WORKING PAPER 177

Learning Algorithms for Price Control in an Internet-Based Dutch Auction

By

K. Ravi Kumar &
Diatha Krishna Sundar

October 2001

Please address all correspondence to:

Dr. K. Ravi Kumar
IBM India Research Laboratory
Block – I, IIT Delhi
Hauzkhas, New Delhi 110 016

Professor Diatha Krishna Sundar
Associate Professor
Indian Institute of Management Bangalore
Bannerghatta Road
Bangalore 560 076
Phone : 080 - 6993276
Fax : 080 – 6584050
E-mail : diatha@iimb.ernet.in

Working Paper

On

Learning Algorithms for Price Control in an Internet-Based Dutch Auction

K. Ravikumar
IBM India Research Laboratory
Block I, IIT Delhi
Hauzkhas, New Delhi 110016
India.

Diatha Krishna Sundar
Associate Professor
Indian Institute of Management Bangalore
Bangalore 560076
India

Abstract

In this paper, we consider a multi-unit Dutch auction over the Internet where auctioneer gradually decrements per unit price of the item during the course of the auction. We investigate an optimal price control problem of the auctioneer, that is, the problem of finding a decrementing price sequence that maximizes his total expected revenue in the presence of uncertainty with regard to arrival pattern of bidders and their individual price-demand curves.

We start with an analysis of an *analogous* multi-unit *pay-your-bid* auction in a discrete setting and provide a characterization of mixed strategy equilibrium. Since it is difficult to arrive at a pure strategy equilibrium, we deviate from game theoretic consideration and model the above decision problem in a Dutch auction as a single-agent Reinforcement Learning in an uncertain non-stationary auction environment wherein the auctioneer (or his agent) uses its experience interacting with the environment to improve his (its) pricing strategies.

Over the Internet, auctioneer always has an option of concealing information pertaining to dynamics of the ongoing Dutch auction from bidders. In this situation, it can be assumed that each bidder values the items independently of other bidders. For this case we develop a finite horizon Markov Decision Process (MDP) model with undiscounted returns and propose a Q-learning algorithm for generating a decrementing price sequence for optimal revenue.

In a more general setting where bidders can observe the ongoing auction, the state-space representation will lead to the domain appearing non-Markov to the reinforcement learning agent. In a particular case where history provides the needed missing state information, we investigate the applicability of direct reinforcement learning and contrast the temporal difference based RL with actual return based RL. We show that direct application of temporal difference based reinforcement learning algorithms will in general fail to learn even deterministic optimal policies in a Dutch auction environment. Furthermore, our analysis suggests that actual return based reinforcement learning algorithms be used instead.

1. Introduction

A Dutch auction is a form of a multi-unit single item first price auction widely used for sale of perishable commodities such as flowers and also for disposal of inventory of items with low salvage value. This auction form is particularly recommended when time factor plays a crucial role because of declining intrinsic value of the item or because of high maintenance (holding) cost of the item. The auction mechanism is detailed below.

Multiple units of an item are placed on auction for sale and the number of units, the duration of auction are announced before the auction begins. A Dutch auction begins with the auctioneer announcing a very high price per unit. This price is then lowered gradually until a buyer calls a halt. The first buyer to do so acquires the desired number of objects at the price outstanding when he or she intervened and the auction continues with the remaining number of items. There are two variations on how the initial price for the remaining units is reset once a set of units is acquired by a buyer. In a decreasing type of Dutch auction, the starting price for the remaining units is the price at which the previous buyer gets the items and the auction restarts following the same rules. In the other form of the auction, this price is reset to some price higher than that of the last sale. In this paper, we analyze the decreasing type of Dutch auction in a discrete framework and derive price control strategies for the auctioneer that will maximize his expected revenue. But the analysis presented here can be extended to the second form of the auction.

The above auction is analogous to a multi-unit first price sealed-bid auction but for some differences in strategy space of bidders. Contrary to the sealed-bid scenario, the bidders in this auction will get to know about bids of winners who bought items. In the sealed-bid case, multi-unit first-price auctions have been analyzed by Menezes and Monteiro [1995] and it is shown that game theoretic equilibrium in continuous strategy space exists only in a very few special cases. We give characterization for mixed strategy equilibrium in a discrete domain. Since it is very hard to derive conditions under which pure strategy equilibrium exists in such a discrete model, later in this paper, we depart from the game theoretic domain and look at the problem from a macroscopic perspective, *i.e.*, where bidders' demand is aggregated into a price-demand variational function over time, in a discrete setting. Later, we analyze the applicability of Reinforcement Learning based algorithms in the absence of a proper model of bidders' strategies. Recently, Reinforcement Learning (RL) has received increasing attention in the e-commerce domain. Kephart and Tesauro [1999, 2000], for instance, focus on dynamic pricing issues in oligopoly markets using RL framework in the underlying stochastic games. Convergence of RL in general sum games is an active area of research. (See Hu and Wellman [1998])

In contrast with the above studies, our development deviates from game theoretic perspective and rather views auctioneer's decision problem as a *single -agent learning* in an uncertain environment. We state a few reasons for adhering to such an approach.

In an Internet –enabled Dutch auction, auctioneer, while designing the auction, may choose not to reveal all the information to bidders who access his auction page. For example, the auctioneer may hide information pertaining to quantity remaining from a bidder upon his arrival. In such cases, the auctioneer's price control problem can be appropriately modeled as a stochastic control problem in non-stationary Markovain domains. The non-stationarity stems from two considerations. Firstly, if the items are of perishable nature, then intrinsic value of the items decrease with time and hence one would find variations in bidding patterns. Besides this, as mentioned in Kagel [1995] bidders will derive positive utility from suspense playing odds at the gambling table. This leads to non-stationary evolution as auction progresses.

At the other extreme, the bidder or his agent may be allowed to observe complete dynamics during the course of the auction which may lead to a strategic game among participants. In such cases, bidders update their estimates of other bidders' valuations per unit of item as offering price goes down. Even in these cases, the dynamics can still be modeled as a control problem for the auctioneer in a Non-Markovian domain. In the absence of a proper model which can predict the outcomes, our aggregated view will allow us to exploit rich theory underlying learning in Markov Decision Processes (or in general in sequential decision processes), namely the reinforcement learning, to learn in a disguised way the price-demand variational function over time while concurrently attempting to act optimally in such domains. While in Markovian domain, model-free algorithms such as Q-learning (Watkins and Dayan [1992]) provably converge to optimal policies, in non-Markovian domains no such satisfactory theory appears in literature. However, there have been quite a few attempts in this direction. See Singh et al [1994] and Jaakkola et al

[1995] for studies on application of RL methods in Partially Observable Markov Decision Process (POMDP) domains.

Contributions of our paper are as follows: We analyze an approximately strategically equivalent multi-unit pay-your-bid auction in a discrete framework where strategic price-quantity pairs are assumed to come from a finite grid. In this setup, we provide characterization for the structure of mixed strategy equilibrium. However, it is very difficult to find conditions under which equilibrium in pure strategies exist. At this point we deviate from game theoretic d^main and discuss price control problem in Dutch auctions using single agent learning. Here again we discuss the decision problem in two different scenarios. In the first scenario, a bidder upon his arrival can have the knowledge of only the ongoing price. The bidder may choose to bid at the price or may want to come back to the system after a random time interval in which case he will be treated as a fresh arrival into the system. In the second case, a bidder is allowed to have knowledge about prices at which items were previously bought in the ongoing auction and also about the quantity remaining in the inventory. In the former case, the dynamics can be modeled using Markov Decision Process framework since a bidder's decision is dependent only on the observed state of the auction. However in the latter case, the bidder can effectively use the information avaliable, such as rate of quantity movement and also winners' quotes, to act strategically. In such situations, the control problem falls in non-Markovian domain. In both the cases, it is fairly difficult to build a model for auction dynamics. That is, it is difficult to arrive at transition probability structure that will capture the dynamic behavior of the system. In such situations, the Reinforcement Learning (RL) framework proposed in this paper, offers a very simplistic adaptive control framework that learns an optimal strategy from its interactions with auction environment instead of building suitable models. The auctioneer can devise an RL agent which can autonomously act on his behalf and take price control decisions in an adaptive fashion. Since RL uses a simple updating procedure, intelligence to the agent can be incorporated with minimal effort.

