Simple Reinforcement Learning Agents: Pareto Beats Nash in an Algorithmic Game Theory Study

Steven O. Kimbrough
Ming Lu
University of Pennsylvania
Draft: 9 November 2003

1 Background

Contexts of strategic interaction (CSIs) appear in nearly every social situation. They are characterized by interdependent decision making: two or more agents have choices to make and the rewards an individual receives in consequence of its choices depend, at least in part, on the choices made by other agents. Such contexts, when abstracted and formalized in certain ways, are the subject of game theory, which seeks to "solve"—predict and explain the outcomes of—games (i.e., of CSIs abstracted and formalized in certain stylized fashions).

Any solution theory for CSIs (or games) must make and rely upon two kinds of assumptions:

1. *SR (Strategic Regime) assumptions*. There are assumptions about the representation and structure of the CSI (or game), including the rules of play and the payoffs to the players. Typically, these assumptions are expressed as games in strategic form, games in extensive form, characteristic function games, spatial games, and so on.

2. *SSR assumptions*. These are assumptions about the Strategy Selection Regimes (SSRs) employed by the agents, or players, in the game. Classical game theory makes two kinds of SSR assumptions, which typically apply to all players (Luce & Raiffa, 1957; Shubik, 1982):

    a. *Ideal rationality assumptions*. It is normally assumed that agents are 'rational' and that Rational Choice Theory in some form (e.g., Savage's Subjective Expected Utility theory) characterizes this kind of (ideal) rationality. Roughly, agents are assumed to have utilities and to be maximizers of their utilities.

b. *Knowledge assumptions*. It is normally assumed that agents are omniscient with respect to the game. The agents know everything about the game, common knowledge obtains among all the players, and all agents have unlimited computational/ratiocination powers.

In what follows, we report on a series of experimental investigations that examine play in games under non-standard SSR assumptions, at least as judged by the classical game theory literature. We investigate a series of games that are well recognized in the classical literature and that have been extensively studied. Our game—Strategic Regime—Assumptions are conventional, although we focus on repeated or iterated (aka: staged) games.

It is, and has always been, recognized that the classical SSR assumptions (as we call them) are unrealistic. The original experimental work on Prisoner's Dilemma, among other games (Flood, 1952), was motivated by such concerns. Even so, they—and the consequences they engender—are interesting. The assumptions often afford tractability, allowing games to be 'solved'. Because they capture the notion of a certain plausible kind of ideal rationality, it is interesting to determine how well they describe actual human behavior. Even if they are inaccurate, they have value as a normative benchmark. And given the considerable powers of human cognition and institutions, it is not prima facie implausible that classical SSR assumptions will often yield accurate predictions.

This is all well and good, but the story is not over. There are certain puzzles or anomalies associated with the classical SSR assumptions. Famously in the Prisoner's Dilemma game, and in other games, the Nash Equilibrium (NE) outcome is not Pareto efficient. Classical theory sees the NE as the solution to the game, yet many observers find it anomalous and experiments with human subjects often indicate support for these observers (Luce & Raiffa, 1957; Rapoport & Guyer 1976). Further, the NE need not be unique, posing thereby a challenge to the classical theory, which often struggles, or has to be stretched, to predict equilibrium outcomes that seem natural and that are reached by human subjects easily. In short, the classical theory has often proved to be a poor—weak and inaccurate—predictor of human behavior (Roth & Erev, 1995).

Besides the well-known puzzles and anomalies, there is another category of reasons to study games under variations of the classical SSR assumptions. Rational Choice Theory and omniscience may be plausible assumptions for experienced humans in certain favorable institutional settings (e.g., well-established markets). They are often not plausible assumptions for games played by birds, bees, monkeys up in trees, bacteria, and other similarly less cognitively well-endowed creatures. It is, simply put, scientifically interesting to investigate the play and outcomes in games in which the SSR assumptions of classical game theory are relaxed sufficiently to be capable of describing these kinds of more limited agents. Equally so, this is interesting from a practical, applications-oriented perspective. Adaptive artificial agents, e.g. fielded for purposes of electronic commerce, will inevitably resemble the lower animals more than their creators, at least in their cognitive powers.

With these motivations principally in mind, we investigated repeated play by simple, adaptive agents in a number of well-known games. Any such investigation, however, faces an immediate and urgent theoretical problem: There are indefinitely many ways to relax the classical SSR assumptions; how does one justify a particular alternative? We choose with a number of criteria in mind.

1. *Simple*. There are few ways to be ideally rational and indefinitely many ways not to be. In examining alternatives it is wise to begin with simple models and complexify as subsequent evidence and modeling ambition requires.
2. *New*. Much has been learned about non-ideally rational agents through studies of the replicator dynamic (see Gintis, 2000, for a review). These investigations, however, see *populations* as evolving, rather than individual agents adapting. The individuals are typically modeled as naked, unchanging strategies, rather than adaptive agents, which proliferate or go extinct during the course of continuing play. Agents in some 'spatialized', cellular automata-style games have been given certain powers of state change and adaptation, but these have on the whole been limited in scope (e.g., Epstein & Axtell, 1996; Grim et al., 1998). Experimenting with game-playing agents that are using reinforcement learning is a comparatively under-developed area and the kinds of experiments we report here are, we believe, original.
3. *Theoretically motivated*. Reinforcement learning as it has developed as a field of computational study has been directly and intendedly modeled on learning theories from psychology, where there is an extensive supporting literature. This important class of learning model is a natural first choice for modeling agents in games, because it appears to apply broadly to other areas of learning, because its theoretical properties have been well investigated, and because it has achieved a wide scope of application in multiple domains.
4. *Adaptive*. Agents should be responsive to their environments and be able to learn effective modes of play.
5. *Exploring*. Agents should be able actively to probe their environments and undertake exploration in the service of adaptation; agents face the exploration-exploitation tradeoff and engage in both.

