**Q-learning for the pricing problem: iterative mixed strategy game approach**

**Methodology**

Here we first consider only two prices, and . Instead of using the average price to calculate the payoff as in the previous note, we calculate the expected payoff using the probability of choosing the high/low price, just like in a mixed strategy game. Let be the probability of choosing the high price, ; is thus the probability of choosing the low price . The probability is determined by the Q-matrix using the Boltzmann experimentation model,

I. States and Actions

Parameters:

n: number of products/number of firms (each product is supplied by different firms)

: product quality indexes, is outside good

: index of horizontal differentiation (perfect substitutes if )

: marginal cost

k: the length of memory

T: used in Boltzmann weight

: the discount factor

1. Action space: as mixed strategy is used, there is no specific action.
2. State space: The prices chosen of all the players in the previous k steps. The size of the state space, ; for different price memory patterns and the action, we have a Q matrix element. In this note we consider only k=0 for simplicity (so there is no state, the Q matrix is reduced to Q-vector).
3. The demand for the product i, given all the prices
4. The average reward of firm i is

Here if and if

We can also calculate the conditional average reward of the firm i (conditional on the price of firm i sets, )

So is the average reward when firm i set price ; is the average reward when firm i set price . The average now is over the price choices of all other firms.

By definition,

1. At t = 0,

II. Q-matrix update

Parameter:

: the learning rate

Algorithm:

Initialize

Initialize s randomly

Repeat:

Calculate for each firm and update the Q-matrix

For two prices used, we have

If convergence, break

Check convergence

A very loose criterion --- If the total payoff does not change much in, say 100 iterations, we can stop simulating.

III. Average Profit Gain

: the average per-firm profit at time t

: the profit under full collusion

: the profit in the Bertrand-Nash static equilibrium

In this study, we would like to check the time series of . Is the evolution of smooth or with intermittent volatility?

**Comments**

1. If we choose ; the payoff matrix will satisfy the condition for pure prisoners’ dilemma game. The payoffs are

T(chooses , the other chooses

R(both choose )=0.333

P(both choose )=0.234

S(chooses , the other chooses )=0.106

So T>R>P>S and R+P > T+S

1. Generalization: this framework can be easily generalized the cases of more than 2 prices.
2. We can also do the pure strategy simulation by sampling the price using . The Q-matrix is updated using

,

The one-step mixed-strategy version can be thought of the time average of the pure strategy version with a fixed q-matrix. The simulation of the mixed-strategy version should be equivalent to the pure strategy version with slowly varying q-matrix.