Learning to Respond
The Use of Heuristics in Dynamic Games

Mike Shor

July 2004
Games 2004
Approaches to Modeling Learning

Various disciplines mean quite different things by “learning”

Computer Science

- Most efficient algorithms for solving complex problems

Psychology

- Most accurate representation of the heuristics at work in human learning, even at the expense of closed form (or even mathematical) expressions

Economics

- Simple functional forms which predict actual play
- Simple functional forms which permit theoretical results

Goals:

- Consider how basic learning models incorporate heuristics
- Evaluate experimentally how well they explain behavior in a very simple environment
Heuristics

All learning models incorporate the Law of Effect (Thorndike, 1898)

*An action with a positive reinforcement will be used more often
Poorly-performing actions will be used less often*

- History (memory, forgetfulness, Markov learning properties)
  Law of Exercise (Thorndike, 1898; Watson, 1914, Blackburn, 1936)
  *Actions used more often carry stronger reinforcement*
  - Propensity-based models
  - One-step (memoryless) models
  - Tests of model fitness

- Reference Points (aspirations, payoff scaling)
  *Whether a payoff is a positive or negative reinforcement depends on how the payoff compares to a point of reference*
  - Fixed reference point models
  - Evolving reference point models
  - Aspirations / satisficing models

- Experimentation (exploration)
  *Tradeoff between acquiring environmental information and taking advantage of information already acquired*
  - Undirected experimentation methods
  - Directed experimentation (Sampling, Recency)
  - Diminishing vs. bounded probability of experimentation
Experimental Design

Play a simple monopoly “game”

$$\Pi_t = aq_t - bq_t^2$$

But, in a very low-information environment:

- **Limited Information**
  - The nature of the game, the number of players, and the payoff matrix are all unknown to the players

- **Continuous Time**
  - Players may change strategies at any time
  - Payoffs received every second

- **Structural Changes**
  - The underlying game is subject to structural changes without warning (e.g. Demand Shocks)
  - Optimal strategy changes from 40 to 60 at seven minutes
MONOPOLY RESULTS  
(Friedman, et. al. GEB 2004)

Experimentation

- Subjects experiment 15% of the time

- Experimentation is autocorrelated
  - Probability of Experimentation:
    following a period of experimentation: 0.69
    following a period of near-equilibrium play: 0.05

Arrhythmic Heartbeat Patterns

- Experimentation increases when the “world” changes
Convergence in Dominance-Solvable Games?

Two-player Cournot games without information

![Graphs showing convergence in dominance-solvable games for Player 1 and Player 2, along with BR1 and BR3 strategies.](image-url)
Subjects are responsive
They react quickly to changes in their environment

- Median Play
- Inner Quartile
- Payoff Change
Models

Propensity-Based Models

(Roth and Erev, 1995; Erev and Roth, 1998)

Propensities:

\[ \rho_t(i) = (1-\gamma)\rho_{t-1}(i) + (1-\varepsilon)(\pi_t(i)-\alpha_{t-1}) \]
\[ \rho_t(j) = (1-\gamma)\rho_{t-1}(i) + (\varepsilon/S)(\pi_t(i)-\alpha_{t-1}) \quad j \neq i \]

\( \rho_0 \) is the initial propensity
\( \gamma \) is a forgetfulness or recency parameter, and
\( \varepsilon \) is the probability of experimentation

Reference Points:

\[ \alpha_t = \lambda \alpha_{t-1} + (1-\lambda)\pi_t(i) \]

\( \alpha_0 \) is the initial reference point
\( \lambda \) is the persistence of reference points

Probability of Play:

\[ p_t(i) = \rho_{t-1}(i) / \sum \rho_{t-1}(j) \]

Variants:

We consider six formulations

*History accumulates over time*
*Learning becomes sluggish*
Models

Responsive Learning Automata

(Friedman and Shenker, 1996)

Probability of Play:

\[
p_{t+1}(i) = p_t(i) + \varepsilon \beta \pi_t(i) \sum_{j \neq i} \omega_t(j) p_t(j)
\]

\[
p_{t+1}(j) = p_t(i) - \varepsilon \beta \pi_t(i) \omega_t(i) p_t(i) \quad j \neq i
\]

where

\[
\omega_t(j) = \min \left[ 1, \frac{p_t(j) - \varepsilon / S}{\varepsilon \beta \pi_t(i) p_t(j)} \right]
\]

\(\beta\) captures the speed of learning, and
\(\varepsilon\) is the probability of experimentation

Probabilty of most recently played strategy increases in proportion to its payoff

World-Resetting

Roth-Erev Basic Model with a stability criterion:
If the environment changes, propensities are reset

Allows history to be discarded when it is no longer an accurate representation of the world
Models

Aspirations-based models

(Selten, 1991; Karandikar, et. al., 1998; Borgers and Sarin, 1995)

