Non-Rational Discrete Choice Based On Q-Learning And The Prospect Theory

Gustavo Kuhn Andriotti
University of Würzburg
Department of Artificial Intelligence (Lehrstuhl VI)
Am Hubland, D-97074 Würzburg
Germany/Detuschland
andriotti@informatik.uni-wuerzburg.de

ABSTRACT

When modelling human discrete choice the standard approach is to adopt the rational model. This has been shown, however, to fail systematically under some conditions, which makes evident the need for a better approach. The choice model is however only part of the problem because it does not say how to deal with uncertainty, where learning is necessary. In this regard, some evidences support the claim that the brain employs a reinforcement learning strategy and the Q-Learning is a model that copes with it. The Q-Learning, nevertheless, is based on the rational choice defined by Bernoulli principle and von Neumann's axioms. Then to cope with a non-rational learning algorithm, for modelling human discrete choice, a novel modification to the Q-Learning algorithm is here presented.

1. INTRODUCTION

In lay terms, the discrete choice problem is to choose an option from a set of those that, to the individual, the best is. According to the Utility Theory [20] (UT) it is possible to model such human process by assigning a numerical value to each of those options. Then, if those values are correctly assigned, the observed preferences (in the human subjects) are represented by the numerical relations established by those values.

One of the firsts tries to model this behaviour has been proposed by Bernoulli [11, 12], which says that humans seek to maximise the accumulated total wealth. Based on this principle von Neumann and Morgenstern [40, 41] proposed a formal model, called Expected Utility Theory (EUT), and since then it is widely accepted to be the definition of rationality. Besides being the model for rationality it is also accepted as the model for the human discrete choice behaviour, where the criticisms arise. Among the critics, Simon [37, 38] became the most notorious but McFadden [29] and Kahneman [22] have expressed their concern with the rational model as well.

Moreover, not recently the rational assumption had been demonstrated wrong by Allais [1, 2] and Ellsberg [19], to cite only the most well-known. For this reason the Allais paradox (after Allais [1, 2]) is the focus in this text – formally presented in Sec. 5.3. This paradox shows that the rational assumption systematically fails to reproduce the human preferences over a simple discrete choice problem. Fortunately a mathematical model has been developed by Kahneman and Tversky [24] to cope with the Allais paradox, it is called the Prospect Theory (PT) – presented in Sec. 5.

The PT, however, does not cope with uncertainty, i.e., environments where the exactly nature of the choices is unknown. Examples of such uncertain environments are stock markets and route choice. One approach to model such environments is to suppose a Markovian Decision Process [27, 8, 7] (MDP), where the agents are neither aware of the transition nor the reward functions. This way, for the agent, the reward for each state is uncertain and to tackle this issue reinforcement learning is an alternative. Among the several algorithms the $\mathcal{Q}$-Learning [45, 46] is the algorithm adopted here.

The problem with the $\mathcal{Q}$-Learning is that it is based on the rational principle, i.e., it is a rational model. This leads to the theme of this text: modifying the $\mathcal{Q}$-Learning algorithm to reproduce the human non-rational behaviour using the PT as its basis (instead of the EUT). It is important to make it clear that it is not a general purpose decision model but a model to be used when modelling human decision making for discrete choice problems. This means that this modified $\mathcal{Q}$-Learning algorithm does only apply for modelling human agents and in particular for MDP environments.

2. STATE-OF-THE-ART

Rational models have been developed since von Neumann [40, 41] in several directions and in the microeconomics are the most impressive developments. The EUT is a too restricted model because it demands from the individuals perfect knowledge and unrestricted computational power. Then a new class of models, the Random Utility Model [26] family (RUM), emerges to cope with the perfect knowledge, allowing a partial awareness of the options. This development led to several other models that have, as common feature, a structure to explicitly model the correlation among the options. Examples of such models are: the Multi Nominal Logit [28], with no correlation; Probit [13, 14]; Nested-Logit [9]; Cross-

Nested-Logit [42]; Path Size Logit [10]; and Mixed Logit [30].

These are, nevertheless, macroscopic and analytical models that, as far as the knowledge of the author goes, have not counterparts in the Artificial Intelligence (AI). In AI much more effort has been put into developing the learning algorithms. Which, put in other words, is to say that besides the common ground, microeconomics and AI have grown apart. For the learning mechanisms several options are available but the focus here is in the reinforcement learning algorithms.

The adoption of a reinforcement learning algorithm comes from the fact that biological evidences support such assumption. In the experiments reported in [33] it is shown that when rats were investigated about learning behaviour it was observed a positive reinforcement process based on dopamine. It was also observed a correlation between learning performance and dopamine stimulation, i.e., reinforcement on the neuronal activity, accelerating the learning process. In [32] the human brain activity was also investigated for learning processes. It was observed that in the beginning of the learning process the medial temporal lobe (MTL) activity is low, but as the learning process is intensified the activity becomes higher for the MTL. This indicates reinforcement based learning process.

Moreover, the $\mathcal{Q}$-Learning algorithm is a well known and studied method with several extensions and applications. Among them is the work made by Littman and colleges [25, 31, 15].

For the adoption of the PT as the model for non-rationality (therefore, the adopted theory to model human decision process) is because it is similar to the EUT model. Second, it reproduces the results observed in the Allais paradox. Another reason is in the evidences shown in [18, 23] that support the PT as a valid hypothesis for the human decision behaviour. Moreover, after the original model has been proposed [24], some improvements were developed. Among them are the Cumulative Prospect Theory [39, 43], the Prospect Theory for continuous distributions [34], and the Continuous Cumulative Prospect Theory [16].

