

A Note on Q-Learning in the Cournot Game^{*†}

Steven O. Kimbrough
Ming Lu
University of Pennsylvania
kimbrough@wharton.upenn.edu
milu@wharton.upenn.edu

WeB2003, Seattle, 12/13/03

*File: cournot-seattle-foils.tex/pdf.

†Thanks, too, to Ann Kuo.

Goals for the Presentation

- On beyond the paper: context, subsequent work and results. All very quickly.
- Context/background: Four ways to study contexts of strategic interaction (CSIs or games)
 1. *A priori* — classical game theory; analytic study and results; rational choice theory presumed
 2. *In vivo* — “games in the wild” Natural history of games? (needed)
 3. *In vitro* — behavioral game theory; experimental economics; biology, too
 4. *In silico* — or algorithmic game theory; our way

Algorithmic Game Theory

- What happens when well-defined algorithmic agents meet in CSIs? Nash? Pareto?
- Why does what happens happen?
- How do smarts pay off (if at all)? Learning? What sorts of learning? *et cetera* . . .
- Important, essential for fielding artificial agents in CSIs, as in e-business
- Our focus: Agents that learn (not merely adapt) in CSIs.

Focus: Repeated Games

- *Supergame* — a game composed of games (called subgames)
- *Repeated game* — a special kind of supergame: subgames all the same; these are called *stage games*
- Begin: repeated games with 2×2 stage games.
- Then: Cournot supergame. Players chose quantities to produce, receive rewards, and cycle. Can be thought of as a repeated game, but better as a supergame, since multiple states obtain.

2×2 Games

- Previous work: “Simple Reinforcement Learning Agents: Pareto Beats Nash in an Algorithmic Game Theory Study” by Kimbrough and Lu, forthcoming in *Information Systems and e-Business Management*.
- Subsequent work (in NetLogo; in Java (thanks to Ann Kuo)).

Learning Regime

1. Alternative (or consideration) set of strategies: $\mathcal{A} = \{0, 1\}$ for each player.
2. Attractiveness estimation: linear updating rule for A^i , $i \in \mathcal{A}$:
$$A_{t+1}^i = A_t^i + \alpha\{r_t^i - A_t^i\}$$

NewEstimate = CurrentEstimate + StepSize{reward - CurrentEstimate}
3. Choice/exploration policies.
Softmax. ϵ -greedy

Softmax

$$\Pr(A_t^i) = \frac{e^{A_t^i/\tau}}{\sum_j e^{A_t^j/\tau}}$$

$\tau \longrightarrow 0$ as $n \longrightarrow \infty$

Essential Findings

- When the numbers are right, agents tend to find Pareto outcomes, even at the expense of Nash outcomes, in terms of the stage game. (E.g. in Prisoner's Dilemma, chicken)
- When Nash and Pareto outcomes coincide and multiple Nash, agents tend (when the numbers are right) to find Pareto-optimal Nash outcomes. (E.g., Stag Hunt)
- Results sensitive to actual payoffs (in contravention to classical game theory)
- In any event, players tend to extract more wealth than would otherwise be predicted.

