Reinforcement Learning with Preferences

Johannes Feldmaier
Chair for Data Processing
Department of Electrical and Computer Engineering
Technische Universität München
johannes.feldmaier@tum.de

Hao Shen
Chair for Data Processing
Department of Electrical and Computer Engineering
Technische Universität München
hao.shen@tum.de

Dominik Meyer
Chair for Data Processing
Department of Electrical and Computer Engineering
Technische Universität München
dominik.meyer@tum.de

Klaus Diepold
Chair for Data Processing
Department of Electrical and Computer Engineering
Technische Universität München
klaus.diepold@tum.de

Abstract

In this work, we propose a framework of learning with preferences, which combines some neurophysiological findings, prospect theory, and the classic reinforcement learning mechanism. Specifically, we extend the state representation of reinforcement learning with a multi-dimensional preference model controlled by an external state. This external state is designed to be independent from the reinforcement learning process so that it can be controlled by an external process simulating the knowledge and experience of an agent while preserving all major properties of reinforcement learning. Finally, numerical experiments show that our proposed method is capable to learn different preferences in a manner sensitive to the agent's level of experience.

Keywords: Preference learning, prospect theory, affective state, reinforcement learning.

Acknowledgements

This work has been partly supported by IGSSE - International Graduate School of Science and Engineering, Technische Universität München, Germany.

1 Introduction

The expected utility hypothesis is commonly used to model human's decision making behavior in scenarios with uncertain outcomes, such as gambling. However, situations where preferences of individuals among the same choices are important, are not handled properly by the classic expected utility theory of Bernoulli. The Prospect Theory (PT), proposed by Kahneman and Tversky [7] introduces the concept of reference point to the expected utility theory of Bernoulli. This reference point enables to model preferences of individuals among same choices. It means that an internal reference point for a specific decision is essential to model the decision making behavior of people. Meanwhile, in psychology, a similar concept of affective states plays an important role in describing human's behavior. The human affect system is responsible to regulate the perception and assessment of events. It is able to assign rapidly emotions to occurring situations. This affective representation is then used to influence the decision making process, cf. [6].

Recently, it is considered that both affective and cognitive systems are essential in future smart systems, cf. [8]. The work in [2] also hypothesizes that rational decisions of humans are primarily influenced by so-called *somatic markers*, i.e. the positive or negative feeling towards a specific situation. Therefore, it is vitally important to consider both affective component and cognitive component, in order to design an autonomous and rational decision making agent. Because of these inseparable components involved in the human decision making system, we propose to integrate an externally controlled affective state into autonomous decision making. This results in a simulated emotional and rational experience of a reinforcement learning agent. There are examples (c.f. Section 3) where the level of experience influences the outcome of a task.

2 Related Works and Motivations

There are numerous works about the integration of affective and subjective components into Reinforcement Learning (RL) agents. We selected those, which have influenced our idea of integrating an internal affective state into RL.

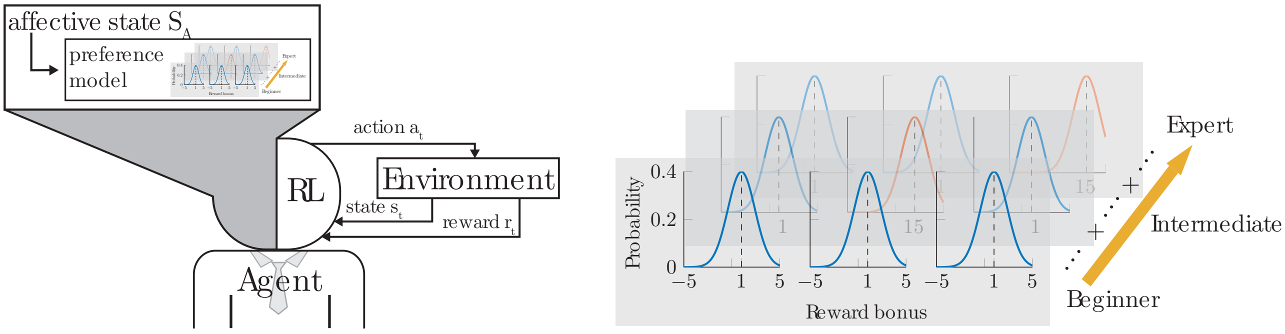
First of all, the work of Kenji Doya describes neuromodulatory systems and the (global) signals that regulate the (reinforcement) learning mechanisms of the human brain [3]. He argues that specific signals control and regulate the metaparameters (like randomness, action selection, reward prediction error, speed of memory update) for the reinforcement learning process. But there is no clear hypothesis regarding how the brain generates those signals. It seems to be a process separated of the actual learning running in different brain areas. This suggests that the brain has the capability of dynamically adjusting these metaparameters towards new or dynamically changing environments.