For the other non-rational models some issues with them made the PT more appropriated. One of such models is the Fast and Frugal Way [21]. This model is based in a hierarchical decision process. It is assumed that the decision problem can be modelled by establishing a sequence of binary criteria (with *yes/no* answers). The problem with this theory is that it is only possible to have binary choices and only binary criteria are allowed (no numerical comparison is considered). Another model was also taken into account: the Small Feedbacks [5], but it presented some drawbacks – it was behavioural unstable, which were reported to the authors [4].

3. THE DISCRETE CHOICE PROBLEM

This work is focused in the single-person discrete choice problem. It is, according to von Neumann [40, 41], formulated as: to choose the best lottery (option) $\mathbf{x}$ in a lottery set (choice set) $\mathbf{X}$. A lottery is an option that has some utility associated with it, i.e., an option can be associated with a numerical value (called utility). The modelling problem is to find a function $u : \mathbf{X} \mapsto \mathbb{R}$ that reproduces the observed preferences, i.e., if $\mathbf{x} \succ \mathbf{y}$ then $u(\mathbf{x}) > u(\mathbf{y})$. This means that the utility value only says if an option is better or worse than another. When this function is found the problem can be modelled and analysed.

A lottery can be simple, i.e., the expected outcome of that choice is deterministic, for instance a 100U$ bill has always a utility equivalent to 100U$. A mixed lottery, on the other hand, has an associated stochastic element, e.g., a stock option, whose outcome may vary. This type of lottery is expressed as a combination of simple lotteries in a probability distribution function fashion. For now on, when lottery $\mathbf{x}$ is mentioned it refers to a list of pairs $\langle x_i, p_i \rangle$, where $x_i \in \mathbb{R}$ is the ith outcome and $p_i \in (0, 1]$ is its corresponding probability ($\sum_{i \in \mathbf{x}} p_i = 1$). In a lottery no two pairs have the same outcome, i.e., $\mathbf{x} = \{\forall i \in \mathbf{x}, \nexists j \in \mathbf{x} \mid x_i = x_j \wedge i \neq j\}$. The utility for its turn must associate each lottery to single numerical values that is used to reproduce the observed lottery/option ranking.

4. RATIONALITY

Rationality here refers to the axioms defined by von Neumann and Morgenstern [40, 41], which is also known as the Expected Utility Theory (EUT) and its four axioms are: Completeness, Transitivity, Convexity, and Independence. For space reasons only the Independence axiom is discussed (for further information please refer to [41]). The Independence axiom, in lay terms, says that if two lotteries are equivalent (a person expresses no preference regarding them, $\mathbf{x} \sim \mathbf{y}$) then they can be combined using any probability $p \in [0, 1]$: $p\mathbf{x} + (1 - p)\mathbf{y} = (1 - p)\mathbf{x} + p\mathbf{y}$. A particular derived statement is important here: the region (in the probability interval) where the combination occurs does not interfere in the person's preference. This is exactly the point where the PT "breaks" with the rational model (EUT).

After von Neumann, simple lotteries can be combined (associating a probability to each one) and the resulting utility is: $u(\mathbf{x}) = \sum_{i \in \mathbf{x}} x_i p_i$. (Hereafter, the EUT utility function is referred as $eut(\bullet)$.) A strong restriction of the original model is that the option set is totally ordered (from the Completeness and Transitivity axioms), i.e., $\mathbf{X} = \{\forall \mathbf{x}, \mathbf{y} \in \mathbf{X} | \mathbf{x} \sim \mathbf{y} \oplus \mathbf{x} \succ \mathbf{y} \oplus \mathbf{x} \prec \mathbf{y}\}$.

5. PROSPECT THEORY

The name Prospect Theory [24] (PT) comes from calling lotteries prospects, differentiating from the EUT. Prospects differ from lotteries by introducing a reference point (called *status quo*) that says which outcomes are gains and which are losses. This new theory is an answer to the Allais paradox [2, 1] that cannot be explained/modelled using rationality. The Allais paradox, for its turns, exploits the Independence axiom to show that the rational model does not apply to human decision makers. This is not the only paradox that violates the rational model, another one is the Ellsberg paradox [19]. But it is not covered by the PT and considered out of the scope of this discussion.

5.1 Prospect Theory: Overview

In this short section the fundamentals underlying the PT are presented, emphasising the differences between it and the

EUT. The PT introduces some elements that are not presented in the EUT and they are the *status quo*, perception distortion, and lottery editing. The first represents the point where the outcomes are classified into gains and losses and its value can have different meanings. Among such meanings are the current wealth of the individual (her/his bank account balance) or the expected outcome (such as expecting to receive an increase of 5% in some stock options and any outcome below it is considered loss and above gain).

The second element (perception distortion) depends on the *status quo* and two types of distortion are present: outcome and probability perception distortion. It is said that in average a person tends to be conservative in gains, i.e., less sensitive to gains and risky in losses, i.e., more sensitive to the outcome variations. For this reason the outcome perception distortion function differentiates between gains and losses. This difference is represented by having the derivative in the losses higher than in the gains. (It can be visually appreciated in Fig. 1.)

