# Pseudo-convergent Q-Learning by Competitive Pricebots

**Jeffrey O. Kephart** KEPHART@US.IBM.COM
**Gerald J. Tesauro** TESAURO@WATSON.IBM.COM
IBM Thomas J. Watson Research Center, 30 Saw Mill River Rd., Hawthorne, NY 10532 USA

## Abstract

We study novel aspects of multi-agent Q-learning in a model market in which two identical, competing "pricebots" strategically price a commodity. Two fundamentally different solutions are observed: an exact, stationary solution with zero Bellman error consisting of symmetric policies, and a non-stationary, broken-symmetry pseudo-solution, with small but non-zero Bellman error. This "pseudo-convergent" asymmetric solution has no analog in ordinary Q-learning. We calculate analytically the form of both solutions, and map out numerically the conditions under which each occurs. We suggest that this observed behavior will also be found more generally in other studies of multi-agent Q-learning, and discuss implications and directions for future research.

## 1. Introduction

Within the next few years, we expect electronic commerce to be an important multi-agent domain in which reinforcement learning will find numerous applications. One such application is automated dynamic pricing by software agents (Greenwald & Kephart, 1999). Suppose that each seller agent individually attempts to maximize profits through judicious setting of prices and other product parameters. Even if the seller agents do not communicate with one another directly, market forces may strongly couple their actions, resulting in a highly dynamic multi-agent system. Since decision making in markets and economies benefits greatly from one's ability to forecast economic trends and opponents' strategies, reinforcement learning is likely to be an essential component of decision making by economically-motivated software agents.

Unfortunately, from a theoretical perspective, the issue of what happens when multiple interacting agents simultaneously adapt, using RL or other approaches, is largely an open question. This stands in contrast to the case of single-agent RL: in stationary Markov Decision Problems, a solid theoretical understanding has been provided by research on algorithms such as Dynamic Programming and Q-learning. Various theorems establish that global convergence to a unique optimal value function and optimal policy will always be obtained. However, these theorems do not apply in the multi-agent case, as adapting agents provide effectively non-stationary environments for other agents.

Some progress has been made in analyzing certain special-case multi-agent problems. For example, cooperative teams of agents sharing a common goal or utility function have been studied in (Stone & Veloso, 1999), among others. The purely competitive case of zero-sum utility functions has been studied in (Littman, 1994), where an algorithm called "minimax-Q" was proposed for two-player zero-sum games, and shown to converge to the optimal value function and policies for both players. Simultaneous Q-learning by two players in the Iterated Prisoner's Dilemma game was studied empirically in (Sandholm & Crites, 1995), who found that the learning procedure frequently converged to stationary solutions. An important first step in analyzing Q-learning for arbitrary-sum two-player games was recently taken in (Hu & Wellman, 1998). This algorithm assumes that the players follow Nash equilibrium policies. Issues remaining to be addressed include the "equilibrium coordination" problem (i.e. how the agents choose from amongst multiple Nash equilibria) and verification that the policies implied by the learned Q-functions are consistent with the initially assumed Nash policies.

In our previous work (Tesauro & Kephart, 1999), we studied simultaneous Q-learning by two price-setting agents in three simple models, showing that in all cases Q-learning by one or both of the agents raised both of their profits substantially from what they would obtain using myopic best-response pricing. Simultaneous convergence of both sellers to stationary policies was found in some but not all cases, depending

on the model's payoff functions and on the discount parameter $\gamma$. In other cases, we observed "pseudo-convergence," i.e. near-convergence to a slightly unstable solution with non-zero Bellman error—an intriguing phenomenon that has no analog in ordinary single-agent Q-learning. Moreover, in the shopbot model, the pseudo-convergent policies of the two agents were *asymmetric* even though the two agents were identical in every respect, including their payoffs.

In this paper, we delve into the nature and behavior of simultaneous Q-learning by two interacting agents. As a vehicle for our study, we focus on the "shopbot" model originally introduced in (Greenwald & Kephart, 1999). In any given training run, we had observed (Tesauro & Kephart, 1999) that either of two solutions could be obtained: symmetric and stable, or asymmetric and pseudo-convergent. We also found that the asymmetric solution was increasingly favored as the discount parameter $\gamma$ was increased. Having observed pseudo-convergence under a wide range of experimental conditions—even in entirely different agent models—we conjecture that it is somehow endemic to the problem of pricing in multi-agent domains. By studying these phenomena in the context of the shopbot model, we hope to develop more broadly useful analytic and experimental tools and insights that will bring us to a broader understanding of the nature of multi-agent Q-learning, particularly in competitive, non-zero-sum settings like markets and economies.

The rest of this paper is organized as follows. Section 2 provides some background on the shopbot model and our previous work on simultaneous Q-learning in that context. Section 3 presents an analysis of both the symmetric and asymmetric solutions, and the dynamic behavior of Bellman error vs. time in the latter case. Section 4 analyzes the conditions under which each solution is chosen, and the final section gives concluding remarks and suggested directions for future work.