In addition, the SSRs should be realizable in sense that they specify definite procedures that simple agents could actually undertake. It is here, perhaps, that the present approach, which we label *algorithmic game theory*, differs most markedly from classical game theory and its assumption of ideal rationality, irrespective of realizability constraints.

We turn now to a discussion of the elements of reinforcement learning needed as background for our experiments.

2 Reinforcement learning

2.1 Simple Q-learning

Our experimental agents used a simple form of Q-learning, itself a variety of reinforcement learning. Detailed description of Q-learning is easily found in the open literature (e.g., Watkins, 1989; Watkins & Dayan, 1992; Sutton & Barto, 1998). We limit ourselves here to a minimal summary for the purposes at hand.

The Q-learning algorithm works by estimating the values of state-action pairs. The value $Q(s,a)$ is defined to be the expected discounted sum of future payoffs obtained by taking action a in state s and following an optimal policy thereafter. Once these values have been learned, the optimal action from any state is the one with the highest Q-value. The standard procedure for Q-learning is as follows. Assume that $Q(s,a)$ is represented by a lookup table containing a value for every possible state-action pair, and that the table entries are initialized to arbitrary values. Then the procedure for estimating $Q(s,a)$ is to repeat the following loop until a termination criterion is met:

1. Given the current state s choose an action a. This will result in receipt of an immediate reward r, and transition to a next state s'. (We discuss below the policy used by the agent to pick particular actions, called the exploration strategy.)

2. Update $Q(s, a)$ according to the following equation:

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

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

In the context of repeated games, a reinforcement learning (Q-learning) player explores the environment (its opponent and the game structure) by taking some risk in choosing actions that might not be optimal, as estimated in step 1. In step 2 the action that leads to higher reward will strengthen the Q-value for that state-action pair. The above procedure is guaranteed to converge to the correct Q-values for stationary Markov decision processes.

In practice, the exploration policy in step 1 (i.e., the action-picking policy) is usually chosen so that it will ensure sufficient exploration while still favoring actions with higher value estimates in given state. A variety of methods may be used. A simple method is to behave greedily most of the time, but with small probability, ε, choose an available action at random from those that do not have the highest Q value. For obvious reasons, this action selection method is called ε-greedy (see Sutton & Barto, 1995). Softmax is another commonly used action selection method. Here again, actions with higher values are more likely to be chosen in given state. The most common form for the probability of choosing action *a* is

$$\frac{e^{Q_t(a)/\tau}}{\sum_{b=1}^{n} e^{Q_t(b)/\tau}} \qquad (2)$$

where τ is a positive parameter and decreases over time. It is typically called the temperature, by analogy with annealing. In the limit as $\tau \to 0$, Softmax action selection becomes greedy action selection. In our experiment we investigated both ε-greedy and Softmax action selection.

2.2 *Implementation of Q-learning for 2 by 2 games*

A Q-learning agent does not require a model of its environment and can be used on-line. Therefore, it is quite suited for repeated games against an unknown co-player (especially an adaptive, exploring co-player). Here, we will focus on certain repeated 2 by 2 games, in which there are two players each having two possible plays/actions at each stage of the game. It is natural to represent the state of play, for a given player, as the outcome of the previous game played. We say in this case that the player has memory length of one. The number of states for 2 by 2 game is thus 4 and for each state there are two actions (the pure strategies) from which the player can choose for current game. We also conducted the experiments for the case that players have memory length of two (the number of states will be 16) and obtained broadly similar results. The immediate reward a player gets is specified by the payoff matrix.

For the Softmax action selection method, we set the decreasing rate of the parameter τ as follows.

$$\tau = T * \Theta^n \qquad (3)$$

T is a proportionality constant, *n* is number of games played so far. Θ, called the annealing factor, is a positive constant that is less than 1. In the implementation, when *n* becomes large enough, τ is close to zero and the player stops exploring. We use Softmax, but in order to avoid cessation of exploration, our agents start using ε-greedy exploration once the Softmax progresses to a point (discussed below) after which exploration is minimal.

3 Experiments

3.1 Motivation

Repeated 2 by 2 games are the simplest of settings for strategic interactions and are a good starting point to investigate how outcomes arise under a regime of exploring rationality versus the ideal rationality of classical game theory. The Definitely Iterated Prisoner's Dilemma, involving a fixed number of iterations of the underlying game, is a useful example. Classical game theory, using a backwards induction argument, predicts that both players will defect on each play (Luce & Raiffa, 1957). If, on the other hand, a player accepts the risk of cooperating, hoping perhaps to induce cooperation later from its counter-player, it is entirely possible that both players discover the benefits of mutual cooperation. Even if both players suffer losses early on, subsequent sustained mutual cooperation may well reward exploration at the early stages.

Motivated by this intuition, we selected 8 games and parameterized their payoffs. The players are modeled as Q-learners in each repeated game. In 5 of the games the Pareto optimal (socially superior, i.e., maximal in the sum of its payoffs) outcome does not coincide with a Nash Equilibrium. The remaining 3 games, which we included to address the multi-equilibrium selection issue, each have two pure-strategy NEs.

3.2 The games and the parameterization

We parameterized each of our 8 games via a single parameter, δ, in their payoff matrices. In the payoff matrices below, the first number is the payoff to the row player and the second is the payoff to the column player. We mark the Nash Equilibria with # and the Pareto efficient outcomes with *. Pareto optimal (socially superior) outcomes are labeled with **. C and D are the actions or pure strategies that players can take on any single round of play. The row player always comes first in our notation. Thus, CD means that the row player chose pure strategy C and column player chose pure strategy D. So there are four possible outcomes of one round of play: CC, CD, DC, and DD.

The first two games are versions of Prisoner's Dilemma (PD). The value of δ ranges from 0 to 3. When its value is 2 (see Table 1), it corresponds to the most common payoff matrix in the Prisoner's Dilemma literature.

Place Table 1, “Prisoner’s Dilemma, Pattern 1”, approximately here.

