

# Exploring Games and Rationality\*

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

CMU, Heinz School, 2/13/04

---

\*File: cmu-20040213-foils.tex/pdf.

## **Abstract: Exploring Games and Rationality**

The classical (or received) account of rationality, as embedded in game theory and rational choice theory, is attended by a number of well-known problems, including various paradoxes and limitations. Even so, the received account is founded upon seemingly unassailable principles, such as avoidance of dominated choices. This talk is a report on work in progress that examines behavior of artificial agents in strategic contexts (games). Experimental/computational results are described that, in conjunction with certain conceptual moves presented in the talk, suggest an alternative perspective on strategic rationality. The experimental/computational results compare two kinds of reinforcement learning by agents, distinguished by the objects of learning, either states or policies. It is demonstrated for a variety of games that policy learners are quite effective in reaching Pareto optimal outcomes.

# Outline

A report on a project in progress. Thanks to Ming Lu and Ann Kuo. NSF.

⇒ Setting & Setup

- Abstraction
- Paradox
- Data: 1
- Data: 2
- Comments & Conclusion

## Focus of the Project

- 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 focus

# 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*...
- Essential for fielding artificial agents in CSIs, as in e-business
- Not to be neglected: play by non-ideal agents (chickens, spiders, ... even humans)
- Our focus: Agents that learn (not merely adapt) in CSIs.

## Focus: Composite Games

- *Composite game* — a game composed of a sequence of games (element games). (Contrast with supergame and its subgames, a special case.)
- *Repeated game* — a special kind of composite game: elements all the same; these are called *stage games*.
- Begin: repeated games with  $2 \times 2$  stage games.
- Then: Cournot composite game. Players chose quantities to produce, receive rewards, and cycle.

## At Issue: Problems of Ideal Rationality in Strategic Interaction

- Received views of rationality: Avoid dominated options; maximize expected utility; behave consistently. What's not to like?
- Well-known problems/concerns, including: unrealistic (impossible?) cognitive and computational requirements; sensitivity to assumptions; challenges to the coherence of the idea; puzzles and paradoxes; predicts outcomes of games, but not how to play them; too many equilibria; off-equilibrium play often predominates; too many assumptions/specifications; neglect of procedure; etc.
- Is there something to learn “from the ground up”? (Axelrod &c.)

# Outline

- Setting & Setup

⇒ Abstraction

- Paradox
- Data: 1
- Data: 2
- Comments & Conclusion

## A (mostly) Terminological Move

- Allow that there may be many quite different yet valid concepts of rationality.
- Likened to etiquette. No universal system; many particular systems.
- A move to avoid begging of questions.
- Game-theoretic rationality, utility theory, etc. seen as a particular system of rationality.
- Will speak of rationality (etiquette) as attendance to goals (politeness), without putting too fine a point on it.

## At Least One Economist...

From Amartya Sen, *Rationality and Freedom*, “Introduction: Rationality and Freedom”

Rationality is interpreted here, broadly, as the discipline of subjecting one’s choices—of actions as well as of objectives, values and priorities—to reasoned scrutiny. Rather than defining rationality in terms of some formulaic conditions that have been proposed in the literature (such as satisfying some prespecified axioms of “internal consistency of choice,” or being in conformity with “intelligent pursuit of self-interest,” or being some variant of maximizing behavior), rationality is seen here in much more general terms as the need to subject one’s choices to the demands of reason.

# Outline

- Setting & Setup

- Abstraction

⇒ Paradox

- Data: 1

- Data: 2

- Comments & Conclusion

## How to Do Philosophy

“A certain body of indefinable entities and indemonstrable propositions must form the starting-point for any mathematical reasoning; and it is this starting-point that concerns the philosopher. When the philosopher’s work has been perfectly accomplished, its results can be wholly embodied in premisses from which deduction may proceed. Now it follows from the very nature of such inquiries that results may be disproved, but can never be proved. This disproof will consist in pointing out contradictions and inconsistencies; but the absence of these can never amount to proof. All depends, in the end, upon immediate perception; and philosophical argument, strictly speaking, consists mainly of an endeavour to cause the reader to perceive what has been perceived by the author. The argument, in short, is not of the nature of proof, but of exhortation.” —Bertrand Russell, *The Principles of Mathematics*, 1902, XV, 124

(See also, “The Ways of Paradox” by W.V. Quine.)

# The Surprise Examination Paradox

- Aka: Surprise Hanging Problem
- “There will be a surprise exam given in one of the next 6 meetings of the class.”
- Reasoning by backwards induction...

## **From Grim et al., *The Philosophical Computer*, page 163**

The similarity of this reasoning to that of the argument for dominant defection throughout a series of known finite length is worth noting because of course the Surprise Examination is treated standardly in the philosophical literature as a *paradox*, thought to hide some fallacious piece of logical legerdemain. That the same form of reasoning is thought of as valid in the theoretical economics literature, though perhaps inapplicable in some practical sense, indicates that important work remains to be done in bridging the two bodies of work.

## First Question on the Exam

1. Explain the fallacy in the reasoning that led you to believe it impossible for me to give you a surprise exam as announced.
- Can the teacher give a surprise exam *and* speak truly in saying that there will be a surprise exam?
  - Will reconstruct and present a proper way of reasoning about this. Disclaimer: no unique solution.
  - Begin with the one-shot problem: “There will be a surprise exam tomorrow.”