## 2. Background

A simple shopbot model was first proposed in (Greenwald & Kephart, 1999) to study price dynamics that may occur when a large fraction of consumers have ready access to price information. It is postulated that increased price awareness on the part of buyers will induce sellers to be more responsive in their pricing, and that this will lead to increased reliance on automated price-setting agents, or "pricebots". It is therefore of interest to develop automated pricing algorithms that yield high profits for sellers, and to explore the price dynamics that ensue when various pricing algorithms (Greenwald et al., 1999) are employed.

Briefly, the model supposes that a commodity is offered for sale by $S$ sellers. Buyers generate purchase orders at random times: they identify the lowest-priced seller from among a subset of from 1 to $S$ prices, and they purchase from that seller if the price is less than the buyer's valuation. The buyer population can therefore be described in terms of the distribution of valuations, and the distribution of search strategies, denoted $\vec{w}$, where $w_i$ represents the fraction of buyers that compare the prices of $i$ randomly selected sellers. At random intervals, individual sellers will decide unilaterally to revise their prices as they wish.

If the sellers are perfectly informed about the distributions of buyer valuations and search strategies, and if they know their competitors' current prices, they can employ a myopic best-response strategy. This entails choosing the price that will maximize the seller's expected profit, assuming no further changes to the price vector. (Of course, this assumption is violated when other sellers change their prices.) If all sellers use this "myoptimal" strategy, the market experiences cyclical price wars. The price war cycle begins with each seller charging $p_m$, the price that a profit-maximizing monopolist would charge. Each seller then competes for market share by lowering its price, and the prices fall linearly until a critical threshold $p^*$ is reached. At this point, the prices immediately jump back up to $p_m$, and the cycle begins anew. The upward jump in price occurs when ownership of a large market share fails to compensate for the very low profit margin. Instead, it is better for an agent to give up on market share and simply cater to the much smaller segment of buyers that only check a single price.

Despite its obvious flaws, the myoptimal algorithm tends to perform well in head-to-head competition against other simple pricing algorithms, including one based on a one-shot game-theoretic analysis.[1] However, there is clearly room for improvement. If a seller could foresee that undercutting would invite quick retaliation by its competitors, it might be less keen to undercut. If all sellers were to reason in this manner, one would expect higher average prices and profits. Previously, we have verified that a Q-learning approach *can* raise average profits in a two-seller market, not just for the shopbot model, but also for two other models of agent-mediated markets in which myoptimal pricing leads to price war cycles (Tesauro & Kephart, 1999).

To understand how this improvement comes about,

[1] Application of one-shot game-theoretic analysis is equivalent to assuming that the sellers have no access to (or ignore) one another's prices; this is why other algorithms that *do* use such information can yield higher profits.

it is helpful to first consider myoptimal pricing. For simplicity, we assume that there are just $S = 2$ sellers ($A$ and $B$), each with production cost $c = 0$, and that all consumers have the same valuation $v = 100$. Then, provided that $p_B$ does not exceed $v$, $B$'s expected profit per sale systemwide is simply:

$$\pi_B(p_A, p_B) = \begin{cases} (1 - \frac{w_1}{2})p_B & \text{if } p_B < p_A \\ \frac{1}{2}p_B & \text{if } p_B = p_A \\ \frac{w_1}{2}p_B & \text{if } p_B > p_A \end{cases} \quad (1)$$

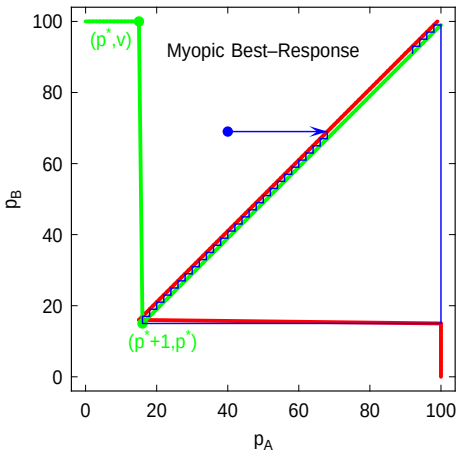


*Figure 1.* Cross plot of myopic response functions $R_B(p_A)$ and $R_A(p_B)$. Thin zigzag curve represents price trajectory starting from the initial condition (0.40,0.69).

The myoptimal pricing policy of seller $B$ is obtained by choosing the price $p_B$ that maximizes $\pi_B(p_A, p_B)$. This can be represented as a response function $R_B(p_A)$:

$$R_B(p_A) = \arg\max_{p_B} \pi_B(p_A, p_B) \quad (2)$$

The myoptimal response function $R_B(p_A)$ is depicted in Fig. 1. The analogous and symmetric myoptimal reponse function $R_A(p_B)$ is rotated and superimposed. Starting from any given initial price vector, one can successively apply the response functions $R_B$ and $R_A$. Graphically, this is represented by making alternate vertical and horizontal movements to the $R_B$ and $R_A$ curves. As illustrated in Fig. 1, this leads to the cyclical price war discussed above. Here and throughout the remainder of the paper, we shall assume that prices are quantized to integers.

