Computing Science

Evaluating the Effect of Using Restriction-free Policy in Reinforcement Learning for Dynamic Pricing within the Taxi Domain

Todor Davchev

No portion of the work contained in this document has been submitted in support of an application for a degree or qualification of this or any other university or other institution of learning. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

Signed:

Date: December 11, 2013

CS4040 Report

Department of Computing Science
University of Aberdeen
King's College
Aberdeen AB24 3UE

December 2013

Evaluating the Effect of Using Restriction-free Policy in Reinforcement Learning for Dynamic Pricing within the Taxi Domain

Todor Davchev

Department of Computing Science
University of Aberdeen

December 11, 2013

Abstract

The concept of reinforcement learning (RL) affecting profit maximisation is important for nowadays world since the ability to set costs appropriately and increase profit by learning from past experience can lead to great profit maximisation strategies. In this paper I evaluate the effect of applying reinforcement learning with a policy which allows full freedom for the agent when achieving the goal set. I formalise the notion of dynamic pricing within the taxi domain utilising strategies which make use of knowledge about a destination, discounting prices to those locations containing greater chance of providing a new customer and increasing prices to destinations which does not. I compare dynamic pricing policies with the restriction-free policy when it defines a taxi being able to take decisions on its own without taking into consideration any constraints. Additionally, I compare it with the currently existing approach for distance based charging.

1 Introduction

Taxi drivers base their charging fees on distance travelled only, without taking into consideration other influences such as the possibility of getting a new customer on dropping the current one at the destination acquired. The technique currently undertaken is known to be not as profitable as any driver would want it to be. The ability to adjust a price in order to maximise profit has challenged a large number of people and businesses throughout years. The reason for this are the various constraints and externalities that influence a given environment which makes fixing a price as challenging. A properly utilised RL strategy can lead to 100 000 times greater profit compared to the currently used techniques.

I examine the effect of using reinforcement learning with optimal policy in comparison to other pricing policies in order to better encapsulate the benefits of following the approach utilised. The solution is described within the taxi domain providing the opportunity to apply it within any other scenario depending on the situation. The research interest focuses on formalising the notion of dynamic taxi pricing strategies which make use of knowledge about areas in order to discount prices to locations

containing many potential passengers and increases fares to those areas with few potential such [7].

The rest of the paper is divided into 8 sections. In (Section 2), I provide a brief overview of the background and related work to the research interest of this paper. In (Section 3), I define the model for the environment that is required in order to describe the experiment. This will lead to (Section 4), where the research question is being described. (Section 5), presents the software overview- where a short description of the software used is defined. (Section 6) is described the experminetal design In (Section 7), I present the results from the experiments followed by a discussion (Section 8). Finally, (Section 9) involves a conclusion of the research question and experiment.

2 Background and related work

The area of reinforcement learning is an important field within AI and is proved to be highly applicable and widely used within a large number of domains [10] [3] [9] [2]. It is used for agents to learn optimal or near-optimal plans in the absence of labelled examples of what to do. The interest of this paper is to examine the performance of the so called active reinforcement learning within the taxi domain and to represent the extent to which the aforementioned type of learning contributes in increasing profit in comparison to other approaches. In this scenario an active agent must decide what actions to undertake based on a given policy. The basic idea utilises Bellmans equation [6, p. 652]

$$U(s) = R(s) + \gamma max_{\alpha} \sum_{s'} P(s'|s,a)U(s') \tag{1}$$

Profit maximisation is a topic which has been also researched in the field of AI. As an example [1] examine how efficient fare structures can influence the profit made by a given transport provider. The solution they suggest ”demonstrates that variable pricing mechanisms can significantly improve the efficiency of transport systems in general and rural transport in particular. I have undertaken similar to their approach in designing the experiment for this paper.Additionally, [2] describes a problem where an agents goal is to get the passenger and deliver it to the destination without running out of fuel and hence maximise the drivers profit per tank of fuel. He introduces a MaxQ function which decomposes the task into smaller ones and hence decomposes the reward given per subtask.The work on this paper has been inspired by his findings. Moreover, the proposed future research from the discussion is based on decomposition of tasks and utilisation of different RL policies.This is achieved by using hierarchical reinforcement learning [4]. In order to fully understand the concept of RL using policies, one must have a good understanding of what policies are themselves within the basic Bayesian reinforcement learning [6]. Such RL assumes a probability P(h) for each hypothesis h about what the true model is. P(h—e) is then obtained using Bayes rule given the observations made so far.Then, if the agent has decided to stop learning, the optimal policy is the one that gives the highest expected utility. If U_h^{π} is the expected utility, obtained by executing policy in model h, then