Besides these neurophysiological findings, the Prospect Theory shows the psychological influence of external and internal signals on the decision making behavior of human. In the Prospect Theory, a value function is described which is sensitive to deviations of the outcome (reward) according to a reference point in case of a risky choice. The reference point thereby is set by the current decision problem and depends rather on the losses and gains than on the final net asset value. This is also the reason why the framing of a choice problem becomes critical. The framing of the problem results in an external shift of the reference point which alters obviously the decision behavior [10].

A concrete combination of Reinforcement Learning and Prospect Theory is described by Ahn [1]. He has extended the conventional framework of RL for Markov decision processes (MDPs) with PT-based subjective value functions to model experienced-utility and predicted-utility functions. Furthermore, these functions vary dynamically according to the affective state of the decision maker (agent). This enables the agent to choose an action according to different risk attitudes and action tendencies on the basis of subjectively evaluated previous outcomes of decisions. The performance of his algorithm appears to be very good in the selected domains, but are difficult to reproduce due to the parameter dependence (which were additionally optimized for each domain). Moreover, the reference point of the PT value function is set automatically by the algorithm and an external control is not intended. Both, the external control as well as a strict separation of the reference point from the learning process seem to be important, if we consider the framing hypothesis and neurophysiological findings.

A third topic which has influenced our idea were preferences and biases. As preferences are fundamental for the human choice behavior, they are also an important component in learning. Human decision making is based on both, an objective and a subjective component [2]. Both components mainly depend on the emotional experience of a person. That means that decisions are evaluated in terms of objective and subjective rewards. The subjective rewards base on experienced feelings and emotions of previous outcomes of actions and decisions. Over time, specific situations and their past outcomes are associated with particular emotions (and their corresponding bodily changes). During decision making, these emotion-situation pairs are used as physiological signals (or *somatic markers*) to bias decision making towards certain policies while avoiding others. The whole set of *somatic markers* can be seen as the (emotional) experience of a human and is gathered during life. Regarding the development of artificial life-long learning agents such an emotional component is still a side issue. Therefore, the introduction of an experience-driven learning agent with specific preferences and aversions is vitally important to build more human like agents [12, 11].

A concrete approach combining RL with preferences is described by Fürnkranz et al. [5], where they combine preference learning [4] with RL. They learn a preference model from qualitative feedback and use it for ranking different policies



(a) (b)

Figure 1: (a) Reinforcement Learning framework with integrated preference model and affective state regulation. (b) Preference model using Gaussian distribution functions to add a reward bonus to a specific state or action according to the experience level S_A. In the depicted model there are three different experience levels. The actual preference in each level is highlighted (orange) and results in a reward bonus with a higher mean than the average. The different reward boni in each level are added up.

(fixed trajectory of a Markov process). The main drawback of this approach is the qualitative feedback which is used to evaluate already learned policies. This external qualitative feedback is given by human experts. Also, the learning process is directly interrelated with the preference model and there is no control how much the preference model affects the decision.