How can sellers introduce foresight into their pricing decisions? One method, introduced in (Tesauro & Kephart, 1999), optimizes cumulative future-discounted profit instead of immediate profit. One of the sellers, say $B$, regards any price that it may set plus $A$'s anticipated response to it as a single timestep, and adds the expected profits from its move and $A$'s countermove. Projecting further into the future, $B$ computes its expected future-discounted profit as an infinite weighted sum of expected profits in future timesteps. More compactly, we define

$$\begin{aligned} Q_B(p_A, p_B) &= \pi_B(p_A, p_B) + \pi_B(R_A(p_B), p_B) \quad (3) \\ &\quad + \gamma Q_B(R_A(p_B), R_B(p_A)) \end{aligned}$$

where $\gamma$ is a discount parameter, ranging from 0 to 1, and $R_B$ now represents the policy that optimizes $Q_B$:

$$R_B(p_A) = \arg\max_{p_B} Q_B(p_A, p_B) \quad (4)$$

It is well established that, if $R_A$ represents *any* fixed policy (myoptimal or otherwise), then a reasonable updating procedure will yield a unique, stable future-discounted profit landscape $Q_B(p_A, p_B)$ and associated response function $R_B(p_A)$. However, if $A$ similarly computes *its* $Q_A(p_B, p_A)$ and associated response $R_A(p_B)$, then a unique fixed point is not guaranteed.

In previous work (Tesauro & Kephart, 1999) we found empirically that there are two basic solutions: a symmetric solution in which $R_A(p_B)$ and $R_B(p_A)$ have the same functional form, and an approximate broken-symmetry solution in which $R_A$ and $R_B$ are quite different from one another and slightly unstable. These were discovered by a standard Q-learning updating procedure in which, starting from initial functions $Q_A$ and $Q_B$, a random seller and a random price vector were chosen, the right-hand side of Eq. 3 was evaluated for that seller and price vector using Eq. 4, and the Q-value for that seller and price vector was moved toward this computed value by a fraction $\alpha$ that diminished gradually with time.

Regardless of the details of the buyer valuation and search-strategy distributions, and regardless of the details of the initial conditions or updating procedure of the Q-learning procedure, we have found to our surprise that only these two solutions are ever observed, and that they always have exactly the same qualitative form. This has motivated a deeper analysis of their form and nature, which we now present.

## 3. Symmetric and asymmetric solutions

### 3.1 Symmetric solution

For all values of $\gamma$ in the range $0 < \gamma < 1$, the symmetric best-response policy $R(p)$ is observed to have a functional form that depends on just two parameters $\eta$ and $\theta$:

$$R_{A,B}(p) = \begin{cases} v & \text{if } p \le \theta \\ \theta & \text{if } \theta < p < \eta \\ p - 1 & \text{if } \eta \le p \end{cases} \quad (5)$$

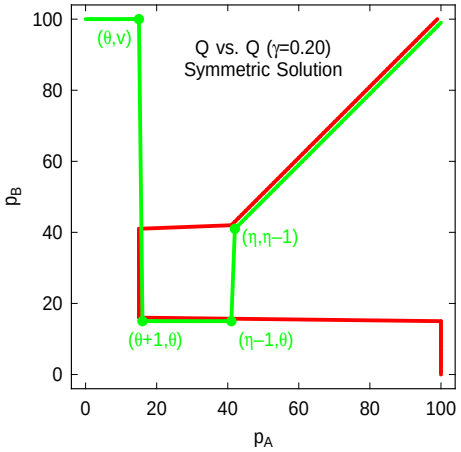


*Figure 2.* Cross plot of symmetric response function solutions for Q-learning with $\gamma = 0.2$.

Figure 2 illustrates the symmetric $R(p)$ for the case $w_1 = 0.25$ and $\gamma = 0.2$. This differs from the myoptimal policy of Fig. 1 in that undercutting does not continue all the way down to $p^*$. Instead, when the price gets down to $\eta$, the agent aggressively drops its price all the way down to a value $\theta$. The opponent's best response to $\theta$ is to set the price back up to $v$. While this aggressive price lowering decreases the agent's immediate profit, it proves advantageous in the long run for at least two reasons. First, both agents avoid the lower portion of the price-war cycle, and so the average price over the course of a cycle is increased. Second, when the competitor responds with price $v$, the agent can then undercut at the price $v-1$, yielding a relatively high profit.

To calculate the parameters $\eta$ and $\theta$, consider first that $\theta$ is the price below which no agent will price because the advantage of increased market share is outweighed by the very small profit margin. The price $\theta$ can be determined by noting that the discounted reward must be just marginally higher than that of choosing price $v$. Taking price quantization into account, $\theta$ must be the smallest integer such that $Q(\theta+1,\theta) > Q(\theta+1,v)$.