# Formalizing

- The teacher's utterance,  $u$ , is an action, a speech act. It is veridical ("truthy") iff there is an exam tomorrow,  $E$ , and that exam is a surprise to the students,  $S$ .

$$V(u) \leftrightarrow (E \wedge S)$$

- The exam is a surprise only if (and if?) the exam is held and the probability of its being held is below some critical level,  $l$ .

$$(P(E) < l \wedge E) \rightarrow S \quad (\text{or even } (P(E) < l \wedge E) \leftrightarrow S)$$

## A One-Shot Surprise Exam Argument

1.  $V(u) \leftrightarrow (E \wedge S)$

2.  $(P(E) < l \wedge E) \rightarrow S$  (or even  $(P(E) < l \wedge E) \leftrightarrow S$ )

3.  $l = \frac{2}{5}$  (or any value  $> 0$ )

4.  $P(E) < l$

$$\models (E \wedge S \wedge V(u)) \vee (\neg E \wedge \neg V(u))$$

And  $(E \wedge S \wedge V(u))$  with probability  $P(E)$ .

## Achieving Surprise (and Speaking Veridically)

- Observe a two-outcome (H/T) random event whose  $P(H) < l$ . If  $H$  occurs, give the exam; else, don't.
- The teacher's risk of speaking falsely,  $R_T$ , is  $1 - P(H) \geq 1 - l$ .
- If the teacher is willing to bear the risk and  $H$  occurs, then the teacher speaks with veridicality and there is a surprise exam.

Even the teacher will be surprised.

## Why Should the Teacher Want to Bear the Risk?

- There are five possible outcomes:
  1.  $u$  occurs and  $(E \wedge S \wedge V(u))$
  2.  $u$  occurs and  $(\neg E \wedge \neg V(u))$
  3.  $u$  does not occur and  $E \wedge \neg S$
  4.  $u$  does not occur and  $\neg E$
  5.  $u$  does not occur and  $E \wedge S$
- Given  $l$ , the teacher prefers to gamble on 1 and 2, versus either 3 or 4 or 5 with certainty.
- On balance the teacher's risk-reward tradeoff favors taking the risk.

## Or for the n-Shot Surprise Exam

- Select the day of the exam uniformly among the  $n$  possible days:  
 $P(E_i) = \frac{1}{n}$ .
- Let  $d_r =$  number of days remaining. Find  $d^* = \arg \min_{d_r} 1/d_r < l$
- Set  $R_T > 0$ , the maximum risk the teacher is willing to bear.
- Set  $n$  such that

$$\frac{d^*}{n} \leq R_T$$

## Upshot

- If the teacher is willing to bear some risk of speaking falsely, there is some minimal  $n$  at which the teacher is willing to announce a surprise exam.
- The teacher can give a surprise exam, just not with certainty. And the uncertainty may be made arbitrarily low.
- The students' mistake is in assuming that in making a risk-return tradeoff on the announcement, the teacher would take no risk whatsoever.

## Question 2 on the Exam

2. In a 100-shot Repeated Prisoner's Dilemma game, played between the teacher and an unknown, but fully competent human subject, the teacher announces that she will gain the reward from mutual coöperation at least 2 times, net. That is, if  $P$  is the penalty for mutual defection and  $R$  is the reward for mutual coöperation, the teacher is asserting that she will get at least  $98 \cdot P + 2 \cdot R$  points from the 100 trials. Can this assertion be plausibly justified? Why or why not?

## The One-Shot Prisoner's Dilemma Game

The (one-shot) Prisoner's Dilemma game involves two players each with two strategies: C (coöperate) and D (defect). In strategic form the game is:

	C	D
C	R	S
D	T	P

with the requirement that  $T > R > P > S$  and that  $2 \cdot R > T + S$ . Typically, even usually, in experiments  $T = 5$ ,  $R = 3$ ,  $P = 1$ , and  $S = 0$ . Since  $T > R$  and  $P > S$ , there is only one equilibrium point (EP): both players play  $D$ . The dilemma, of course, is that if both players could play  $C$ , both would be better off, since  $R > P$ .

## Of Course...

- In the one-shot case there is one Nash equilibrium: both defect.
- In the definitely-repeated case there is one subgame perfect equilibrium: all defect.
- In the indefinitely-repeated case, just about anything can be a Nash equilibrium.
- Extensive human experiments support the teacher's claim.

## Experiments say the teacher is right

It is interesting, and significant, that in the first human experiment with repeated prisoner's dilemma the human subjects were asked to record their thoughts as the game was being played. Comments such as

- “Perverse!”
- “Oh ho! Guess I’ll have to give him another chance.”
- “In time he could learn, but not in ten moves so:”
- “What’s he doing?!!”

- “I’m completely confused. Is he trying to convey information to me?”  
and
- “This is like toilet training a child—you have to be very patient.”

appear throughout the 100 iterations of the game. Even so, the two subjects jointly cooperated in 60 of the 100 iterations. By the lights of classical game theory this was a remarkably rewarding triumph of irrational behavior.<sup>1</sup>

---

<sup>1</sup>These results are not inconsistent with subsequent empirical findings.