$$\pi^* = argmax_\pi \sum_\pi P(h|e) U_h^\pi \tag{2}$$

In the discussion this paper suggests further research in attempt to contribute in businesses and current economies. The reason for this is that profit maximisation influenced by past experience is a widely studied area in the field of Economics. It finds its application when it comes to price setting with rational expectations [5] where current cost settings are based on past experience. Similar concepts could also be inferred within the concept of price discrimination[8] and price differentiation [**?**]. Those techniques are usually known for charging different prices based on various conditions such as age of buyer, quantity of bought products and others. Those are also defined as constrains and externalities and are vital on consideration of applying the researched findings of this paper within other domains, as proposed further down.

3 Modelling the Environment

In order to better represent the environment for my research I use a non-scalable weighted graph. Denoting G = (N, E, W) with nodes N representing locations, E defining the edges E: N x N and W being the weight, also known as the time taken to travel between nodes. The shortest path between two nodes is denoted as $|n_i, n_j|, (n, j) \in N$.The model undertaken utilises a discrete model of time in which at each time a node has a probability of generating a passenger where this passenger wants to travel to another node j. [7] represents this quantity as $p_i^j \in [0,1]$. Additionally, passengers are defined by a demand function $df \in [0,1]$ which is used to return the probability that the passenger will accept a price offered by the taxi driver. While a universal df is currently being considered, in future work a family of such could be considered instead by depending on the starting and ending nodes.

It is also important to note that at any point in time a taxi T=(l,td,e) can either occupy a node (defining $l \in N$), or be driving between two points (in which case $l \in E$). Moreover, a taxi always has its destination point td and could also be either empty (e = T) or with a passenger (e = $\perp$).

Algorithm 1 [7]

while true **do**
  drive to the destination node(or wait if at destination node)
  if *taxiisoccupied* **then**
    drop off passenger (taxi is empty)
    pick a new destination node according to S
  else if *apassengerisatthisnode* **then**
    propose a price based on strategy S
    if *priceisaccepted* **then**

pick up passenger and set destination node to the passenger's destination node
end if
end if
end while

4 Research question

In comparison to the distance based strategy proposed by [7] a dynamic pricing strategy takes into consideration the probability of finding another customer when calculating a price to propose to a passenger. Intuitively, taking a passenger to a destination which has a lower chance of providing a new customer should cost more than one which guarantees acquisition of a new fare. According to the description provided in the previous section, a profit maximising agent should be able to increase its income over all trips that it has already had rather than taking into consideration only the current one. The research question as follows asks how can reinforcement learning utilising restriction-free policy be used towards maximising ones profit within the taxi domain and how good is it compared to other approaches and policies? In order to better justify the question I will refer to [7], who claims that the net income for a taxi (I) is defined by the expected income for the taxi to obtain when starting at a given node j, defining $x - c(i, j) + I_j$. In this scenario the agent thus seeks to find an x which maximises

$$EI(x,i,j) = df(x,|i,j|)\,(x - c(i,j) + I_j) + (1 - df(x,|i,j|))\,I_i \tag{3}$$

Examining the income for the taxi at a given node i, leads to the conclusion that it is solely dependent on the possibility of a passenger to appear at that node who wish to travel to another node j, or the taxi moving to another node k and obtaining income from there.