Example: Prisoner's Dilemma

	C	D
C	$(3,3)^{**}$	$(0, 3+\delta)^*$
D	$(3+\delta, 0)^*$	$(\delta, \delta)\#$

Table 1: $\#$ =Nash; $*$ =Pareto

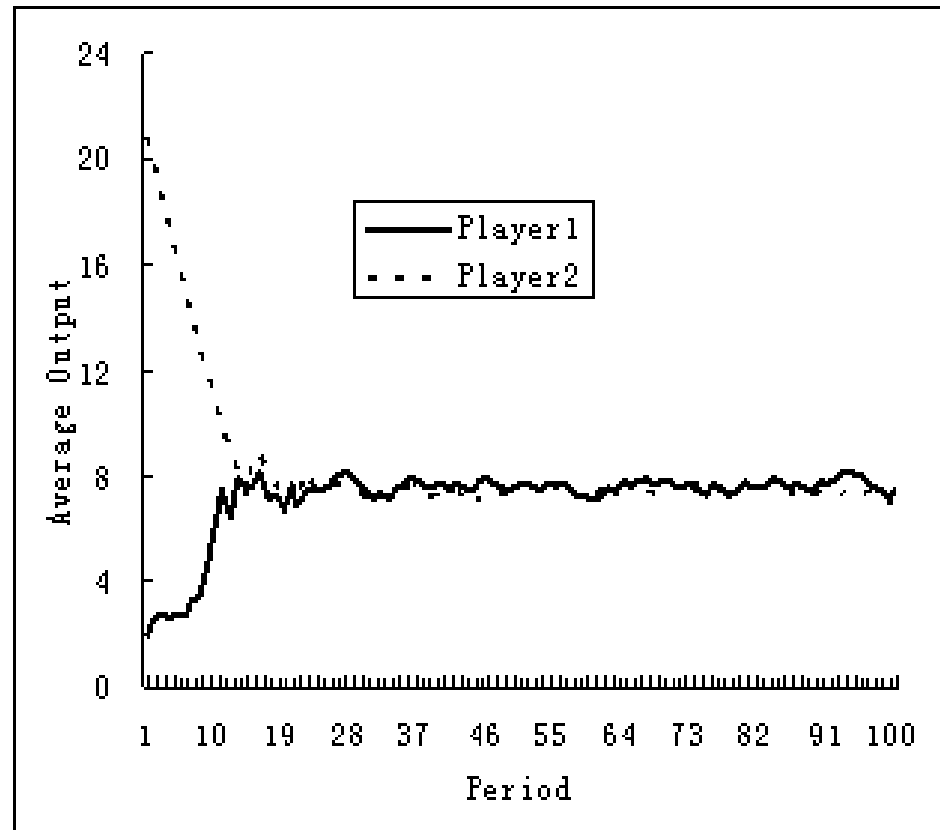
Summary of Results

ϵ -greedy selection				action		Softmax selection			
CC	CD	DC	DD	δ	CC	CD	DC	DD	
9422	218	183	177	0.05	9334	302	285	79	
9036	399	388	150	0.5	9346	294	220	140	
5691	738	678	2693	1	7537	954	1267	242	
3506	179	275	6040	1.25	8203	542	994	261	
1181	184	116	8519	1.5	7818	767	775	640	
2	98	103	9797	1.75	4685	270	422	4623	
97	114	91	9698	2	1820	217	220	7743	
0	100	92	9808	2.5	0	77	117	9806	
2	96	94	9808	2.95	0	90	114	9796	

WeB2003 Paper: Holt's Cournot Game

- $\pi(x, y) = (12 - 0.5(x + y))x$ and similarly for $\pi(y, x)$
- Competitive outcome: $x + y = 12/0.5 = 24$, 12 each for a profit each of 0.
- Monopoly outcome: $x + y = 12$, 6 each for a profit each of 36.
- Cournot/Nash outcome: $x + y = (2 \cdot 12)/(3 \cdot 0.5) = 16$, 8 each for a profit each of 32.
- Holt's findings: human subjects produce slightly less than 8 each on average.

Our Agents in the Holt/Cournot Supergame



Atomic versus Molecular Strategies

- Heretofore (us and others): agents only learn strategies for the stage game. Atomic strategies.
- These are m-0, memory of 0, strategies.
- What if they learned strategies defined over more than one subgame? Molecular strategies.
- We looked at m-1, memory of 1 previous subgame, molecular strategies.

Molecular Learning Regime in 2×2 Games

1. Alternative (or consideration) set of strategies: $\mathcal{A} = \{00, 01, 10, 11\}$ for each player. Form: ab: play a if last time counter-player played 0, play b if last time counter-player played 1. E.g., in Prisoner's Dilemma, 01 is TIT FOR TAT.

All else the same:

2. Attractiveness estimation: linear updating rule for A^i , $i \in \mathcal{A}$:

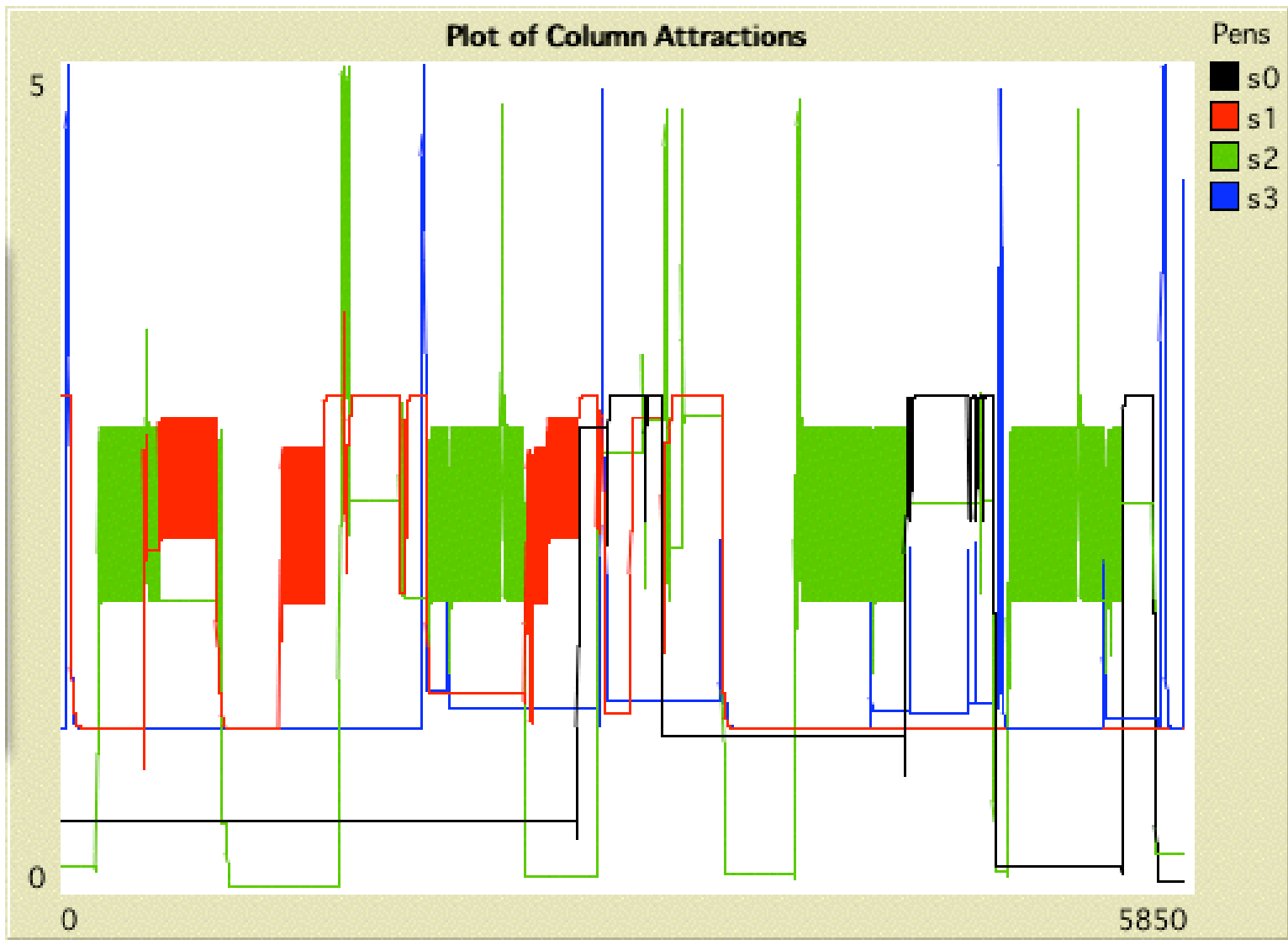
$$A_{t+1}^i = A_t^i + \alpha \{r_t^i - A_t^i\}$$

NewEstimate = CurrentEstimate + StepSize{reward - CurrentEstimate}

3. Choice/exploration policies. Softmax. ϵ -greedy

Example: Standard Prisoner's Dilemma

	D	C
D	1,1	5,0
C	0,5	3,3



Row Points Per

2.197

Col Points Per

2.255

Essential Findings on 2×2 Games

As before, but the results are not very sensitive to the actual payoffs.

- Agents tend to find Pareto outcomes, even at the expense of Nash outcomes, in terms of the stage game. (E.g. in Prisoner's Dilemma, chicken)
- When Nash and Pareto outcomes coincide and multiple Nash, agents tend to find Pareto-optimal Nash outcomes. (E.g., Stag Hunt)
- Actual payoffs do matter (in contravention to classical game theory)
- In any event, players tend to extract much more wealth than would otherwise be predicted.

4 Molecular Strategies in Holt's Cournot Game

1. G-TFT. If $y_{t-1} > y_{t-2}$, then $x_t = x_{t-1} + \delta$; else $x_t = x_{t-1} - \delta$
2. BESTRESPONSE. $x_t = 12 - 0.5y_{t-1}$.
3. S-TFT.
 - (a) If $x_{t-1} < y_{t-1}$ and $y_{t-2} \leq y_{t-1}$, then $x_t = x_{t-1} + \delta$.
 - (b) If $x_{t-1} > y_{t-1}$ or $x_{t-2} = x_{t-1} = y_{t-1} = y_{t-2}$, then $x_t = x_{t-1} - \delta$.
 - (c) Else, $x_t = x_{t-1}$.
4. COPYCAT. $x_t = y_{t-1}$.

