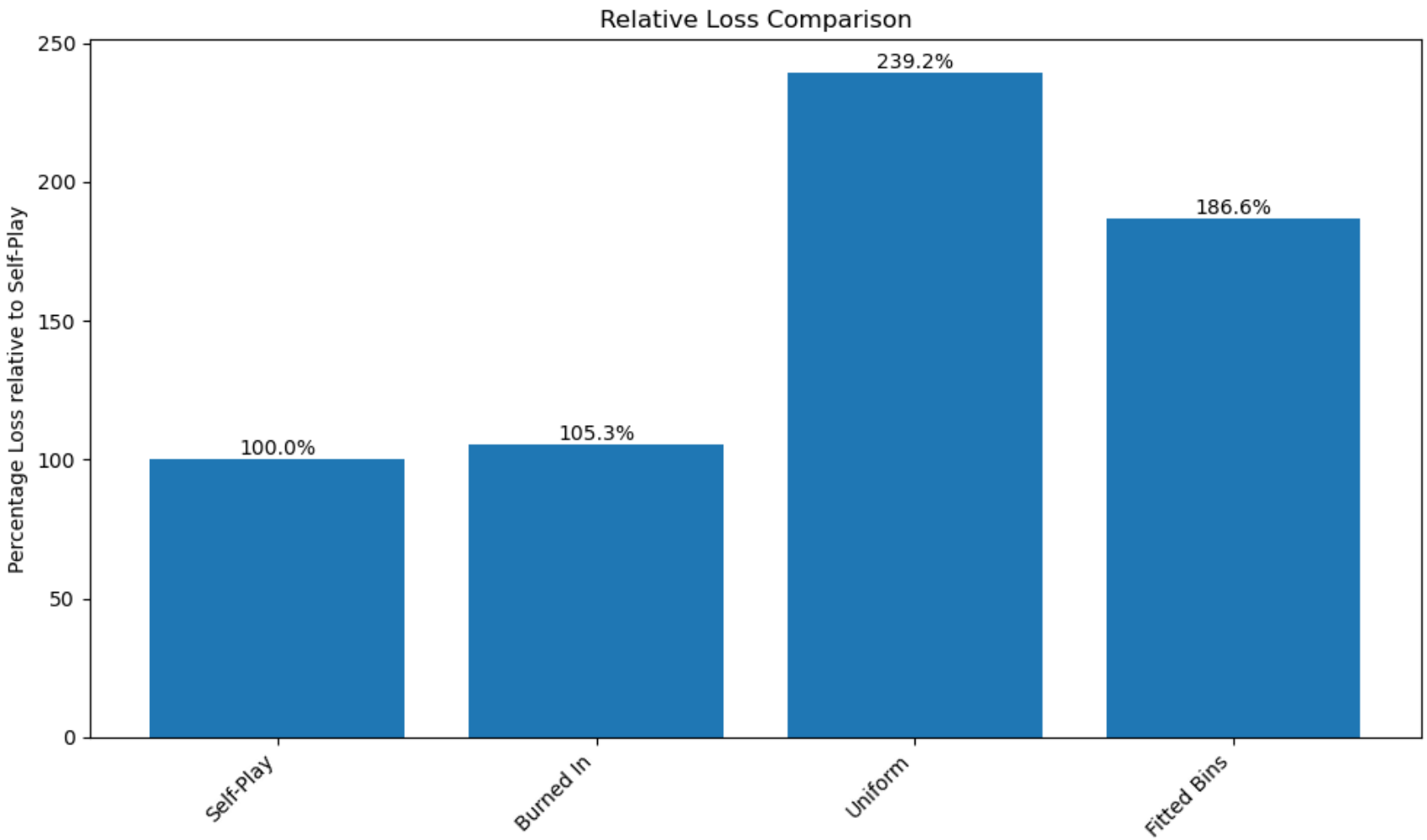
- D_j is either the training data (when fitting) or evaluation data (when evaluating performance).
- s is a session in D_j.
- y_t is an observed choice in data
- $\hat{y}_t(\theta, r)$ is a simulated choice made in the AMB using parameters θ and runs r
- α is the weight placed on variance. $\alpha = 0.05$.

3 - The Data Revisited

White - Training data, Gray - Evaluation data

Session	p	N	Agg(.)	Rounds
1	0.5	15	Median	4
2	0.5	15	Median	4
3	0.5	13	Median	4
4	0.5	13	Median	10
5	0.5	16	Mean	4
6	0.5	14	Mean	4
7	0.5	15	Mean	4
8	0.5	14	Mean	10
9	0.5	15	Max	4
10	0.5	15	Max	4
11	0.5	15	Max	4
12	0.5	15	Max	10

3 - The Results



Conclusion

Simulated Self-Play can be used with nearly any learning model to generate priors in a theoretically coherent and parsimonious way. The parameter it introduces, number of self-play simulation b, also has meaning across all games, making its use appealing over some alternatives from the perspective of understanding human decision-making.

Based on the preliminary runs, I also see some evidence that Simulated Self-Play performs about as well or better empirically than alternatives at matching human play out of sample.

Next Steps

- Larger runs.
- Explore other decision-theoretic learning models (Case-Based Reasoning, etc.).

Any suggestions would be greatly appreciated!