## Consider the one-shot PD in this form, Pattern 1

		C	D
	R		T
C	R	S	
D	T	P	

 $\implies$ 

		C	D
	$B$		$B + \epsilon$
C	$B$	$S$	
D	$B + \epsilon$	$B - \epsilon$	

( $\epsilon > 0$ , let  $S = 0$ , assume  $B - \epsilon > S$ )

- Note:  $2 \cdot R > T + S$ , but possibly  $T + S > 2 \cdot P$ .
- Fix  $B$ , make  $N$  (the number of plays) be smallish, make  $\epsilon$  head towards 0. Are you willing to risk cooperation?

## Now this version, Pattern 2

		C	D
C	R	S	T
D	T	P	P

 $\implies$ 

		C	D
C	$B$	$B + \epsilon$	$S$
D	$B + \epsilon$	$S$	$S + \epsilon$

$(\epsilon > 0, \text{ let } S = 0)$

- Let  $B$  and  $N$  increase arbitrarily, and  $\epsilon$  decrease arbitrarily. Are there no values at which you would risk cooperation?

## Summing up on Definitely IPD

- To succeed, the backwards induction argument for DIPD must recommend ALL DEFECT in both patterns (above), regards of parameter values.
- It is plausible and not unreasonable for the teacher to have a risk/return tradeoff allowing her to offer a surprise exam.
- It is plausible and not unreasonable for both players in DIPD to have risk/return tradeoffs allowing them to try some cooperation.
- Think of TIT FOR TAT as a simple reinforcement schedule.

## A Game Theorist's Counter, with Response

C: Nothing new here. Of course if your opponent is irrational it may be rational to try some cooperation. That's old news.

R: Name calling merely evades the issue. You can insist on your definition of rationality, but you can't thereby make the concept interesting, descriptively sound, broadly useful, and free of paradox.

C: What positive concept do you have, that is more interesting, descriptively sound, broadly useful, and free of paradox?

R: A fair question.

## A beginning of an answer

- *adaptive (learning) agent*  $\approx$  responds to experience and information, and modifies its behavior. Recognized in the literature.
- *exploring agent*  $\approx$  takes risks to obtain experience and information (and then is adaptive).
- Metaheuristics (metastrategies), procedures for finding solutions (strategies).

## A beginning of an answer (con't.)

Metaheuristics (metastrategies):

1. Local search (incestuous) methods, e.g., hill climbing, simulated annealing, Q-learning
2. Population-based (promiscuous) methods, e.g., Learning Classifier Systems, memetic algorithms, evolutionary computation

All: not merely adaptive, but also exploring.

Explosion of attention and innovation. They are proving interesting, broadly useful, often descriptively sound.

And where are the paradoxes?

## Counter and Reply, again

C: But can't this be incorporated into a Bayesian framework, so that we're back where we started?

R: In principle, yes, but this requires huge computational power. Also:

1. There is a premium on simpler explanations. For 'lower' species (animals and plants and bacteria, etc.). To understand just how far we can get with simpler models.
2. Even if we have (super) humans with big computers, the computational power needed is often way beyond reach. Beyond intractability we even have undecidability.

## Counter and Reply, and again

C: Then haven't we just discovered bounded rationality again?

R: Yes, and the *Mona Lisa* is just oil paint on canvas.

1. Exploring rationality may be seen as a special kind of bounded rationality. Compare with satisficing.
2. Exploring rationality is appropriate even with essentially unlimited computational bounds, due to intractability and undecidability.
3. Nothing in the arguments I gave required bounded rationality. Riskless rationality leads to paradoxes and Pareto-inferior solutions. A risky—or *exploring*—rationality will often serve us better even when unbounded. What is true in the finite case need not be false in the infinite case.

## What Are We to Make of This?

- Are people irrational? Are they smart to be irrational, rather than be “rational fools” (Sen)?
- Are people classically rational (more or less), but playing a different game?
- Is something else going on? (Answer: yes, I think so)

## Consider the Much-Replicated Light-Guessing Experiments

- Two lights—red and green—flash randomly in turn.  
 $Pr(\text{red})=1-Pr(\text{green})=x$ .
- Subjects observe, guess, and are rewarded for being correct.
- After a time, subjects estimate  $x$  rather accurately, then probability match, guessing red with probability  $\hat{x}$  and green with probability  $1 - \hat{x}$ .
- Why do they behave non-optimally, given that they have the requisite information?

## Two (Kinds of) Explanations

...in addition to the people-are-just-stupid explanation.

### 1. Classically rational with hidden goals.

Subjects are (approximately) rational in the classic sense; it's just that they impose other attributes and values (e.g., they get bored).

A popular, received, and entrenched explanation. Still, worrisome. Addition of parameters. Brittleness and generalization. &c.

How do you get evidence for this explanation, independently of the data for the experiment?

## Two (Kinds of) Explanations

### 2. Transfer of heuristic.

Subjects are (wisely or not, but more or less reasonably) employing a heuristic in the new situation that has worked well in other contexts.

What might that be and how would you get evidence for it, independently of the data for the experiment?

Replace lights with resources and add competition, then probability matching is an ESS. Going exclusively with the higher probability is not.

Haverford duck pond experiments.

## The upshot

- In the Surprise Exam, the students failed to recognize that the teacher faces a risk/return tradeoff. The teacher can give a surprise exam if she is willing to undertake some risk of speaking falsely.
- Like the students, classical, game-theoretic (riskless) rationality takes an extreme, limiting position on the risk/return tradeoff. It may be wise (and rational) to explore with a counter-player at the risk of some loss.
- This leads to a notion of an exploring rationality, for which there is a rich and fruitful body of algorithms instantiating the concept.
- Agent-based modeling is a natural and entirely appropriate tool for exploring exploring rationality.