To this end, we consider bidders as part of the environment and the auctioneer by way of executing price decrementing action learns from responses (or immediate reinforcement sequences) from the environment. In the Markovian case, we propose a finite horizon Markov decision process model in finite state and action spaces and present modifications necessary to model the problem as a learning problem over infinite horizon. Later, we devise a Q-learning algorithm for the auctioneer, which can be proven to converge to an optimal strategy using standard stochastic approximation techniques.

In the Non-Markovian case, intuition suggests that any Temporal Difference based learning algorithm such as Q-learning should be able to learn a best deterministic observation based policy. We design some realistic and illustrative Dutch auction examples to show that such a hypothesis fails. Also, we further articulate that actual-return based learning methods (eg. Monte-Carlo based) may offer superior performance compared to TD based learning methods in these domains.

Later, we discuss simulation experiment for Dutch auctions that models bidders and their strategies. In some simple test cases where optimal policies can be directly obtained, our experiments show that the proposed learning algorithms converge to optimal policies within a reasonable number of iterations.

This paper is organized as follows. In the next section, we develop the Markovian Model and provide details of our Q-learning based algorithm for the same. Section 3 discusses the non-Markovian setting and analyses performance of TD methods using a few illustrative realistic examples. In Section 4, we give a brief sketch of our simulation design. Our sample simulation experiments show that Q-learning algorithms perform well even in cases where Markovian assumption is marginally violated. In a completely non-Markivan setting, trace- based (actual-return) mechanisms give higher expected rewards compared to single-step Q-learning. Since our experimentation is still in progress, we will discuss the early qualitative results in this paper. In Section 5, we provide some pointers to future research.

2. Multi-Unit Pay-Your-Bid Auction

Consider auctioning of items in inventory of an auctioneer using a Pay-your-bid auction in sealed bid format. This auction is generally used to dispose off items with low salvage value. No strategic equivalence is known to exist between the pay-your-bid auction considered here and a multiunit Dutch auction. However, we the analysis presented below can be used as first level approximation to Dutch auctions.

A multiunit pay-your-bid auction is a discriminatory price auction for multiple identical units of an item where price-quantity orders of bidders are arranged in the decreasing order of price and the auctioneer accepts quantities up to the amount he is selling. In this auction format, each bidder pays the amount equal to his bid. In this document, we consider a discrete version of the above auction in which bids are chosen from a finite grid of price-quantity pairs.

Assume that a maximum quantity of N units of an item are being sold in a discriminatory price auction. To make our development simple, we assume that there are only two bidders participating in the auction. Each bidder has an associated demand function $D_i : [0, \hat{p}] \rightarrow \mathcal{R}_+$, $i = 1,2$ which is assumed to be continuous and strictly decreasing. Further assume that $\sum_i D_i(0) > N$ and $\sum_i D_i(\hat{p}) = 0$ (**Assumption A1**). From the assumptions on the demand functions it follows that there exists a unique market -clearing price p^* which satisfies $\sum_i D_i(p^*) = N$. Bidders are allowed to bid a price from $[0, \hat{p}]$ and a quantity less than N.

The strategy of bidder i is represented by a price-quantity pair (p_i, q_i) where these pairs are selected from the grid $\mathcal{G}$ defined as follows:

$\mathcal{G} =: \mathcal{P} \times \mathcal{Q} = \{m\alpha : m = 1,2,\ldots,\hat{m}\} \times \{n\beta : n = 1,2,\ldots,\hat{n}\}$ where $\hat{m}$ and $\hat{n}$ are such that $\alpha = \frac{\hat{p}}{\hat{m}}$ and $\beta = \frac{N}{\hat{n}}$ for pre-determined unit grid values, α and β of price and quantity respectively. Here $\mathcal{P}$ is the one- dimensional price grid and $\mathcal{Q}$ is the one-dimensional quantity grid. This discretization provides a realistic framework since in multi-unit auctions of commodities auctioneer will accept quantity and price bids in multiples of small units. In the following we make an assumption that α and β are such that $D_1(2\alpha) + D_2(2\alpha) > N$. This condition says that (discretized)(true) demand functions of the bidders are such that price as low as 2α rather than zero price used in Assumption

A1 is enough to sell off the units in the inventory. It is always possible to choose an α that approximates this behaviour.

ln the continuous counterpart of this auction, the items are allocated to the two bidders as follows. Say (p_1,q_1) and $(p_2,q_2)$are the bids placed by bidder 1 and bidder 2 respectively. If $p_1 > p_2$, then bidder 1 gets his quantity and the bidder 2 gets the remaining quantity, $N-q_1$ or q_2 whichever is smaller. If $p_2 > p_1$, then the allocation is analogous with bidders' roles reversed. If $p_1 = p_2$, then both the bidders will get their respective quantities whenever $q_1+q_2 \leq N$. If $q_1+q_2 > N$, each bidder is rationed where rationing is proportional to their respective quantity bids. Formally, the allocations $Z_1(p_1,q_1,p_2,q_2)$ and $Z_2(p_1,q_1,p_2,q_2)$ for bidder 1 and bidder 2 respectively are as given by:

$$\begin{aligned} Z_1(p_1,q_1,p_2,q_2) &= q_1 \wedge N, && \text{if } p_1 > p_2 \qquad (1)\\ &= \frac{q_1}{q_1+q_2}N, && \text{if } p_1 = p_2\\ &= q_1 \wedge \{N-q_2\}, && \text{if } p_1 < p_2 \end{aligned}$$

and

$$\begin{aligned} Z_2(p_1,q_1,p_2,q_2) &= q_2 \wedge N, && \text{if } p_2 > p_1 \qquad (2)\\ &= \frac{q_2}{q_1+q_2}N, && \text{if } p_2 = p_1\\ &= q_2 \wedge \{N-q_1\}, && \text{if } p_2 < p_1 \end{aligned}$$

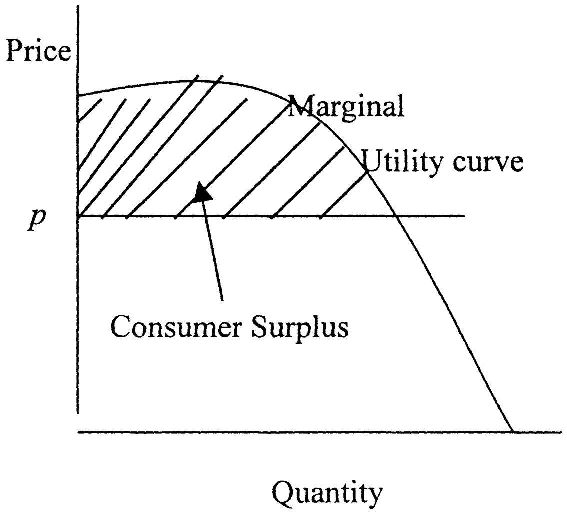
The pay-offs for the bidders (consumer's surplus or consumer's rent) are given respectively by:

$$\pi_1(p_1,q_1,p_2,q_2) = \int_o^{z_1} (D_1^{-1}(s) - p_1)ds \qquad (3)$$

and

$$\pi_2(p_1,q_1,p_2,q_2) = \int_o^{z_2} (D_2^{-1}(s) - p_2)ds. \qquad (4)$$

One way to interpret any price-demand function is as follows: the price p corresponding to any quantity q on a bidder's demand function is the price a bidder is willing to pay for an extra unit of the item given that he has currently q units of the item for his consumption. Thus, any bidder's demand function can also be treated as his marginal utility function. The bidder's surplus at any price p is the area of his marginal utility curve above the price p as shown in the figure below. Hence (3) and (4) follow.



