ReScience

[Re] Speed/accuracy trade-off between the habitual and the goal-directed processes

Guillaume Viejo[1], **Benoît Girard**[1], **and Mehdi Khamassi**[1]

1 Sorbonne Universités, UPMC Univ Paris 06, CNRS, Institute of Intelligent Systems and Robotics (ISIR), F-75005 Paris, France

guillaume.viejo@isir.upmc.fr

Editor
Nicolas P. Rougier

Reviewers
Julien Vitay
Georgios Detorakis

Received Jan, 20, 2016
Accepted Feb, 9, 2016
Published Feb, 10, 2016

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

Article repository

Code repository

A reference implementation of

→ Speed/accuracy trade-off between the habitual and the goal-directed processes, M. Keramati, A. Dezfouli, P. Piray, PLoS computational biology, 7, 2011

Introduction

This study is a reference implementation of Keramati, Dezfouli, and Piray [2] that proposed an arbitration mechanism between a goal-directed strategy and a habitual strategy, used to model the behavior of rats in instrumental conditionning tasks. The habitual strategy is the Kalman Q-Learning from Geist, Pietquin, and Fricout [1]. We replicate the results of the first task, i.e. the devaluation experiment with two states and two actions. The implementation is in python with numpy, scipy and matplotlib library. The authors couldn't provide the original implementation and we are not aware of other implementations elsewhere.

Methods

We used the description of the model from the original article except for the implementation of the Kalman Q-Learning which we took from Geist, Pietquin, and Fricout [1]. We used the same parameters as the original article except for the update rate of the transition function ϕ, the initialization of the covariance matrice and an uncentered transform parameter κ that were not mentionned in the original article. The largest uncertainty about the model concerned the devaluation procedure. Besides setting the reward r to null, the authors stated that "For modeling the devaluation of the outcome in the first two simulations, $R(S_1, EM)$ is set to -1." As this notation $(R(S_1, EM))$ is not defined in the rest of the article, we assumed that it is $\hat{R}(S_1, EM)$ updated by equation (14) in the original article.

The parameters are as follows :

Name	Description	Value
σ	Updating rate of the average reward	0.02
η	Variance of evolution noise	0.0001
P_n	Variance of observation noise	0.05
β	Rate of exploration	1.0
ρ	Update rate of the reward function	0.1
γ	Discount factor	0.95

Name	Description	Value
τ	Time step of graph exploration	0.08
depth	Depth of search in graph exploration	3
ϕ	Update rate of the transition function	0.5
init cov	Initialisation of covariance matrice	1.0
κ	Unscentered transform parameters	0.1

We describe the algorithm of our implementation in details. The process of action selection and reward update are separated for clarity.

Initialization

$Q(s,a)^{Goal-Directed} = \{0, \ldots\}$

$Q(s,a)^{Habitual} = \{0, \ldots\}$

\# Covariance matrix

$$\Sigma = \begin{pmatrix} cov \times \eta & 0 & \ldots & 0 \\ 0 & cov \times \eta & \ldots & \vdots \\ \vdots & \ldots & \ddots & 0 \\ 0 & \ldots & 0 & cov \times \eta \end{pmatrix}$$

$R(S1, EM) = 1$ \# Reward value

$\bar{R} = 0$ \# Reward rate

$\hat{R}(s,a) = \{0, \ldots\}$ \# Reward function

Main Loop

FOR $i = 1 : T$

$s_t = S_0$ \# Initial state

IF $i = T_{devaluation}$ \# Moderate / Extensive training

$R(S1, EM) = 0$

$\hat{R}(S1, EM) = -1$

WHILE $s_t \neq S1 \bigwedge a_t \neq EM$

$a_t =$ **Selection**(s_t)

$r_t = R(s_t, a_t)$

$s_{t+1} = transition(s_t, a_t)$

Update(s_t, a_t, s_{t+1}, r_t)

Selection

\# Sort the Q-values in descending order

$\{a_1, \ldots, ai, \ldots\} \leftarrow sort(Q(s_t, a_i))$

\# VPI : Value of Precise Information

$VPI(s_t, a_1) = (Q(s_t, a_2)^H - Q(s_t, a_1)^H) P(Q(s_t, a_1)^H < Q(s_t, a_2)^H) + \frac{\sigma(s_t,a_t)}{\sqrt{2\pi}} e^{-\frac{(Q(s_t,a_2)^H - Q(s_t,a_1)^H)^2}{2\sigma(s_t,a_t)^2}}$

$VPI(s_t, a_i) = (Q(s_t, a_i)^H - Q(s_t, a_1)^H) P(Q(s_t, a_i)^H > Q(s_t, a_1)^H) + \frac{\sigma(s_t,a_t)}{\sqrt{2\pi}} e^{-\frac{(Q(s_t,a_1)^H - Q(s_t,a_i)^H)^2}{2\sigma(s_t,a_t)^2}}$

FOR $i \in \{a_1, a_2, \ldots, a_i, \ldots\}$

IF $VPI(s_t, a_i) \geq \tau \bar{R}$

Q-Value from Goal-directed system is evaluated

$Q(s_t, a_i) = \hat{R}(s_t, a_i) + \gamma \sum_{s'} p_T(\{s, a\} \rightarrow s') \max_{b \in A} Q(s', b)^{Goal-directed}$