	C	D
C	(3, 3)**	(0,3+δ)*
D	(3+δ, 0)*	(3-δ, 3-δ)#

Table 1: Prisoner’s Dilemma, Pattern 1

Place Table 2, “Prisoner’s Dilemma, Pattern 2”, approximately here.

	C	D
C	(3,3)**	(0, 3+δ)*
D	(3+δ, 0)*	(δ, δ)#

Table 2: Prisoner’s Dilemma, Pattern 2

While the Prisoner’s Dilemma, in its usual form, is a symmetric game (see Tables 1 and 2), the following three games, adapted from Rapoport and Guyer (1976), are asymmetric. The value of δ ranges from 0 to 3 in our experiments with these games. Note that as in Prisoner’s Dilemma, in Games #47, #48, and #57 (Tables 3—5) the Nash Equilibrium does not coincide with the Pareto optimal outcome.

Place “Table 3: Game #47” approximately here.

	C	D
C	(0.2 ,0.3)#	0.3+δ, 0.1*
D	0.1, 0.2	(0.2+δ, 0.3+δ)**

Table 3: Game #47

Place “Table 4: Game #48” approximately here.

	C	D
C	(0.2, 0.2)#	(0.3+δ, 0.1)*
D	(0.1, 0.3)	(0.2+δ, 0.3+δ)**

Table 4: Game #48

Place “Table 5: Game #57” approximately here.

	C	D
C	(0.2 ,0.3)#	(0.3+δ, 0.2)*
D	(0.1, 0.1)	(0.2+δ, 0.3+δ)**

Table 5: Game #57

For games with two NE, the central question is which equilibrium (if any) is most likely to be selected as the outcome. We choose three examples from this class of game. The game of Stag Hunt has a Pareto optimal solution as one of its two NE. The game of Chicken and the game of Battle of Sexes are coordination games. In Battle of the Sexes the two coordination outcomes (CC and DD) are NEs and are Pareto optimal. In Chicken, the coordination outcomes (CD and DC) may or may not be NEs, depending on δ. The value of δ ranges in our experiments from 0 to 3 for Stag Hunt and Bottle of sexes. For Chicken, the range is from 0 to 2.

Insert “Table 6: Stag Hunt” approximately here.

	C	D
C	(5,5)**#	(0,3)
D	(3,0)	(δ, δ)#

Table 6: Stag Hunt

Insert “Table 7: Battle of the Sexes” approximately here.

	C	D
C	(δ, 3-δ)**#	(0,0)
D	(0,0)	(3-δ, δ)**#

Table 7: Battle of the Sexes

Insert “Table 8: Chicken” approximately here.

	C	D
C	(2,2)*	(δ, 2+δ)*#
D	(2+δ, δ)*#	(0, 0)

Table 8: Chicken. . $0 \leq \delta < 2$. CC is ** for $\delta \leq 1$. CD and DC are ** for $\delta \geq 1$.

3.3 Settings for the experiments

We set the parameters for Q-learning as follows. Learning rate, α = 0.2 and discount factor, γ = 0.95. We ran the experiment with both Softmax action selection and ε-greedy action selection**.** For Softmax action selection, T is set to 5 and the annealing factor Θ = 0.9999. When τ is less than 0.01, we began using ε-greedy action section. We set ε to 0.01. We note that these parameter values are typical and resemble those used by other studies (e.g., Sandholm and Crites 1995). Also, our results are robust to changes in these settings.

Each game was iterated 200,000 times in order to give the players enough time to explore and learn. For each setting of the payoff parameter δ, we ran the repeated game 100 times. We recorded the frequencies of the four outcomes (CC, CD, DC and DD) every 100 iterations. The numbers usually become stable within 50,000 iterations, so we took frequencies of the outcomes in the last 100 iterations over the 200,000 iterations to report, unless noted otherwise.

The summary of results tables, below, all share a similar layout. In the middle column is the payoff parameter δ. On its left are the results for ε-greedy action selection. The results for Softmax action selection are on the right. Again, the numbers are frequencies of the four outcomes (CC, CD, DC and DD) in the last 100 iterations, averaged over 100 runs.

3.4 Results

It is generally recognized as disturbing or at least anomalous when classical game theory predicts that a Pareto inferior Nash Equilibrium will be the outcome, rather than a Pareto optimal solution (Flood, 1952; Luce & Raiffa, 1957; and ever since). This is exactly what happens in our first five games, in which the unique subgame perfect Nash Equilibrium is never the Pareto optimal (or even a Pareto efficient!) outcome. Will the outcomes be different if agents use adaptive, exploring SSRs, such as reinforcement learning? More specifically, can players learn to achieve a Pareto optimal solution that is not a Nash Equilibrium? Among competing NEs, will players find the Pareto efficient outcome? Our results indicate a broadly positive answer to these questions.

Consider Table 9, "Summary of Results for Prisoner's Dilemma, Pattern 1" (summarizing results for the parameterized PD game in Table 1). If δ is close to zero, the two players choose to defect most of the time. (That is, see above, during the final 100 rounds of 200,000 iterations, they mostly play DD. The entries in Table 9, and in similar tables report counts out of 100 rounds × 100 runs = 10,000 plays.) We note, by way of explanation, that there is not much difference in rewards between mutual defection and mutual cooperation: 3-δ and 3, with δ small. The Pareto optimal outcome does not appear to provide enough incentive for these players to risk cooperation. But as δ gets larger, we see more cases of mutual cooperation. The last row in Table 9 has an interesting interpretation: The players have incentive to induce each other's cooperation so as to take advantage of it by defecting. This is always the case in Prisoner's Dilemma, but exacerbated here (final row of Table 9) because the temptation for defection in the presence of cooperation is unusually large. Consequently, we see many CDs and DCs, but less mutual cooperation (CC). Notice that CC is maximized and DD minimized somewhere in the range of [1.75, 2] for δ. (Softmax and ε-greedy results are, here and elsewhere, in essential agreement.) When δ is low the benefit of mutual cooperation is too low for the agents to find the Pareto optimal outcome. When δ is very high, so is the benefit of defection in the face of cooperation, and again the agents fail to cooperate jointly. In the middle, particularly in the [1.75, 2] range, the benefits of mutual cooperation are high enough and the temptation to defection is low enough that substantial cooperation occurs.

