[Re] Speed/accuracy trade-off between the habitual and the goal-directed processes
Guillaume Viejo, Benoît Girard, Mehdi Khamassi

To cite this version:
Guillaume Viejo, Benoît Girard, Mehdi Khamassi. [Re] Speed/accuracy trade-off between the habitual and the goal-directed processes. The ReScience journal, GitHub, 2016, 2 (1), 10.5281/zenodo.45852 . hal-01524285

HAL Id: hal-01524285
https://hal.archives-ouvertes.fr/hal-01524285
Submitted on 17 May 2017
[Re] Speed/accuracy trade-off between the habitual and the goal-directed processes

Guillaume Viejo¹, Benoît Girard¹, and Mehdi Khamassi¹

1 Sorbonne Universités, UPMC Univ Paris 06, CNRS, Institute of Intelligent Systems and Robotics (ISIR), F-75005 Paris, France
guillaume.viejo@isir.upmc.fr

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 conditioning 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 others 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 $\phi$, the initialization of the covariance matrix and an uncentered transform parameter $\kappa$ that were not mentioned 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 $R(S_1, EM)$ updated by equation (14) in the original article.

The parameters are as follows:

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma$</td>
<td>Updating rate of the average reward</td>
<td>0.02</td>
</tr>
<tr>
<td>$\eta$</td>
<td>Variance of evolution noise</td>
<td>0.0001</td>
</tr>
<tr>
<td>$P_{\nu}$</td>
<td>Variance of observation noise</td>
<td>0.05</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Rate of exploration</td>
<td>1.0</td>
</tr>
<tr>
<td>$\rho$</td>
<td>Update rate of the reward function</td>
<td>0.1</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>Discount factor</td>
<td>0.95</td>
</tr>
</tbody>
</table>
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

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

\[ R(S1, EM) = 1 \# \text{Reward value} \]
\[ \hat{R} = 0 \# \text{Reward rate} \]
\[ \hat{R}(s, a) = \{0, \ldots\} \# \text{Reward function} \]

Main Loop

FOR \(i = 1 : T\)

\(s_t = S_0 \# \text{Initial state}\)

IF \(i = T_{devaluation} \# \text{Moderate / Extensive training}\)

\[ R(S1, EM) = 0 \]
\[ \hat{R}(S1, EM) = -1 \]

WHILE \(s_t \neq S1 \land a_t \neq EM\)

\(a_t = \text{Selection}(s_t)\)

\(r_t = R(s_t, a_t)\)

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

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

Selection

# Sort the Q-values in descending order

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

# VPI : Value of Precise Information
\[
VPI(s_t, a_1) = (Q(s_t, a_2)^H - Q(s_t, a_1)^H) P(Q(s_t, a_1)^H < Q(s_t, a_2)^H) + \frac{\sigma(s_t, a_1)}{\sqrt{2\pi}} e^{-\frac{(Q(s_t, a_2)^H - Q(s_t, a_1)^H)^2}{2\sigma(s_t, a_1)^2}}
\]
\[
VPI(s_t, a_2) = (Q(s_t, a_1)^H - Q(s_t, a_2)^H) P(Q(s_t, a_1)^H > Q(s_t, a_2)^H) + \frac{\sigma(s_t, a_1)}{\sqrt{2\pi}} e^{-\frac{(Q(s_t, a_2)^H - Q(s_t, a_1)^H)^2}{2\sigma(s_t, a_1)^2}}
\]

FOR \( i \in \{a_1, a_2, \ldots, a_i, \ldots\} \)

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

# Q-Value from Goal-directed system is evaluated
\[
Q(s_t, a_i) = \hat{R}(s_t, a_i) + \gamma \sum_{s'} \mathcal{P}_T([s, a] \rightarrow s') \max_{b \in A} Q(s', b)_{Goal-directed}
\]

ELSE

# Q-Value from Habitual system is retrieved
\[
Q(s_t, a_i) = Q(s_t, a_i)^{Habitual}
\]
\[
a_t \leftarrow \text{SoftMax}(Q(s_t, a), \beta)
\]

Update
\[
\hat{R} = (1 - \sigma)\hat{R} + \sigma r_t \# \text{Reward Rate}
\]
\[
\hat{R}(s_t, a_t) = (1 - \rho)\hat{R} + \rho r_t \# \text{Reward function}
\]
\[
\mathcal{P}_T(s_t, a_t, s_{t+1}) = (1 - \phi)\mathcal{P}_T(s_t, a_t, s_{t+1}) + \phi \# \text{Probability of transition}
\]

Specific to Kalman Q-Learning

# Sigma-points sampling
\[
\Theta = \{\theta_j, 0 \leq j \geq 2|S.A|\}
\]
\[
\mathcal{W} = \{w_j, 0 \leq j \geq 2|S.A|\}
\]
\[
\hat{R} = \{\hat{r}_j = \theta_j(s_t, a_t) - \gamma \max_{b \in A} \theta_j(s_{t+1}, b), \ 0 \geq j \geq 2|S.A|\}
\]
\[
r_{predicted} = \sum_{j=0}^{2|S.A|} w_j \hat{r}_j
\]

# Covariance computation
\[
\mathcal{P}_{\theta_j r_j} = \sum_{j=0}^{2|S.A|} w_j(\theta_j - Q_t^{Habitual})(\hat{r}_j - r_{predicted})
\]
\[
\mathcal{P}_r = \sum_{j=0}^{2|S.A|} w_j(\hat{r}_j - r_{predicted})^2 + P_n
\]
\[
K_t = \mathcal{P}_{\theta_j r_j} \mathcal{P}_{\hat{r}_j}^{-1} \# \text{Kalman gain}
\]
\[
\delta_t = r_t - r_{predicted} \# \text{Reward-prediction error}
\]
\[
Q_t^{Habitual} = Q_t^H + K_t \delta_t
\]
\[
P_{t+1}^H = P_t^H - K_t P_{\Sigma} K_t^T
\]
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,

\[ p(s_0, pl) \approx 0.7 \]

and

\[ p(s_0, em) \approx 0.3 \]

before devaluation. In the original article,

\[ p(s_0, pl) \approx 0.6 \]

and

\[ p(s_0, em) \approx 0.4 \]

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

![Figure 1](image.png)

**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