Paired Strategies: Profits

	G-TFT	BESTRESPONSE	S-TFT	COPYCAT
G-TFT	(36,36)	(33.138,28.165)	(36,36)	(36,36)
BR	(28.165, 33.138)	(32,32)	(32,32)	(32,32)
S-TFT	(36,36)	(32,32)	(36,36)	(36,36)
CC	(36,36)	(32,32)	(36,36)	(23.166,23.166)

Play with all 4 Strategies

What will happen if the agents engage all four strategies simultaneously, under the above learning regime?

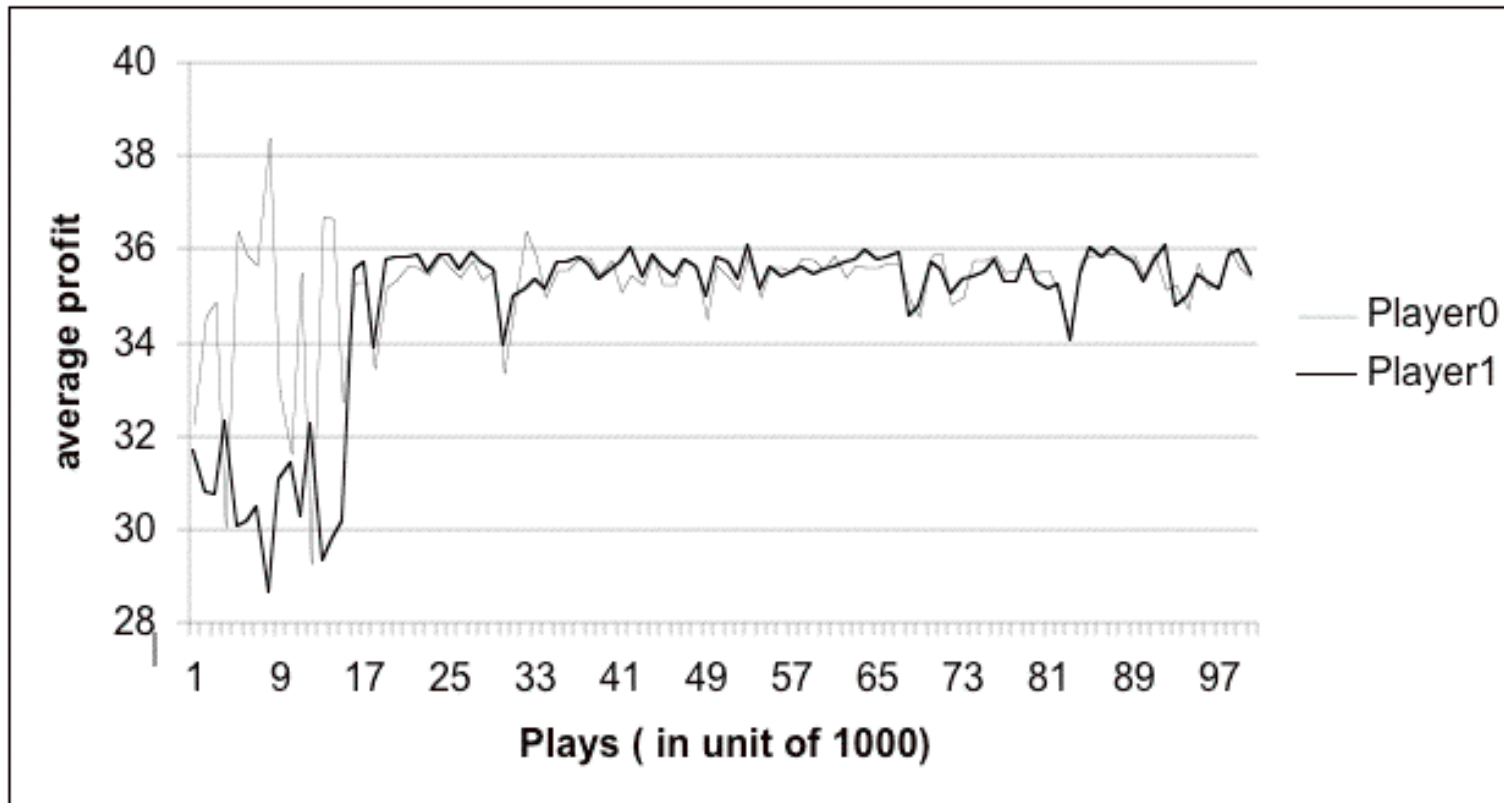
$$\mathcal{A} = \{\text{G-TFT}, \text{BR}, \text{S-TFT}, \text{CC}\}$$

Results over 100 runs of 100,000 (rounds) plays, averaged over the last 1000 rounds of play.

