Utility as an order relation

Yizhe Huang
School of Intelligence Science and Technology
Peking University
szhyz@pku.edu.cn

Abstract

Utility, as a type of basis for decision-making, serves to establish the ranking relationship between the optimal action and other actions. Adopting the perspective of order relations, this essay delves into an analysis of utility in both scalar and preference forms, elucidating their appropriateness while also addressing their limitations. The essay concludes by exploring potential pathways for the evolution of utility forms and highlighting alternative decision-making methodologies.

1 Introduction

Utility generally measures the quality of all consequences of decision-making and serves as the basis for decision-making [6, 7]. In economics, it is often assumed that rational decision-makers will make choices that maximize their utility [4]. However, cognitive limitations may lead individuals to misjudge their utility function. Despite this, individuals still strive to make optimal decisions based on their estimated utility function, a process known as bounded rationality [17]. Furthermore, there exist philosophical and psychological theories that support the idea that human behavioral decision-making is often utilitarian, particularly in interpersonal or moral decision-making scenarios [12, 5, 3, 10].

The representation of utility has traditionally been categorized into two types: scalar form and preference form [20]. Prior to comparing these two forms, it is essential to establish the role of utility, which is to guide decision-making among several possible choices. This necessitates the establishment of some form of order relationship for each decision, even if it is a partial order relationship. For instance, in the context of Markov decision processes (MDP), assuming the need to determine the next action, denoted as a, when in state s, the optimal decision can be represented as a_{OPT}. This implies the requirement that

$$u(s, a_i) \preceq u(s, a_{OPT}), \forall i, \tag{1}$$

where $x \preceq y$ describes y is more favored than x.

2 Scalar utility

For the current state s, a utility function $u(s)$ is employed to map the state to real numbers. This is the utility in scalar form. This approach is commonly utilized in economic contexts, where decision quality is primarily evaluated based on net profit, a scalar quantity. Implicitly, this method assumes that any gain or loss can be adequately captured by a scalar value. Does this assumption hold true in other fields?

2.1 Is a scalar utility enough?

This issue has caused a lot of discussion, particularly following the publication of [16]. Some articles have emerged to response to this viewpoint [9], or to critiqued it [18].

When utilizing a scalar to represent a utility function, this function effectively maps $(s, a)_{s\in S, a\in A}$ to the real number set $\mathbb{R}$, which comprises the total order relationship. Thus, this utility function

establishes a total order relation for the set $(s,a)_{s\in S,a\in A}$. Such a utility function can obviously meet our requirements for utility functions (Eq. (1)). In fact, it also provides a lot of order relationships that are not needed for our decision-making.

For instance, comparing $u(s_1,a_1)$ and $u(s_2,a_2)$ where $s_1 \neq s_2$ is meaningless in decision-making, indicating that $(s,a)_{s\in S,a\in A}$ ought not to possess a total order relationship. However, this does not imply that scalar utility is flawed. On the contrary, it suggests that the expressivity of scalar utility exceeds the requirements.

Another criticism of scalar utility is that it cannot handle multiple optimization goals, or multiple value dimensions. However, we can combine these value dimensions to derive a scalar value for comparison. Suppose we have N utility functions $u_1(s,a), u_2(s,a), \cdots, u_N(s,a)$, where $u_i : S \times A \to \mathbb{R}, \forall i$. Employing multiple value dimensions for decision-making typically implies that

$$(s,a_i) \preceq (s,a_j) \leftrightarrow (u_1(s,a_j), u_2(s,a_j), \cdots, u_N(s,a_j)) \preceq (u_1(s,a_j), u_2(s,a_j), \cdots, u_N(s,a_j)).$$

It can be proved that there exists a utility function $U_{all} : R^N \to R$, such that

$$\begin{aligned}&(u_1(s,a_j), u_2(s,a_j), \cdots, u_N(s,a_j)) \preceq (u_1(s,a_j), u_2(s,a_j), \cdots, u_N(s,a_j))\\ \leftrightarrow\ &U_{all}(u_1(s,a_i), u_2(s,a_i), \cdots, u_N(s,a_i)) \leq U_{all}(u_1(s,a_j), u_2(s,a_j), \cdots, u_N(s,a_j)).\end{aligned} \tag{2}$$