# Outline

- Setting & Setup
  - Abstraction
  - Paradox
- ⇒ Data: 1
- Data: 2
  - Comments & Conclusion

# Reinforcement Learning

- Old and obvious idea.
- Associative learning: state-action pairs. Those with favorable outcomes get reinforced; those without don't.
- Embodied in modern machine learning as Q-learning:

$$Q(s, a) = Q(s, a) + \alpha[r + \gamma \max_b Q(s', b) - Q(s, a)] \quad (1)$$

where  $\alpha$  is the learning rate parameter and  $Q(s,a)$  on the left is the new, updated value of  $Q(s,a)$ .

$Q(s,a)$  is a table of estimated values for state-action pairs.

# Reinforcement Learning in Games

- In a stable environment it is often possible to prove convergence to correctness for Q-learning (etc.)
- Strategic environments are another matter, yet there has been some, modest success in explaining human data with reinforcement learning models.
- Modest to disappointing results with artificial agents.
- Will now present results of several experiments.

## 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*.

# Learning in State Space

repeat forever:

1. Observe the current state,  $s_t$ .
2. Select the current action,  $a_t$ , from  $Q(s, a)$ .
3. Take action  $a_t$  and obtain reward  $r_t$ .
4. Update  $Q(s, a)$  based on  $r_t$ .

loop

Figure 1: Pseudo-Code for Q-Learning in Games

# Learning Regime

1. Alternative (or consideration) set. In the  $2 \times 2$  case, conditioning on the last play by the counter-player,  $\mathcal{A} = \{[0, 0], [0, 1], [1, 0], [1, 1]\}$  for each player.
2. Attractiveness estimation. E.g., 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}  
NB. Or close variants.
3. Choice/exploration policies.  
Softmax.  $\epsilon$ -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

CC	CD	DC	DD	$\delta$	Row's % CC
9422	218	183	177	0.05	0.963
9036	399	388	150	0.5	0.963
5691	738	678	2693	1	0.931
3506	179	275	6040	1.25	0.972
1181	184	116	8519	1.5	0.930
2	98	103	9797	1.75	0.805
97	114	91	9698	2	0.735
0	100	92	9808	2.5	0.839
2	96	94	9808	2.95	0.986

Table 2: Summary of Results for Prisoner's Dilemma.  $\epsilon$ -greedy action selection. Totals for the last 100 rounds of 100 series of 10,000 plays.

## Row Chooser's Wealth Extraction

Softmax		(DC)		(CC)	
Delta	Total WE	Pmax	WE-Q:Pmax	Pgmax	WE-Q:Pgmax
0.05	28875	3.05	0.947	3	0.963
0.50	28878	3.50	0.825	3	0.963
1.00	27921	4.00	0.698	3	0.931
1.25	29160	4.25	0.686	3	0.972
1.50	27902	4.50	0.620	3	0.930
1.75	24150	4.75	0.508	3	0.805
2.00	22046	5.00	0.441	3	0.735
2.50	25159	5.50	0.457	3	0.839
2.95	29577	5.95	0.497	3	0.986

# Stag Hunt

	C	D
C	$(5,5)^{**\#}$	$(0,3)$
D	$(3,0)$	$(\delta, \delta)^{\#}$

## Stag Hunt: Summary of Results

$\epsilon$ -greedy selection				action		Softmax selection			
CC	CD	DC	DD	$\delta$	CC	CD	DC	DD	
9390	126	122	362	0	9715	108	109	68	
9546	91	108	255	0.5	9681	120	121	78	
9211	112	125	552	0.75	9669	111	101	119	
8864	119	110	907	1	9666	98	102	134	
8634	115	132	1119	1.25	9598	139	134	129	
7914	122	130	1834	1.5	9465	99	109	327	
7822	122	104	1952	2	9452	126	126	296	
5936	87	101	3876	2.5	8592	116	89	1203	
5266	121	106	4507	3	3524	111	115	6250	

