Q Learning and Fees Relating to The Santa Fe Bar Problem

Julie Farago
Department of Computer Science
Harvard University
Cambridge, MA 02138
farago@fas.harvard.edu

Amy Greenwald **Keith Hall**
Department of Computer Science
Brown University, Box 1910
Providence, RI 02912
[amy|kh]@cs.brown.edu

1 Abstarct

In this paper we investigate learning in the Sante Fe bar problem(SFBP). The Sante Fe bar problem is paradoxical due to the incompatibility of rationality and predictability. We experiment with a Reinforcement Learning algorithm, Q-Learning, and present the results. Additionally, we experiment with a variation of the SFBP in which a fee mechanism imposes fairness. We find Q-Learning algorithms learn near-optimal solutions in the second game.

2 Introduction

The *Santa Fe Bar Problem* (SFBP) was introduced by Brian Arthur, an economist at the Santa Fe Institute. Arthur challenges the behavioral predictions of rational choice theory, since, in the SFBP, *rationality precludes learning*.

> N [(say, 100)] people decide independently each week whether to go to a bar that offers entertainment on a certain night . . . Space is limited, and the evening is enjoyable if things are not too crowded – especially, if fewer that 60 [or, some fixed but perhaps unknown capacity c] percent of the possible 100 are present . . . a person or agent goes if she expects fewer than 60 to show up and stays home if she expects more than 60 to go, Choices are unaffected by previous visits; there is no collusion or prior communication among the agents; the only information available is the number who came in the past weeks.[1]

Following Arthur, let us analyze SFBP under the standard assumption of *rationality: i.e.* utility-maximization with respect to one's beliefs. Define an *uncrowded* bar as one in which attendance is less than or equal to the capacity, c, and define a crowded bar as one in which attendance is strictly greater than c. Let the utility of going to an uncrowded bar be $+1$ and the utility of going to a crowded bar be -1; the utility of staying home is 0, regardless of the state of the bar. In this setup, rational choice theory dictates the following behavior: if an agent believes that the bar will be uncrowded with a probability p, then his rational choices are to go to the bar if $p > 1/2$ and to stay home if $p < 1/2$. (In the case where p=1/2, rational agents are indifferent between attending the bar and staying home and may have arbitrarily: *e.g.* randomize.) The probability p that the bar is uncrowded is determined by the agents (possibly randomized) strategies. If all the agents learn the correct value of p, then we stumble upon the fundamental contradiction. Suppose the agents learn that the bar will be uncrowded with probability $p < 1/2$; then in fact the bar will be empty with probability 1. On the other hand, suppose the agents learn that the bar will be uncrowded with probability $p > 1/2$; then the bar will be empty with probability 0.[1,2] Rational agents cannot learn in SFBP: *rationality precludes learning*.

Thus far, we have considered rationality from the view of individual agents. One might also be interested in the behavior of the collective. The symmetric Nash equilibrium in the SFBP, as described above, is inefficient. If N agents go to the bar with probability c/N, where c is the capacity of the bar, the bar will be uncrowded approximately half the time, yielding positive utilities for the agents. Yet, approximately half the time the bar is crowded, yielding negative utilities for the agents. The expected utility for the collective is zero. A simple solution is to devise learning algorithms by which agents learn (asymmetric) pure strategy Nash equilibrium: *i.e.*, c agents go to the bar and $N - c$ agents stay home, *forever*[2]. This solution may be efficient, but is unfair.

A *fair* market mechanism in the SFBP rectifies the inherent inefficiency of the symmetric equilibria. In the SFBP we can charge an entrance fee to those agents that attend the bar and redistribute this fee equally among those agents who remain at home[5]. This idea is related to the economic literature on user fees and public goods. Note that the fee mechanism does not change the Nash equilibria of SFBP.

The Santa Fe Bar Problem is a repeated game. In Arthur's original paper, he demonstrated via simulations that certain types of *boundedly rational* agents are capable of learning to center their collective attendance around the capacity of the bar[1]. Arthur's approach is based on complex modeling of cognitive aspects of inductive reasoning. In Greenwald [4], boundedly rational, no-regret learning agents are simulated and shown to converge to the neighborhood of the symmetric mixed strategy Nash equilibrium in which all agents attend

[1] Schelling [9] refers to phenomena of this kind as self-negating prophecies.