Aspirations:

\[
\alpha_t = \lambda \alpha_{t-1} + (1-\lambda)\pi_t(i) \quad \text{with probability } (1-\varepsilon)
\]

\[
\alpha_t \sim U(\Pi) \quad \text{with probability } \varepsilon
\]

\(\lambda\) is the persistence of reference points, \(\varepsilon\) is the tremble probability

Probability of Play:

\[
p_{t+1}(i) = \left[1 + \beta (\alpha_t - \pi_t)\right]^{-1} p_t(i)
\]

\[
\pi_t < \alpha_t: \quad p_{t+1}(j) = \frac{p_t(j)}{1 - p_t(i)} \left[1 - \left[1 + \beta (\alpha_t - \pi_t)\right]^{-1} p_t(i)\right] \quad j \neq i
\]

\[
\pi_t \geq \alpha_t: \quad p_{t+1}(i) = 1, \quad p_{t+1}(j) = 0 \quad j \neq i
\]

Satisficing model: Evolving-Aspirations model:

\[
p_{t+1}(i) = \left[p_t(i) + \beta (\pi_t - \alpha_t)\right] \left[1 + \beta (\pi_t - \alpha_t)\right]^{-1}
\]

\[
p_{t+1}(j) = p_t(j) \left[1 + \beta (\alpha_t - \pi_t)\right]^{-1} \quad j \neq i
\]

\(\beta\) captures the speed of learning

A strategy is likely to be repeated if it earns a higher than expected (aspired to) payoff
<table>
<thead>
<tr>
<th>Model</th>
<th>History</th>
<th>Reference Points</th>
<th>Experimentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reinforcement Learning</td>
<td>Yes</td>
<td>No</td>
<td>Undirected</td>
</tr>
<tr>
<td>Basic</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forgetfulness</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Experimentation</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Full model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RL with Reference Points</td>
<td>Yes</td>
<td>Fixed Evolving</td>
<td>Recency</td>
</tr>
<tr>
<td>Fixed reference</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evolving reference</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Two-stage World Resetting</td>
<td>Can be Reset</td>
<td>No</td>
<td>Recency Sampling</td>
</tr>
<tr>
<td>Responsive Learning Automata</td>
<td>No</td>
<td>No</td>
<td>Undirected</td>
</tr>
<tr>
<td>Aspirations Models</td>
<td>No</td>
<td>Evolving</td>
<td>Sampling</td>
</tr>
<tr>
<td>Satisficing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evolving Aspirations</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Method

• In-sample estimation
  *Which learning models best describe ("fit") the data?*
  - Individual fit:
    Simulate the play of each model ten times at each of 2000 sets of parameters, selecting for the parameters that minimize mean squared deviation (MSD)
  - Aggregate fit:
    Similarly, find best aggregate parameters (all subjects)
  - Horse race

• Qualitative comparison
  *Do the models “look” like our human subjects?*

• Out-of-sample prediction
  - How well do the models predict behavior in a slightly different learning task?
## RESULTS

### Model Fits (MSD Scores)

<table>
<thead>
<tr>
<th>Model</th>
<th>Params</th>
<th>Fit to Individual Data</th>
<th>Fit to Aggregate Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reinforcement Learning</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basic</td>
<td>$\rho_0$</td>
<td>0.817</td>
<td>0.852</td>
</tr>
<tr>
<td>Forgetfulness</td>
<td>$\rho_0, \gamma$</td>
<td>0.675</td>
<td>0.849</td>
</tr>
<tr>
<td>Experimentation</td>
<td>$\rho_0, \varepsilon$</td>
<td>0.725</td>
<td>0.838</td>
</tr>
<tr>
<td>Full model</td>
<td>$\rho_0, \gamma, \varepsilon$</td>
<td>0.612</td>
<td>0.836</td>
</tr>
<tr>
<td><strong>RL with Reference Points</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fixed reference</td>
<td>$\rho_0, \alpha$</td>
<td>0.692</td>
<td>0.824</td>
</tr>
<tr>
<td>Evolving reference</td>
<td>$\rho_0, \alpha_0, \gamma$</td>
<td>0.672</td>
<td>0.794</td>
</tr>
<tr>
<td><strong>Two-stage World Resetting</strong></td>
<td>$\alpha_0$</td>
<td>0.618</td>
<td>0.829</td>
</tr>
<tr>
<td><strong>Responsive Learning Automata</strong></td>
<td>$\varepsilon, \beta$</td>
<td>0.606</td>
<td>0.820</td>
</tr>
<tr>
<td><strong>Aspirations Models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Satisficing</td>
<td>$\gamma, \beta, \varepsilon$</td>
<td>0.545</td>
<td>0.750</td>
</tr>
<tr>
<td>Evolving Aspirations</td>
<td>$\gamma, \beta, \varepsilon$</td>
<td>0.567</td>
<td>0.788</td>
</tr>
<tr>
<td><strong>Benchmark Models</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Random Choice</td>
<td></td>
<td>0.900</td>
<td>0.900</td>
</tr>
<tr>
<td>Equilibrium</td>
<td></td>
<td>1.026</td>
<td>1.026</td>
</tr>
</tbody>
</table>

**QUESTION:**

Do people differ in *learning styles* or in *rates of learning*?
(Do the same heuristics work for most people?)
The Horse Race

Each number reflects the proportion of subjects for whom the model in the row predicted at least as well as the model in the column.