As we have seen the human learning process depends on external signals like framing effects and on internal signals generated by neurophysiological systems, preferences, and (emotional) experience. These are the influencing factors we want to use in our algorithm to model an agent with experience based preferences. Therefore, we propose a framework which should illuminate the role of an internal affective state (like an experience value) in the context of RL. We constructed this affective state so that it fulfills the Markov property, enabling the external control of the policy together with a preference model. In parallel, the affective state could also be used to model reference points as proposed in the prospect theory, enabling the agent to subjectively evaluate outcomes according to a (subjective) reference point. There are possible applications in the field of artificial intelligence, human-computer interfaces, and decision-support systems (recommender systems) where preferences of policies are needed and they are an essential aspect of rational autonomous agents.

3 Reinforcement Learning with Preferences

Basic Reinforcement Learning assumes a scenario in which an agent acts in a (finite) state space by performing different actions. A reward signal gives the agent feedback about its actions. The goal of RL is defined as maximizing the expected total sum of rewards. The basic formulation of a RL problem builds on the notation of a Markov decision process [9]:

- A set of states $S \in \{s_1, s_2, \ldots, s_i\}$ and a set of actions $A(s_i) \in \{a_1, a_2, \ldots, a_j\}$ which the agent can perform in a particular state s_i.
- Transition probabilities $\mathscr{P}^a_{ss'} = Pr\{s_{t+1} | s_t, a_t\}$ which denote the probabilities that an action a_t in state s_t leads to state s_{t+1}.
- A reward function $r(s_t, a_t)$ giving the agent reward according to the state and action performed.

We extended this basic RL framework with an additional state called affective state S_A (Figure 1a). According to this state the agent uses a preference model (Figure 1b) which gives a reward bonus to specific actions or states. In our experiment we use a one-dimensional state representation for controlling the preference model. For representing more complex affective states (e.g. general mood states) it would be possible to extend the state representation to a multi-dimensional vector controlling different preference models.

The basic idea underlying our preference model are various reward functions which are selected according to an affective state S_A. The additional reward functions correspond to reward facets which an agent can only perceive with increased experience or external feedback. The difference between our framework and approaches for multi-dimensional or dynamic reward environments surfaces in scenarios, where the agent first has to learn how the reward process works. For example, as a novice in cooking someone tells you to cook fried eggs and gives you eggs, a pan, and all other necessary equipment. You will start frying eggs until it looks like a fried egg. Now someone tastes and tells you that the yolk has to be cooked through. Up to now, as a beginner in cooking you only judged fried eggs according to the appearance, but now a new dimension is added: how the yolk has to be cooked. Next time cooking fried eggs (with an increased level of experience), this dimension is also considered and evaluated.

Another example, more complex especially for machines is the taste of coffee. As a beginner in drinking coffee you only judge your coffee according to the overall taste of bitterness or sweetness and probably the temperature. After drinking coffee regularly you start to taste different flavors within a specific kind of coffee. Consequently, your experience in drinking coffee has added new dimensions of possible rewards to your preference model. As a coffee expert you choose your coffee according to a variety of different tastes and ways of preparation.

So, our preference model consists of several levels of reward functions which are only visible or accessible to the agent in a specific affective state (like a specific level of experience). We additionally introduce the constraint, that each level of experience can only increase, hence the additional reward functions are always added and can never be removed again (no-go-back-policy).

The signal generation representing the level of experience seems to be a substantial question. There are several solutions conceivable, like a monotonically increasing function according to the action taken by the agent or more advanced approaches considering the gathered knowledge and external feedback. Furthermore, according to neurophysiology a strict separation of the control signals for the learning process is desirable. That means, the internal affective signal (in the actual case the level of experience) must be externally controlled and should only be based partially on the results of the learned actions and outcomes. Ideally, it should base on past experiences, temporal effects, external feedback, and cognitive biases (like framing effects). Therefore, we decided to use in this stage of the development a simple stepwise function which increases the level of experience in three steps after a specific number of actions and does not interfere directly with the reinforcement learning process. This might be a oversimplification but allows a clear and concise formulation of the experiment.