$$I_i = \gamma max\left(\sum(p_i^j argmax_k(I_k - c(i,k)))\right) \tag{4}$$

Referring to the reinforcement learning described by [6], denotes a discounting factor which needs to be set between zero and one. In the aforementioned case 1 needs to be excluded in order to prevent unbounding the equation above. The softwares aim is to reflect how much the agent values current income over future income [7]. Note that equation (Equation 4) must only consider neighbouring nodes when evaluating k. As a proof consider the following scenario: A taxi is at a given node i and wants to move to a given node j. In order to do as described the taxi passes through a third node k. A question to ask oneself is whether another taxi starting at node k would also wish to move to node j to maximise its profits or would wish to move elsewhere. In the former case the taxi would have to only consider the neighbouring nodes whereas in the latter one a taxis strategy becomes a lot more

complex. The aforementioned conclusion from this paragraph has been denoted and proved by [7] by defining the following theorem:

Theorem 1. *If c(i, j) is proportional to —i, j—, then for any set of nodes k_1, k_2... along the minimum cost path between i and j, if an empty taxi at node i would wish to move to node j, then an empty taxi at nodes k1, would wish to move to node j.*

The main hypothesis undertaken by my research is whether reinforcement learning using restriction-free policy would lead to a more profitable outcome for any taxi driver than any other policy adopted mechanism. In order to achieve this I use a finite MDP denoted by[7] in which df uses integers as parameters and assumes that there is a maximum cost to the demand function for any distance such that the probability of acceptance by a customer would be 0 for that price or greater.

5 Software Overview

[7] define a graph with a finite number of nodes. They consider two types of actions move actions, and bid and move actions. A move action allows a taxi to move from node i to node j or stay at the same node without moving. Whereas a bid and move action consists of proposing a bid to potential customers and if they accept move to the destination required and drop them off there. With regards to size, in the worst case, there are $O(n^2 + b^2)$ states in the MDP, where n is the number of nodes and b is the number of possible bids [7]. The worst case occurs when the graph is fully connected. I can now use a standard finite MDP solving technique to identify an optimal strategy for a taxi given a specific road network. It is important to note that the approach described above cannot be used for large road networks and therefore a different solution and further research has been proposed further in the discussion section.

6 Experimental Design

In evaluating the contributions of my approach, I test the following hypothesis:

Hypothesis: Utilising reinforcement learning with restriction-free policy can significantly improve a taxi drivers income in a non-large world.

There are four null hypotheses. The first one compares π_1 with the distance based strategy (str). There are four null hypotheses. The first one compares π_1 with the distance based strategy (str). Here I formulate that $\mu_{\pi_1} = \mu_{str}$ or alternatively $\mu_{\pi_1} = 0$. The alternative hypothesis I accept is that $\mu_{\pi_1} < \mu_str$ and hence $\mu_{\pi_1} < 0$. The other three hypotheses compare π_1 with the other three policies as follows: $H_0 : \mu_{\pi_1} = \mu_{\pi_{2,3,4}}; H_1 : \mu_{\pi_1} < \mu_{\pi_{2,3,4}}$.

In order to analyse the benefits of using reinforcement learning with restriction-free policy and better encapsulate the conditions under which those benefits are valid, I apply four different policies to reinforcement learning and review their performance with regards to profit making. Policy one, denoted π_1, represents the restriction-free policy, where the taxi has a complete freedom for its behaviour. Policy 2, π_2, defines such where the taxi bids the optimal value only, however it does not have to always

bid. π_3 denotes a taxi which bids the optimal value only and has to bid if possible. Finally, π_4 a taxi bids the optimal value only, has to bid if possible and stays where it is if no one is around, whereas the other three policies enable the taxis to decide whether going to another destination will or will not increase their profit. I apply different treatments to the group of units described above in order to define a taxi within the aforementioned world and apply the four policies in comparison. The closer to zero my results will be the less extra profitable the approach is. In order to simplify the computation, I assume that the world described has nodes $n = 15$, known as destination hubs, and each node will have edges $e \leq 4$, defining that each hub will have at most 4 neighbours. The edge weight, that is distance between two nodes stays within the following boundaries: $w \in [0, 100]$.Additionally, I am assuming that the base cost is solely dependent on the distance between the source and the target, and the probability to generate a passenger p is based on the random function implemented in Ruby. This scenario defines a non-large network which provides an opportunity for testing the differences between using the four polices and will highlight the optimal conditions to obtain maximum profit. The experiment will allow me to evaluate the benefits of using π_1 and also to specify the best environment for obtaining optimal profit by doing so. I conduct 100 runs of the experiment, with 20 different modifications of the weights and 5 repetitions for each of them. The experiment consists of the following: I will first overview the profit changes in the system when the substitution of the maximum distance with the minimum distance is constant and equal to 10, or $maxDist - minDist = 10$. I will then set the *minDist* to a constant value equal to 1 and observe the profit behaviour when modifying the *maxDist*. All observations will be made with regards to the null and alternative hypotheses. This will then lead to a discussion and suggestions for further research.