For low to moderate values of $\gamma$, very accurate approximations to both $Q$ values can be computed. First, consider $Q_B(\theta+1,\theta)$. As the undercutter, $B$'s expected profit according to Eq. 1 is $(1-w_1/2)\theta$. Next, Agent $A$ will respond with $R_A(\theta) = v$. Since $B$ is still the undercutter, its profit will be another $(1-w_1/2)\theta$. $B$ will then respond by undercuting $A$ with price $v-1$. Therefore, using Eq. 3,

$$Q_B(\theta+1,\theta) = (2-w_1)\theta + \gamma Q_B(v, v-1) \quad (6)$$

Now consider $Q_B(\theta+1,v)$. Since $B$ is undercut by $A$, its expected reward is $(w_1/2)v$. $A$ then responds with $v-1$; this too undercuts $B$, and thus $B$ again receives $(w_1/2)v$. At its next opportunity, $B$ undercuts Agent $A$ with price $v-2$. Thus the Q-value in this case is

$$Q_B(\theta+1,v) = (w_1)v + \gamma Q_B(v-1, v-2) \quad (7)$$

To compute $Q_B(v, v-1)$ and $Q_B(v-1, v-2)$, note that these price vectors are at or near the beginning of the price war cycle. Until the price drops down to $\eta-1$, $B$ will alternately be the undercutter and the undercuttee. Thus, when $B$ sets its price to $p$, the expected profits from its move and $A$'s countermove will simply be $(1-w_1/2)p + (w_1/2)p = p$. At $B$'s next turn, it will set its price to $p-2$, and so on. Therefore,

$$\begin{aligned} Q_B(p+1,p) &= \sum_{i=0}^{i=\lfloor (p-\eta)/2 \rfloor} (p-2i)\gamma^i \quad (8) \\ &\approx \left[ \frac{p}{1-\gamma} - \frac{2\gamma}{(1-\gamma)^2} \right] \end{aligned}$$

where the approximation comes about because the finite arithmetico-geometric series is being approximated by an infinite one; this is valid to the extent that $\gamma^{(p-\eta)/2} << 1$ — i.e. it assumes that $B$ places negligible weight on events after the end of the price-war cycle.

Noting that $\theta$ must be the integer just greater than the value obtained by equating Eqs. 6 and 7, and using Eq. 8, we obtain

$$\theta \approx \left\lceil \frac{w_1 v - \frac{\gamma}{1-\gamma}}{2-w_1} \right\rceil \quad (9)$$

where the approximation is quite accurate provided that the following condition is satisfied:

$$\gamma^{(v-\eta)/2} << 1. \quad (10)$$

Similarly, $\eta$ must be the smallest price for which $Q_A(\eta,\eta-1) > Q_A(\eta,\theta)$, i.e. the point at which the future discounted reward of slightly undercutting $\eta$ is just barely higher than that of agressive undercutting to $\theta$. In the first scenario, the price sequence leading into the beginning of the price-war cycle is $(\eta,\eta-1) \rightarrow (\theta,\eta-1) \rightarrow (\theta,v) \rightarrow (v-1,v) \rightarrow (v-1,v-2)$, i.e. $B$'s immediate profit will be higher but it will then be undercut twice in a row by $A$. In the second scenario, the price sequence leading into the price-war cycle will be $(\eta,\theta) \rightarrow (v,\theta) \rightarrow (v,v-1)$, i.e. $B$'s immediate profit will be lower but it will undercut $A$ twice in a row. The Q functions for these two scenarios can be computed:

$$Q_B(\eta,\eta-1) = \eta - 1 + \gamma w_1 v + \gamma^2 Q_B(v-1,v-2) \quad (11)$$

and

$$Q_B(\eta,\theta) = Q_B(\theta+1,\theta) = (2-w_1)\theta + \gamma Q_B(v, v-1) \tag{12}$$

and equated to yield

$$\eta \approx \left\lceil (2-w_1)\theta + 1 - \gamma w_1 v + \gamma\left(v - 1 - \frac{\gamma}{1-\gamma}\right)\right\rceil \tag{13}$$

which is accurate provided that Eq. 10 holds. A simpler but less accurate expression can be derived by substituting Eq. 9, ignoring the integer restrictions, and neglecting terms of $O(1)$:

$$\eta \approx [(1-\gamma)w_1 + \gamma]\, v \tag{14}$$

Figure 3 plots the values of $\eta$ and $\theta$ as a function of $\gamma$ for $w_1 = 0.25$. The solid circles represent measurements taken by running the Q algorithm to convergence, while the curves represent the theoretical approximations obtained from Eqs. 13 and 9 up to the point where Eq. 10 becomes seriously violated.