4 Experiment

We conducted an experiment to investigate the properties of this extended reinforcement learning framework demonstrating that additional reward functions which are controlled externally can introduce preferences to the learned policy. In the experiment we simulated a three-armed bandit with Gaussian distributed rewards. At the beginning, each arm returns the same Gaussian distributed reward r_{ext} with $\mu_j = 1$ and $\sigma_j = 1$ for all arms $j = 1, 2, 3$. The experiment was repeated independently 100 times and 300 trials were played. Afterwards, the results where averaged. In this experiment, we used a simple affective state, which can be compared to a general level of experience. The state was generated externally and was modelled as an exponential increasing process simulating a continuously increasing level of experience.

We have used Q-learning with ϵ-greedy exploration [9] to learn the optimal policy in this experiment ($\epsilon = 0.05$, $\alpha = 0.8$, $\gamma = 0.4$). The results depicted in Figure 2 show that with increasing experience level the policy changes. At the first level (beginner, $0 \leq S_A \leq 0.5$), the agent is not able to differentiate the decisions and each action a_j is taken equally. After gaining some experience (intermediate, $0.5 < S_A \leq 0.95$) a second layer or dimension of reward functions is added to the learning process (second layer of Figure 1b). Now, the agent receives the original reward of the bandit process and an additional reward bonus (in the current setting also Gaussian distributed) which is added to the actual reward. In the intermediate level a preference model (or reward dimension) with $\mu_{2,int} = 15$, $\mu_{1,int} = \mu_{3,int} = 1$ and $\sigma_{1,int} = \sigma_{2,int} = \sigma_{3,int} = 1$ is added. So, a clear preference for action a_2 is introduced at this level of experience and the agent starts to prefer this action. In the expert level ($0.95 < S_A \leq 1$), a preference model is added which assigns a reward bonus, slightly higher than the one for action a_2, to action a_3 ($\mu_{3,exp} = 15$, $\mu_{1,exp} = \mu_{2,exp} = 1$ and $\sigma_{1,exp} = \sigma_{2,exp} = \sigma_{1,exp} = 1$). This results in a bias for action three which is therefore most frequently selected while the probability for action one and two decreases. The total reward r for updating the Q-function can be denoted as

$$r(s_t, a_t, S_A) = r_{ext} + \sum_{j \in A, k \in S_A} X_{j,k}, \qquad X_{j,k} \sim \mathcal{N}(\mu_{j,k},\, \sigma^2_{j,k}), \tag{1}$$

where S_A denotes the set of experience levels, A the set of possible actions, and r_{ext} the external reward given by the three-armed bandit. $X_{j,k}$ is the reward bonus for a given action j in a specific affective state k which is in this example a sample of a normal distribution $\mathcal{N}(\mu_{j,k},\, \sigma^2_{j,k})$.

5 Results

The results of the experiment as depicted in Figure 2 are straight forward and meet our expectations. But they show that the extension of the conventional RL framework with various reward process dimensions controlled by an external affective state can introduce preferences to the learned policy. In the introduction we motivated this extension by psychological and neurophysiological findings. This enables developers of learning agents to use it for developing agents with preferences, specific tastes and risk dispositions. The additional reward dimensions could be used e.g. for integrating prospect theory value functions besides conventional (like Gaussian) reward functions to simulate rational componentes of decision making (like risk aversion and attraction) while preserving the main underlying reward process maximizing the expected utility.