2.1 Allocations in Discrete Model

Now observe that in the continuous price and quantity strategy space, if both the bidders bid the same price, the rationed allocation is proportional to the quantity he bids. But in a discrete setting, it may happen that this quantity so apportioned may assume fractional values and hence may not fall on the grid. In such cases, the actual allocation is rounded off to the maximal point on the quantity grid which lies below the actual value. In other words, the new allocations are according to the following rules:

Let

$$z_1(p,q_1,p,q_2) := \max\{q \in \mathcal{Q} : q \le \frac{q_1}{q_1+q_2} N\}$$

and analogously define $z_2(p,q_1,p,q_2)$. Now the allocations are redefined as follows:

$$\begin{aligned} \widetilde{Z}_1(p_1,q_1,p_2,q_2) &= q_1 \wedge N \ , && \text{if} \quad p_1 > p_2 \\ &= q_1 \wedge z_1(p,q_1,p,q_2), && \text{if } p_1 = p_2 \\ &= q_1 \wedge \{N - q_2) \ , && \text{if} \quad p_1 < p_2 \end{aligned}$$

Define $\widetilde{Z}_2(p_1,q_1,p_2,q_2)$ analogously. The pay-offs are found replacing the Z_1 and Z_2 in (3) and (4) by $\widetilde{Z}_1$ and $\widetilde{Z}_2$ respectively.

Since any bidder is allowed to choose his strategies from the grid $\mathcal{G}$, his true (continuous) demand function needs to be suitably transformed to take care of the discretization effect. If at any price p, the quantity demanded against a bidder's original continuous demand function does not fall on the nodes of the grid $\mathcal{G}$, but cuts the quantity grid at an intermediate point between two successive quantity grid nodes, then this quantity needs to be adjusted to one of these nodes. This is done according to consumer's surplus against each such adjustment and the node which gives maximum surplus is selected for consideration. That is, the modified demand function looks as follows:

If $D_i(p) < \beta$ at any price p, then set $D_i(p) = \beta$. If $D_i(p) > N$, set $D_i(p) = N$. For any p such that $\beta \le D_i(p) \le N$ define $\breve{D}_i(p)$ and $\hat{D}_i(p)$ as follows:

$\breve{D}_i(p) = \max\{q \in \mathcal{Q} : q \le D_i(p)\}$ and $\hat{D}_i(p) = \min\{q \in \mathcal{Q} : q \ge D\ (p)\}$

Then the modified $D_i(p)$ is as follows:

$$D_i(p) = \breve{D}_i(p) \text{ if } \int_0^{\breve{D}_i(p)} (D_i^{-1}(s) - p)ds > \int_0^{\hat{D}_i(p)} (D_i^{-1}(s) - p)ds$$

$$= \hat{D}_i(p) \text{ if } \int_0^{\hat{D}_i(p)} (D_i^{-1}(s) - p)ds > \int_0^{\breve{D}_i(p)} (D_i^{-1}(s) - p)ds$$

With the above discrete representation, we analyze the auction game in the ensuing sections.

2.2 Auction Game in Mixed Strategies

We consider an extension of the above game in which bidders may decide to play randomized strategies. Consider the collection, $\mathcal{F}$, of probability mass functions on $\mathcal{G}$. That is, $f \in \mathcal{F} \Leftrightarrow f(p,q) \ge 0$ and $\sum_{\mathcal{G}} f(p,q) = 1$. Let $f_1, f_2 \in \mathcal{F}$ be any mixed strategies of bidder 1 and bidder 2 respectively. Then, their individual expected pay-offs against the strategy profile (f_1, f_2) are given respectively by

$$\pi_1'''(f_1, f_2) = \sum_{(p_1,q_1)\in\mathcal{G}} \sum_{(p_2,q_2)\in\mathcal{G}} \pi_1(p_1, q_1, p_2, q_2) f_1(p_1, q_1) f_2(p_2, q_2)$$

$$\pi_2'''(f_1, f_2) = \sum_{(p_1,q_1)\in\mathcal{G}} \sum_{(p_2,q_2)\in\mathcal{G}} \pi_2(p_1, q_1, p_2, q_2) f_2(p_2, q_2) f_1(p_1, q_1)$$

A strategy profile (f_1^*, f_2^*) is a Nash equilibrium for the game if

$$f_1^* = \arg\max_{f_1} \pi_1'''(f_1, f_2^*) \text{ and } f_2^* = \arg\max_{f_2} \pi_2'''(f_1^*, f_2).$$

The following proposition is a direct consequence of Nash's theorem for finite action games.

Proposition 2.1 *A Nash equilibrium in mixed strategies exists for the discrete auction game*

Below we give a characterization for mixed strategy equilibria and later develop some conditions that a pure strategy equilibrium should satisfy. Interestingly, in the equilibrium of the auction game considered here, the price supports of individual players are similar. Or more precisely,

Proposition 2.2 *Let* (f_1^*, f_2^*) *be a Nash equilibrium. Further, let* $P_1^* := \{p : \exists q \ni f_1^*(p,q) > 0\}$ and $P_2^* := \{p : \exists q \ni f_2^*(p,q) > 0\}$. *Then the following hold:*

$$\forall p_1 \in P_1^* \exists p_2 \in P_2^* \ni p_2 \in \{p_1 - \alpha, p, p + \alpha\} \quad \text{--(5)}$$

$$\forall p_2 \in P_2^* \exists p_1 \in P_1^* \ni p_1 \in \{p_2 - \alpha, p_2, p_2 + \alpha\} \quad \text{-- (6)}$$

Proof (by contradiction):

Suppose that (5) and (6) do not hold. That is, suppose that there exists a $p_1^* \in P_1^*$ such that $P_2^* \cap \{p_1^* - \alpha, p_1^*, p_1^* + \alpha\} = \phi$. We analyze the situation in two disjoint cases.

<u>Case 1</u>: Suppose $\exists p_2^* \in P_2^*$ such that $p_2^* > p_1^* + \alpha$.

In this case, the bidder 2 can move the probability mass on p_2^* to $p_1^*+\alpha$ and increase his expected pay-off. To see this, lets redefine the second bidder's strategy f_2^* as follows:

$$\begin{aligned} \tilde{f}_2^*(p,q) &= f_2^*(p,q), && \text{if } p \neq p_2^* \text{ or if } p \neq p_1^*+\alpha \\ &= 0, && \text{if } p = p_2^* \\ &= f_2^*(p_2^*,q)+f_2^*(p_1^*+\alpha,q), && \text{if } p = p_1^*+\alpha \end{aligned}$$

Now it is easy to see that the difference in bidder 2's expected pay-off due to the above change in randomization of his pure strategies is:

$$\pi_2'''(f_1^*,\tilde{f}_2^*)-\pi_2'''(f_1^*,f_2^*)=$$
$$\sum_{(p_1,q_1)\in G}\sum_{q_2\in Q} f_1^*(p_1,q_1)f_2^*(p_2^*,q_2)[\pi_2(p_1,q_1,p_1^*+\alpha,q_2)-\pi_2(p_1,q_1,p_2^*,q_2) \text{ --(7)}$$

Now,

$$\pi_2(p_1,q_1,p_1^*+\alpha,q_2)=\int_o^{q_2\wedge N}(D_2^{-1}(s)-(p_1^*+\alpha))ds \text{ and}$$

$$\pi_2(p_1,q_1,p_2^*,q_2) \quad = \quad \int_o^{q_2\wedge N}(D_2^{-1}(s)-p_2^*)ds$$

Since by assumption $p_2^* > p_1^*+\alpha$, $\pi_2(p_1,q_1,p_1^*+\alpha,q_2) > \pi_2(p_1,q_1,p_2^*,q_2)$ and hence, $\tilde{f}_2^*$ is a better response to f_1^* than f_2^* contradicting the equilibrium property.

Case 2: $\forall p_2^* \in P_2^*, p_2^* < p_1^* - \alpha$.

The above condition implies that $\max_p\{p : p \in P_2^*\} =: p_{\max} < p_1^* - \alpha$. In this case, it can be argued proceeding along the above lines that bidder 1 can improve his expected profit by moving the probability mass $f_1^*(p_1^*,q_1)$ to the point $(p_{\max},q_1)$ in response to bidder 2's strategy f_2^* contradicting the equilibrium property.