For the probability perception function the behaviour is similar in gains and losses, what changes is in how accentuated are the characteristics of the function (more acute in the gains than in the losses). This function maps the following characteristics. It was observed that people perceive probabilities in a distorted way, when confronted with its numerical value. The first distortion is the hope when facing low probabilities, i.e., despite of low probability values the individuals behave as if the probability is actually higher (the value is overcompensated with "hope"). The second phenomenon is towards high probabilities where a person tends to perceive the value lower than it actually is, i.e., the individuals show prejudice against the real value (they under compensate with "scepticism/prejudice"). This means that the probability perception function returns a value higher than the actual probability if the value is low and lower if the probability is high. (Fig. 2 depicts these characteristics.)

The last element is lottery editing. This concept was not formalised in the original model and to this day lacks a standard definition. Nevertheless the guide-lines are informally stated as to group similar outcomes aggregating their probabilities. This means that the outcome domain must be analysed beforehand to establish what can be considered similar and then apply a suitable "editing" algorithm. For these reasons a lottery must first be edited (becoming a prospect) before it is suitable for an evaluation by the PT. The used editing algorithm is presented in Sec. 6.1 when presenting the implemented $\mathcal{Q}$-Learning. The formal presentation of the PT is made in the next section.

5.2 Prospect Theory: Formalisation

The PT utility function is shown in Eq. 1 where the difference to the EUT is evident. In EUT the utility function is $eut(\mathbf{x}) = \sum_{i\in\mathbf{x}} x_i p_i$ and in PT $pt(\mathbf{x}) = \sum_{i\in\mathbf{x}} v(x_i)\pi(p_i)$. This means that the distortion functions are explicit in the utility function – $v(x_i)$ for the outcome and $\pi(p_i)$ for the probability distortion functions. (For simplicity it is assumed that x_i is already adjusted according to the *status quo*. If that is not case it would be: $v(x_i - sq)$, where sq is the *status quo* value.) These functions "distort" the real values to map the features discussed in the previous section. The formal definition of these functions are in Eq. 2 and 3 and to have a visual impression of the functions' behaviour please refer to Fig. 1 and 2.

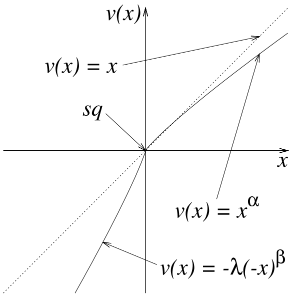

Figure 1: $v(x)$ with the original parameters

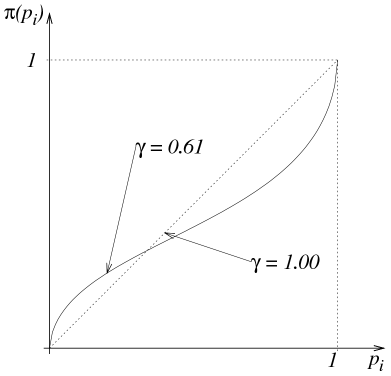

Figure 2: $\pi(p)$ with different γ values

The parameters were calibrate (in [24]) for a modified version of the Allais paradox (Tab. 1 from Sec. 5.3). In the Eq. 2 the parameters' values are: $\alpha = \beta = 0.88$ and $\lambda = 2.25$. For the Eq. 3 the values are $\gamma^+ = 0.61$ and $\gamma^- = 0.69$.

$$pt(\mathbf{x}) = \sum_{i\in\mathbf{x}} v(x_i)\pi(p_i) \tag{1}$$

$$v(x_i) = \begin{cases} x_i^\alpha & x_i \geq 0 \\ -\lambda(-x_i)^\beta & x_i < 0 \end{cases} \tag{2}$$

$$\pi(p_i) = \frac{p_i^\gamma}{(p_i^\gamma + (1-p_i)^\gamma)^{1/\gamma}} \begin{cases} \gamma = \gamma^+ & x_i \geq 0 \\ \gamma = \gamma^- & x_i < 0 \end{cases} \tag{3}$$

The function in Eq. 2, as explained, just expresses a higher sensibility in the losses than in the gains ($\frac{d}{dx}v(x_i^+) < \frac{d}{dx}v(x_i^-)$, where x_i^+ represents the positive values and x_i^- the negative). Remember that it can be problematic to establish which value must the *status quo* must assume. It could be 0, the current wealth, or the expected outcome (taking some obvious candidates).

The probability distortion function $\pi(\bullet)$, in Eq. 3, is characterised by its inverted "S" shape (for $\gamma < 1$). This shape overestimates the low and underestimates the high probabilities; it, as explained before, says that individuals tend to compensate the low probabilities with "hope" and the

Table 1: Modified Allais problem

Opt.	Outcome	Prob.	$eut(\bullet)$	$pt(\bullet)$	Pref. (in %)
C	2500	0.33	825	326.7	83
	0	0.67			
D	2400	0.34	816	320.1	17
	0	0.66			
A	2500	0.33	2409	806.5	18
	2400	0.66			
	0	0.01			
B	2400	1.0	2400	943.2	82

high with "prejudice" (some external factor might interfere). These features are better seen with an example, which is the subject of the next section.

5.3 The Allais Paradox

The Allais paradox is rather well-known in economics and shows that people systematically violate the rational behaviour. But instead of presenting the original version of it the modified version is presented – used for the results in [24] and for which the PT was calibrated.

The decision problem is to choose first between two options/prospects (A and B) and then repeat the process for other two options (C and D). In Tab. 1, extracted from [24], the problem is reproduced with the corresponding results and analysis.