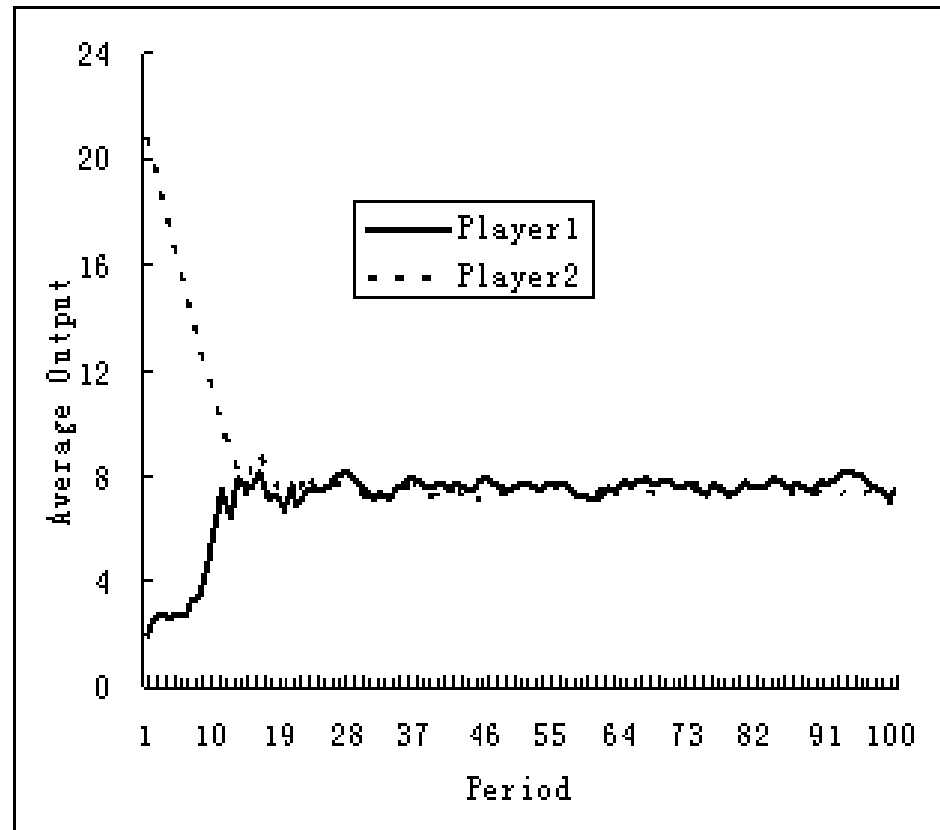
## Row Chooser's Total Wealth Extracted in Stag Hunt (Softmax)

$\delta$	Total WE	Pgmax (CC)	WE-Q:Pgmax
0.0	48902	5	0.978
0.5	48807	5	0.976
0.8	48737	5	0.975
1.0	48770	5	0.975
1.3	48553	5	0.971
1.5	48143	5	0.963
2.0	48230	5	0.965
2.5	46235	5	0.925
3.0	36805	5	0.736

## 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



## Characteristic Function Games: 3 Players

- Cooperative: agreements are enforced.

## Our Games

Game:	<u>I</u>	<u>II</u>	<u>III</u>	<u>IV</u>	<u>V</u>
$v(AB)$	95	115	95	106	118
$v(AC)$	90	90	88	86	84
$v(BC)$	65	85	81	66	50
<u>Quotas</u>					
$\omega_A$	60	60	51	63	76
$\omega_B$	35	55	44	43	42
$\omega_C$	30	30	37	23	8

Table 3: Characteristic Function and Quota Solutions by Game

# Summary of Results

Absolute, Percentage, and Total Absolute Deviation of Reward From Quota Solutions for 5K Episodes by Game (N=200)

	<u>I</u>	<u>II</u>	<u>III</u>	<u>IV</u>	<u>V</u>
<b>Softmax</b>					
A	0.621 (-1.04%)	0.178 (-.3%)	0.255 (-.5%)	0.62 (-.1%)	1.217 (-1.6%)
B	0.337 (.96%)	0.143 (-.26%)	0.047 (.11%)	0.099 (.23%)	0.49 (1.17%)
C	0.388 (1.29%)	0.378 (1.26%)	0.229 (.62%)	0.679 (2.95%)	1.308 (16.4%)
Total	1.346	.699	.531	1.398	3.015
<b>Greedy</b>					
A	0.487 (-.82%)	0.428 (-.71%)	0.19 (-.37%)	0.612 (-.97%)	1.015 (-1.34%)
B	0.181 (.52%)	0.258 (-.47%)	0.045 (-.10%)	0.192 (-.47%)	0.23 (.55%)
C	0.376 (1.25%)	0.777 (2.59%)	0.218 (.59%)	0.933 (4.06%)	1.159 (14.49%)
Total	1.044	1.463	.453	1.737	2.404
<b><math>\epsilon</math>-greedy</b>					
A	0.505 (-.84%)	0.411 (-.69%)	0.152 (-.30%)	0.615 (-.98%)	1.366 (-1.80%)
B	0.173 (.49%)	0.158 (-.29%)	0.004 (.01%)	0.034 (.08%)	0.411 (.98%)
C	0.404 (1.35%)	0.658 (2.19%)	0.158 (.43%)	0.712 (3.09%)	1.5 (18.75%)
Total	1.082	1.227	.314	1.361	3.277
<b>Human Subjects</b>					
A	2.57 (-4.28%)	3.00 (5%)	2.73 (5.36%)	.93 (-1.48%)	4.40 (-5.79%)
B	3.90 (11.14%)	.60 (-1.09%)	.60 (-1.36%)	2.20 (5.12%)	3.70 (8.81%)
C	.47 (-1.57%)	3.07 (-10.23%)	2.33 (-6.30%)	3.73 (-16.22%)	10.17 (127.13%)
Total	6.94	6.67	5.66	6.86	18.27

# Outline

- Setting & Setup
- Abstraction
- Paradox
- Data: 1
- ⇒ Data: 2
- Comments & Conclusion

# Learning in Policy Space

repeat forever:

1. Select a policy  $\pi_i \in \Pi$ , where  $\Pi$  is the consideration set of policies.
2. Pick a length of play,  $l$ , for policy  $\pi_i$ .
3. Play the next  $l$  rounds of the game using  $\pi_i$ .

Note: At each round,  $\pi_i$  will observe the current state,  $s_t$ , take an action  $a$  and obtain a reward  $r_t$ .