[2] Yogi Berra also observed this phenomenon. He said, "Nobody goes there anymore; it's too crowded."

the bar with probability $\sim c/N$.

In the remainder of this paper we present the results of Q-learning agents acting in the SFBP. We discuss different models of the SFBP and the results of the Q-learning agents in these models. Finally, we report the results of Q-learning agents in the fee-based SFBP.

3 On learning a fair outcome

We address the issue of learning efficient and fair outcomes in the SFBP. A *fair* outcome is characterized by all agents attending the bar an equal number of times. An *efficient* outcome implies each agent has maximized her cumulative utility.

The expected cumulative utility after J iterations of play is computed as (given the rewards are $+1$ for being accepted into the bar and -1 for being rejected from the bar):

$$\begin{aligned} E_J &= J\left[(1-\pi)p - \pi p\right] \\ &= Jp(1-2\pi) \end{aligned} \tag{1}$$

Where π is the probability of the bar being uncrowded. The π can be computed as a cumulative binomial distribution:

$$\pi = \sum_{i=c}^{N-1} \binom{N-1}{i} p^i (1-p)^{N-1-i} \tag{2}$$

Where N is the potential bar attendees, c is the capacity of the bar, and p is the probability each agent goes to the bar.

p	Cumulative Utility	p	Cumulative Utility
0	0.0	.54	41.9
.1	10.0	.56	33.0
.2	20.0	.58	18.8
.3	30.0	.6	0.5
.4	40.0	.7	-67.5
.46	45.8	.8	-80.0
.48	47.2	.9	-90.0
.5	47.8	1.0	-100.0
.52	46.5		

Table 1: The cumulative utility per game when the agents probability of going to the bar is set to p. When this value is set to c/N, or .6, the expected utility converges to about 0.

Table 1 shows the cumulative utility for different values of p computed by simulation. When each agent's probability of going to the bar, p, is set to .6, mean bar attendance becomes 60 with a variance of 24 (standard deviation is thus about $\pm$ 5). The agents attend the bar equally, but the average utility converges to around 0. This solution is therefore fair but inefficient.

3.1 Q-Learning

A Markov Decision Process(MDP) is a modeling tool that allows us to study non-deterministic state-space search problems. Recall Bellman's equations that characterize the optimal state and action-values for a single agent and an MDP: for $0 \leq \gamma < 1$,

$$Q^*(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) V^*(s') \tag{3}$$

$$V^*(s) = \max_{a \in A(s)} Q^*(s,a) \tag{4}$$

Where $r(s,a)$ is the reward for choosing action a from state s, $P(s'|s,a)$ is the Markov probabilistic transition function that describes state transitions, and γ is a future reward discount factor. $Q^*(s,a)$ is the expected reward for choosing action a at state s and $V^*(s)$ value associated with state s, given we are behaving optimally.

The value $Q^*(s,a)$ equals the immediate reward obtained at state s for taking action a plus the discounted expected value of the future rewards obtained by following the optimal policy thereafter. The value function $V^*(s)$ at state s is defined as the value that maximizes $Q^*(s,a)$ over all actions a. The actions that maximize $Q^*(s,a)$ at each state s describe the (deterministic) optimal policy.

$$Q'(s,a) = (1-\alpha)Q'(s,a) + \alpha\left[r(s,a) + \gamma V^*(s')\right] \tag{5}$$

Q-learning is a dynamic programming technique that uses Equation 5 as the update equation[10]. As the number of instances where action a is chosen at state s increases, Equation 5 converges to Equation 3 without explicitly knowing the probabilistic transition function $P(s'|s,a)$.

3.2 Bar models

We consider two Markov Decision Process (MDP) models of the SFBP. This first model captures the SFBP as stated, where each player is either at the bar or at home. The second model introduces a *wait* state where the player has expressed intent to enter the bar, but is not yet let in due to the capacity constraints.