The problem is so presented: first choose between A and B and then between C and D. To the participant only the columns "Outcome" and "Prob." were presented (representing the monetary outcomes and the corresponding probabilities). The survey results can be seen in the column "Pref. (in %)" (the preferences in percentage). If the utility values given in the "$eut(\bullet)$" are compared with the actual preference it can be seen that it corresponds to the preferences for C and D, i.e., $C \succ D \Rightarrow eut(C) > eut(D)$. But looking at A and B it is observed that $A \prec B \not\Rightarrow eut(A) < eut(B)$, which is not right, i.e., it invalidates u as a valid utility function – it does not reproduce the actual preferences. A solution is to make monetary outcomes to have a negative utility, which first does not make any sense and second transfers the problem to the choice between C and D. On the other hand, if the "$pt(\bullet)$" column in Tab. 1 (the PT evaluation) is observed it yields all preferences: $A \prec B \Rightarrow pt(A) < pt(B)$ and $C \succ D \Rightarrow pt(C) > pt(D)$.

6. Q-LEARNING BASED IN THE PT

The $\mathcal{Q}$-Learning algorithm optimises the total discounted accumulated reward, the rational utility in the original model [45, 46]. To better explain how the modified version works, the original formulation is presented in Eq. 5, where $V_n(s)$ returns an action a for interaction n and state s and this action tends to be the best choice, based on the accumulated knowledge about the received rewards (stored in the table Q from Eq. 4). In Eq. 4, r_n is the immediate reward (received in interaction n), γ a discount factor, and α_n the decreasing learning factor.

$$\begin{aligned} Q_n(s,a) &= (1-\alpha_n)Q_{n-1}(s,a) \qquad (4)\\ &\quad +\alpha_n\left[r_n + \gamma V_{n-1}(s^{'})\right]\\ V_n(s) &\equiv \operatorname*{argmax}_{a\in\mathcal{A}}\left[Q_n(s,a)\right] \qquad (5)\end{aligned}$$

It is shown, but here omitted, that Q converges to the mathematical expectation of the reward and this is equivalent to $eut(\bullet)$ value. This means that the original $\mathcal{Q}$-Learning algorithm is rational and to change it to be non-rational (based in the PT) means to have Q converging to the $pt(\bullet)$ instead of the $eut(\bullet)$ value.

A naive implementation would be to keep a reward frequency table (assuming that the reward is a discrete probability distribution). This is impractical (linear complexity growth in time and space) and also implies that having this table makes the $\mathcal{Q}$-Learning algorithm unnecessary (since the $eut(\bullet)$ or $pt(\bullet)$ value can be direct calculated and no learning is necessary). Another problem of this approach is that it still does not include the editing phase (necessary for the $pt(\bullet)$ function).

6.1 Editing Phase

To cope with the problem of complexity growth, the editing phase comes handy. The main idea is to aggregate similar outcomes into a single value that receives a representing probability. Then inspired by [17], the following method is used. An aggregation threshold ϵ is given so that rewards closer than ϵ to each other must edited (aggregated). This implies that the reward domain is more or less known, i.e., the aggregation threshold is known or given. (It is known that this means that to the agent an advantage is given, because it is not "blind" but sees some shades of "grey" from the reward function.) An example would be: if the reward is the gain per share in the stock market then using quarters of dollars as the threshold might suit it. This example illustrates that the agents have some idea of the reward range (because it must be fairly divided) and also that this method depends on the problem domain.

Having the threshold ϵ is not enough to solve the problem. Because holding all outcomes and their probabilities is not an option (discussed previously) it is necessary to build the prospect in an "on-demand" way. The method used to edit the lottery was a centroid based clustering algorithm that builds the clusters as they appear.

The algorithm is basically formed by a set of centroid-and-counter pairs $\mathbf{C}$. This set is defined by $\mathbf{C} = \{\mathbf{c} \in \mathbf{C} \mid \mathbf{c} = \langle c, a\rangle \wedge c \in \mathbb{R} \wedge a \in \mathbb{N}^*\}$, where c is the centroid value and a the counter/accumulator. Moreover, $\mathbf{C} = \{\forall\ \mathbf{c} \in \mathbf{C}, \nexists\ \mathbf{d} \in \mathbf{C} \mid c = d\}$, i.e., no pair in $\mathbf{C}$ may share the same centroid value. The complete, non-optimised, algorithm is in Algo. 1.

Algorithm 1: Clustering algorithm: *cluster*()

```
Data: ⟨centroid, accumulator⟩ set $\mathbf{C}$, the threshold $\epsilon$, and
      the reward $r$ to be included
Result: Updated set $\mathbf{C}$
$\mathbf{c}_{min} \leftarrow NIL$ ; /* $\mathbf{c}_{min}$ is the closest centroid to $r$ */
$\Delta d \leftarrow \infty$ ; /* $\Delta d$ is the distance between $r$ and $\mathbf{c}_{min}$ */
/* finds the closest centroid to the reward $r$ */
forall pair $\boldsymbol{c} \in \boldsymbol{C}$ do
    if distance($c$, $r$) $< \Delta d$ then
        $\Delta d \leftarrow$ distance($c$, $r$);
        $\mathbf{c}_{min} \leftarrow \mathbf{c}$;
    end
end
/* checks if the centroid is suitable for
   aggregation */
if $\Delta d \leq \epsilon$ then
    $c_{\mathbf{c}_{min}} \leftarrow \frac{r + a_{\mathbf{c}_{min}} c_{\mathbf{c}_{min}}}{1 + a_{\mathbf{c}_{min}}}$ ; /* new centroid value */
    $a_{\mathbf{c}_{min}} \leftarrow a_{\mathbf{c}_{min}} + 1$ ; /* increments
    counter/accumulator */
else/* it is a new centroid */
    $\mathbf{C} \leftarrow \langle r, 1 \rangle$ ; /* adds the new centroid */
end
```