Now, we present in the form of two propositions, how a pure strategy equilibrium should look if it exists.

Proposition 2.3 Let $(p_1^*, q_1^*, p_2^*, q_2^*)$ be a Nash equilibrium in pure strategies. Then the following should hold.

(i) $p_1^* = p_2^*$ or

(ii) $(p_1^*, p_2^*) = (\alpha, 2\alpha)$ or $(2\alpha, \alpha)$

Proof: From Proposition (2.2), it follows that $|p_1^* - p_2^*| \leq \alpha$. Hence, only the following three possible cases can´occur in equilibrium in our discrete model.

$p_1^* = p_2^*$ or $p_1^* = p_2^* + \alpha$ or $p_1^* = p_2^* - \alpha$. --- (8)

Suppose $p_2^* > 2\alpha$. In this case, we show that player 1's best response to player 2 cannot have a price component $p_1^* = p_2^* - \alpha$. Using a symmetric argument, one can show that if , $p_1^* > 2\alpha$, player 2's best response on price cannot be $p_1^* - \alpha$. In view of (8) above, it follows that if any player quotes a price $p > 2\alpha$, then the other player also should quote the same price p in his strategy.

Now consider the case when $p_2^* > 2\alpha$. In this case, we show that $p_2^* - \alpha$ cannot be a best response of Player 1. First, note that

$$\pi_1(\, p_2^* - \alpha, q_1^*, p_2^*, q_2^*) = \int_0^{q_1^* \wedge \{N - q_2^*\}} (D_1^{-1}(s) - (\, p_2^* - \alpha))ds$$

$$\leq \int_0^{q_1^* \wedge \{N - q_2^*\}} (D_1^{-1}(s) - \alpha)ds$$

$$\leq \int_0^{\tilde{D}_1(\alpha)} (D_1^{-1}(s) - \alpha)ds$$

$$= \pi_1(\alpha, \widetilde{D}_1(\alpha), p_2^*, q_2^*)$$

if $(q_1^* \wedge N - q_2^*) \neq 0$. Hence $(\alpha, \widetilde{D}_1(\alpha))$ is a better strategy than $(p_2^* - \alpha, q_1^*)$ whenever the term $(q_1^* \wedge N - q_2^*)$ contradicting the equlibrium property. In the case when $(q_1^* \wedge N - q_2^*) = 0$, since $q_1^* \geq \beta > 0$, it follows that $q_2^* = N$. But in such a scenerio, player 2 can in fact improve his profit by using the strategy $(2\alpha, N)$ in stead of (p_2^*, N).

Thus, if either $p_1^* > 2\alpha$ or $p_2^* > 2\alpha$, then it follows from (8) that, $p_1^* = p_2^*$ should hold always.

In the case when $p_1^* \leq 2\alpha$ or $p_2^* \leq 2\alpha$, then it follows that one of the above two conditions (i) and (ii) should hold by our discreteness assumption on price grid. Hence the proof.

Proposition 2.4 Let $(p_1^*, q_1^*, p_2^*, q_2^*)$ is a Nash equilibrium in pure strategies. Then $\widetilde{D}_1(p_1^*) + \widetilde{D}_2(p_2^*) > N$.

Proof: From Proposition (2.3), it follows that $p_1^* = p_2^*$ or $(p_1^*, p_2^*) = (\alpha, 2\alpha)$ or $((2\alpha, \alpha)$.

In the cases when (i) $p_1^* = p_2^* = \alpha$ or (ii) $p_1^* = \alpha(2\alpha)$ and $p_2^* = 2\alpha(\alpha)$, the result is from our assumption on demand functions.

Consider the case when $p_1^* = p_2^* > \alpha$. Assume that $\widetilde{D}_1(p_1^*) + \widetilde{D}_2(p_2^*) < N$. Then the quantity components of equilibrium strategy should be $q_1^* = \widetilde{D}_1(p_1^*)$ and $q_2^* = \widetilde{D}_2(p_1^*)$. But then, player 1 (or player 2) can get the same quantity with a lower price component and improve his pay-off: that is, for player 1, $(\alpha, \widetilde{D}_1(p_1^*))$ is a better strategy than $(p_2^*, \widetilde{D}_1(p_2^*)$) against player 2's strategy $(p_2^*, \widetilde{D}_2(p_2^*)$)(and a similar statement holds for player 2 also) contradicting the equilibrium property of $(p_1^*, q_1^*, p_2^*, q_2^*)$. Hence the proof.

Since it is fairly difficult to derive conditions under which equilibrium in pure strategies exist, in the following we depart from game theoretic analysis and analyse the Dutch auction problem as a single agent learning in uncertain environments. We present below two independent models for Internet-based Dutch auctions and provide details of various learning procedures.

3. A Dutch Auction when Bidders cannot Observe Dynamics

Consider a decreasing type Dutch auction over a fixed duration T where Q identical units of an item are auctioned out for sale. Bidders who arrive before T will observe the ongoing price and decide on whether to purchase the items at the offered price or to revisit the system after a random duration.

To make our model realistic, we assume that the auctioneer decides on price-setting only at a prespecified decision epochs, $0 = t_0 < t_1 < .. < t_N = T$. For notational simplicity, we use $\Im := \{0,1,2..,N\}$ to denote the set of decision epochs. Demand observed between two successive decision epochs is the aggregated demand from all the bids that occur during that specific decision interval.

Let the auctioneer begin the auction by asking an initial price, say, P_0 and at a decision epoch, $t \in \Im$ decides to choose a price decrement action a_t from a finite set of available actions, $\mathcal{A} := \{a^1, a^2, ..., a^M\}$. An action $a_t = a^j$ is interpreted as decreasing the current offered price by an amount a^j or equivalently, setting the ongoing price at time t, P_t at $P_t \; - a^j$ and holding the price at that level until the next decision epoch $t+1$. State of the auction at any time t, is represented by a pair: $x_t =: (q, P)$ where q is the remaining quantity for sale at time t and P is the offered price at time t. We use S to denote the state space. The action a_t in state x_t results possibly in a change in the state of the auction which is assumed to follow the probabilistic transition structure defined by:

$$\Pr(x_{t+1} = \mathbf{y} \mid x_t = \mathbf{x}, a_t = a; \mathbf{x}_{t-1}, a_{t-1}, .., \mathbf{x}_0, a_0)$$
$$= P^t_{\mathbf{xy}}(a)$$

Observe that the above conditional probability law follows Markovian evolution. This follows since if a bidder upon his arrival cannot get information regarding quantity sold upto his arrival epoch and also if information regarding other successful bids is hidden, then it follows that his bidding behaviour is only dependent on his own private value for the item and his own demand-price function which will possibly vary with time. Hence the above non-stationary Markovian transition law captures the evolution of the auction in such a scenario fairly well.

Also, a price setting action by the auctioneer will result in an immediate reward amounting to revenue earned from sale of items happeing during the interval $(t, t+1]$. At $t = N (= T)$ the auction ends and the rewards accumulate at the end of next time unit say, $N+1$. Formally, we can say that all the actions taken at $N+1$ will lead to a zero reward.

Let $r_t := r(\mathbf{x}_t, a, \mathbf{y}_t)$ denote the immediate reward that accumulates over two successive decision epochs $(t, t+1]$ and is given by

$r(\mathbf{x}, a, \mathbf{y}) = (q_1 - q_2)(P_1 - a)$ where $\mathbf{x} = (q_1, P_1)$ and $\mathbf{y} = (q_2, P_2)$.

With the above notation, the following Markov decision process model is in order:

A (non-stationary Markovian) deterministic policy π is an associative mapping:

$$\pi : (t, x) \rightarrow \pi(t, x) \qquad (1)$$

which associates with each time - state pair an admissible action from $\mathcal{A}_{\mathbf{x}} \subseteq \mathcal{A}$.

Note that by the nature of our decreasing type model set of admissible actions in any state $\mathbf{x} = (q, P)$ should be such that the price set thereon should lie within $[R, P]$ where R is the seller's reserve price. Thus, available action set is state dependent. Let $\prod$ denote the collection of such Markovian deterministic polices. By the finiteness assumption on admissible actions, $\prod$ is of finite cardinality.