7 Results

Referring to the first null hypothesis and after obtaining the sample data by setting a medium distance- with $minDist = 5$ and $maxDist = 15$, I calculate the mean $\overline{x_A} = 108278$ and assume that H_0 is *true*. I then estimate the distribution of mean scores for 10 possible samples where $maxDist - minDist = 10$ for π_1 which leads to $\mu_{\pi_1} = 71251.25$. Considering that the destination based strategy is always equal to 0, I am able to reject the null hypothesis in favour of π_1. Considering that the closer to 100 *minDist* goes the greater the chance of obtaining a negative result by using reinforcement learning with restriction-free policy is. (Figure 3) shows that when $minDist \approx 90$ there is about 40% chance for the result to become negative- this has been estimated when 4 out of 10 runs produced negative outcome.

This will denote that for less than 10% of the possible scenarios within the range of $w \in [1, 100]$ there is about 40% chance for them to be negative- summing up to $< 5\%$ in overall. This proves that utilising reinforcement learning with restriction-free policy only guarantees profit if the distances are not too large. Therefore, other approaches need to be considered for larger distances.

Observing the results from (Figure 1) and (Figure 2) I deduced that the group of the three polices- π_2

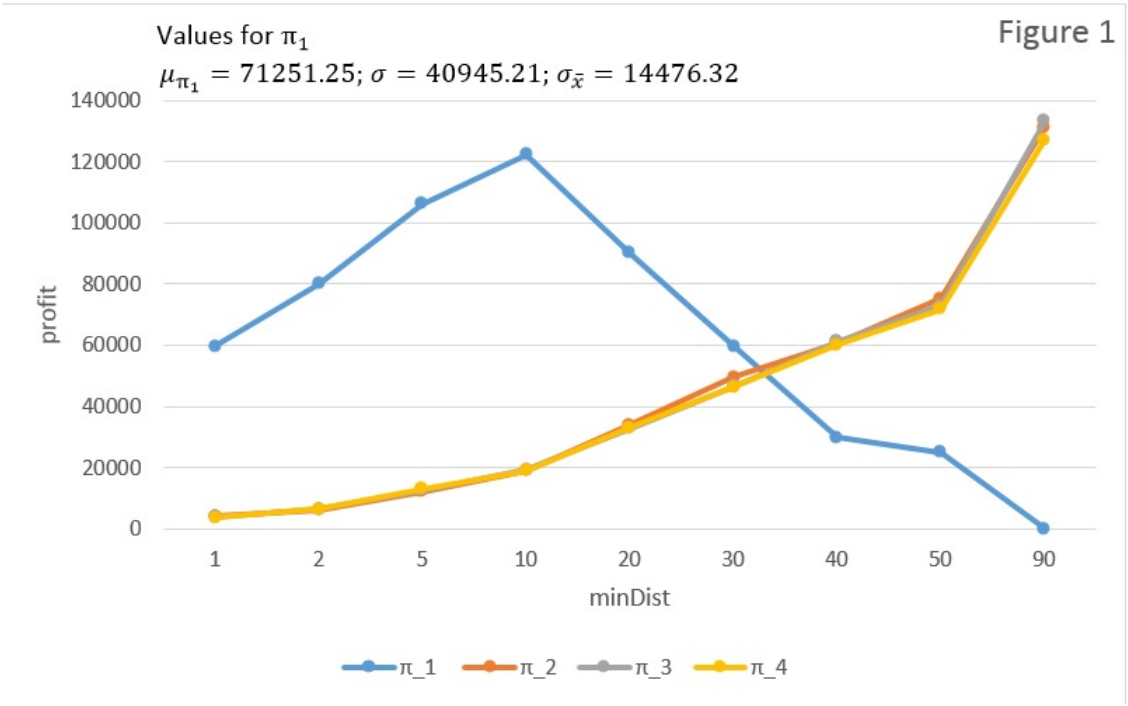

Figure 1: Results when the difference between distances is constant and is equal to 10, while the distances themselves change

to π_4 have similar results among each other and they all differ π_1 with alike ratios. It can also be seen, in both figures that after a certain distance the profit from restriction-free policy starts decreasing until it reaches 0 while the profit from the other three policies starts increasing.