This algorithm avoids linear complexity growth for any reward function (assuming the function's codomain being a limited interval in $\mathbb{R}$) and also edits the prospect (assuming that the threshold ϵ was properly chosen). Furthermore, this algorithm is rather simple and its complexity independent from the amount of rewards (learning horizon), the proof is omitted but informally it is: the clustering complexity depends on the amount of clusters. Then since the amount of clusters is constant (because of the fixed ϵ and limited reward codomain) its complexity is too.

6.2 Modified Q-Learning

Assuming that $\mathbf{C}_n \leftarrow cluster(\mathbf{C}_{n-1}, r_n)$ represents the centroid set at step n and the function $cluster(\bullet)$ is the Algo. 1, then the modified $\mathcal{Q}$-Learning algorithm is given by Eq. 6, 7, 8, and 9; where $v(\bullet)$ and $\pi(\bullet)$ are the same as in Eq. 2 and 3. It is worth noticing that the $\mathcal{Q}$-Learning suffers no complexity increase. This is so because the only modifications are the inclusion of the set $\mathbf{C}$ and the $pt(\bullet)$ function. Since the $pt(\bullet)$ function depends on the size of $\mathbf{C}$ and function $cluster(\bullet)$ (which also depends on the size of $\mathbf{C}$). Then the complexity is independent from the learning horizon.

$$
\begin{aligned}
pt(\mathbf{C}) &= \sum_{\mathbf{c} \in \mathbf{C}} v(c_{\mathbf{c}}) \pi \left(\frac{a_{\mathbf{c}}}{\sum_{\mathbf{c} \in \mathbf{C}} a_{\mathbf{c}}} \right) & (6) \\
\mathbf{C}_n &= cluster(\mathbf{C}_{n-1}, r_n) & (7) \\
Q_n(s,a) &= (1 - \alpha_n) Q_{n-1}(s,a) & (8) \\
&\quad + \alpha_n \left[pt(\mathbf{C}_n) + \gamma V_{n-1}(s^{'}) \right] & \\
V_n(s) &\equiv \operatorname*{argmax}_{a \in \mathcal{A}} \left[Q_n(s,a) \right] & (9)
\end{aligned}
$$

7. RESTRICTIONS TO THE USE OF THE PT

In this section the "worthiness" of using the PT is discussed, more specifically the PT modified version of $\mathcal{Q}$-Learning. The PT has some particularities concerning its deviation from rationality. The first of them is that if all outcomes have sufficiently close probabilities (or frequencies) the result is the same as in the EUT. Second, the *status quo* can be tricky to establish (commented previously), specially if it is not known how people really establishes it for the modelled problem domain. A last criticism is concerning the calibration parameters. The original values are set for monetary lotteries in the specific scenario of the modified Allais paradox, in [24]. This means that they should be calibrated and checked for scenarios that deviate from its original scenario.

These criticisms, as already said, concern the cases where the PT deviates from the EUT and in the worst case it reduces to the EUT. That is it, if the issues mentioned in the previous paragraph are not observed the PT model might turn out to be a "fancy" EUT. This leaves the question: "when the PT is needed?"

8. REASONS TO USE THE PT

The first reason to use the PT is to model human decision processes involving prospects (the options have a stochastic nature). But only the stochastic nature of the outcome does not fulfil all requirements. After editing the prospects, whenever the proposed method is used or not, the outcomes must end up having distinct probabilities/frequencies. Being more specific, some of these probabilities must be located on at least one of the "bumps" of the $\pi(\bullet)$ function (see Fig. 2). If this does not occur then no ranking (given by the utility function) will deviate from rationality and the PT degenerates to an EUT model ($pt(\bullet) \sim eut(\bullet)$).

Regarding the outcomes rather than the probabilities, the *status quo* must be self-evident and shall not imply in having outcome mixture (positive and negative outcomes assuming the *status quo* as the reference). Another issue is to avoid outcomes near the *status quo*, whose behaviour is not well-known (regarding the human behaviour). But these concerns are also true for the EUT (they are just grossly ignored). Regarding the ranking, the function $v(\bullet)$ (in Eq. 2) does not account for the major deviation from EUT, i.e., if the model does not include the $\pi(\bullet)$ function it will behave exactly as an EUT model (even though the numerical utility values are different) because what the $v(\bullet)$ function does is more or less as an Affine transformation, not a "distortion". (It is not in consideration the case of outcome mixture, i.e., options with positive and negative outcomes.)

The main concern then is the distribution of the reward function r that must have the "features" described on the first paragraph of this section. If that is not present then the PT is not needed.