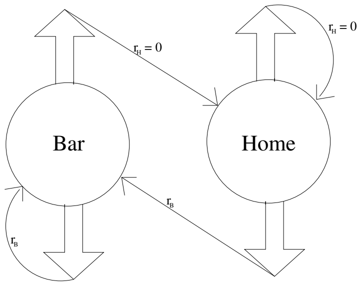

Figure 1: A graphical representation of the Q state/action transitions

In the first model, shown in Figure 1, the SFBP is represented as a two-state MDP. At any time t the agent is either at home or at the bar. From either state, the agent may either try to enter the bar or stay home for the evening.

When thinking about the best possible outcome (a fair game in which the overall utility of the game is maximized) it is natural to imagine a certain amount of coordination between agents. In order for the agents to coordinate, a third action was added, in this case the *wait* action.

A model with the *wait* state is shown in Figure 2. The *wait* state is like the line outside the bar. At time t an agent that wishes to enter the bar will to the *wait* state. No reward is given prior to entering the *wait* state. At time $t+1$ the agent

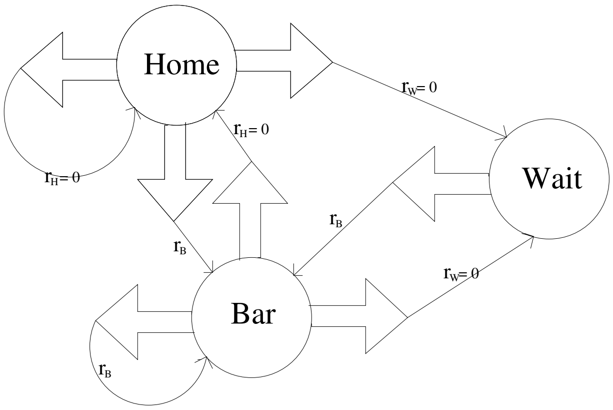

Figure 2: The Three State Q Learner setup.

in the *wait* state will automatically choose to enter the bar. The reward for entering the bar is dependent on the current capacity of the bar.

In both models, the reward for attempting to enter the bar is based on the bar status. If the bar is crowded, then the reward is -1. Otherwise, if the bar is uncrowded, the reward is $+1$.

3.3 Experimental results

We experimented with homogeneous Q-learning environment in both SFBP models. The Q-learning algorithm requires two externally controlled parameters, learning-rate α and exploration rate ϵ. The learning-rate controls how quickly an agent estimates the a Q-values. Similarly, the exploration rate determines how often a random action will be selected instead of the action prescribed by the maximal expected reward (i.e. the action with the maximal

In these experiments we decayed α according to the following schedule[3].

$$\alpha_n = \alpha_0 \left(\frac{n_0 + 1}{n_0 + n} \right)^3 \tag{6}$$

Additionally, we decay the ϵ according to:

$$\epsilon_n = \epsilon_0 \left(\frac{n_0 + 1}{n_0 + n} \right)^4 \tag{7}$$

Our experiments succeed in generating fair solutions, but only accomplished this goal at the cost of the overall utility. The set of Q-learners converges to the mixed Nash equilibrium (as seen in in Figures 3 and 3.3), and, as would be expected, the average utility approaches 0.

When analyzing the Q-learners and their actions, one must notice that the results being displayed are *not converging*. Because Alpha decays so slowly, the result tends to stay closer to the Nash Equilibrium. This happens because there agents place a lot of weight on their most recent results without paying much head to the past. Thus, their actions are always

[3]In our experiments α_0 was set to 1, n_0 was set to n (the number of rounds)/10.

[4]In our experiments ϵ_0 was set to 1, n_0 was set to n (the number of rounds)/100.