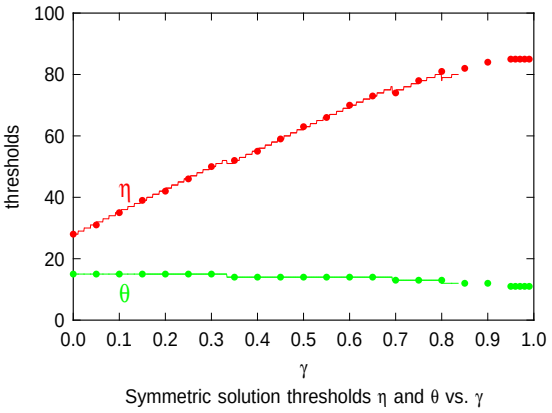


*Figure 3.* Symmetric solution: theoretical and observed $\eta$ and $\theta$ as a function of $\gamma$. Wiggles in theoretical curve are due to integer ceiling functions in Eq. 9 and 13.

### 3.2 Asymmetric solution

The asymmetric solution is observed to always have the form illustrated in Figure 4. It can be described using one parameter ($\zeta$) for one of the agents (say Agent $A$) and three parameters ($\phi$, $\chi$, and $\psi$) for the other agent (say Agent $B$). The functional forms of the response curves are described by:

$$R_A(p) = \begin{cases} v & \text{if } p < \zeta \\ p-1 & \text{if } \zeta \le p \end{cases} \tag{15}$$

$$R_B(p) = \begin{cases} \chi & \text{if } p < \psi \\ p-1 & \text{if } \psi \le p \le \chi \\ \chi & \text{if } \chi < p < \phi \\ p-1 & \text{if } \phi \le p \end{cases} \tag{16}$$

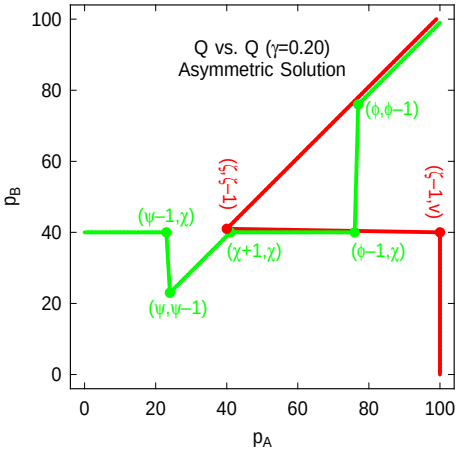


*Figure 4.* Cross plot of asymmetric response function solutions at $\gamma = 0.2$.

(Not all of these parameters are independent: $B$'s price $\chi$ is just below $A$'s price-war threshold $\zeta$, i.e. $\chi = \zeta - 1$.) We can derive approximations to the values of $\phi$, $\chi$, $\psi$ and $\zeta$ that are accurate for sufficiently small $\gamma$. First, consider the determination of $\zeta$, the lowest price that $A$ is willing to undercut. At this value, $A$ is just on the verge of preferring to opt out of a price war and set its price up to $v$. Therefore, temporarily disregarding the fact that $\zeta$ must be an integer, we seek $\zeta$ such that $Q_A(\zeta,\zeta-1) = Q_A(\zeta, v)$. These two Q-values can be computed by following the price trajectories up to the point where they join the price-war trajectory:

$$\begin{aligned} Q_A(\zeta,\zeta-1) &= \zeta - 1 + \gamma w_1 v + \gamma^2 Q_A(v-1, v-1) \\ Q_A(\zeta, v) &= w_1 v + \gamma Q_A(v-1, v-2) \end{aligned} \tag{17}$$

Equating the right-hand sides of these equations and rounding up to the nearest integer, we obtain

$$\zeta \approx \left\lceil (1-\gamma)w_1 v + \gamma\left(v - 2 - \frac{2\gamma}{1-\gamma}\right) + 1\right\rceil \tag{18}$$

The parameter $\phi$ is the value of $p_A$ at which Agent $B$ decides to set its price aggressively low. This is (approximately) the point at which $Q_B(\phi,\phi-1) = Q_B(\phi,\chi)$. Following the price trajectories up to the point where they join the standard price-war trajectory, we find:

$$\begin{aligned} Q_B(\phi,\phi-1) &= \phi - 1 + \gamma(2-w_1)\chi + \gamma^2 Q_B(v-1, v-1) \\ Q_B(\phi,\chi) &= (2-w_1)\chi + \gamma Q_B(v, v-1) \end{aligned} \tag{19}$$

from which we obtain

$$\phi = \left\lceil \chi(2-w_1)(1-\gamma) + 1 + \gamma\left(v - 1 - \frac{2\gamma}{1-\gamma}\right)\right\rceil \tag{20}$$

