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 layman's 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 [9] (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 [6], which says that humans seek to maximise the accumulated total wealth. Based on this principle von Neumann and Morgenstern [18] 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 [17] became the most notorious but McFadden [16] and Kahneman [11] have expressed their concern with the rational model as well.

Moreover, not recently the rational assumption had been demonstrated wrong by Allais [1] to cite only the most well-known. For this reason the Allais paradox (after Allais [1]) is the focus in this text – formally presented in Sec. 5.2. 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 [13] 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 [4] (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 [19] 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.

2. STATE-OF-THE-ART

Rational models have been developed since von Neumann [18] 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 [14] 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. The Multi Nominal Logit [15], with no correlation, and Nested-Logit [5] are examples of such models.

These are, nevertheless, macroscopic and analytical models

that, as far as the knowledge of the author goes, have no 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.

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 [8, 12] that support the PT as a valid hypothesis for the human decision behaviour.

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 [10]. 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 [3], but it presented some drawbacks – it was behavioural unstable, which were reported to the authors [2].

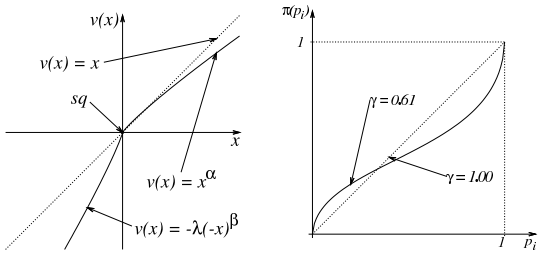
3. THE DISCRETE CHOICE PROBLEM

This work is focused in the single-person discrete choice problem. It is, according to von Neumann [18], 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.

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 (possible) outcome and $p_i \in (0, 1]$ is its corresponding probability ($\sum_{i \in \mathbf{x}} p_i = 1$). 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 [18], 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 [18]). The Independence axiom, in layman's 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). Hereafter, the EUT utility function is referred as $eut(\bullet)$.



(a) $v(x)$ with the original parameters (b) $\pi(p)$ with different γ values

Figure 1: Prospect Theory functions

5. PROSPECT THEORY

The name Prospect Theory [13] (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 [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.

5.1 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. 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. 1a and **??**.

The parameters were calibrate (in [13]) for a modified version of the Allais paradox (Tab. 1 from Sec. 5.2). 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 $\gamma^- = 0.69$.

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

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

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

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).

The probability distortion function $\pi(\bullet)$, in Eq. 3, is char-

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

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

5.2 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 [13] 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 [13], 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 [19]. 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.

$$Q_n(s,a) = (1-\alpha_n)Q_{n-1}(s,a) + \alpha_n\left[r_n + \gamma V_{n-1}(s')\right] \quad (4)$$

$$V_n(s) \equiv \underset{a\in\mathcal{A}}{\operatorname{argmax}}\left[Q_n(s,a)\right] \quad (5)$$

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.

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 [7], the following method is used. An aggregation threshold ϵ is given so that rewards closer than ϵ to each other must edited (aggregated).

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 ϵ, and the reward r to be included
Result: Updated set $\mathbf{C}$

```
1  c_min ← NIL ; /* c_min is the closest centroid to r */
2  Δd ← ∞ ; /* Δd is the distance between r and c_min */
   /* finds the closest centroid to the reward r */
3  forall pair c ∈ C do
4  |  if distance(c, r) < Δd then
5  |  |  Δd ← distance(c, r);
6  |  |  c_min ← c;
7  |  end
8  end
   /* checks if the centroid is suitable for
      aggregation */
9  if Δd ≤ ε then
10 |  c_{c_min} ← (r + a_{c_min} c_{c_min}) / (1 + a_{c_min}) ; /* new centroid value */
11 |  a_{c_min} ← a_{c_min} + 1 ; /* increments
   |  counter/accumulator */
12 else/* it is a new centroid */
13 |  C ← ⟨r, 1⟩ ; /* adds the new centroid */
14 end
```

This algorithm avoids linear complexity growth for any re-

ward 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.

$$pt(\mathbf{C}) = \sum_{\mathbf{c}\in\mathbf{C}} v(c_{\mathbf{c}})\pi\left(\frac{a_{\mathbf{c}}}{\sum_{\mathbf{c}\in\mathbf{C}} a_{\mathbf{c}}}\right) \quad (6)$$

$$\mathbf{C}_n = cluster(\mathbf{C}_{n-1}, r_n) \quad (7)$$

$$Q_n(s,a) = (1-\alpha_n)Q_{n-1}(s,a) + \alpha_n\left[pt(\mathbf{C}_n) + \gamma V_{n-1}(s')\right] \quad (8)$$

$$V_n(s) \equiv \underset{a\in\mathcal{A}}{\operatorname{argmax}}\left[Q_n(s,a)\right] \quad (9)$$

7. 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. **??**). 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 ($pt(\bullet) \sim eut(\bullet)$).

8. 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.

9. REFERENCES

[1] 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.

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

[3] 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.

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

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

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

[7] 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.

[8] 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.

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

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

[11] 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.

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

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

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

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

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

[17] H. Simon. *Models of Bounded Rationality*. MIT Press, 1982.

[18] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*. Princeton University Press, Princeton University, 60th anniversary edition edition, March 2007. 1st Edition: 1944.

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