The key to the proof is simple. The cardinal of $\mathbb{R}^N$ is equal to the cardinal of $\mathbb{R}$:

$$|\mathbb{R}^N| = \aleph_1^N = (2^{\aleph_0})^N = 2^{N\aleph_0} = 2^{\aleph_0} = \aleph_1 = |\mathbb{R}|,$$

so U_{all} doesn't have to be a compressed mapping that breaks the order. Consequently, I believe that the expressivity of a scalar is sufficient to represent the order relationship among a limited number of optimization objectives.

Based on the aforementioned considerations, I believe that a scalar utility is theoretically adequate for decision-making. However, this does not imply that the corresponding utility function is easily attainable in practice. For instance, in decisions involving multiple value dimensions, linear weighting is often employed to address the optimization of multiple value dimensions, which may not always satisfy the conditions outlined for U_{all} in Eq. (2) [1]. In a related study, [1] conducts experiments to assess the capability of rewards to obtain "a set of acceptable behaviors", "a partial ordering over behaviors" and "a partial ordering over trajectories".

2.2 Learn a scalar utility

Notice that I avoided using the definition of value in reinforcement learning (RL) when discussing utility. I still have doubts about the consistency between utility and value or cumulated reward. For some discussion, please refer to [13].

However, it is noteworthy that the majority of state-of-the-art advancements in scalar utility arise within the RL paradigm. A plethora of related algorithms and literature exist, and detailed information can be accessed in [2, 19, 21].

The primary training challenge in RL lies in its demand for an extensive volume of data interaction with the environment, a process that evidently diverges from human learning and decision-making. Additionally, RL's ability to generalize across tasks is limited, and it responds slowly to changes in the environment. Presently, there is no pretrained model like GPT that can offer robust performance across all RL tasks.

In general, the learning efficiency of the RL paradigm is hindered by the scarcity of information and the sparse nature of scalar rewards. One potential remedy involves incorporating denser supervision signals to facilitate the learning of utility functions, akin to imitation learning. Another approach entails integrating common knowledge to enhance the prior of utility functions.

[1] Despite this, it is worth noting that linear weighting is adequate for reflecting Pareto optimality, suggesting that it may also suffice in numerous scenarios.

3 Preference utility

The approach offered by preference utility presents a more direct solution. Since Eq. (1) necessitates providing the order relations of the specified state-action pairs, we can directly train a model that outputs the order relationship between two state-action pairs, that is, a preference model.

This approach differs somewhat from preference-based reinforcement learning (PbRL), which primarily centers on feedback given in the form of preferences rather than rewards. PbRL may also make decisions by learning scalar utility functions or directly learning policies.

The utilization of preferences appears to align more closely with people's daily decision-making processes. In everyday decision-making, individuals may seldom explicitly convert the consequences of each decision into quantitative measures. However, it is debatable whether people directly output preferences through a model in many cases. Nevertheless, in decision-making scenarios that have been repeated frequently and have become "muscle memory," such a preference model may exist.

The preference model is expressive enough. Compared with Eq. (1), it may also additionally provide the order relationship between two non-optimal actions.

3.1 Some issues for a preference utility

One potential issue with the preference model is its potential failure to maintain the consistency of the ordering. While it is generally assumed that actions in a specific state s always adhere to a transmissive order relationship, this may not always be guaranteed through the preference model alone. The classic example illustrating this challenge is the Condorcet paradox [14].

Using preference as a supervision signal circumvents potential inconsistencies between reward and training goals, particularly in complex tasks where determining a reasonable reward can be challenging. However, it also raises issues related to temporal credit assignment and how to handle trajectory preferences.