Place Table 9, "Summary of Results for Prisoner's Dilemma, Pattern 1" approximately here.

ε−greedy action selection					Softmax action selection			
CC	CD	DC	DD	δ	CC	CD	DC	DD
3	87	82	9828	0.05	0	106	101	9793
0	92	105	9803	0.5	0	90	94	9816
52	110	111	9727	1	1	111	111	9777
51	110	93	9746	1.25	2475	338	358	6829
1136	160	198	8506	1.5	3119	526	483	5872
1776	245	381	7598	1.75	4252	653	666	4429
3526	547	413	5514	2	789	883	869	7549
848	766	779	7607	2.5	496	2276	2368	4860
544	2313	2306	4837	2.95	539	2821	2112	4528

Table 9: Summary of Results for Prisoner's Dilemma, Pattern 1

Further insight is available by considering Table 10, the Wealth Extraction Report for Table 9. The Total Wealth Extracted (WE) by an agent is simply the number of points it obtained in playing a game. Table 10 presents the Total WE for the row chooser in PD, Pattern 1. (Results are similar for the column chooser; this is a symmetric game played by identically endowed agents.) WE-Q:Pmax is the ratio (quotient, Q) of (a) Total WE and (b) 100 iterations × 100 runs × Pmax, the maximum number of points row chooser could get from outcomes on the Pareto frontier. Pmax = ($3+\delta$) and is realized when DC is played. WE-Q:Pgmax is the ratio of (a) Total WE and (b) 100 iterations × 100 runs × Pgmax, the maximum number of points row chooser could get from outcomes on the Pareto frontier whose total rewards are maximal (among Pareto efficient outcomes). Here, Pgmax = 3 and is realized when CC is played. WE-Q:Pgmax might be called the "wealth extraction quotient for socially optimal outcomes." Each of these measures declines as δ increases. Our agents have progressively more difficulty extracting available wealth. This hardly seems surprising, for at $\delta=0.05$ the game is hardly a PD at all and the reward 3 for mutual cooperation is a paltry improvement over the 'penalty' for mutual defection, 2.95. As delta increases, however, strategy selection becomes more and more of a dilemma and the agents become less and less successful in extracting wealth from the system. Note that these trends are more or less monotonic (see Table 10), while the actual outcomes change rather dramatically (see Table 9). From the perspective of classical game theory, changes in delta should not matter. Each of these games is a PD and should produce identical outcomes, all DD. Note that had the players played DD uniformly when delta=2.95, the row (and similarly the column) player would have extracted a total wealth of 10,000 × 0.05 = 500. In this light, extracting 14,410 is a considerable achievement.

Place Table 10, “Row Chooser’s Total Wealth Extracted in Prisoner’s Dilemma, Pattern 1” approximately here.

Softmax		(DC)		(CC)	
delta	Total WE	Pmax	WE-Q:Pmax	Pgmax	WE-Q:Pgmax
0.05	29197	3.05	0.957	3	0.973
0.50	24869	3.50	0.711	3	0.829
1.00	20001	4.00	0.500	3	0.667
1.25	20897	4.25	0.492	3	0.697
1.50	20339	4.50	0.452	3	0.678
1.75	21456	4.75	0.452	3	0.715
2.00	14261	5.00	0.285	3	0.475
2.50	16942	5.50	0.308	3	0.565
2.95	14410	5.95	0.242	3	0.480

Table 10: Row Chooser’s Total Wealth Extracted in Prisoner’s Dilemma, Pattern 1

Consider now the parameterized family of Prisoner’s Dilemma Pattern 2 games (see Table 2). Here, the players stand to lose almost nothing by trying to cooperate when δ is close to zero. Exploration seems to help players reach the superior (“socially superior”) Pareto optimal outcome (CC) and as we can see from Table 11, mutual cooperation happens 94% of time. Consider the scenario with δ close to 3. Note first, there is not much incentive to shift from the Nash equilibrium (DD) to the socially superior Pareto outcome (CC), since there is not much difference in payoffs; second, the danger of being exploited by the other player and getting zero payoff is much higher. Indeed, the players learn to defect most of the time (98%).

Place Table 11, “Table 11: Summary of Results for Prisoner’s Dilemma, Pattern 2” approximately here.

ε-greedy action selection					Softmax action 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

Table 11: Summary of Results for Prisoner’s Dilemma, Pattern 2

The Wealth Extraction Report, Table 12, for Pattern 2 corresponds to Table 10 for Pattern 1. We see that our row chooser is able to extract a roughly constant amount of wealth from the game, even as delta and the strategy choices vary drastically. Note further that WE-Q:Pgmax is approximately constant (mostly over 90%) even though the players are mostly not playing CC at all.

Place Table 12, "Row Chooser's Total Wealth Extracted in Prisoner's Dilemma, Pattern 2" approximately here.

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

Table 12: Row Chooser's Total Wealth Extracted in Prisoner's Dilemma, Pattern 2

We now turn to games #47, #48, and #57, which are asymmetric games having a common feature: the row player has a dominant strategy C. Thus a fully rational row player will never choose D. What will happen if players are able to explore and learn? Tables 13—15 tell us that it depends on the payoffs. If δ is close to zero, the outcome will be the Nash equilibrium (CC) almost always. As δ increases, however, the incentives favoring the socially superior Pareto outcome (CC) concomitantly increase, drawing the players away from CC (Nash) to DD (socially superior Pareto). We note that row chooser would prefer CD to DD, yet in all three games (see Tables 13—15) we see a similar pattern of CD play as delta increases.

Place “Table 13: Game #47” approximately here.