9. EXPERIMENTS AND RESULTS

To verify the relevance of such theory for MASim, a traffic scenario inspired in the El Farol Bar Problem [3] is adopted. This scenario is a minority game instance, where, given two routes one has a lower capacity than the other, i.e., the

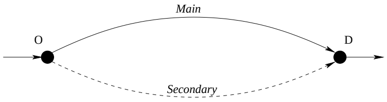

Figure 3: El Farol traffic scenario

option is only attractive if a minority of the agents choose that option. The traffic scenario used is depicted in Fig. 3. There two routes are offered to go from O (origin) to D (destination). These routes are named *Main* and *Secondary* where the latter has half the capacity of the first (meaning that *Main* as twice the amount of lanes as *Secondary*).

In this scenario the optimal split is 1/3 for the *Secondary* and 2/3 for the *Main* route, i.e., with this distribution the User Equilibrium [44] (UE) is reached. A similar scenario is also used in [6]. The objective of the experiments were to verify if and when the PT diverges from the rational choice, i.e., under which conditions the PT becomes a worth using model for human decision. If the PT does not diverge from the EUT then it is not relevant for the investigation of traffic assignment models. But that is not what the results shown.

The experiments done were to verify, first, if the PT behaviour is stable across different simulation horizons and agent amount. (Each experiment was repeated 100 times and the results aggregated into the mean and standard deviation.) For this particular case (agent amount), it is stable for an agent population of 50 or higher. In the horizon experiment it shows that the behavioural stability is reached with 300 or more iterations (for this scenario). The third experiment was to verify if the PT deviates from rationality, i.e., when the PT is relevant.

Before showing the results it is necessary to explain how the travel-time (the metric for the utility function) is calculated. In each iteration all agents are asked about their decisions and then each route is "burden" with the corresponding amount of agents. After the occupation for each route is determined, the travel-time for each route is calculated. To calculate the travel-time a simple function is used (based in [35] and [36]). This function determines the occupational density and from this value it derives the vehicle flow, which gives the corresponding travel-time. This relation between density and flow is called the fundamental diagram (depicted in Fig. 4).

Resuming the discussion of the final experiment, the scenario was simulated with different target densities, to verify the validity of the PT. A target density was determined and this means that it represents the density expected for the UE, i.e., when both routes have the same travel-time (the same density) and the density expected in both routes for the optimal split. The results are shown in Tab. 2. These values (Tab. 2) are the resulting occupation for the *Main* route over 100 repetitions, using 100 agents. The column $PLAIN$ is the result for agents using the standard Q-Learning, $C-EUT$ is the clustered version of it (to verify if the clustering method is biased), and $C-PT$ the PT modified Q-Learning.

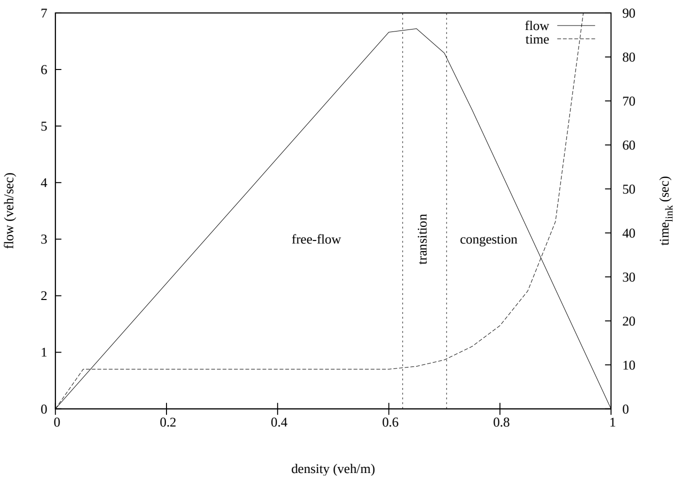

Figure 4: Fundamental Diagram

The first issue concerns the use of the clustering method. As it can be seen in Tab. 2 the values of $PLAIN$ and $C-EUT$ are fairly similar (which is an evidence for saying that the clustering method has no bias). Then it can also be shown that the $C-PT$ column has a consistent divergence from the other two at densities 0.6 and 0.8.

The apparent determinism of the first values (densities from 0.1 to 0.4) is due to the fact that all agents are accommodated by *Main* route under the free-flow regime (see the corresponding travel-time region in Fig. 4). The determinism is apparent because the variation is too low to be captured with only 100 agents – see the row identified by "0.3 with 500 agents". Another point is why all agents use the *Main* and not the *Secondary.* This is caused by the route internal order, with a different order another distribution is shown – see the row "0.1 with inverse route presentation".

The reason why at density 0.5 the $C-PT$ does not differ from the rational model is due to the lack of variability. For the PT to diverge from EUT the outcome probabilities must be located at the $\pi(\bullet)$ function "bumps". The next value (target density 0.6) shows a disproportional occupation at the *Main* route. This is a consequence of the first "bump", where most of the (bad) outcomes of the *Secondary* are located. This means that the high outcomes are overestimated (and as higher the travel-time as worse for route evaluation) and therefore the *Secondary* route is avoided by the agents. The outcomes at density 0.7 could not be explained with satisfactory evidences[1] and the reasons can only be speculated. The strongest speculation is that the outcomes are located at both "bumps" of $\pi(\bullet)$, therefore cancelling each other.

For the density 0.9 the reason is simple. Because the scenario is overloaded no much room is left for deviations.

The optimal density, where the PT behaviour can be better appreciated is 0.8. This density is a consequence of the fundamental diagram used (in Fig. 4) that forces the scenario to

[1] The other results are a consequence of the prospects, which were omitted but support the argumentation.