Tests of significance were applied to the results and was found out that the results were statistically significant by t-test with $p < 0.05(5\%)$. Overall, the profit was considerably higher in a world which consists of neighbours with a smaller distance between each other. Moreover, it can be seen that worlds which consist of neighbours with distances from 1 to 30 offer the best environment for the approach defined. Additionally, a world can provide a profit increase when using bigger range of possible distance between neighbours (e.g. 1 to 30 instead of 20 to 30 or 10 to 20). It can be then concluded that the larger the range between neighbours is, as long as the bigger proportion of the neighbours are within smaller distance, the profit will be increasing rather than decreasing towards 0. Therefore, the optimal strategy is to use restriction-free policy with RL in non-large networks which allow at most 4 neighbours at a distance between 1 and 30 from each other.

8 Discussion

The results from my experiment prove that restriction-free policy can increase profit, however it can only be used in smaller worlds. The behaviour of the other three policies provide better performance for larger distances. Therefore, they might be considered as replacement for the

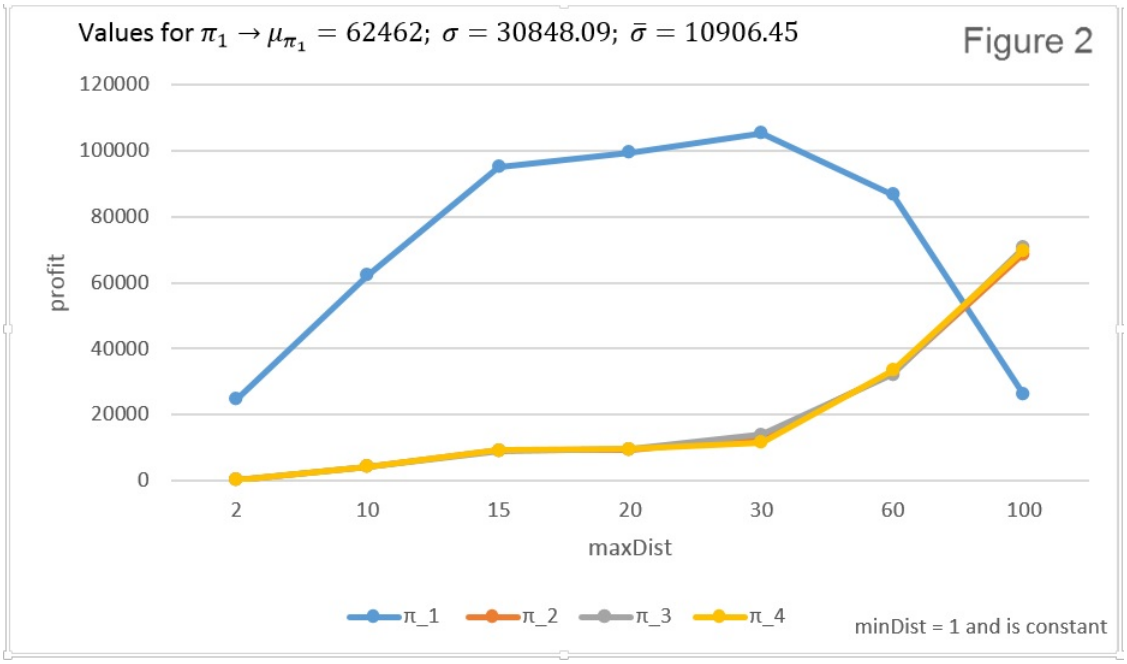

Figure 2: Results when minDist is set to 1 and maxDist changes

restriction-free policy for worlds and situations which have larger distances between the neighbours. Although it is highly dependent on the policies used, I have learned that reinforcement learning might lead to profit increase in many situations.