4. Update  $V^{\pi_i}$  based on the individual-round rewards,  $r_t$ s, obtained during the  $l$  rounds of play of policy  $\pi_i$ .

loop

Figure 2: Pseudo-Code for Policy-Space-Learning in Games

## Policy Learning Regime in 2×2 Games

1. Alternative (or consideration) set of policies:

$\mathcal{A} = \{000, 001, 010, 011, 100, 101, 110, 111\}$  for each player. Form: abc: play a the first time; if last time counter-player played 0, play b; if last time counter-player played 1, play c. E.g., in Prisoner's Dilemma, 101 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.  $\epsilon$ -greedy

# Parameterized Prisoner's Dilemma, again

$\delta$	Average Payoff	Modal Strategy	Freq.	Est. Value	Row's % CC
0.05	2.7224	5	0.5319	2.885	0.907
0.5	2.7577	5	0.7571	2.901	0.919
1.0	2.8108	5	0.8731	2.926	0.937
1.25	2.8139	5	0.8623	2.933	0.938
1.5	2.8083	5	0.8381	2.932	0.936
1.75	2.7950	5	0.8011	2.935	0.932
2.0	2.7314	5	0.6604	2.918	0.910
2.5	2.5324	0	0.8164	2.613	0.844
2.95	2.9524	0	0.8643	3.056	0.984

Table 4: Summary of Results for Policy-Space Learning in Prisoner's Dilemma. Average Payoff over 800,000 rounds of play. Modal Strategy=most frequently-played strategy; 5 = TIT FOR TAT, 0 = ALL DEFECT