Similarly, we can compute $\psi$ by setting $Q_B(\psi, \psi-1) = Q_B(\psi, \chi)$:

$$\begin{aligned} Q_B(\psi, \psi-1) &= (2-w_1)\psi + \gamma Q_B(v, v-1) \\ Q_B(\psi, \chi) &= \chi + \gamma Q_B(v, v-1) \end{aligned} \quad (21)$$

from which we obtain

$$\psi = \left\lceil \frac{\chi}{2-w_1} \right\rceil \quad (22)$$

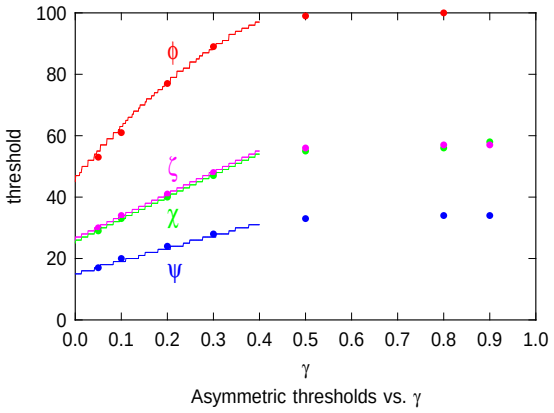


*Figure 5.* Asymmetric solution: theoretical and observed $\phi$, $\chi$, $\psi$, and $\zeta$ as a function of $\gamma$.

Figure 5 plots the values of $\phi$, $\chi$, $\psi$, and $\zeta$ as a function of $\gamma$ for $w_1 = 0.25$. The solid circles represent measurements taken by running the Q-learning algorithm until the Bellman error is minimized, while the solid curves represent the theoretical approximations given by Eqs. 18, 20, and 22, which are valid provided that $\gamma^{(v-\phi)/2} << 1$, along with the relation $\chi = \zeta - 1$.

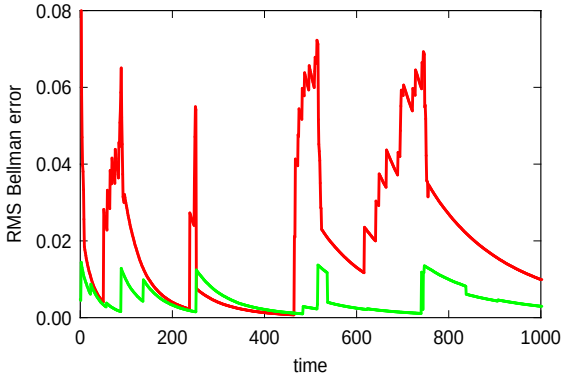


*Figure 6.* Bellman error for simultaneous Q-learning by agents 1 and 2, with $\gamma = 0.1$. Each time unit represents a number of random updates equal to the total number of price pairs.

Interestingly, this solution just barely fails to be fully self-consistent. A clear symptom of inconsistency can be seen in Figure 6, which plots the Bellman error (the discrepancy between the lefthand and righthand sides of Eq. 3) as a function of training time for sellers $A$ and $B$. The Bellman error, defined as the average RMS error weighted equally over all price pairs, comes extremely close to zero, but suddenly shoots up dramatically. The error soon decreases, again dropping nearly to zero but shooting up again, and so the cycle continues unceasingly. For example, at time 464, the policies have the canonical pseudo-solution form, and the Bellman error is just 0.0007 for $A$ and 0.0012 for $B$. However, at time 465, the response curve for $A$ suddenly shifts from $R_A(p_B = 20) = 100$ to $R_A(p_B = 20) = 19$ as one ridge in $Q_A$ at $p_A = 19$ just rises above a ridge at $p_A = 100$. This is manifested as a long finger extending across the crossplot illustrated in Figure 7.

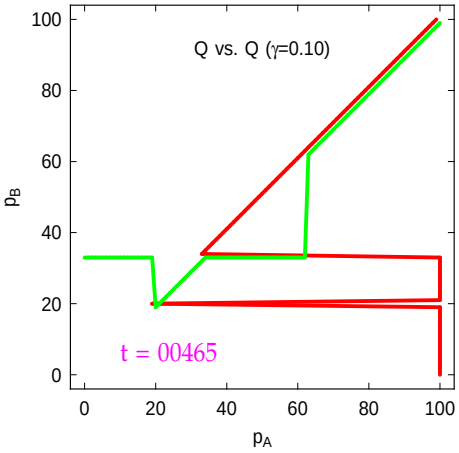


*Figure 7.* Policy crossplot at time $t = 465$ during the Q-learning run of Fig. 6.