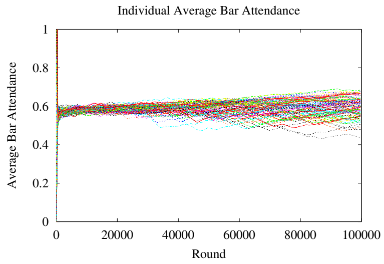

Figure 3: A fair solution is reached by the Q-learner, as the cumulative bar attendance for all of the agents is virtually the same.

htb

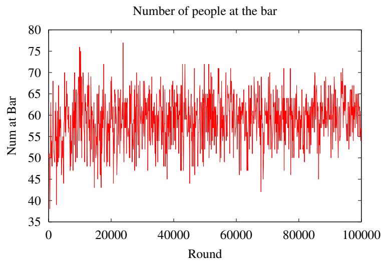

Figure 4: The actual bar attendance at different round, which oscillates around 60.

changing. When looking at the Q matrices and their values one can see this fluctuation more clearly. Different actions take precedence at different rounds, without converging to a set action plan. See Figures 5 and 6.

Currently, our experimental results indicate that adding the *wait* state did not improve the average utility of the game. Unfortunately, the wait state does not induce turn-taking behavior in the outcome.

4 On designing an efficient mechanism

We now consider changing the mechanism of the SFBP to induce fairness. In order to add an incentive to stay home, while not changing the basic structure of the problem, we added a *fee*, χ, for attending the bar. χ varies vary over the range of [0,ρ].[5] Thus, the reward for attending the bar is changed to:

$$\rho' = \begin{cases} \rho - \chi & \text{if } n \leq \\ -\rho - \chi & \text{if } n > c \end{cases} \quad n \Rightarrow \text{number of people at the bar} \tag{8}$$

Those agents who remain home redistribute the sum collected from the fee evenly amongst themselves. Therefore their new reward (for staying home) is :

$$\rho = \frac{n\chi}{N - n} \text{ where } N \text{ is the total number of agents}$$

χ both decreases the incentive to go to the bar all of the time and increases the incentive to stay home. The goal of the fee is to encourage an efficient result (unlike the strictly mixed Nash equilibrium).

[5]Recall that ρ is the reward for attending an uncrowded bar

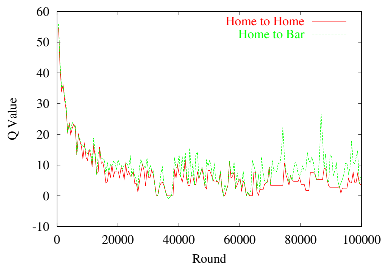

Figure 5: The changing Q values in the home state for the bar and home actions. The action to stay home from the home state is very rarely chosen, as it is rarely the top line (higher value).

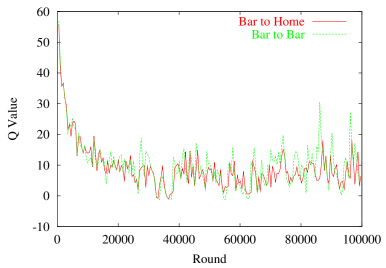

Figure 6: The Q values for the bar state, which clearly switch off between the go home and go to the bar actions.

4.1 Equilibrium Analysis

In order to see how a fee actually effected the predicted equilibrium, we had to solve equation 9 for the different possible values of χ. Equation 9 sets the total possible rewards received for going to the bar (in all the different combinations of players) equal to the total possible rewards for staying home. While the reward function for going to the bar changes depending on the capacity and load, the reward function for staying home remains constant (that is why there are two summations on the left and only one on the right). This equation reflects the reward functions explained about, and are summed over N, as there are N possible situations.

$$(\rho - \chi) \sum_{i=0}^{c} \binom{N}{i} p^i (1-p)^{N-i} + (-\rho - \chi) * \sum_{i=c+1}^{N} \binom{N}{i} p^i (1-p)^{N-i} = \chi \sum_{i=0}^{N} \binom{N}{i} p^i (1-p)^{N-i} \frac{i}{N-i} \tag{9}$$

Since we are dealing with the standard SFBP, let N be 100 and c be 60. When we set ρ to 1 and vary χ over [0:1] (inclusive) the equilibrium decreases slowly from $\chi = 0$ to $\chi = .4$, and then rapidly declines until $\chi = 1$. For a graphical view of this see Figure 7

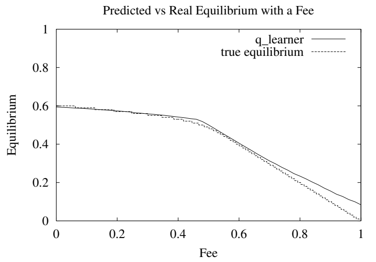

Figure 7: The predicted equilibrium verses the fee.