Associated with every such policy π is a value function:

$$V^{\pi}(x) = E_{\pi}[\sum_{t=0}^{N} r(\mathbf{x}_t, \pi(t,\mathbf{x}_t), \mathbf{y}_{t+1}) \mid \mathbf{x}_0 = \mathbf{x}]$$

where E_{π} is the expectation operator under the policy π. Similarly we can define at any time k, the total expected reward from time k on if policy π is being followed, is denoted by

$$V_k^{\pi}(\mathbf{x}) = E_{\pi}[\sum_{t=k}^{N} r(\mathbf{x}_t, \pi(t,\mathbf{x}_t), \mathbf{y}_{t+1}) \mid \mathbf{x}_k = \mathbf{x}]$$

with $V^{\pi}(\mathbf{x}) = V_o^{\pi}(\mathbf{x})$

Let $V_k^*(\mathbf{x}) = \max_{\pi \in \prod} V_k^{\pi}(\mathbf{x})$. Then V_k^* satisfies the following dynamic programming equations:

$$V_k^*(\mathbf{x}) = \max_{a \in A_{\mathbf{x}}} \{\sum P'_{\mathbf{xy}}[r(\mathbf{x}, a, \mathbf{y}) + V_{k+1}^*(\mathbf{y})]\}$$

$$\forall \mathbf{x}, \mathbf{y} \in S, k \in \{0,1,.., N\} \qquad (3)$$

with $V_{N+1}^*(\mathbf{x}) = 0, \forall \mathbf{x}$

If we denote by $Q_k(x,a)$ the term inside the braces in (3), it follows that

$V_k^*(x) = \max_{a \in A_{\mathbf{x}}} Q_k(\mathbf{x}, a)$ and

$$\pi_k^*(\mathbf{x}) = \arg\max_a Q_k(\mathbf{x}, a) \qquad (4)$$

$Q_k(\mathbf{x}, a)$ represents the expected reward earned by following action a in state $\mathbf{x}$ at time k and proceeding optimally from time $k+1$ onwards.

Note that if the probability structure is known , then (3) can be solved using Backward Induction. Since the immediate rewards are bounded and at any time, depend on the history only through the current state and action and also since the probability of transitions depend only on current state and action an optimal Markovian deterministic policy exists and an optimal policy in $\prod$ can be derived from (4).

When the probability structure is not known apriori, Reinforcement Learning in MDPs can be invoked to learn an optimal Markovian determinstic policy. The procedure is detailed in the next section.

2.2 Q-learning for Optimal Policy

Reinforcement Learning algorithms take a stochastic approximation approach to solve the Dynamic Programming equations that appear in MDPs. Instead of attempting to build the underlying MDP model and then solve the problem using Backward Induction, these algorithms take advantage of underlying dynamic Markovian behaviour and directly learn optimal policies using iterated observations of rewards and state transitions.

Powerful convergence results have been established for many Reinforcement Learning algorithms in the Markovian setting. A majority of these algorithms address infinite horizon stationary problems. Ho'vever, the Dutch auction model described above does not directly fall under the purview of these algorithms by virtue of the non-stationarity and finite horizon nature of the problem. In order to pose the problem in the traditional infinite horizon setting, we suggest the following transformation to our model. Following Bertsekas [1995], we convert our non-stationary finite horizon problem into an equivalent stationary problem by considering time component as part of the state: The new state space will now be $\tilde{S} := \Im \times S$. Further , in $\tilde{S}$ identify a reward free absorbing set, $\tilde{S}_0$. A state $s = (t, x), t \in \Im, x \in S$ is in $\tilde{S}_0$ if either $t = N + 1$ or $x = (0, P)$ for any $P \leq P_0$. The motivation underlying this transformation is that the Dutch auction ends either when the stipulated end period is reached (possibly with a full or partial sale of items in the inventory) or when all the quantity is sold out. For space reasons and also since it is fairly a simple exercise, we do not reformulate the model with the suggested transformation.

Further, note that it is not possible to directly invoke learning mechanism in our case since the sample paths (auction traces) terminate before or at $N + 1$ and learning requires repeated occurrence of all the states. In order to be able to learn, we suggest that whenever the auction enters the absorbing set $\tilde{S}_0$, a new sample path is initiated from state $\mathbf{x}_0 = (Q, P_0)$, that is, the same starting state is used for all sample paths. This can be viewed as learning from different complete auction traces that are stitched together to form an infinite horizon sample trace. For online learning, this is a realistic framework when the auctioneer repeatedly conducts such Dutch auction with the same start inventory and start prices or when the auctioneer can categorize auctions happening at his site according to some similarity features and then uses proper normalization to make them qualify as exactly similar auctions. Discussion on such classification is beyond the scope of this paper. For offline purposes, the auctioneer can always simulate a Dutch auction and start an actual auction with the simulated experience. In fact in Section 5, we devise a simulation experiment using some realistic models of bidders and their strategies.

Q-learning is an RL algorithm which iteratively generates $Q(\mathbf{x},a)$ values described in (3) for each state-action pair $(\mathbf{x},a)$ using online or simulated experience. The algorithm proceeds as follows: At each iteration n, it updates the estimation $\hat{Q}_n$ of Q from the current observations $(\mathbf{x}_n, a_n, \mathbf{y}_n, r(\mathbf{x}_n, a_n, \mathbf{y}_n))$ obtained from actual execution of action a_n in $\mathbf{x}_n$ in an online or simulated experiment and replacing the conditional average appearing in (3) by evaluation of the actual transition. It moves only incrementally in the desired direction by taking a weighted linear combination of previous estimate and the value suggested by current observation to ensure stability of updating procedure. For details of the Q-learning procedure see Bertsekas and Tsitsiklis [1996] or Sutton and Barto [1998].

For the Dutch auction problem, the Q-learning algorithm is as follows:

Algorithm

Initialize $Q(\tilde{\mathbf{x}},a)$ for all $\tilde{\mathbf{x}} \in \tilde{S}$ and $a \in \mathcal{A}_{\tilde{\mathbf{x}}}$ to zero value.

For $n \geq 0$, observe $\langle \tilde{\mathbf{x}}_n, a_n, \tilde{\mathbf{y}}_n, r(\tilde{\mathbf{x}}_n, \tilde{\mathbf{y}}_n) \rangle$. Let $l = n \bmod (N+1)$

Update Q-value estimates by:

$$Q_{n+1}(\tilde{\mathbf{x}},a) = Q_n(\tilde{\mathbf{x}},a) + \beta_n(\tilde{\mathbf{x}},a)\Delta_n \qquad \text{--- (5)}$$

where Δ_n is given by:

$$= r_n + \max_b Q_n(\tilde{\mathbf{y}}_n, b) - Q_n(\tilde{\mathbf{x}},a), \text{ if } (\tilde{\mathbf{x}},a) = (\tilde{\mathbf{x}}_n, a_n) \text{ and } l < N$$

$$= r_n - Q_n(x,a), \text{ if } (\tilde{\mathbf{x}},a) = (\tilde{\mathbf{x}}_n, a_n) \text{ and } l = N$$

$$= 0 \text{, otherwise}$$

Further, set $\tilde{\mathbf{x}}_{n+1} = \tilde{\mathbf{y}}_n$ if $\tilde{\mathbf{x}}_n \notin \tilde{S}_0$ and $\tilde{\mathbf{x}}_{n+1} = \tilde{\mathbf{x}}_0$, otherwise. The learning parameters $\beta_n(\tilde{\mathbf{x}},a)$ are small learning rates decreasing with time.

Theorem 2.1: If each pair $(\tilde{\mathbf{x}},a) \in \tilde{S}$ *is visited infinitely often and if* $\sum_{n} \beta_n(\tilde{\mathbf{x}},a) = \infty$, $\sum_{n} \beta_n^2(\tilde{\mathbf{x}},a) < \infty$, *then* $Q_n(\tilde{\mathbf{x}},a) \to Q(\tilde{\mathbf{x}},a) \forall (\tilde{\mathbf{x}},a) \in \tilde{S}$ *w.p.1*

Since proof of convergence follows from a straight forward application of stochastic approximation algorithms , we omit details.