The location of the finger suggests that the problem lies in a part of $R_A(p_B)$ that is only relevant during transients before the price-war cycle has begun — somewhere around $\psi < p_B < \chi$. In fact, detailed analysis reveals that, in this region, $Q_A(p_B, \psi - 1)$ very slightly exceeds $Q_A(p_B, v)$ if all other $Q$ values are taken as described in Eqs. 15 and 16. As the Q function gradually becomes more self-consistent and accurate, it finally reaches the point where, for some value of $p_B$ in the critical range, $A$'s best response shifts from $v$ to $\psi - 1$. Analogous fingers may develop for other $p_B$ in this range as well. $B$ soon discovers that, simply by shifting its threshold from $\psi$ to $\psi - 1$, it can undercut $A$. Interestingly, analysis and observation demonstrate that $A$ cannot retaliate by extending its finger a little further to the left; instead it retreats back to playing $v$. After a while, $B$ shifts its threshold back up to $\psi$, and the *policy cycle* is ready to begin anew. The dramatic and cyclical shift in the policies translates into large cyclical spikes in the Bellman er-

ror for both players. The irregularity in amplitude and frequency is due to the randomness of the Q algorithm, and the gradual lengthening of the period is due to the cooling of the $\alpha$ parameter.

## 4. Which solution will occur?

Here we explore the conditions that determine whether the symmetric or asymmetric solution is obtained. First, we start by adding gaussian noise, with mean 0 and amplitude $\sigma$, to the Q functions of the symmetric solution. We then allow Q-learning to run, and observe whether the symmetric or asymmetric solution is obtained. An example illustrating the initial perturbed policies at $\sigma = 2.0$ and $\gamma = 0.2$ is shown in Fig. 8; this particular initial condition happened to evolve to the symmetric solution.

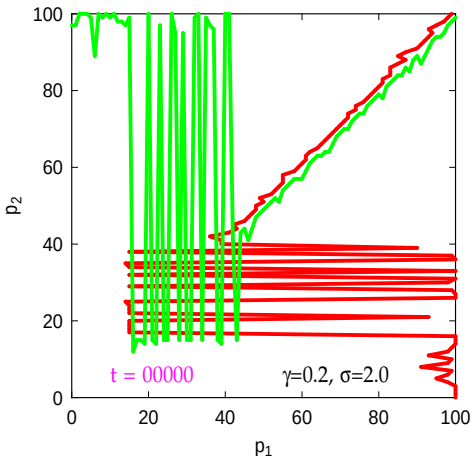


*Figure 8.* Example initial perturbed policies obtained by starting from the symmetric solution at $\gamma = 0.2$ and adding noise with amplitude $\sigma = 2.0$. Final response functions are symmetric; most other trials at this noise amplitude reached the asymmetric solution.

As the noise amplitude is increased, the initial policies tend to be further from the symmetric solution, and the Q functions tend to evolve more often towards the asymmetric solution. For 16 different choices of noise amplitude ranging from 0.01 to 50.00, 100 trials were conducted. At $\sigma = 0.01$, the noise is so slight that the policies are usually unchanged. At $\sigma = 50.0$, the initial response functions are essentially random.

The percentage of trials that yielded the symmetric solution as a function of the noise amplitude for $\gamma = 0.2$ and $\gamma = 0.5$ is given by Fig. 9(a). The same data can be viewed in a different way, by plotting the probability of obtaining the symmetric solution as a function of the total Manhattan distance of the two noisy initial response curves from the ideal noise-free symmetric solution. This is shown in Fig. 9(b). (The probability is obtained by averaging over 100 trials centered around each distance.)

In addition to the randomness of the starting state, the random exploration dynamics of Q-learning also influences the resulting final state. We have performed experiments starting many trials from a specific random starting state, and found that some trials converged to the symmetric solution while other trials went to the asymmetric solution.

Fig. 9 supports the conceptual interpretation of two-player Q-learning dynamics in terms of a basin of attraction around the symmetric solution, delineated by a distance parameter in either policy space or in Q-function space. In both spaces, there is a small region around the symmetric solution such that virtually any starting state within the region will invariably converge to the symmetric solution. For low $\gamma$, the symmetric solution can be reached even when the starting state is well outside this region—roughly a 30% chance when $\gamma = 0.2$. For moderate to high $\gamma$, the symmetric solution cannot be reached if the starting state lies beyond this region. This explains why we only observed the asymmetric solution for moderate to high $\gamma$ in our earlier studies (Tesauro & Kephart, 1999).

## 5. Conclusions

In this paper and in our earlier work (Tesauro & Kephart, 1999), we have observed interesting and unexpected phenomena when using simultaneous Q-learning to attempt to obtain optimal pricing policies for competing pricebots. These findings include:

(1) In the shopbot model studied here, and in the two additional models studied in (Tesauro & Kephart, 1999), there do exist exact optimal Q-learning solutions, in which the two players' policies are simultaneously optimal vs. each other. Furthermore, the Q-learning procedure is able to find these solutions. In general, for arbitrary payoff functions, there is no reason *a priori* to expect that there will always exist a pair of response curves such that $R_A$ is optimal vs. $R_B$ and vice versa. The fact that such solutions exist in three different economically motivated models suggests that the structure of the payoff functions of naturally occuring real-world problems may be sufficiently "nice" to lead to such exact solutions, even though they cannot be expected for arbitrary problems.