Table 2: Occupation results of the *Main* route for density experiments

Density value	$\mu_{PLAIN}(\sigma_{PLAIN})$	$\mu_{C-EUT}(\sigma_{C-EUT})$	$\mu_{C-PT}(\sigma_{C-PT})$
0.1	100.0(0.0)	100.0(0.0)	100.0(0.0)
0.2	100.0(0.0)	100.0(0.0)	100.0(0.0)
0.3	100.0(0.0)	100.0(0.0)	100.0(0.0)
0.4	100.0(0.0)	100.0(0.0)	100.0(0.0)
0.5	76.6732(3.7392)	61.3960(2.6924)	63.6039(3.5406)
0.6	67.3762(2.2814)	67.4653(2.1413)	91.6831(5.0206)
0.7	67.1782(2.2663)	63.7326(2.7463)	66.7029(4.3156)
0.8	67.5346(2.3570)	67.7326(3.9823)	71.9504(2.8297)
0.9	67.1485(2.4550)	68.6435(2.7198)	68.2376(2.2345)
0.1 with inverse route presentation	0.0(0.0)	0.0(0.0)	0.0(0.0)
0.3 with 500 agents	499.9900(0.0990)	499.9900(0.0990)	499.9801(0.1393)
0.8 with inverse $\pi(\bullet)$	67.0099(2.2755)	68.2277(3.4808)	34.4059(4.3239)
User equilibrium	$66.\overline{6}$		

deal with constant congestion (where the highest variation in the travel-time occurs). The other experiments (horizon and agent population variation) where performed under this conditions and the results are consistent the results depicted in Tab. 2 for the density 0.8.

One point left open is why the *Main* route is the stressed one and not the *Secondary*. The reason is on the shape of the $\pi(\bullet)$ function and because the *Secondary* route is the most "sensible" of them, i.e., a variation of one agent in the occupation has a stronger influence in the travel-time there than for the *Main*. This means that high and seldom travel-times are located at the first "bump" and therefore penalising the *Secondary* route while frequent and not optimal travel-times in the *Main* route are tolerated. If this argument is true then "inverting" the shape[2] of $\pi(\bullet)$ function would also invert this distribution. This is exactly what happens – see the row "0.8 with inverse $\pi(\bullet)$" in Tab. 4.

10. CONCLUSION AND FUTURE WORK

In this article a modified $\mathcal{Q}$-Learning algorithm based on the Prospect Theory was presented for modelling non-rational human behaviour. Model and implementation issues were discussed emphasising the drawbacks in the use of the PT. It was so done to give a fair and critical view of the implications in using these algorithms. A point that is recurrent in the text, which is here mentioned once more, is that the algorithm will reduce to the normal $\mathcal{Q}$-Learning in the worst case. It was also informally shown that the use of the PT does not imply in computational complexity growth (the big-O complexity stays the same).

In a not so far future this algorithm will be used for urban traffic modelling, it is actually already in use but the analysis is preliminary. This means that a complete framework is already implemented and is being used.

For the experiments, the main conclusion is that the PT is worth investigating. It is a valid alternative for the EUT and may be a better approach for modelling real discrete choice problems such as traffic assignment, stock markets, and shop/consumption behaviour.

[2] This can be made by making γ higher then 1.0 in Eq. 3.

11. REFERENCES

[1] M. Allais. Le comportement de l'homme rationnel devant le risque: Critique des postulats et axiomes de l'École américaine. *Econometrica*, 21(4):503–546, October 1953. Translated and reprinted in [2].

[2] M. Allais and G. M. Hagen. *Expected Utility Hypotheses and the Allais Paradox: Contemporary Discussions of the Decisions Under Uncertainty with Allais' Rejoinder*, volume 21 of *Theory and Decision Library*. Kluwer Academic, September 1979.

[3] W. B. Arthur. Inductive reasoning and bounded rationality. *The American Economic Review*, 84(2):406–411, May 1994. Papers and Proceedings of the Hundred and Sixth Annual Meeting of the American Economic Association.

[4] G. Barron. personal communication, September 12th 2007.

[5] G. Barron and I. Erev. Small feedback-based decisions and their limited correspondence to description-based decisions. *Journal of Behavioral Decision Making*, 16(3):215–233, July 2003.

[6] A. L. Bazzan, R. H. Bordini, G. K. Andriotti, R. Vicari, and J. Wahle. Wayward agents in a commuting scenario (personalities in the minority game). In E. Durfee, editor, *Proceedings of the Fourth International Conference on Multi-Agent Systems (ICMAS'2000),10–12 July, Boston*, pages 55–62, Los Alamitos, CA, 2000. IEEE Computer Society.

[7] R. Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, dover paperback (2003) edition, 1957.

[8] R. Bellman. A markovian decision process. *Journal of Mathematics and Mechanics*, 6, 1957.

[9] M. Ben-Akiva. *The structure of travel demand models*. PhD thesis, MIT, 1973.

[10] M. Ben-Akiva and M. Bierlaire. Discrete choice methods and their applications to short-term travel decisions. In R. W. Hall, editor, *Handbook of Transportation Science*, Kluwer's International Series Operations Research Management Science, book 2, pages 5–34. Kluwer, 1999.

[11] D. Bernoulli. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperiales Petropolitanae*, 5:175–192, 1738. Translated

and reprinted in [12].

[12] D. Bernoulli. Exposition of a new theory on the measurement of risk. *Econometrica*, 22(1):23–36, January 1954.

