ReScience

[Re] Adaptive properties of differential learning rates for positive and negative outcomes

Sophie Bavard[1] **and Héloïse Théro**[1]

1 Laboratoire de Neurosciences Cognitives Computationnelles (ENS - INSERM), Département d'Études Cognitives, École Normale Supérieure, PSL Research University, 29 rue d'Ulm, 75005 Paris, France

sophie.bavard@gmail.com, thero.heloise@gmail.com

Editor
Olivia Guest

Reviewers
Xavier Hinaut
Benoît Girard

Received Feb, 20, 2018
Accepted Jun, 14, 2018
Published Jun, 14, 2018

Competing Interests:
The authors have declared that no competing interests exist.

Article repository

Code repository

A reference implementation of

→ Cazé, R. D., & van der Meer, M. A. (2013). Adaptive properties of differential learning rates for positive and negative outcomes. Biological cybernetics, 107(6), 711-719. https://doi.org/10.1007/s00422-013-0571-5

Introduction

Reinforcement learning represents a fundamental cognitive process: learning by trial and error to maximize rewards and minimize punishments. Current and most influential theoretical models of reinforcement learning assume a unique learning rate parameter, independently of the outcome valence (Sutton and Barto [14], O'Doherty et al. [10], Behrens et al. [1]). However human participants were shown to integrate differently positive and negative outcomes (Frank, Seeberger, and O'Reilly [3], Frank et al. [4], Sharot, Korn, and Dolan [13]). This motivated the reference article to implement a modified version of the reinforcement learning model, with two distinct learning rates for positive and negative outcomes (Cazé and Meer [2]).

They have shown that although differential learning rates shifted reward predictions and could thus be seen as a maladaptive bias, this model can outperform the classical reinforcement learning model on tasks with specific outcome probabilities. Following Cazé and Meer [2]'s predictions, a subsequent empirical article have modeled human behavior on these specific tasks (Gershman [7]). The question is still an active research area, as various articles have further investigated the difference learning rates bias (Garrett and Sharot [5], Moutsiana et al. [9], Shah et al. [12], Garrett and Sharot [6], Lefebvre et al. [8], Palminteri et al. [11]).

A link to the pdf version of the reference article was posted on the last author's laboratory website (http://www.vandermeerlab.org/publications.html), but the corresponding code was not available (https://github.com/vandermeerlab/papers/tree/master/Caze_vanderMeer_2013). We believe that an openly available code repository replicating the results of Cazé and Meer [2]'s paper can be helpful to the scientific community. We therefore implemented the model and analysis scripts using Python, with numpy, random and matplotlib libraries.

Methods

We first implemented our scripts on Matlab, as we were more familiar with this language, and then adapted them on Python.

We used the modeling description of the reference article to implement our replication. They used standard Q-learners with a softmax action selection rule (Sutton and Barto [14]), and their precise description enabled us to implement them with low difficulty. But we found four ambiguities in the simulation procedure.

First, the authors described their analytical results to be valid for "$Q_0 \neq \{-1, 1\}$" in section 2, but did not specify what value of Q_0 they used in all the following simulations. We chose to use $Q_0 = 0$, as this initial value is the middle point between the two possible outcomes (i.e., -1 and 1). As we replicated all the original figures, even the dynamics in the beginning of the learning curves (see Figures 2 A, 3 and 4 B), we believe the reference article must have used similar initial Q-values.

Second, regarding the parameter setting for Figure 1's simulations, the ratio of α^+ over α^- was said to be either 0.25, 1 or 4, but they did not specify what were the exact values of α^+ and α^- used. We thus set them according to the following description of the pessimistic, rational and optimistic agents in section 3, i.e.,:

- $\alpha^+ = 0.1$ and $\alpha^- = 0.4$ for the ratio of 0.25
- $\alpha^+ = 0.1$ and $\alpha^- = 0.1$ for the ratio of 1
- $\alpha^+ = 0.4$ and $\alpha^- = 0.1$ for the ratio of 4

Third, the number of iterations made to generate Figures 3 and 4 were not indicated, and we assumed the authors used the same number as in Figures 1 and 2 (i.e., 5,000 runs).

Finally, in the reinforcement learning framework, the probabilities to choose each action are computed, then used to select an action through a pseudo-random generator. In the reference article, it was sometimes unclear whether the analyses were performed on the probabilities of choice, or rather the proportions of implemented choices. For example Figure 2's legend indicated: "Mean probability of choosing the best arm", suggesting that the probabilities themselves were used. However, when commenting the figure in section 3, the authors appeared to say that the actual choices were rather used: "the optimistic agent learns to take the best action significantly more than the rational agent". For our analyses, we started by using the probabilities of choice, as this would lead to more clear, less noise-corrupted results. However we then obtained very smooth learning curves, and were unable to reproduce the spikiness of the original Figures 2, 3 and 4. We thus computed the proportions of implemented choices for all our figures.