ELSE

Q-Value from Habitual system is retrieved

$Q(s_t, a_i) = Q(s_t, a_i)^{Habitual}$

$a_t \leftarrow SoftMax(Q(s_t, a), \beta)$

Update

$\bar{R} = (1 - \sigma)\bar{R} + \sigma r_t$ # Reward Rate

$\hat{R}(s_t, a_t) = (1 - \rho)\hat{R} + \rho r_t$ # Reward function

$p_T(s_t, a_t, s_{t+1}) = (1 - \phi) p_T(s_t, a_t, s_{t+1}) + \phi$ # Probability of transition

Specific to Kalman Q-Learning

Sigma-points sampling

$\Theta = \{\theta_j, 0 \geq j \geq 2|S.A|\}$

$\check{W} = \{w_j, 0 \geq j \geq 2|S.A|\}$

$\check{R} = \{\check{r}_j = \theta_j(s_t, a_t) - \gamma \max_{b \in A} \theta_j(s_{t+1}, b),\ 0 \geq j \geq 2|S.A|\}$

$r_{predicted} = \sum_{j=0}^{2|S.A|} w_j \check{r}_j$

Covariance computation

$P_{\theta_j \check{r}_j} = \sum_{j=0}^{2|S.A|} w_j (\theta_j - Q_t^{Habitual})(\check{r}_j - r_{predicted})$

$P_{\check{r}_j} = \sum_{j=0}^{2|S.A|} w_j (\check{r}_j - r_{predicted})^2 + P_n$

$K_t = P_{\theta_j \check{r}_j} P_{\check{r}_j}^{-1}$ # Kalman gain

$\delta_t = r_t - r_{predicted}$ # Reward-prediction error

$Q_{t+1}^{Habitual} = Q_t^H + K_t \delta_t$

$P_{t+1}^H = P_t^H - K_t P_{\Sigma_t} K_t^T$

Results

We only reproduced the results of Figure 3 A, B, G, H in a qualitative manner. Results are presented in Figure 1. We can observe the strategy shift (from goal-directed to habitual) after extensive training around 50 time steps. In the original article, the strategy shift occurs after 100 time steps.

However we can observe a difference between the probabilities of action for the goal-directed model. In our implementation,

$$p(s_0, pl) \simeq 0.7$$

and

$$p(s_0, em) \simeq 0.3$$

before devaluation. In the original article,

$$p(s_0, pl) \simeq 0.6$$

and

$$p(s_0, em) \simeq 0.4$$

Nevertheless, the probabilities of action from the Kalman Q-Learning after strategy shifting are equivalent.

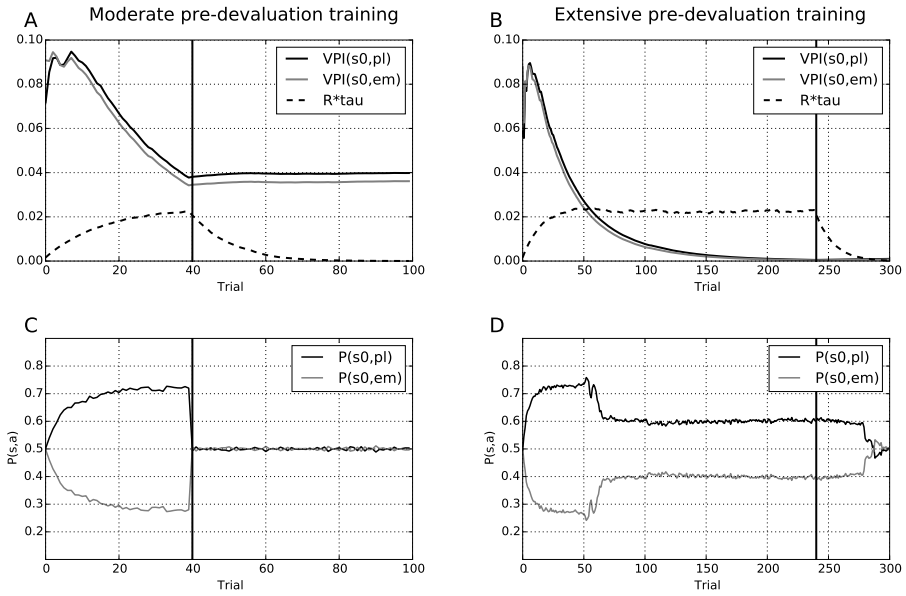

Figure 1: A. Value of Precise Information (full lines) for action press-lever and enter magazine in state S_0 and reward rate (dashed line) in moderate training. Vertical line represents the timing of devaluation. B. In extensive training. C. Probability of actions in state S_0 in moderate training. D. In extensive training.

Conclusion

We were able to qualitatively reproduce the first simulations of the article. Despite the small differences in the exact timing of the strategy shifting and in the probabilities of action, the behavior of our implementation is similar to the original article. Thus, we confirm the correctness of the model presented in the original article.

References

[1] Matthieu Geist, Olivier Pietquin, and Gabriel Fricout. "Kalman Temporal Differences: The deterministic case". In: *IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning* (2009), pp. 185–192.

[2] Mehdi Keramati, Amir Dezfouli, and Payam Piray. "Speed/accuracy trade-off between the habitual and the goal-directed processes". In: *PLoS computational biology* 7.5 (2011), e1002055.