[13] C. I. Bliss. The method of probits. *Science*, 79(2037):28–39, January 1934. Corrected in [?].

[14] C. I. Bliss. The method of probits – a correction. *Science*, 79(2037):409–410, January 1934.

[15] A. Cassandra, M. L. Littman, and N. L. Zhang. Incremental Pruning: A simple, fast, exact method for partially observable Markov decision processes. In D. Geiger and P. P. Shenoy, editors, *Proceedings of the Thirteenth Annual Conference on Uncertainty in Artificial Intelligence (UAI–97)*, pages 54–61, San Francisco, CA, 1997. Morgan Kaufmann Publishers.

[16] G. Davies and S. Satchell. Continuous cumulative prospect theory and individual asset allocation. Cambridge Working Papers in Economics CWPE 0467, Faculty of Economics (formerly DAE), University of Cambridge, April 2005. available at http://www.dectech.org/People_Greg.html.

[17] G. A. Davis and T. Swenson. Collective responsibility for freeway rear-ending accidents?: An application of probabilistic causal models. *Accident Analysis & Prevention*, 2006. In Press, Corrected Proof.

[18] B. De Martino, D. Kumaran, B. Seymour, and R. J. Dolan. Frames, biases, and rational decision-making in the human brain. *Science*, 313(5787):684–687, 2006.

[19] D. Ellsberg. Risk, ambiguity, and the savage axioms. *The Quarterly Journal of Economics*, 75(4):643–669, November 1961.

[20] P. C. Fishburn. *Utility Theory For Decision Making.* Operations Research Society of America. Publications in operations research. Wiley, 1970.

[21] G. Gigerenzer and D. G. Goldstein. Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4):650–669, 1996.

[22] D. Kahneman. Maps of bounded rationality: A perspective on intuitive judgment and choice. In T. Frängsmyr, editor, *The Nobel Prizes 2002*, pages 449–489, Aula Magna, Stockholm University, December 2002. Nobel Foundation. presented by Professor Torsten Persson, Chairman of the Prize Committee.

[23] D. Kahneman and S. Frederick. Frames and brains: elicitation and control of response tendencies. *Trends in Cognitive Sciences*, 11(2):45–46, February 2006.

[24] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–292, March 1979.

[25] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning*, pages 157–163, New Brunswick, NJ, 1994. Morgan Kaufmann.

[26] C. F. Manski. The structure of random utility models. *Theory and Decision*, 8(3):229–254, July 1977.

[27] A. A. Markov. Extension of the limit theorems of probability theory to a sum of variables connected in a chain. In R. Howard, editor, *Dynamic Probabilistic Systems*, chapter Appendix B. John Wiley and Sons, August 1971. reprinted.

[28] D. McFadden. Conditional logit analysis of qualitative choice behavior. *Frontiers of Econometrics*, 1974.

[29] D. McFadden. Rationality for economists? *Journal of Risk and Uncertainty*, 19(1–3):73–105, December 1999.

[30] D. McFadden and K. Train. Mixed mnl models of discrete response. *Journal of Applied Econometrics*, 15:447–470, 2000.

[31] L. M.L. Value-function reinforcement learning in markov games. *Cognitive Systems Research*, 2:55–66, April 2001.

[32] R. A. Poldrack, J. Clark, E. J. Paré-Blagoev, D. Shohamy, J. C. Moyano, C. Myers, and M. A. Gluck. Interactive memory systems in the human brain. *Nature*, 414(6863):546–550, November 2001.

[33] J. N. J. Reynolds, B. I. Hyland, and J. R. Wickens. A cellular mechanism of reward-related learning. *Nature*, 413:67–70, September 2001.

[34] M. O. Rieger and M. Wang. Prospect theory for continuous distributions. *Journal of Risk and Uncertainty*, 2008. to appear.

[35] R. P. Roess, E. S. Prassas, and W. R. McShane. *Traffic Engineering.* Pearson Education, Inc., Upper Saddle River, New Jersey 07458, third (international) edition, 2004.

[36] A. Schadschneider and M. Schreckenberg. Car-oriented mean-field theory for traffic flow models. *J. Phys. A: Math. Gen.*, 30(4):L69–L75, February 1997.

[37] H. A. Simon. A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1):99–118, February 1955.

[38] H. A. Simon. Rational choice and the structure of the environment. *Psychological Review*, 63:129–138, 1956.

[39] A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk And Uncertainty*, 5(4):297–323, October 1992.

[40] J. von Neumann. Zur theorie der gesellschaftspiele. *Mathematische Annalen*, 100(1):295–320, December 1928.

[41] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior.* Princeton University Press, Princeton University, 1944.

[42] P. Vovsha. The cross-nested logit model: Application to mode choice in the tel-aviv metropolitan area. *Transportation Research Record*, 1607:6–15, 1997.

[43] P. Wakker and A. Tversky. An axiomatization of cumulative prospect theory. *Journal of Risk and Uncertainty*, 7(2):147–175, October 1993.

[44] J. G. Wardrop. Some theoretical aspects of road traffic research. In *Proceedings of Institution of Civil Engineers, Part II*, volume 1, pages 325–378, London, 1952.

[45] C. J. C. H. Watkins. *Learning with Delayed Rewards.* PhD thesis, Cambridge University, 1989. Psychology Department.

[46] C. J. C. H. Watkins and P. Dayan. Technical note: Q-learning. *Machine Learning*, 8(3):279–292, 5 1992.