ε−greedy action selection					Softmax action selection			
CC	CD	DC	DD	δ	CC	CD	DC	DD
9790	101	101	8	0	9808	94	98	0
4147	137	156	5560	0.1	9812	94	93	1
3019	123	165	6693	0.15	9799	95	104	2
2188	141	132	7539	0.2	8934	85	109	872
185	355	130	9330	0.5	730	284	208	8778
131	309	135	9425	1	120	532	138	9210
138	288	99	9475	1.5	77	471	103	9349
99	321	131	9449	2	88	441	126	9345
126	172	88	9614	3	64	366	92	9478

Table 13: Game #47

Place “Table 14: Game #48” approximately here.

ε-greedy action selection					Softmax action selection			
CC	CD	DC	DD	δ	CC	CD	DC	DD
9789	102	107	2	0	9787	106	105	2
3173	515	173	6139	0.1	9811	86	101	2
2832	457	207	6504	0.15	8127	256	137	1480
1227	348	141	8284	0.2	2986	755	230	6029
109	627	143	9121	0.5	143	631	146	9080
90	492	139	9279	1	79	1320	126	8475
88	318	134	9460	1.5	117	1076	128	8679
241	236	119	9404	2	62	473	126	9339
76	284	139	9501	3	64	277	128	9531

Table 14: Game #48

Place “Table 15: Game #57” approximately here.

ε-greedy action selection					Softmax action selection			
CC	CD	DC	DD	δ	CC	CD	DC	DD
9767	119	107	7	0	9764	131	105	0
1684	587	175	7554	0.1	9794	106	98	2
531	518	191	8760	0.15	9550	105	105	240
238	543	159	9060	0.2	1048	497	257	8198
126	307	121	9446	0.5	224	852	152	8772
118	520	114	9248	1	113	753	119	9015
104	526	125	9245	1.5	74	538	117	9271
66	225	102	9607	2	57	569	123	9251
123	296	116	9465	3	61	302	125	9512

Table 15: Game #57

The Wealth Extraction Reports for game #47 are also useful for understanding the row versus column power relationship in these games (Tables 16—17). Notice that at the Nash Equilibrium (CC) total WE for row is 2/3 of that for column. See Tables 16—17 for delta=0. As delta increases and CC play decreases both players uniformly increase their WE. At the same time, their WE becomes more and more equal, and by the time delta=2 row chooser is extracting more wealth from the game than column chooser. This occurs even though in more than 92% of the games the play is DD and column chooser extracts more wealth than row chooser! The difference is due to the occasional ‘defection’ by row chooser to play C. Finally, we note that DC is neither Nash nor Pareto in these games. Our agents play DC at a rate that is low and essentially invariant with delta. That rate may be interpreted as a cost consequence of exploration.

Finally, it is instructive to note that when delta = 3 the expected value for column playing D is 3.3-3.2*p*, if row plays C with probability *p*. Similarly the expected value of playing C is 0.2+0.1*p*. Consequently, column should play D so long as $p < 31/33$. These considerations lead us to wonder whether our row chooser agents have not learned to be sufficiently exploitive. They may be too generous to column chooser, although column chooser is not without recourse. However, the fact that C and D for row chooser are so close in value, given that column chooser plays D, may impute stability in this stochastic, noisy, learning context. Note that in all three games CD is more rare when delta = 3 than when delta = 2.

Place "Table 16: Row Chooser's Total Wealth Extracted in Game #47" approximately here.

Softmax	#47	(CD)		(DD)	
Delta	Total WE	Pmax	WE-Q:Pmax	Pgmax	WE-Q:Pgmax
0	2000	0.30	0.667	0.20	1.000
0.1	2010	0.40	0.502	0.30	0.670
0.15	2014	0.45	0.447	0.35	0.575
0.2	2189	0.50	0.438	0.40	0.547
0.5	6539	0.80	0.817	0.70	0.934
1	11781	1.30	0.906	1.20	0.982
1.5	16767	1.80	0.931	1.70	0.986
2	21604	2.30	0.939	2.20	0.982
3	31559	3.30	0.956	3.20	0.986

Table 16: Row Chooser's Total Wealth Extracted in Game #47

Put "Table 17: Column Chooser's Total Wealth Extracted in Game #47", about here.

Column Chooser:

Softmax Delta	#47 Total WE	(DD) Pmax	WE-Q:Pmax	(DD) Pgmax	WE-Q:Pgmax
0	2962	0.30	0.987	0.30	0.987
0.1	2963	0.40	0.741	0.40	0.741
0.15	2961	0.45	0.658	0.45	0.658
0.2	3136	0.50	0.627	0.50	0.627
0.5	7291	0.80	0.911	0.80	0.911
1	12076	1.30	0.929	1.30	0.929
1.5	16909	1.80	0.939	1.80	0.939
2	21577	2.30	0.938	2.30	0.938
3	31342	3.30	0.950	3.30	0.950

Table 17: Column Chooser's Total Wealth Extracted in Game #47

In PD and games #47, #48, and #57, the Nash Equilibrium is not on the Pareto frontier. The Stag Hunt game is thus interesting because its Pareto optimal solution is also one of its two pure strategy NEs. But which one, or which mixture, will be sustained remains a challenging problem for classical game theory. A mixed strategy seems natural in this repeated game for classical game theory. Table 18 shows that the outcomes of for our reinforcement learning agents do not conform to the prediction of a mixed strategy. Say, for example, when delta is equal to 1, the mixed strategy for both players will be choosing action C with probability 1/3 and D with probability 2/3. (Let p be the probability of playing C, then at $5p+0p=3p+(1-p)$ the players are indifferent between playing C or D. This happens at p=1/3.) We should expect to see CC with a frequency less than 33%, while Table 15 shows CC happening at a rate of 88%.

Place "Table 18: Summary of Results for Stag Hunt" approximately here.

ε-greedy action selection					Softmax action selection			
CC	CD	DC	DD	δ	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

Table 18: Summary of Results for Stag Hunt

In Stag Hunt, CC is Pareto optimal but risky, while DD is riskless (on the down side) but Pareto dominated. As delta increases from 0 to 3.0 the risk/reward balance increasingly favors DD. Our agents respond by favoring DD at the expense of CC and in consequence they extract a decreasing amount of wealth. It is as if they were operating with a risk premium, yet we know they are not.