(2) In addition to the exact solutions, we also find in the shopbot model, and in the Information Filtering model studied in (Tesauro & Kephart, 1999), a non-stationary "pseudo-convergent" solution with

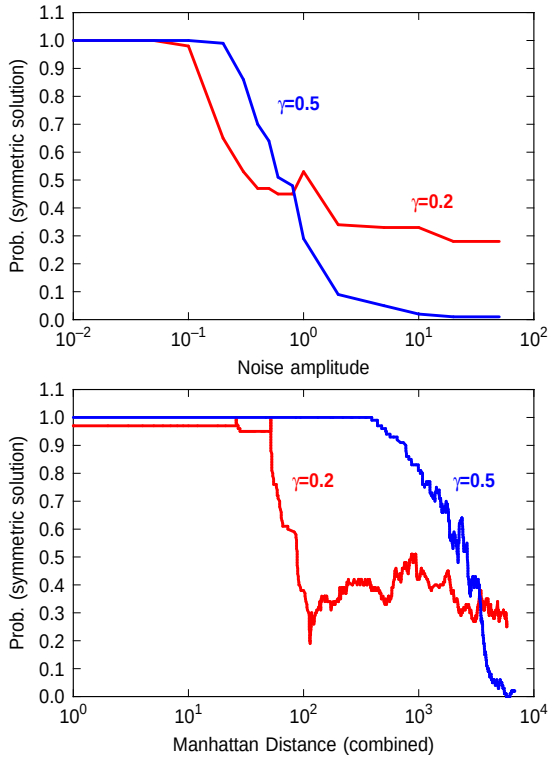


*Figure 9.* Probability of obtaining the symmetric solution, starting from randomly perturbed initial Q-functions, at $\gamma = 0.2$ and $\gamma = 0.5$. a) As a function of noise amplitude $\sigma$. b) As a function of total Manhattan distance of initial policies from the ideal symmetric solution.

small non-zero Bellman error. This is a novel phenomenon that does not occur in ordinary single-agent Q-learning. Here we found that the non-stationarity is inherently cyclic in nature; this had previously been masked by the random exploration of the Q-learning algorithm. The pseudo-solution in the shopbot model is of further interest because it is a broken symmetry solution, whereas the payoff functions for the two players are symmetric. In theoretical physics, broken symmetry solutions to symmetric equations often indicate interesting underlying physics. Broken symmetry solutions may be of similar interest in understanding the dynamics of multi-agent Q-learning.

(3) We find that the exact solution and the pseudo-solution can co-exist, and that the choice of which solution is obtained by Q-learning depends on initial conditions, the discount parameter, and (to a lesser extent) on the randomness of exploration.

This paper has presented analytic and numerical techniques for calculating the forms of the solutions, the dynamics of the learning procedure (how the policies and Bellman errors behave with time), and understanding the conditions under which each solution is chosen. A large-$\gamma$ approximation for the various thresholds describing the policies would be desirable, as would a more detailed characterization of the basins of attraction. More broadly, we would like to develop a deeper and more general understanding of pseudo-convergent solutions and the range of scenarios and conditions under which they occur. From our experience to date, we suspect that they will be found generally in automated price-setting applications, where instabilities generated by undercutting may be quite endemic. But only after conducting a broader analytical and experimental study of multi-agent Q-learning in various applications (most likely aided by techniques from nonlinear dynamical systems theory) will we know whether pseudo-convergence is a phenomenon confined narrowly to non-cooperative non-zero-sum games between two players, or a fundamental characteristic of multi-agent Q-learning.

## References


Greenwald, A., & Kephart, J. O. (1999). Shopbots and pricebots. *Proceedings of Sixteenth International Joint Conference on Artificial Intelligence*

Greenwald, A. R., Kephart, J. O., & Tesauro, G. J. (1999). Strategic pricebot dynamics. *Proceedings of the First ACM Conference on Electronic Commerce* ACM Press.

Hu, J., & Wellman, M. P. (1998). Multiagent reinforcement learning: theoretical framework and an algorithm. *Proceedings of ICML-98*

Littman, M. (1994). Markov games as a framework for multi-agent reinforcement learning. *Proceedings of the Eleventh International Conference on Machine Learning* (pp. 157–163). Morgan Kaufmann.

Sandholm, T. W., & Crites, R. H. (1995). On multiagent Q-learning in a semi-competitive domain. *14th International Joint Conference on Artificial Intelligence (IJCAI-95),, Workshop on Adaptation and Learning in Multiagent Systems, Montreal, Canada* (pp. 71–77).

Stone, P., & Veloso, M. (1999). Team partitioned, opaque transition reinforcement learning. *Proceedings of the Second International Conference on Autonomous Agents*

Tesauro, G., & Kephart, J. (1999). Pricing in agent economies using multi-agent q-learning. *Proceedings of Workshop on Game-Theoretic and Decision-Theoretic Agents*