When the 2 state Q-learning agents were trained on varying fees, their equilibrium closely follows the predicted equilibrium. Again, see Figure 7

Because there are many different possible cases for the general Santa Fe Bar Problem (i.e. maybe 563 agents want to go to the bar, but are unhappy when there are over 467 people there), we could not simply determine a single value as the optimal fee. We use derivative-following algorithm that will converge to the optimal feel.

4.2 An Iterative Mechanism

In Arthur's original formulation of SFBP, it is assumed that all agents have identical preferences, and moreover that these preferences are known. Now suppose that the value to each agent of going to the bar is unknown to a central planner, say the mayor, whose goal is to design a mechanism, the equilibria of which maximize the total utility of the townspeople. Assume each agent i's utility u_i is an endogenous function of the other agents' actions: let u_i be a concave function of the attendance at the bar, say λ, and let μ_i be the point at which agent i's utility peaks: *i.e.*, $u_i(\lambda) = \max\{1 - (\lambda - \mu_i)^2/\sigma_i^2, 0\}$ for $\sigma_i \geq 0$. In this setting, total utility is maximized at the median value of the μ_i's, say μ_i^*, with those agents whose valuations are closest to μ_i^* attending the bar. Thus, if 25 agents' utilities peak at 25, 50 agents' utilities peak at 50, and 25 agents' utilities peak at 75, with $\sigma_i = 50$ for all agents i, total utility is maximized if the 50 agents for which $\mu_i = 50$ go to the bar. Can the mayor elicit this efficient outcome?

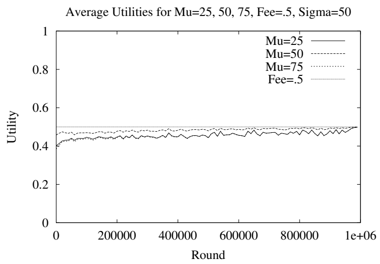

Figure 8: A heterogeneous population of Q-learners learn the equilibrium of the mechanism: *i.e.*, the social choice.

Let us begin our analysis by giving the mayor access to an oracle that informs him of all the agents' peak valuations. In this case, he can compute the median, which is 50 in our example. But can he induce precisely those 50 agents whose peak utilities occur at 50 to attend the bar so as to maximize total utility? The following taxation scheme is designed to achieve this outcome: charge agents that attend the bar an entrance fee, say x, and distribute the proceeds evenly among those agents that do not attend the bar. Suppose λ agents go to the bar. The payoffs to those λ agents are $u_i(\lambda) - x$, if the bar is uncrowded, and $-(u_i(\lambda) - x)$, if the bar is crowded. The payoffs to those $N - \lambda$ agents that do not go to the bar are $\lambda x/(N - \lambda)$. An entrance fee $x = 0.5$ yields an equilibrium in which the 50 agents whose peak valuations are 50 attend the bar, and the remaining 50 agents stay home. All agents achieve utility 0.5 in equilibrium. Allowing Q-learning agents [10] to adapt their behavior to this mechanism, average utility per agent approaches this equilibrium value. See Figure 8.

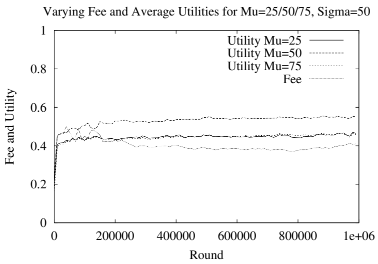

Figure 9: A heterogeneous population of Q-learners learn the equilibrium of the mechanism while a derivative follower simultaneously implements an efficient social choice.