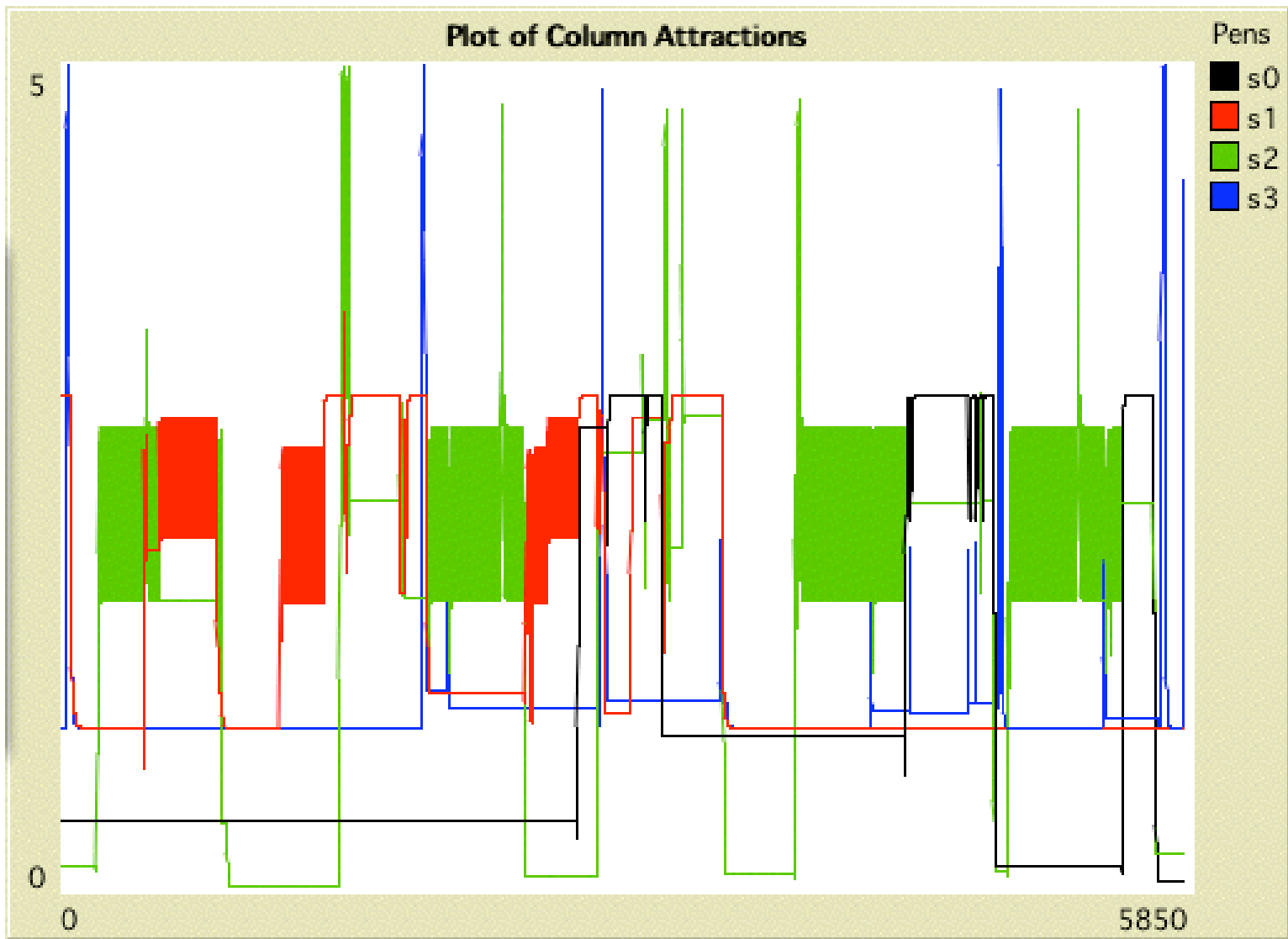
# Parameterized Stag Hunt, again

$\delta$	Average Payoff	Modal Strategy	Freq.	Est. Value	Row's % CC
0	4.402	5	0.4358	4.721	0.8804
0.05	4.4387	5	0.5104	4.705	0.8877
0.5	4.4747	5	0.6312	4.789	0.8949
1.0	4.5805	5	0.8221	4.854	0.9161
1.25	4.5544	5	0.7467	4.812	0.9109
1.5	4.6057	5	0.8757	4.864	0.9211
2.0	4.6263	5	0.8204	4.869	0.9253
2.5	4.6531	5	0.8822	4.884	0.9306
2.95	4.6733	5	0.8832	4.892	0.9347
3.0	4.7334	5	0.9040	4.901	0.9467

Table 5: Summary of Results for Policy-Space Learning in Stag Hunt. Average Payoff over 800,000 rounds of play. Modal Strategy=most frequently-played strategy; 5 = TIT FOR TAT.

## 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 $2 \times 2$ 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?

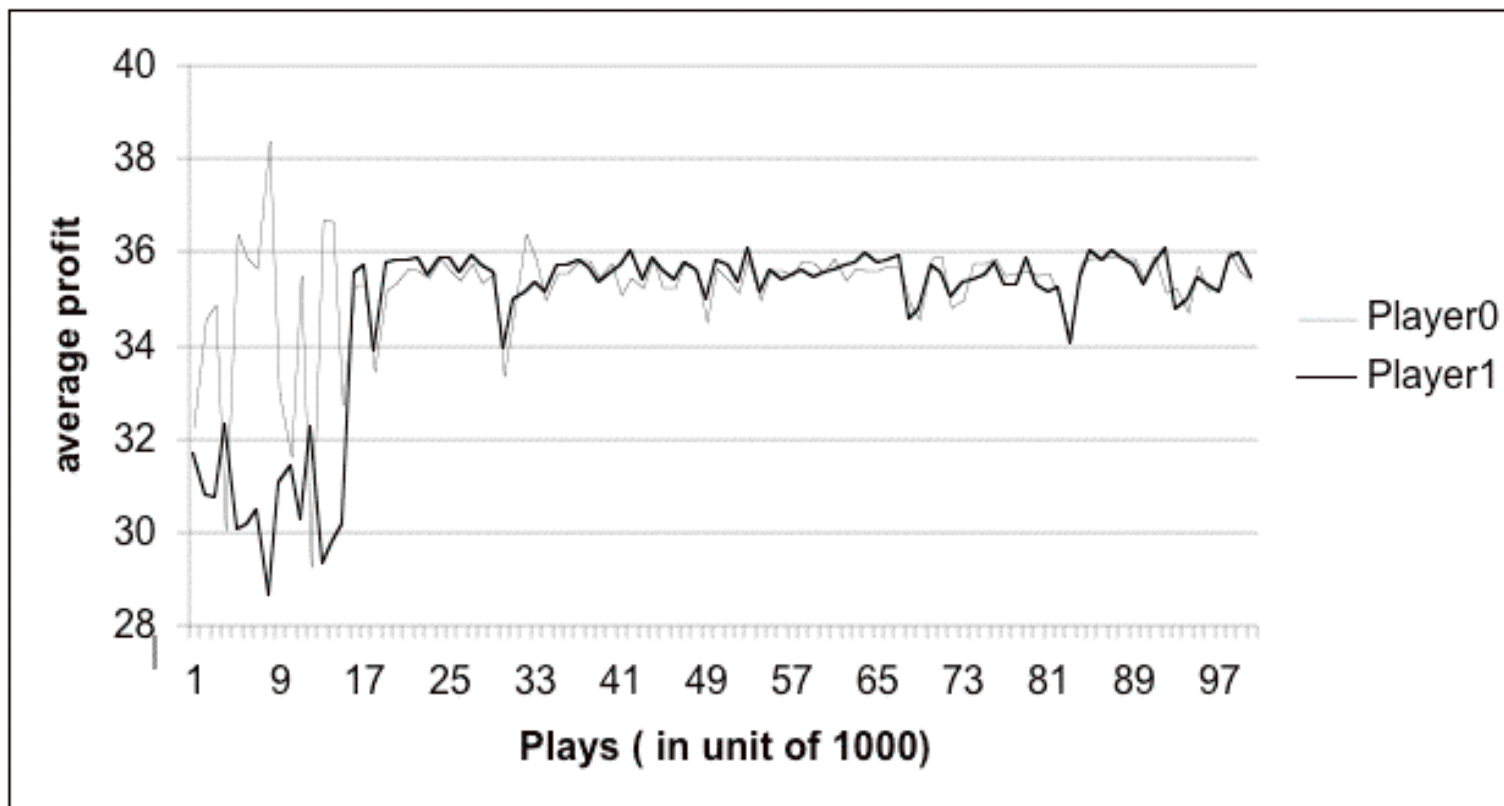
$$\mathcal{A} = \{\text{G-TFT}, \text{BR}, \text{S-TFT}, \text{CC}\}$$