Results

We numbered our figures in the same way as the reference article.

All our figures reproduced the patterns of the original results. We were even able to replicate the fine-grained details of the learning curves, like the early bumps in performance in the high-reward task (Figures 2 A, 3 and 4 B, right panels, around 50-100 trials). In Figure 1, the mean and the variance of the Q-values were also very similar as the ones in the original figure.

The only discrepancy we found was in Figure 4 A. Although the general pattern was replicated, our learning curves appeared smoother than in the reference article. As the number of simulations were not explicitly specified for this figure, we cannot know if this is due to us running a higher number of simulations than the reference article, or from another difference in model implementation.

Conclusion

All the figures in Cazé and Meer [2] have been successfully reproduced with high fidelity, and we confirm the validity of their simulations. Overall the whole replication

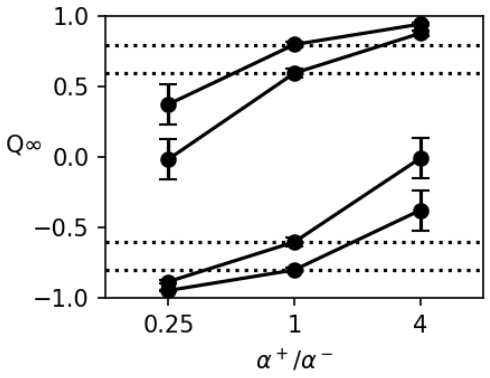

Figure 1: Average estimated Q-values after 800 trials averaged for different ratios of α^+ and α^-. The dotted lines represent the underlying average reward: 0.8, 0.6, -0.6, -0.8. The error bars represent the variance of the estimated Q-values.

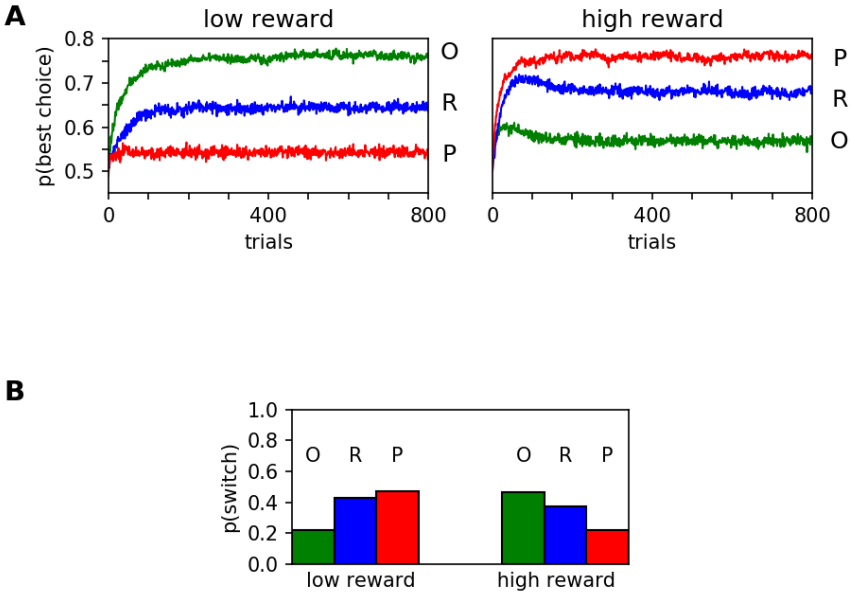

Figure 2: **A.** Performance, i.e. proportion of choices for the best action, for the three agents: Rational (R, $\alpha^+ = \alpha^-$, blue line), Optimistic (O, $\alpha^+ > \alpha^-$, green line) and Pessimistic (P, $\alpha^+ < \alpha^-$, red line). In this figure and the following ones, the left (resp. right) panel corresponds to the low-reward (resp. high-reward) task. **B.** Proportion of action switch after 800 trials for each agent, in the two different tasks.

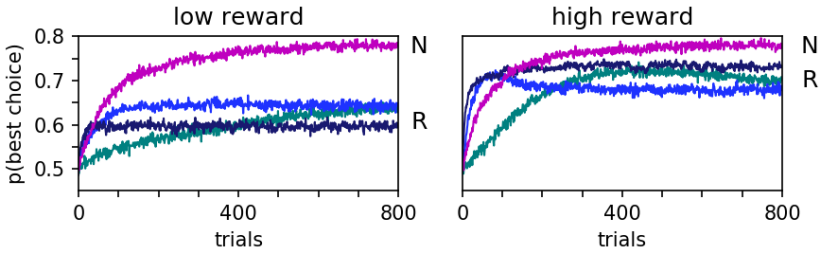

Figure 3: The performances of the Meta-learner (N) are shown in *purple* and those of the Rational agents (R) in different colors of blue (in *teal* for $\alpha = 0.01$, in *royal blue* for $\alpha = 0.1$ and in *navy blue* for $\alpha = 0.4$).