But in fact the mayor does not have access to an oracle that informs him of all the agents' peak valuations. Thus, he cannot compute the median, and he cannot compute the equilibrating fee. His behavior, too, is constrained to be boundedly rational. As both *planner* and *player*, assume the mayor sets a fee according to his beliefs about the population of Santa Fe, and updates the fee weekly according to the trend in utility. One simple algorithm for adjusting the fee is *derivative-following* [6], which experiments with incremental changes, continuing to move in the same direction until average utility decreases, at which point the direction of movement is reversed: given increment $\beta > 0$, the fee $f_{t+1} = f_t + \beta[\text{sign}(f_t - f_{t-1})\text{sign}(u_t - u_{t-1})]$, where u_t is the agents' average utility at time t. Figure 9 depicts simulations not only of boundedly rational agents adapting their behavior to the mechanism, but the planner adjusts the mechanism in response to the agents' collective choices as well. Both the fee and average utilities per agent converge near the equilibrium value of 0.5.

Now suppose the agents' valuation functions are such that $\mu_i = \sigma_i = 60$ for all i. In this setting, it is appropriate to define a social choice not only as an efficient outcome (*i.e.*, utility-maximizing), but among all efficient outcomes, a social choice should be *fair*. Fairness requires that agents with identical preferences be equally likely to attend the bar. Note that the mixed strategy Nash equilibrium satisfies fairness,

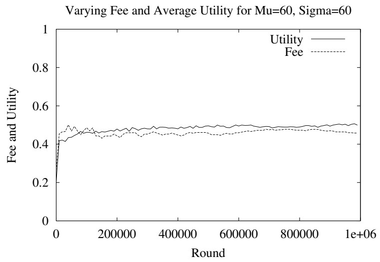

Figure 10: A homogeneous population of Q-learners learn the equilibrium of the mechanism while a derivative follower simultaneously implements a fair and efficient social choice.

but it is inefficient—it yields an average utility near zero—since half the time the bar is overcrowded. Boundedly Rational Q-learning agents together with a planner that abides by the derivative-following algorithm generate fair outcomes in this scenario that approximate efficiency. See Figure 10. This taxation scheme is an example of an iterative mechanism for which a set of boundedly rational agents together with a boundedly rational planner approach equilibrium behavior that implements an approximate social choice.

5 Conclusions

We have explored two approaches to finding *efficient* and *fair* outcomes to the Sante Fe Bar problem. Our results show that Q-Learning is capable of learning an *efficient* outcomes, but is not *fair*. Modifications of the Markov Decision Process model proved to be ineffective in the Q-Learning setting. Additionally, we have explored adding a *fair* market mechanism, in this case a fee. The empirical results show that the fee does induce a *fair* outcome.

References

[1] W.B. Arthur. Inductive reasoning and bounded rationality. *Complexity in Economic Theory*, 84(2):406–411, 1994.

[2] A. M. Bell, W. A. Sethares, and J. A. Bucklew. Coordination failure as a source of congestion in information networks. May 1999.

[3] Jutsin Boyan. *Learning Evaluation Functions For Global Optimizations*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, August, 1998.

[4] A. Greenwald. *Learning to Play Network Games*. PhD thesis, Courant Institute of Mathematical Sciences, New York University, New York, May 1999.

[5] A. Greenwald, Bud Mishra, and Rhoit Parikh. On the impossibility of rational learning in the sante fe bar problem. Working Paper, June 2001.

[6] A.R. Greenwald and J.O. Kephart. Shopbots and pricebots. In *Proceedings of Sixteenth International Joint Conference on Artificial Intelligence*, volume 1, pages 506–511, August 1999.

[7] R.B. Myerson. *Game Theory: Analysis of Conflict*. Harvard University Press, Cambridge, 1991.

[8] M. Osborne and A. Rubinstein. *A Course in Game Theory*. MIT Press, Cambridge, 1994.

[9] T. Schelling. *Micromotives and Macrobehavior*. W.W. Norton and Company, New York, 1978.

[10] C. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge University, Cambridge, 1989.