Place “Table 19: Row Chooser’s Total Wealth Extracted in Stag Hunt”, approximately here.

Row Chooser:

Softmax delta	Stag Hunt Total WE	(CC) Pgmax	WE-Q:Pmax 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

Table 19: Row Chooser’s Total Wealth Extracted in Stag Hunt

The remaining two games are coordination games. We are concerned not only with which NEs are to be selected, but also with a larger question: Is the Nash Equilibrium concept apt for describing what happens in these games? The later concern arises as we observe different behavior in human experiments. Rapport et al. (1976) reported a majority of subjects quickly settling into an alternating strategy, with the outcome changing back and forth between the two Nash coordination points (CD and DC) when playing the game of Chicken.

From Table 20 we can see these two NEs (and coordination points) in Battle of the Sexes are equally likely to be the outcome in most cases since the game is symmetric and these two outcomes are superior to other two, which give both players a zero payoff. In the game of Chicken (Table 21) we see that if the incentive for coordinating is too small (i.e., delta is close to zero), the players learn to be conservative and land on the non-NE (CC) since they cannot afford the loss resulting from DD (getting zero). As delta increases, the game ends up more and more in one of the Nash coordination points (CD or DC).

Insert “Table 20: Summary of Results for Battle of the Sexes” approximately here.

ε-greedy action selection					Softmax action selection			
CC	CD	DC	DD	δ	CC	CD	DC	DD
2641	63	4571	2725	0	2872	73	4477	2578
3842	135	1626	4397	0.1	4615	101	1732	3552
5140	102	90	4668	0.5	4772	102	162	4964
4828	107	94	4971	1	4862	88	89	4961
4122	101	109	5668	1.5	4642	85	102	5171
4983	100	97	4820	2	4623	97	87	5193
3814	111	96	5979	2.5	5139	102	99	4660
4015	1388	107	4490	2.9	4303	1794	118	3785
2653	4921	70	2356	3	2593	4776	58	2573

Table 20: Summary of Results for Battle of the Sexes

Insert “Table 21: Summary of Results for Chicken” approximately here.

ε-greedy action selection					Softmax action selection			
CC	CD	DC	DD	δ	CC	CD	DC	DD
9276	227	347	150	0	9509	165	222	104
9587	143	135	135	0.25	9119	428	320	133
9346	209	223	222	0.5	9375	220	225	180
6485	1491	1858	166	0.75	8759	424	632	185
1663	3532	4706	99	1	1339	4903	3662	96
385	4161	5342	112	1.25	158	5416	4323	103
113	4488	5274	125	1.5	115	4700	5099	86
111	4301	5504	84	1.75	100	4704	5083	113
100	4853	4953	94	2	94	4772	5044	90

Table 21: Summary of Results for Chicken

The Wealth Extraction Report for Chicken, Table 22, is particularly revealing. When delta is small, play is overwhelmingly CC. CC is non-Nash and Pareto and for delta ≤ 1.0 CC is socially superior Pareto. C is less risky for both players than D (both CD and DC are Pareto and Nash outcomes), so when delta is small it stands to reason that our agents should stick with CC. Note in this regard that if the players exactly alternate the CD and DC outcomes, they each will receive a payoff of 1+delta on average. See the column labeled "WE if Perfect Alternation" in Table 22. We see that when delta < 1.0 (i.e., when CC is socially superior), CC play is preponderant and Total WE is greater, often substantially greater, than WE if Perfect Alternation. For delta ≥ 1.0, CD and DC are socially superior Pareto. In the neighborhood of 1.0, play transitions from predominantly CC to predominantly CD and DC. Note that Total WE increases uniformly with delta (except for a slight decline in the neighborhood of delta=1.0, which we attribute to transition-induced error). As delta ranges from 1.0 to 2.0, Total WE closely approximates WE if Perfect Alternation. In short, the agents are impressively effective at extracting wealth. Outcomes are Nash (for the most part) if and only if there is not more money to be made elsewhere.

Insert "Table 22: Row Chooser's Total Wealth Extracted in Chicken" about here.

Row chooser:		Chicken				
Softmax delta	Total WE	(DC) Pmax	WE-Q:Pmax	Pgmax	WE-Q:Pgmax	WE if Perfect Alternation
0.0	19482	2.00	0.974	1.0	1.948	10000
0.3	19065	2.25	0.847	1.3	1.525	12500
0.5	19423	2.50	0.777	1.5	1.295	15000
0.8	19574	2.75	0.712	1.8	1.119	17500
1.0	18567	3.00	0.619	2.0	0.928	20000
1.3	21136	3.25	0.650	2.3	0.939	22500
1.5	25127	3.50	0.718	2.5	1.005	25000
1.8	27493	3.75	0.733	2.8	1.000	27500
2.0	29908	4.00	0.748	3.0	0.997	30000

Note: Pgmax assumes perfect alternation of CD and DC.

Table 22: Row Chooser's Total Wealth Extracted in Chicken

In order to see if players can learn alternating strategies, as observed in human subject experiments, we conducted another 100 trials for these two games with delta set to 1 and with Softmax action selection. For most of the trials the outcomes converge (i.e., settle, Dworman et al., 1995, 1996) to one of the Pareto superior outcomes. But we did observe patterns showing alternating strategies for both games. These patterns are quite stable and can recover quickly from small random disturbances. For the Battle of the Sexes, we observed only one alternating pattern: the players playing the two Nash Equilibria alternately, in sequence. This pattern occurred in 11 out of 100 trials. For Chicken, we observed other kinds of patterns and have summarized their frequencies in Table 23.

Insert “Table 23: Frequencies of different patterns of outcome in the game of Chicken” approximately here.