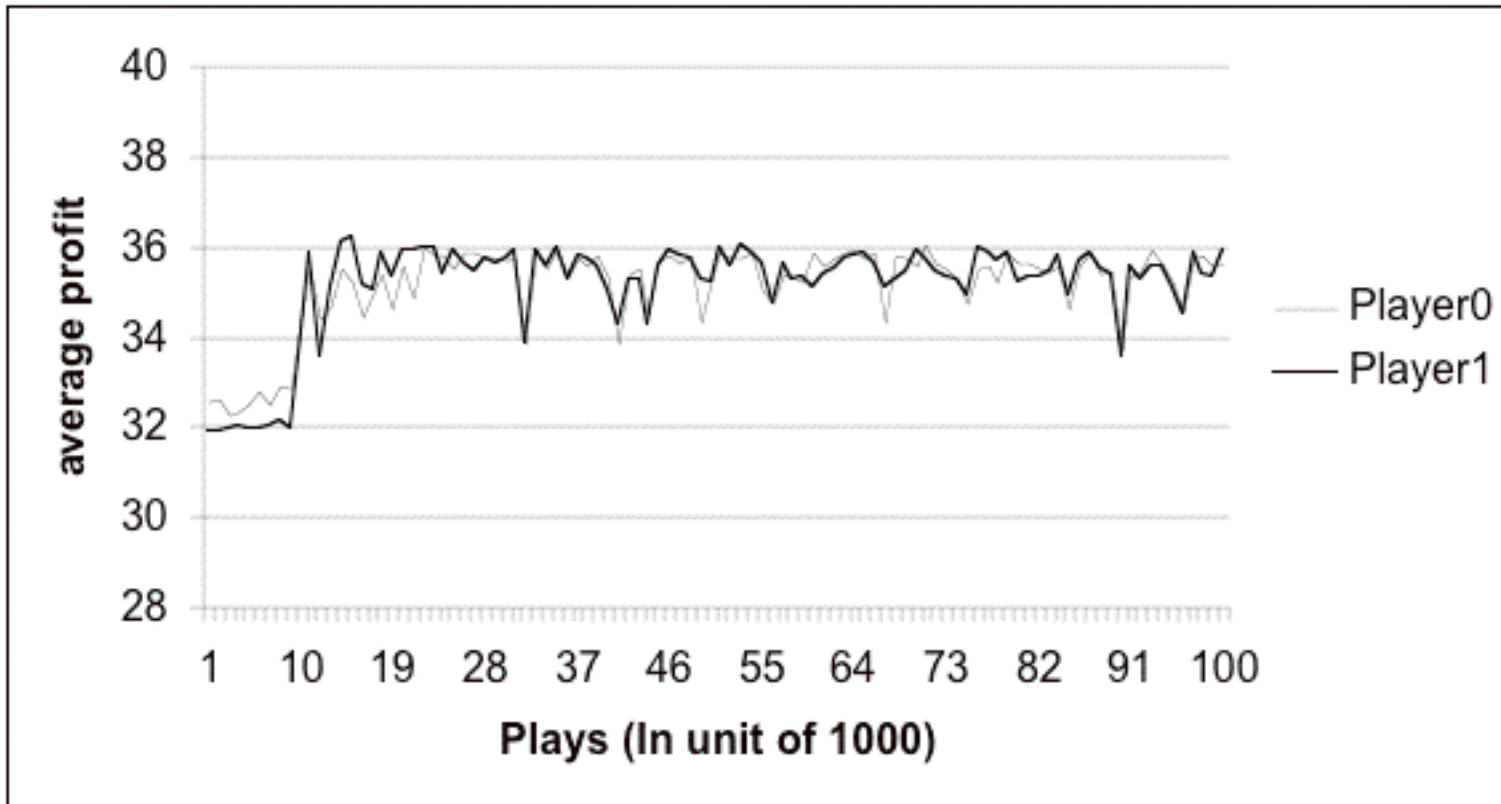
	Average Profit
Softmax action selection	(35.348, 35.323)
ϵ -greedy action selection	(35.487, 35.507)

Two typical runs follow...

Molecular Strategies with Holt's Cournot Game: A



Molecular Strategies with Holt's Cournot Game: B



Conclusion

- Major reward from algorithmic game theory studies: results from molecular strategies.
- Have to ask: If our rather dumb agents can figure out how to get monopoly profits in a Cournot game, why believe the Cournot analysis?
- Much remains to be done. Exciting stuff.

\$Id: cournot-seattle-foils.tex,v 1.3 2003/12/11 16:36:45 sok Exp \$