The technique researched within this paper can contribute to considerably large profit increase within any domain. However, I noticed a correlation between the variables, that is depending on the number of nodes in the world and the ratio to the number of neighbours per node, the ranges might vary accordingly. Regardless, the findings of my experiment are applicable to any scenario and will always lead to the specified causations.

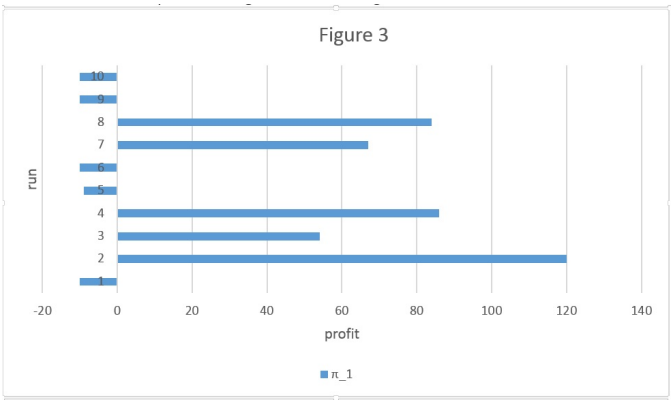

Figure 3: Results when minDist=90 and maxDist=100

Although RL with restriction-free policy enables profit increase within smaller bounds and other policies enable increases within other specifications, a possible future research question might try and combine the benefits from using different policies. A possible approach would be the use of machine learning and utilising profit increase techniques by designing a computational pattern by decomposing

the whole task into smaller subtasks. A further experiment that might be of benefit would be evaluation of the time and processing power consumption of the application of the algorithm. It will most certainly be useful in locating machine learning techniques for generalising the solution and enhancing the performance. Regardless, this technique can benefit and improve the economy of every business and country. In order to utilise the technique of applying RL to profit maximisationto in any economy or business, one should take into consideration the externalities and constraints that might influence the outcomes and try and define them within the bounds of the algorithm. My research proved me that this will most probably lead to difficulties in describing certain variables for the algorithm to act accordingly.

The algorithm provides an evolutionary approach towards using profit maximisation and cost fixation within any potential business and economy.

9 Conclusion

In this paper I have proved that profit can be maximised by applying reinforcement learning while using a restriction-free policy within the taxi domain. I have observed and evaluated the behaviour and variations of the profit made by utilising reinforcement learning. Additionally, I compared it with the distance based strategy as well as with three other policies which included some restrictions. My experiment proved the hypothesis assumed by me while researching the topic.

Acknowledgements

The research undertaken as well as the experiments conducted have been suggested and inspired by Dr. Nir Oren, a Senior Lecturer at The University of Aberdeen.

References

[1] Cheng Zeng C. David Emele, Nir Oren. Agent-driven variable pricing in exible rural transport. http://homepages.abdn.ac.uk/n.oren/pages/publications/emele13agent.pdf.

[2] Thomas G. Dietterich. Hierarchical reinforcement learning with the maxq value. http://www.jair.org/media/639/live-639-1834-jair.pdf.

[3] Yong Duan. Application of reinforcement learning in robot soccer. http://www.sciencedirect.com/science/article/pii/S095219760700005X, 2007.

[4] Mausam. Hrl presentation.

[5] Ondra Kamenik Michael Juillard. Optimal price setting and inflation inertia in a rational expectations model. *Journal of Economic Dynamics and Control*, pages 2584–2621, 2008.

[6] P. Norvig and S.Russell. *Artificial Intelligence a Modern Approach*. Pearson, 3 edition, 2010.

[7] N. Oren and C. Zeng. Dynamic taxi pricing.

[8] Tejvan Pettinger. Airline price discrimination, September 2013.

[9] Adam Coates Pieter Abbeel. An application of reinforcement learning to aerobatic helicopter flight. http://heli.stanford.edu/papers/nips06-aerobatichelicopter.pdf, September 2007.

[10] Vignir Hafsteinsson Yngvi Bjrnsson. Efficient use of reinforcement learning in a computer game. http://www.ru.is/~yngvi/pdf/BjornssonHJJ04.pdf.