The above procedure ensures convergence only in the Markovian setting. However, when a bidder upon her arrival can see the available quantity and also the successful bidders, then her bidding behaviour is dependent on the history of the auction. Hence the auction state transitions are affected by bidders' strategies over time. In such cases, one needs to relook at the above model. In the next section, we try to address the performance of direct temporal difference based Reinforcement Learning algorithms such as 1-step Q-learning discussed above in such non-Markovian domains and contrast them with the trace based mechanisms which are essentially Monte-Carlo based actual return algorithms.

4. A Dutch Auction when Bidders can Observe Dynamics

Consider the decreasing type Dutch auction described in the foregoing in a setting where any bidder can have access to information pertaining to quantity available at any time and can have knowledge of successful bids or in general, can observe dynamics of the auction. In such cases, a bidder plays a strategic game and will possibly derive positive utility by playing odds at the gambling table or a bidder may update his maximum willingness to pay or more generally re-evaluate his price-demand function. In such cases, auction state of our macroscopic model described in the previous case will not in general evolve in a Markovian fashion. The evolution is dictated by the bidding behavior of the participants, which invariably is linked to history of dynamics. Hence, the conditional law with respect to complete history is needed to decide on optimal price control strategy.

Even though there is an empirical evidence that direct application of RL algorithms developed specifically for Markovian environment can even work well in some particular non-Markovian cases [Barto et al, 1983], there is no characterization of such cases as yet. Singh et al [1994] analyze the implications of direct RL in non-Markovian cases and in general POMDPs and show that RL algorithms do not degrade gracefully with the degree of non-Markovianness.

The state (Remaining Quantity, Offered Price) used in the previous section can be regarded as an observation of a hidden state, which evolves in a Markovian fashion. In our Dutch auction model, without loss of generality we assume that at any decision epoch history up to that point is enough to disambiguate observations. In other words, if the complete history or trace of the auction is known, then one can identify states, which evolve in a Markovian fashion. A simple plausible way to handle this is to include history as part of the state space and invoke Q-learning algorithms. But it will turn out to be computationally expensive because of state space explosion. Further, it may even happen that not all states are visited during exploration phase leading to learning algorithm stabilizing on sub-optimal decisions. Instead, using the fact that history is enough to disambiguate observations, we retain the original state space developed in the previous section and accordingly define Q-values using conditional arguments. Even though the (price, quantity) pairs qualify only as *observations*, (following the terminology of POMDPs) in the following discussion, for the sake of clarity and continuity we refer to them as *states* only.

In the next section, we present a few examples to show that in a non-Markovian Dutch auction model Temporal Difference based learning methods may fail to learn optimal policies even in cases where deterministic observation based optimal policies exist. But actual-return or trace-based learning algorithms (of Monte-Carlo type) will successfully be able to learn such policies in the discussed model. We make these notions precise in the sequel.

4.1 Temporal Difference Vs Actual Return based Learning in a non-Markovian Dutch Auction

We construct two illustrative examples that expose some inadequacies of a direct application Q-learning in a non-Markovian Dutch auction.

Assume that a Dutch auction starts with an initial inventory 50 at a price level 100. Now, consider three states x_1, x_2, x_3 representing the tuple - (remaining time, remaining quantity, offered price)[=] - as depicted in Figure 1 below.

[=] State is described according modified representation presented in Section 2.1 but for the difference that remaining time replaces the original time

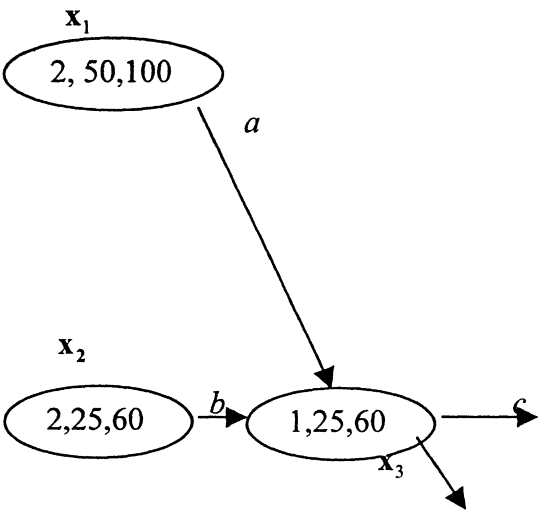


Figure 1

Assume that action a in state $\mathbf{x}_1$ leads to $\mathbf{x}_3$ with a positive probability p_1 and action b in state $\mathbf{x}_2$ leads to state $\mathbf{x}_3$ with a positive probability p_2. The expected payoff from state $\mathbf{x}_3$ following action c(zero decrement or not changing the ongoing price) is dependent on its predecessor states. If the predecessor is $\mathbf{x}_1$, then suppose that this expected payoff is R_1 where as, if the predecessor is $\mathbf{x}_2$ then the expected pay-off from $\mathbf{x}_3$ following action c is R_2 which is orders of magnitudes less than R_1. Recall that Q-value of a state-action pair $(\mathbf{x}, a)$ refers to expected utility gained from taking action a in state $\mathbf{x}$ and proceeding optimally from then on. Given that the state $\mathbf{x}_3$ is reached assume that with a high probability transition must have occurred from $\mathbf{x}_1$. In this case, the auctioneer may consider c to be a profitable action in state $\mathbf{x}_3$. Now if one were to calculate the Q-value for the state-action pair $(\mathbf{x}_3, c)$, then the above expected payoffs need to be unconditioned on history and for ease of argument assume that payoff for $(\mathbf{x}_3, c)$ does not depend on history beyond one step backwards, then it is easy to see that the unconditional expected payoff for $(\mathbf{x}_3, c)$ is approximately equal to $p_1 \cdot R_1$. Now assume that p_1 and p_2 are very high. In the former case last -minute frenzy is being observed; a phenomenon particularly predominant in the Internet context, (See Roth and Ockenfels [2000].) where as in the latter case auction can be viewed as progressing at a relatively smoother pace perhaps in a uniform way. In the light of these interpretations, the conditional rewards mentioned above for action c in $\mathbf{x}_3$ also make a realistic assumption.

However it is interesting to note that in the above scenario if Temporal Difference (TD) based updates are used, the Q-value of $(\mathbf{x}_3, c)$ will settle approximately at $p_1 \cdot R_1$ and as a result action b might even be considered to be performing well by the learning algorithm though the action would always lead to degradation in the average returns from $\mathbf{x}_2$ on if the auctioneer's policy is to select c in $\mathbf{x}_3$. In a non-Markovian Dutch auction the informed buyers try to play a strategic game based on observed dynamics which will affect the total expected revenue of the auctioneer. This dependency of bidders' strategies on previously observed dynamics will have a certain biasing affect on future expected revenues and any RL agent who uses TD based Q-learning updates fails to observe such phenomenon.

The example above has been presented in a very simplified framework with one-step history. The argument can be extended to cases where dependency goes up to the starting point in time.

Now in stead of the TD based Q-value of update (as suggested in (5)) suppose we use the following actual return based update for Q-values: Let $\tilde{\mathbf{x}} = (t, q, P)$ be the state at time t. Also let π be be policy followed from time t on and τ be the number of time steps from t on to reach the absorbing set $\tilde{S}_0$. (Note that $\tau \leq N \;\; a.s$ by our construction).

$$Q_{n+1}(\tilde{\mathbf{x}}, a) \leftarrow (1-\beta_n)Q_n(\tilde{\mathbf{x}}, a) + \beta_n(\tilde{\mathbf{x}}, a)\Delta_n \qquad \text{-- (6)}$$

where Δ_n is given by

$$\Delta_n := \sum_{k=1}^{\tau} r_{t+k}$$

where r_i is the actual immediate return at time step i.

In other words, the update rule (6) is applied only when a termination state is reached and only to the state- action pairs for which no non-policy actions executed subsequently.