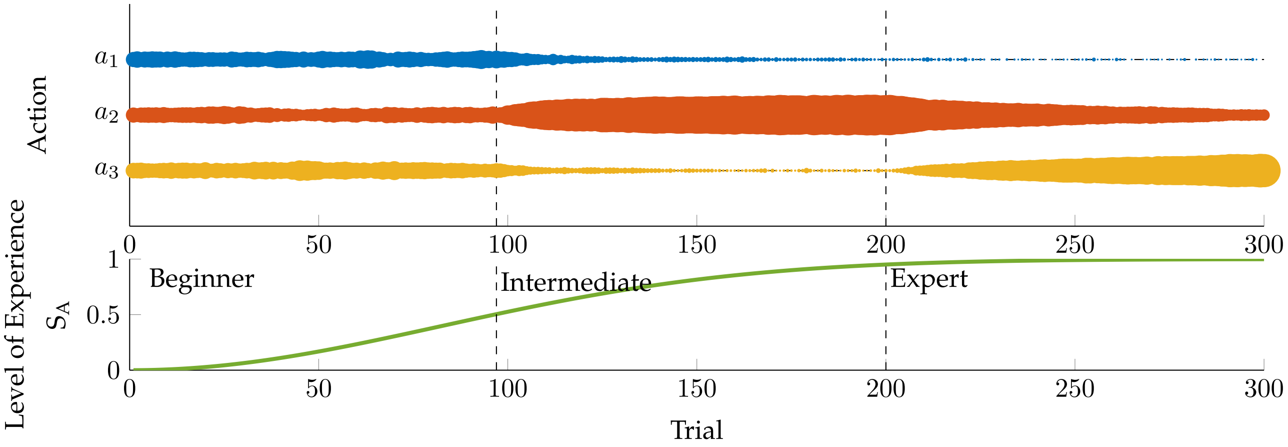


Figure 2: Learned policy for a simulated increasing level of experience. First, the agent acts like a beginner end selects every action equally (the thickness of the bars corresponds to the frequency of selection). At trial 98 the agent enters the intermediate state and can perceive an additional reward model with a preference for the second action (a_2). In the expert state, another reward model is added with a preference for action three (a_3).

6 Conclusion

The basic premise of the paper is that traditional RL can be extended by a preference model which is controlled by a single external state not interfering with the learning process. We also described in this paper the multi-dimensional approach of constructing a preference model with various reward functions as well as the application of it to integrate distinct preferences, biases, or cognitive frames into the framework of reinforcement learning.

In future studies we want to investigate the properties of such a framework regarding uncertainties in the reward process, the exploration and exploitation trade-off, and learning different “flavors or tastes” of polices. The property of an increasing reward function space would also be an interesting topic for further studies. Our vision is to build artificial agents with human-like preferences and sensitivity towards framing effects.

References

[1] Hyung-il Ahn and Rosalind W. Picard. Affective-cognitive learning and decision making: The role of emotions. In *The 18th European Meeting on Cybernetics and Systems Research (EMCSR)*, Vienna, Austria, 2006.

[2] António Damásio. *Descartes' Error: Emotion, Reason, and the Human Brain*. Putnam Publishing, Kirkwood, NY, USA, 1994.

[3] Kenji Doya. Metalearning and neuromodulation. *Neural Networks*, 15(4):495–506, 2002.

[4] Jon Doyle. Prospects for preferences. *Computational Intelligence*, 20(2):111–136, 2004.

[5] Johannes Fürnkranz, Eyke Hüllermeier, Weiwei Cheng, and Sang-Hyeun Park. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine Learning*, 89(1–2):123–156, 2012.

[6] Gerd Gigerenzer, Ralph Hertwig, and Thorsten Pachur. *Fast and Frugal Heuristics - Theory, Tests, and Applications*. Oxford University Press, New York, NY, USA, 2011.

[7] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–291, 1979.

[8] Marvin Lee Minsky. *The Emotion Machine: Commonsense Thinking, Artificial Intelligence, and the Future of the Human Mind*. Simon & Schuster, New York, NY, USA, 2006.

[9] Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, USA, 1st edition, 1998.

[10] Amos Tversky and Daniel Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):453–458, 1981.

[11] Juan D. Velásquez. When robots weep: Emotional memories and decision-making. In *Proceedings of the 15th National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference (AAAI)*, Madison, WI, USA, 1998.

[12] David Vernon, Giorgio Metta, and Giulio Sandini. A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *Transactions on Evolutionary Computation*, 11(2):151–180, 2007.