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

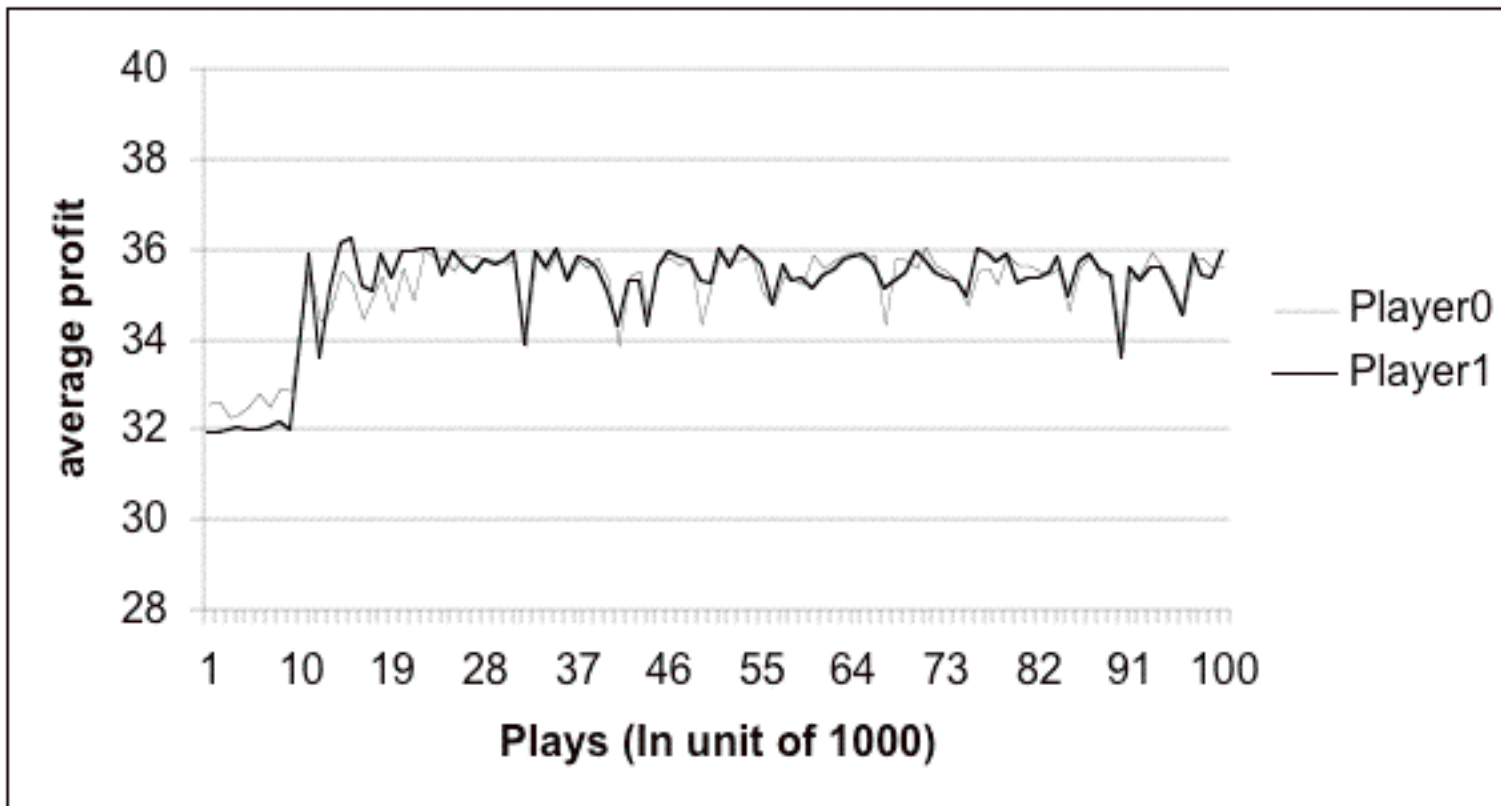
	Average Profit
Softmax action selection	(35.348, 35.323)
$\epsilon$ -greedy action selection	(35.487, 35.507)

Two typical runs follow...

# Molecular Strategies with Holt's Cournot Game: A



# Molecular Strategies with Holt's Cournot Game: B



# Conclusion

Much remains to be done, of course. Provisionally...

- Reinforcement learning, at least in the two modes displayed here, is a useful benchmark—or even solution concept—for repeated games. Predictions/outcomes are precise and reliable. This kind of learning is computable, tractable, and even cognitively plausible. Is paradox banished?
- Findings: the numbers matter; Pareto seems to predict better than Nash; these agents are very effective in extracting wealth.

## Conclusion

- Learning in policy space may be less sensitive to numerical values than learning in state space. We have demonstrated its effectiveness, compared to state-space learning, in a Cournot duopoly game. In addition, policy-space learning affords a number of conceptual and practical advantages.
- 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 in repeated play, as in spot markets?
- Exploring agents in games may realize a rather dynamic stability.

## To Consider

- Much as we accept multiple systems of etiquette as valid, perhaps we should accept multiple systems of rationality.
- Systems will materially differ in how apt they are in particular circumstances.
- Ideal rationality is often a counsel of perfection, insufficiently specific, impossible to realize, or descriptively implausible. An account of the “second best” is thus in order.
- Principled, thorough-going probing, exploring of the environment is a contending feature of any such alternate system of rationality.

## Why, e.g., ever choose a dominated option?

- To explore the environment.
- “The exception that probes the rule.”

\$Id: cmu-20040213-foils.tex,v 1.5 2004/02/16 14:08:40 sok Exp \$