In the above discussed case, if the auctioneer or his agent uses an update rule for Q-values based on actual returns, the agent will in a few iterations would learn that action b is less rewarding than the values derived from Temporal Difference based Q-value updates.

Next we illustrate with an example that in general Temporal Difference based methods may even fail to learn optimal deterministic observation based policies in our non-Markovian Dutch auction.

Consider the dynamics of a Dutch action as shown in Figure 2. Assume that the initial inventory and the start price are as in the previous example. In addition, assume that seller's reserve price is 60.

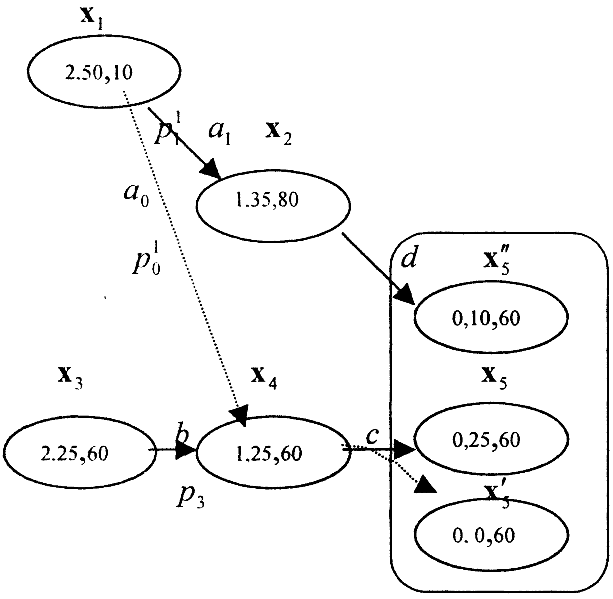


Figure 2

In state $\mathbf{x}_1$, assume that the two actions a_0, a_1 lead to states $\mathbf{x}_4, \mathbf{x}_2$ respectively with high positive probabilities (p_0^1, p_1^1 respectively as indicated). In the other states, the actions and the corresponding probabilities are depicted in the figure. Note that, as in the previous example case, the expected profit from state $\mathbf{x}_4$ for the action c will be $R_1 > 0$ (in equivalent terms, will lead with probability one to $\mathbf{x}_5'$) conditional upon its predecessor state being $\mathbf{x}_2$ and action followed there being a_0. The same action, c can also lead to $\mathbf{x}_5$ with zero reward if the predecessor is $\mathbf{x}_3$ and the action followed is b. Also, consider another path from the state $\mathbf{x}_1$ that terminates in state $\mathbf{x}_5''$ with positive probability via $\mathbf{x}_2$ as shown in the figure.

States $\mathbf{x}_5, \mathbf{x}'_5$ and $\mathbf{x}''_5$ belong to the reward free absorbing set $\widetilde{S}_0$ (see Section 2.2) indicating the terminal stage of the auction.

A realistic way to interpret the above dynamics is as follows: When the remaining time to close the auction is only two units, the auction state can be either $\mathbf{x}_1$ or $\mathbf{x}_3$. The dynamics reflecting transitions: $\mathbf{x}_1 \rightarrow \mathbf{x}_2 \rightarrow \mathbf{x}''_5$ can be seen to result in a loss in expected revenue items (also loss in a possible sale of few more items) which is a fall-out of a sub-optimal action a_1 in $\mathbf{x}_1$. A simple way to explain the above phenomenon from $\mathbf{x}_1$ is to interpret that a significant fraction of the bidders who arrive (or who are already present when the state is $\mathbf{x}_1$) during the last phase have a per unit valuation (or their updated estimates) around 60 (below 80) which is also the reserve price of the seller.

Now consider the case where both the states $\mathbf{x}_1$ and $\mathbf{x}_3$ are equi-probable. (With a suitable modification, the analysis can be extended to cases where these probabilities are arbitrary) Assume that dependencies of rewards on histories do not go beyond one step as in the earlier example. It is easy to see that for any policy π, $Q^{\pi}(\mathbf{x}_4, c)$ equals $\frac{1}{2}\, p_0^1 * R_1$.

Now, let π^1 and π^2 be two (observation- based) policies with the following partial assignments:

$\pi^1(\mathbf{x}_1)= a_0$, $\pi^1(\mathbf{x}_4)= c$ and $\pi^1(\mathbf{x}_3) = b$:

$\pi^2(\mathbf{x}_1)=a_1$ and $\pi^2(\mathbf{x}_2)=d.$

It follows that,

$$Q^{\pi^1}(\mathbf{x}_1, a_0) = r(\mathbf{x}_1, a_0, \mathbf{x}_4) + Q(\mathbf{x}_4, c)$$

and

$$Q^{\pi^2}(\mathbf{x}_1, a_1) = r(\mathbf{x}_1, a_0, \mathbf{x}_2) + Q(\mathbf{x}_2, d)$$

It is possible to identify cases where the values for p_0^1, p_1^1 (recall that both are assumed to have high values as per discussion above) are such that $Q^{\pi^1}(\mathbf{x}_1, a_0) < Q^{\pi^2}(\mathbf{x}_1, a_1)$. (A simple case is with $p_0^1 = p_1^1 = 1$). Hence the one--step Q-learning always picks up action a_1 in state $\mathbf{x}_1$ even though a_0 gives improved performance. Since in the learning phase, weighted linear combinations are used, there is no way that the 1-step Q-learning will learn the improved policy π^1.

Further note that, for the given set of actions shown in the figure, in fact π^1 is a deterministic optimal observation- based policy in the simple case. This shows that the Q-learning sometimes may fail to learn the deterministic optimal policies too in such non-Markovian domains.

Also, interestingly, if in stead of the 1-step Q-learning, actual return updates as suggested in (6) are used to estimate the Q-values above, then action a_0 is always selected in the simple cases referred to above. Also the above argument can be extended to show that any learning rule that uses TD(λ),$0 < \lambda < 1$, type of updates cannot learn the optimal policy π^1 since the methods underlying also use some form of weighted linear combination of corrected truncated returns. For a formal treatment of the above discussion, see Ravikumar and Manish Gupta [2001]

The above arguments allow us to infer that Temporal Difference based methods in general do not converge to optimal deterministic strategies in non-Markovian Dutch auction domains. Since actual return based methods offer superior performance in this context, we suggest that such approaches (of TD(1) genre) be used in this case. Any suitable trace-based learning mechanism may achieve this purpose. For such approaches, see Sutton and Barto [98].

However, actual return based approaches will, in general, be inefficient with regard to implementation and also in terms of speed of convergence. Hence this fact throws open two possible options for learning. The first option is to convert the problem into a Markovian framework by appending in a suitable form all the information that is deemed to be relevant in converting the problem to an MDP and then invoke the simple 1-step Q-learning rule. The second option is to retain the original state space (of the previous sections) and develop efficient trace-based mechanisms to improve speed of convergence.

In Partially Observable domains, Singh et al [1994] and Jaakkola et al [1995] argue that randomized policies may perform consistently better than deterministic polices. Recent developments in actor-critic algorithms (Konda and Borkar [2000]) can give a cue on learning such randomized strategies. We address some of these issues in our future work.

5. Design of Simulation Experiment

The purpose of this simulation study is two fold. Firstly, we design a realistic simulation experiment to validate our analysis of the foregoing sections. Secondly, if the auctioneer has a good model of the auction in terms of the bidders' strategic behavior and their price-demand variational functions, learning algorithms can be run on these simulated models whenever it is very hard to find out optimal policy. The latter case may arise in situations where the state space is formidably large or in the cases where Backward Induction in computationally expensive. In the presence of model, auctioneer can experiment with simulations to train the learning system in an offline manner and can use the offline experience in the actual auction. In such a scenario, our simulation experiment designed below can provide insights in developing an experiment that suits the purpose of the auctioneer. With this motivation behind, we present in the following, our simulation procedure.