“Best” reflects the number of other models the model in the row “beat,” in the sense of predicting better for a majority of individuals.

<table>
<thead>
<tr>
<th>#</th>
<th>Model</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>R</th>
<th>E</th>
<th>best</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Satisficing</td>
<td></td>
<td>.61</td>
<td>.75</td>
<td>.75</td>
<td>.75</td>
<td>.80</td>
<td>.79</td>
<td>.80</td>
<td>.82</td>
<td>.86</td>
<td>.95</td>
<td>1.0</td>
<td>9</td>
</tr>
<tr>
<td>2</td>
<td>Evolving Aspirations</td>
<td>.36</td>
<td>—</td>
<td>.77</td>
<td>.66</td>
<td>.64</td>
<td>.75</td>
<td>.80</td>
<td>.84</td>
<td>.89</td>
<td>.93</td>
<td>.96</td>
<td>.98</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>World Resetting</td>
<td>.25</td>
<td>.23</td>
<td>—</td>
<td>.54</td>
<td>.54</td>
<td>.71</td>
<td>.77</td>
<td>.80</td>
<td>.91</td>
<td>.96</td>
<td>1.0</td>
<td>.91</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>Responsive LA</td>
<td>.25</td>
<td>.34</td>
<td>.46</td>
<td>—</td>
<td>.59</td>
<td>.75</td>
<td>.89</td>
<td>.88</td>
<td>.98</td>
<td>.98</td>
<td>1.0</td>
<td>.86</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>RL Full model</td>
<td>.25</td>
<td>.36</td>
<td>.46</td>
<td>.41</td>
<td>—</td>
<td>.71</td>
<td>.73</td>
<td>.84</td>
<td>.95</td>
<td>.98</td>
<td>1.0</td>
<td>.91</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>RL Evolving reference</td>
<td>.20</td>
<td>.25</td>
<td>.29</td>
<td>.25</td>
<td>.29</td>
<td>—</td>
<td>.57</td>
<td>.54</td>
<td>.86</td>
<td>1.0</td>
<td>1.0</td>
<td>.80</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>RL Forgetfulness</td>
<td>.21</td>
<td>.20</td>
<td>.23</td>
<td>.11</td>
<td>.00</td>
<td>.43</td>
<td>—</td>
<td>.63</td>
<td>.77</td>
<td>.89</td>
<td>1.0</td>
<td>.79</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>RL Fixed reference</td>
<td>.20</td>
<td>.16</td>
<td>.20</td>
<td>.13</td>
<td>.14</td>
<td>.00</td>
<td>.29</td>
<td>—</td>
<td>.64</td>
<td>.91</td>
<td>1.0</td>
<td>.80</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>RL Experimentation</td>
<td>.18</td>
<td>.11</td>
<td>.09</td>
<td>.02</td>
<td>.00</td>
<td>.14</td>
<td>.14</td>
<td>.29</td>
<td>—</td>
<td>.86</td>
<td>1.0</td>
<td>.73</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>RL Basic</td>
<td>.14</td>
<td>.07</td>
<td>.04</td>
<td>.02</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>1.0</td>
<td>.68</td>
<td>0</td>
</tr>
<tr>
<td>R</td>
<td>Random Choice</td>
<td>.07</td>
<td>.04</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>.00</td>
<td>1.0</td>
<td>.59</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>Equilibrium</td>
<td>.00</td>
<td>.02</td>
<td>.09</td>
<td>.14</td>
<td>.09</td>
<td>.20</td>
<td>.21</td>
<td>.20</td>
<td>.27</td>
<td>.32</td>
<td>.41</td>
<td>—</td>
<td></td>
</tr>
</tbody>
</table>

Models clearly ordered
All memoryless models “bested” all models with memory
Quantitative Results may be Insufficient

Actual Play
Quantitative Results may be Insufficient
MSD = 0.794

MSD = 0.829

MSD = 0.820
Satisficing Does Well

Actual Play

Satisficing
Subjects who converge at “equilibrium” will be unaware that a better payoff exists.

Yet subjects still respond well.
Conclusions

• What we observe:
  • A change in the state of the world causes change in play
  • History is abandoned if no longer appropriate
  • Subjects evaluate both current payoffs relative to benchmarks and the worthiness of benchmarks
  • Experimentation is directed, and autocorrelated

• Questions for learning models:
  • How to incorporate experimentation?
    • Random “shock”
    • Aspirations tremble
    • Directed, separate process
  • Reference Points
    • How are they formed?
    • How are they abandoned or reset?

• Generalization:
  • Reinforcement partly attributed to nearby strategies
  • Up to now, ignored by economists
  • How to incorporate this?