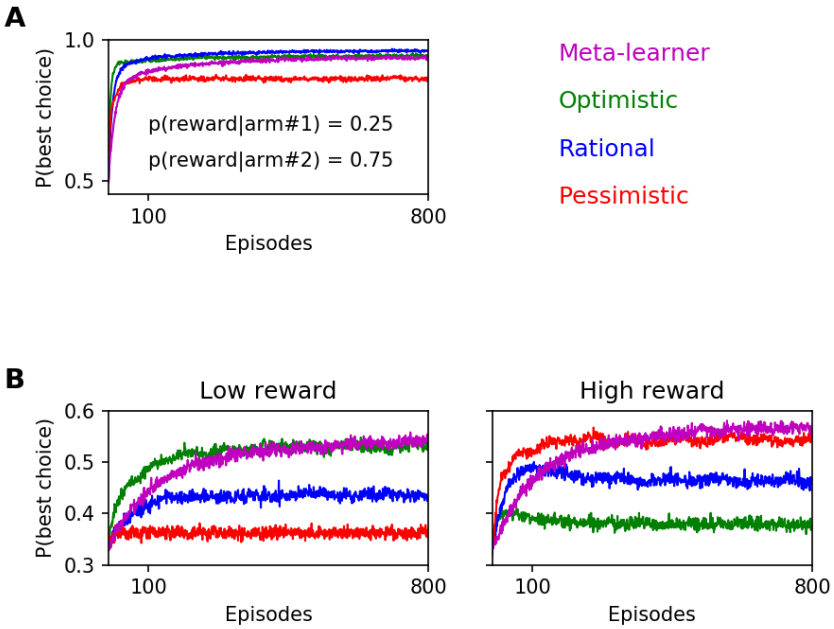

Figure 4: The performances of the Meta-learner, Optimistic, Rational and Pessimistic agents **A.** in a task where the probabilities of reward are 0.75 and 0.25 for the two choices. **B.** in a "three-armed bandit" task.

procedure was smooth: the models were implemented with low difficulty, and the simulations were quite straightforward apart from a few obscure details. We hope this replication can foster future research in the domain.

References

[1] Timothy EJ Behrens et al. "Learning the value of information in an uncertain world". In: *Nature neuroscience* 10.9 (2007), p. 1214.

[2] Romain D Cazé and Matthijs AA van der Meer. "Adaptive properties of differential learning rates for positive and negative outcomes". In: *Biological cybernetics* 107.6 (2013), pp. 711–719.

[3] Michael J Frank, Lauren C Seeberger, and Randall C O'Reilly. "By carrot or by stick: cognitive reinforcement learning in parkinsonism". In: *Science* 306.5703 (2004), pp. 1940–1943.

[4] Michael J Frank et al. "Genetic triple dissociation reveals multiple roles for dopamine in reinforcement learning". In: *Proceedings of the National Academy of Sciences* 104.41 (2007), pp. 16311–16316.

[5] Neil Garrett and Tali Sharot. "How robust is the optimistic update bias for estimating self-risk and population base rates?" In: *PLoS One* 9.6 (2014), e98848.

[6] Neil Garrett and Tali Sharot. "Optimistic update bias holds firm: Three tests of robustness following Shah et al." In: *Consciousness and cognition* 50 (2017), pp. 12–22.

[7] Samuel J Gershman. "Do learning rates adapt to the distribution of rewards?" In: *Psychonomic bulletin & review* 22.5 (2015), pp. 1320–1327.

[8] Germain Lefebvre et al. "Behavioural and neural characterization of optimistic reinforcement learning". In: *Nature Human Behaviour* 1.4 (2017), p. 0067.

[9] Christina Moutsiana et al. "Human frontal–subcortical circuit and asymmetric belief updating". In: *Journal of Neuroscience* 35.42 (2015), pp. 14077–14085.

[10] John O'Doherty et al. "Dissociable roles of ventral and dorsal striatum in instrumental conditioning". In: *science* 304.5669 (2004), pp. 452–454.

[11] Stefano Palminteri et al. "Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing". In: *PLoS computational biology* 13.8 (2017), e1005684.

[12] Punit Shah et al. "A pessimistic view of optimistic belief updating". In: *Cognitive Psychology* 90 (2016), pp. 71–127.

[13] Tali Sharot, Christoph W Korn, and Raymond J Dolan. "How unrealistic optimism is maintained in the face of reality". In: *Nature neuroscience* 14.11 (2011), p. 1475.

[14] Richard S Sutton and Andrew G Barto. *Introduction to reinforcement learning*. Vol. 135. MIT Press Cambridge, 1998.