We assume that the auctioneer has knowledge of the market price. say C of the item currently being auctioned. In the absence of such knowledge, the auctioneer can develop/use a shopbot (See Kephart and Tesauro 1999] discover the current market price of the item. Further, since geographically different consumers participate in the auction, we assume that initially each consumer values the item independently and we assume all bidders are symmetric on their first visit to the auction site and draw their valuations randomly from a uniform distribution over $[0, C]$. We used uniform distribution mainly for reasons of analytical tractability and also since it always has finite support. The upper-limit is justified since buyers can as well use a shopbot before they enter the auction to know about the prevailing market price. In our simulation experiment, we allow each bidder to update later their upper-limit as auction progresses.

The bidders are classified into two categories. The Category 1 bidders (referred to as *definitive bidders* henceforth) are risk averse in the sense that if the current going price of the item is found to be below their valuation, they will immediately call a halt and buy a unit of the item. On the contrary, the Category 2 bidders will wait for the price to fall further below their valuation. the duration being dependent on their estimation of competitive dynamics. So in the simulation experimentation, a fixed proportion of the bidders have been assumed to belong to Category 1. (If there is a *statistical* way to identify such bidders, the relative proportions can be estimated from actual data and can be used in simulation)

A bidder belonging to Category 1 will bid immediately (with probability 1) when the going price at his time of arrival is below his valuation of the item. If not, he will revisit the site (or may even place a (zero)-intelligent agent at the auctioneer's site). In contrast, a bidder of Category 2 will observe the ongoing price and waits for a duration and revisits the site at regular intervals to watch the competitive dynamics in order to estimate the rate of purchase of items. He then uses this estimate to find out the expected number of visitors in the remaining course of auction according to the observed rate and bids probabilistically. In our simulation, we use Bernoulli distribution, with success probability p as a function of his estimated valuation of other bidders, the quantity remaining in the inventory and the expected number of arrivals. In fact, the procedure involved in deriving the parameter for non-definitive bidders, in a simple case where every bidder approaches auction with unit demand and every bidder knows this fact, is as follows: Whenever a bidder has to choose an option of bidding at that hour , he observes the current remaining quantity, say q and accordingly finds the existing rate of inventory movement. Further, he uses this estimate to compute the future expected demand until the end of the auction. By our unit demand assumption, this is equivalent to N the number of future arrivals.

He uses this information, to find out the probability that $q-th$ order statistic in N uniformly distributed valuations over $[0,C]$ exceeds the current offered price. This probability is used as success probability in the above mentioned case. In the general framework, since it is fairly hard to replicate this procedure, we allow each bidder to make a random guess on the future demand and to use this in deciding on his bidding probability. Note that, in the presence of only Category 1 bidders, the simulation provides an approximately Markovian environment where as in the presence of a mix of categories, it provides a non-Markovian environment.

In our simulation studies, we experimented with some simple cases where all the bidders are assumed to be of Category 1 and further, optimal policy can be found explicitly. In many of these cases, the Q-learning algorithm converges to optimal policy in a few iterations. In the general case which includes a mix of definitive and non-definitive bidders, we use a trace based mechanism to update Q-values. That is, a policy is followed through out until the absorbing set is reached and then Q-values of the visited states and actions are updated based on actual return received as in (6). In terms of average performance which measures average expected returns over a sequence of auction traces, the above trace-based procedure has consistently out-performed ordinary one-step Q-learning algorithm supporting our intuition. However, our simulation experiments are still in progress and a thorough and comprehensive study needs to be carried out for additional insights.

6.Future Research

In this paper we analyzed applicability of Reinforcement Learning algorithms for price control in a Dutch auction in two independent cases where bidders can and cannot observe auction dynamics. Even though, the learning algorithms suggested are simple, our table look-up based approaches may turn out to be computationally expensive both for on-line and off-line learning as the state-space grows. For off-line learning extensive simulation studies are required including models of buyers and their demand patterns. For online learning, even though no such models are required per se, the learner or auctioneer must endure losses during the exploration phase of the learning procedure. In view of these facts, there is need for development of function approximation techniques that can expedite the learning process.

In the case of non-Markovian Dutch auction the learning problem is similar to learning in Partially Observable Markovian domains. In such domains, since randomized strategies can outperform deterministic counter parts, there is a need for development of algorithms that can learn such mixed strategies.

Another interesting plausible extension to our work is to analyze a situation where intrinsic value of items decreases with time during the course of the auction.

7. Conclusions

In this work. we have investigated application of Reinforcement Learning for price control in a Dutch auction. In the case when the auctioneer conceals information pertaining to ongoing dynamics of the auction from bidders, we proposed a Q-learning algorithm for generating an optimal price sequence that maximizes seller's expected revenue. In the case when bidders can observe the dynamics of the auction. we have shown that actual-return based Reinforcement Learning algorithms perform well compared to the Temporal Difference based algorithms.

From an implementation perspective since Reinforcement Learning offers a simple adaptive control framework, we hope that scope of its application will further extend to electronic marketplaces where autonomous agents play an active role in decision- making.

References

1. Barto, A.G., Sutton, R.S. and Anderson, C. W. *Neuron-like Elements that can Solve Difficult Learning Control Problems*, IEEE Trans. On Systems, Man and Cybernetics, 13, 835-846, 1983.
2. Bertsekas, D, *Dynamic Programming and Optimal Control, Vol II*, Athena Scientific, Massachusetts, 1995
3. Bertsekas, D and Tsitsiklis, J. N. *Neuro-Dynamic Programming*, Athena Scientific, Belmont, Massachusetts, 1996
4. Duflo, M. *Random Iterative Models*, Springer-Verlag, Heidelberg, 1997
5. Greenwald, A. R and Kephart, J. O., *Shopbots and Pricebots*, Proceedings of the International Joint Conference on Artificial Intelligence, *IJCAI* 99, pp: 506-511, 1999
6. Hu, J and Wellman, M. P., *Multi-agent Reinforcement Learning. Theoretical Framework and An Algorithm*, Proceedings of the Fifteenth International Conference on Machine Learning, *ICML*-98, Madison, WI; Morgan Kaufmann, 1998.
7. Jaakkola, T, Singh, S. P and Jordan, M.I, *Reinforcement Learning Algorithm for Partially Observable Markov Decision Problems*, In Advances in Neural Information Processing Systems 7, Morgan Kauffman, 1995.
8. Konda, V. M and Borkar, V. S. *Actor-Critic Type Learning Algorithms for Markov Decision Processes*, SIAM Journal on Control and Optimization, 38(1), pp 94-123, 1999

9. Menezes, F and Monteiro, P. K, *Existence of Equilibrium in a Discriminatory Price Auction,* Mathematical Social Sciences, 30, pp 285-292, 1995

10. Roth, A. and Ockenfels, *A Last minute bidding and the rules of ending second-price auction: Theory and evidence from experiment from a natural experiment on the Internet,* Technical Report, Harvard University, 2000

11. Ravikumar, K and Manish Gupta, *Reinforcement Learning Algorithms for Price Control in a Dutch auction*, Technical Report, IBM India Research Laboratory, New Delhi, India, 2001.

12. Singh, S.P, Jaakkola, T and Jordan, M, *Learning without State-Estimation in Partially Observable Markov Decision Processes*, In W. Cohen and H. Hirsch (Eds.) Machine Learning, Eleventh International Conference, *ICML-94*, New Brunswick, New Jersey, Morgan Kaufmann, 1994

13. Singh, S. P and Sutton, S. R., *Reinforcement Learning with Replacing Eligibility Traces*, Machine Learning, 22 (1-3), pp 123-158, 1996.

14. Sutton, R S and Barto, A. G., *Reinforcement Learning: An Introduction*, MIT Press, Cambridge, Massachusetts, 1998

15. Tesauro, G and Kephart, J. O. *Pricing in Agent Economies using Multi-agent Q-learning*, Proceedings of the Workshop on Game Theoretic and Decision Theoretic Agents, 1999.

16. Tesauro. G and Kephart, J. O. *Pseudo-Convergent Q-learning in Competitive Pricebots*. Proceedings of the Seventeenth International Conference on Machine Learning, *ICML 2000*. Stanford University. June 2000.

17. Watkins, C. J. C. H and Dayan, P. *Q-learning*, Machine Learning, 8, pp 279-292, 1992.