4 Discussion

We can examine utility through the lens of order relations. From this standpoint, both scalar and preference forms of utility exhibit sufficient expressivity to represent utility and can generate order relations that surpass what we need. While this over-qualification brings practical convenience, it is important to investigate whether it amplifies the complexity of model learning. Is it conceivable for us to devise a simplified form of utility that only needs to satisfy the minimum requirements (Eq. (1))? Could this utility be easier to learn?

I believe that utility is not the sole basis for human decision-making. Its prominence in current decision-making intelligence research stems from its amenability to computational modeling. Other approaches, such as deontology[8, 5] or contractualism [15, 11], merit further exploration.

References

[1] David Abel, Will Dabney, Anna Harutyunyan, Mark K Ho, Michael Littman, Doina Precup, and Satinder Singh. On the expressivity of markov reward. *Advances in Neural Information Processing Systems*, 34:7799–7812, 2021. 2

[2] Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38, 2017. 2

[3] Craig Boutilier, Ioannis Caragiannis, Simi Haber, Tyler Lu, Ariel D Procaccia, and Or Sheffet. Optimal social choice functions: A utilitarian view. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pages 197–214, 2012. 1

[4] John Broome. Utility. *Economics & Philosophy*, 7(1):1–12, 1991. 1

[5] Paul Conway and Bertram Gawronski. Deontological and utilitarian inclinations in moral decision making: a process dissociation approach. *Journal of personality and social psychology*, 104(2):216, 2013. 1, 3

[6] Ward Edwards. The theory of decision making. *Psychological bulletin*, 51(4):380, 1954. 1

[7] Peter C Fishburn, Peter C Fishburn, et al. *Utility theory for decision making*. Krieger NY, 1979. 1

[8] Samuel Freeman. Utilitarianism, deontology, and the priority of right. *Philosophy & Public Affairs*, 23(4):313–349, 1994. 3

[9] David Israel. Response to 'reward is enough'–this is not a review; it's a response. *Artificial Intelligence*, page 103977, 2023. 1

[10] Julian Jara-Ettinger, Hyowon Gweon, Laura E Schulz, and Joshua B Tenenbaum. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8):589–604, 2016. 1

[11] Sydney Levine, Nick Chater, Joshua Tenenbaum, and Fiery Cushman. Resource-rational contractualism: A triple theory of moral cognition. 2023. 3

[12] George F Loewenstein, Leigh Thompson, and Max H Bazerman. Social utility and decision making in interpersonal contexts. *Journal of Personality and Social psychology*, 57(3):426, 1989. 1

[13] Vlad Mikulik. Utility $\neq$ reward, September 2019. URL `https://www.lesswrong.com/posts/bG4PR9uSsZqHg2gYY/utility-reward`. 2

[14] Jean Antoine Nicolas et al. *Essai sur l'application de l'analyse à la probabilité des decisions rendues à la pluralité des voix. Par m. le marquis de Condorcet,...* de l'Imprimerie Royale, 1785. 3

[15] Thomas M Scanlon, Amartya Sen, Bernard Williams, et al. Contractualism and utilitarianism. 1982. 3

[16] David Silver, Satinder Singh, Doina Precup, and Richard S Sutton. Reward is enough. *Artificial Intelligence*, 299:103535, 2021. 1

[17] Herbert A Simon. Bounded rationality. *Utility and probability*, pages 15–18, 1990. 1

[18] Peter Vamplew, Benjamin J Smith, Johan Källström, Gabriel Ramos, Roxana Rădulescu, Diederik M Roijers, Conor F Hayes, Fredrik Heintz, Patrick Mannion, Pieter JK Libin, et al. Scalar reward is not enough: A response to silver, singh, precup and sutton (2021). *Autonomous Agents and Multi-Agent Systems*, 36(2):41, 2022. 1

[19] Xu Wang, Sen Wang, Xingxing Liang, Dawei Zhao, Jincai Huang, Xin Xu, Bin Dai, and Qiguang Miao. Deep reinforcement learning: a survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2

[20] Christian Wirth, Riad Akrour, Gerhard Neumann, Johannes Fürnkranz, et al. A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research*, 18 (136):1–46, 2017. 1

[21] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021. 2