The outcomes	Frequency in 100 trials
Alternating between CD and DC	10
Cycle through CD-DC-CC or CD-CC-DC	13
Converge to one of the three: CC, CD or DC	76
No obvious pattern	1

Table 23: Frequencies of different patterns of outcome in the game of Chicken

At 23% (10+13 of 100), the proportion of alternating patterns cannot be said to be large. Note first that we have used payoffs different from Rapport et al. (1976) and this may influence the incentive to form alternating strategies. Second, our players do not explicitly know about the payoff matrix and can only learn about it implicitly through play. Finally, there certainly are some features of human adaptive strategic behavior that are not captured in our current Q-learning model but that are important for human subjects to learn such alternating strategies. The main point, however, is how irrelevant the Nash Equilibrium concept seems for describing the outcomes of the repeated coordination games—Chicken and Battle of the Sexes—as played by our agents.

4 Summary

Wealth extracted (WE) is the proper measure of an agent’s performance in a game. When the game is a repeated one, it may well be to an agent’s advantage to explore, taking different actions in essentially identical contexts. Our simple reinforcement learning agents do exactly this. They present perhaps the simplest case of an adaptive, exploring rationality. In utter ignorance of the game and their co-players, they merely seek to maximize their WE by collecting information on the consequences of their actions, and playing what appears to be best at any given moment. This is tempered by a tendency to explore by occasionally making what appear to be inferior moves.

Remarkably, when agents so constituted play each other and NEs are distinct from more rewarding Pareto outcomes (Prisoner's Dilemma, games #47, #48, and #57, Chicken with $\delta < 1$), Pareto wins. The drive to maximize WE succeeds. Similarly, a Pareto superior Nash Equilibrium will trump a Pareto inferior NE (Stag Hunt). Finally, in the presence of Pareto outcomes that are socially superior but unequally advantageous to the players, the players learn to extract an amount of wealth close to the maximum available (Battle of the Sexes, Chicken).

Outcomes that are neither Pareto efficient nor Nash Equilibria are rarely settled upon. Nash outcomes give way to Pareto superior outcomes when it pays to do so. A bit more carefully, in the case that there is one sub-game perfect NE, these results violate that as a prediction. In the case that the repeated games are seen as open-ended, there are (viz., the Folk Theorem) a very large number of NEs, but there is also insufficient theory to predict which will in fact occur. Again, our agents defy this as a prediction: they rather effectively maximize their Total WE. To sloganize, "It's not Nash that drives the results of repeated play, it's Pareto."

5 Discussion of Related Work

Reinforcement learning in games has become an active area of investigation. A systematic treatment of the literature would require a rather lengthy review paper of its own. Instead, we shall confine ourselves to brief discussions of certain especially apt works. We begin with several papers describing investigations into reinforcement learning in games by artificial agents.

Hu and Wellman (1998) essay a theoretical treatment of general-sum games under reinforcement learning. They prove that a simple Q-learning algorithm will converge for an agent to a Nash Equilibrium under certain conditions, including uniqueness of the equilibrium. When these conditions obtain the Nash equilibrium is, in effect, also Pareto optimal or dominant for the agent.

Claus and Boutilier (1998) investigate reinforcement learning (Q-learning) agents in coordination (aka: common interest) games. (Claus and Boutilier refer to these as cooperative games, which are not to be confused with cooperative game theory; the games played here are non-cooperative.) The paper studies factors that influence the convergence to Nash equilibrium under the setting of repeated play when using Q-learning. The empirical results show that whether the agent learns the action values jointly or individually may not be critical for convergence and that convergence may not be generally obtainable for more complicated games. The paper also proposes use of a myopic heuristic for exploration, which seems promising to help convergence to optimal (Pareto dominant) equilibrium. However, because the games tested in the paper are restricted to two particular coordination games, the results are somewhat limited in scope.

Bearden (2003) examines two Stag Hunt games, one with ‘high’ risk and one with ‘low’ using reinforcement learning and a genetic algorithm to discover parameter values for the agents’ learning schedules. His results are not easily comparable with ours, since his two games are effectively parameterized differently than our series of games (as delta changes). Broadly, however, our results are in agreement. Bearden’s ‘high’ risk game is closest to our game with delta = 2 or 2.5, while his ‘low’ risk game roughly corresponds to our case with delta = 0.75 or 1. In both studies, there is considerably more joint stag hunting (cooperation) in the ‘low’ risk case and considerably more joint hare hunting in the ‘high’ risk case.

Mukherjee and Sen (2003) explore play by reinforcement learning agents in four carefully-designed 3x3 games, in which the ‘greedy’ (i.e., Nash) outcome is Pareto inferior to the ‘desired’ (by the authors) outcome. Besides the different games, the experimental treatment involves comparison of two play revelation schemes (by one or both players) with straight reinforcement learning. It is found, roughly, that when the ‘desired’ outcome is also a Nash Equilibrium (NE) the revelation schemes are effective in promoting it. This kind of investigation, in which the effects of institutions upon play are explored, is, we think, very much in order, especially in conjunction with further investigation of learning regimes.

Reinforcement learning, in a related sense, has become popular in behavioral economics. A rather extensive series of results finds that reinforcement learning models, often combined with other information, perform well in describing human subject behavior in games. See Camerer (2003) for an extensive and up to date review. In part as a consequent of the experimental results, there has been theoretical interest by economists in reinforcement learning in games. Burgos (2002), for example, tries to use reinforcement learning models to explain subjects’ risk attitudes, which are one aspect of choice theory. The setting is pairwise choices between risky prospects with the same expected value. Two models are used for the simulation; one is from Roth and Erev (1995) and Erev and Roth (1998), the other from Börgers and Sarin (1997, 2000). The paper demonstrates a possible explanation of risk aversion as a side-effect of the learning regime. This raises the important question of whether risk aversion, risk seeking, and even individual utilities could be emergent phenomena, arising from simple underlying learning processes.

Finally, Bendor et al. (2001) study long run outcomes when two players repeatedly play an arbitrary finite action game using a simple reinforcement learning model. The model resembles that in Erev and Roth (1998). A distinguishing feature of the model is the adjustable aspiration level, which is used as an adaptive reference point to evaluate payoffs. Aspirations are adjusted across rounds (each round consists of a large number of plays). They define and characterize what they call Pure Steady States (either Pareto-efficient or Protected Nash equilibrium of the stage game), and the convergence to such states is established under certain conditions. The model limits itself to selection of particular action, thus does not allow mixed strategy or trigger strategy such as “Tit for Tat” in Prisoner’s Dilemma. In this simple, but general case, the authors prove that “*convergence to non-Nash outcomes is possible under reinforcement learning in repeated interaction settings*” (emphasis in original).

The results original to this paper are consistent with and complementary to the results reported in the above papers and other extant work. Further analytic and simulation results can only be welcomed. The experimental technique, however, has allowed us to discover hypotheses that merit continued investigation. In particular, our suggestion is that for agents playing games, and learning, wealth extraction (or some variant of it) is a key indicator for understanding system performance. Agents, we suggest, respond to rewards, but do so imperfectly and in a noisy context. If the reward signals are sufficiently clear, the agents will largely achieve Pareto optimal outcomes. If the signals are less clear, the outcomes obtained represent a balance between risk and reward. In either case, it is far from clear what causal contribution, if any, is made by the Nash Equilibrium. Resolution of these issues awaits much more extensive investigation.

6 References

1. Bearden, Neil J. (2003). “The Evolution of Inefficiency in a Simulated Stag Hunt,” http://www.unc.edu/~nbearden/Papers/staghunt.pdf, accessed October 2003.
2. Bendor, Jonathan, Dilip Mookherjee, and Debraj Ray (2001). “Reinforcement Learning in Repeated Interaction Games,” *Advances in Theoretical Economics*, vol. 1, no. 1, article 3.
3. Börgers, T. and R. Sarin (1997). “Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty,” Journal of Economic Theory, vol. 77, pp. 1—14.
4. Börgers, T. and R. Sarin (2000). “Naïve Reinforcement Learning with Endogenous Aspirations,” *International Economic Review*, vol. 41, pp. 921—950.
5. Alberta Burgos (2002). “Learning to Deal with Risk: What Does Reinforcement Learning Tell Us About Risk Attitudes?” *Economics Bulletin*, 4(10) pp. 1-13
6. Camerer, Colin F. (2003). Behavioral Game Theory: Experiments in Strategic Interaction, Princeton, NJ, Princeton University Press.
7. Claus, Caroline and Craig Boutilier (1998). “The Dynamics of Reinforcement

Learning in Cooperative Multiagent Systems," *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, Menlo Park, CA, AAAI Press/MIT Press, pages 746—752.

8. Dworman, Garett O., Steven O. Kimbrough, and James D. Laing (1995). "On Automated Discovery of Models Using Genetic Programming: Bargaining in a Three-Agent Coalitions Game." *Journal of Management Information Systems*, **12**: 97-125.
9. Dworman, Garett O., Steven O. Kimbrough, and James D. Laing (1996). "Bargaining by Artificial Agents in Two Coalition Games: A Study in Genetic Programming for Electronic Commerce" in *Genetic Programming 1996: Proceedings of the First Annual Genetic Programming Conference, July 28-31, 1996, Stanford University*, John R. Koza, David E. Goldberg, David B. Fogel, and Rick L. Riolo, eds., Cambridge, MA: MIT Press, 54-62.
10. Epstein, Joshua M., and Robert Axtell (1996). *Growing Artificial Societies: Social Science from the Bottom Up*. Cambridge, MA: MIT Press.
11. Erev, Ido and Alvin E. Roth (1998). "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," American Economic Review, vol. 88, no. 4, pp. 848—881.
12. Flood, Merrill (1952). "Some Experimental Games." RAND Corporation Research Memorandum RM-789, Santa Monica: CA.
13. Gintis, Herbert (2000). *Game Theory Evolving*. Princeton, NJ: Princeton University Press.
14. Grim, Patrick, Gary Mar, and Paul St. Denis (1998). *The Philosophical Computer: Exploratory Essays in Philosophical Computer Modeling*. Cambridge, MA: MIT Press.
15. Hu, Junling and Michael P. Wellman (1998). "Multiagent Reinforcement Learning: Theoretical Framework and an Algorithm," *Proceedings of the 15th International Conference on Machine Learning.*
16. Luce, R. Duncan and Howard Raiffa (1957). *Games and Decisions*. New York: Dover.
17. Mukherjee, Rajatish and Sandip Sen (2003). "Towards a Pareto-Optimal Solution in General-Sum Games," *Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems*, pp. 153—160.
18. Rapport, Anatol; Guyer, Melvin J. and Gordon, David G.. (1976). *The 2x2 Game*. Ann Arbor, MI: University of Michigan Press.
19. Roth, Alvin E. and Erev, Ido. (1995). "Learning in Extensive-Form games: Experimental Data and Simple Dynamic Models in the Intermediate Term." *Games and Economic Behavior* **8**:164-212.
20. Sanholm, Thomas W. and Crites, Robert H.(1995) "Multi-agent Reinforcement Learning in Iterated Prisoner's Dilemma", *Biosystems*, **37**:147-166.
21. Shubik, Martin (1982). *Game Theory in the Social Sciences*. Cambridge, MA: MIT

Press.
22. Sutton, R. and Barto, A. (1998) *Reinforcement learning: An Introduction*. MIT Press.
23. Watkins and Dayan, 1992 Watkins, C. J. C. H. and Dayan, P. (1992). “Q-Learning”, *Machine Learning*, **8**:279-292.
24. Watkins, C. (1989). *Learning from Delayed Rewards*. PhD thesis, King's College, Oxford.

This material is based upon work supported by, or in part by, NSF grant number SES-9709548. We wish to thank an anonymous referee for a number of very helpful suggestions.

File: simple-rl-agents-final.doc. Date: 2003/11/9.