Meta-Reinforcement Learning in Homeostatic Regulation

Naoto Yoshida (yoshida.naoto.8x@kyoto-u.ac.jp)
Kyoto University, Kyoto, Japan

Abstract

The homeostasis of internal bodily states is essential for animal survival. In computational neuroscience, Homeostatically Regulated Reinforcement Learning (HRRL) has been proposed as a theoretical framework for modeling the learning of behavior in agents that maintain homeostasis through trial and error. HRRL assumes the existence of the dynamics within the agent and defines rewards based on its internal state. However, it remains unclear what kinds of behavioral learning are enabled by such internally defined rewards. In this study, we hypothesized that when dealing with such internally defined rewards, agents can acquire meta-reinforcement learning (meta-RL) capabilities by incorporating multimodal inputs and recurrent connections into the policy network architecture. Numerical experiments suggested that the proposed architecture enable the HRRL agent to acquire exploratory behaviors in the environment, indicating that meta-learning abilities comparable to those found in previously known meta-RL approaches can be achieved using different architectures.

Keywords:Homeostatic Regulation; Reinforcement Learning; Meta Reinforcement Learning; Deep Reinforcement Learning

Introduction

Maintaining internal bodily states within an appropriate range—*homeostasis*—is a critical ability for animal survival (Hull, 1943; Ashby, 1952). While reinforcement learning (RL) models in decision-making serve as powerful frameworks for modeling cognitive processes, *rewards* in these models are typically introduced as free parameters. How such rewards are grounded in autonomous agents, such as animals, remains an ongoing topic of discussion (Juechems & Summerfield, 2019; Yee, 2024). *Homeostatically regulated RL* (HRRL) is a framework proposed in computational neuroscience that models the learning of animal behavior based on motivation derived from homeostatic regulation (Keramati & Gutkin, 2011). As in standard RL, HRRL models the interaction between the agent and its environment as a partially observable Markov decision process (Kaelbling, Littman, & Cassandra, 1998). The agent receives a (possibly multimodal) observation x from the environment and then selects an action a according to its policy π. This action causes the environment to transition to a new state, after which the agent receives a new observation x' and a scalar reward $r \in \mathbb{R}$. The objective of RL is to obtain an optimal policy π^* that maximizes the expected cumulative future reward $\sum_{t=0}^{\infty} \gamma^t r_t$, based on the agent's experience of interacting with the environment. Here, $\gamma < 1$ is a positive discount factor that reduces the weight of future rewards. Unlike standard RL, HRRL assumes that the agent possesses a body and receives *interoceptive* input x^i, which reflects its internal bodily states. Rewards in HRRL are computed based on the time series of x^i. A drive function $D(x^i) > 0$ is introduced to evaluate the desirability of internal states. This function is typically defined based on the agent's physiological characteristics and measures the distance between a fixed *set point* x^i_* and the current interoception x^i. Accordingly, HRRL defines the reward to be proportional to the change in the drive function: $r_{t+1} = k\left[D(x^i_t) - D(x^i_{t+1})\right]$, where k is a positive constant.

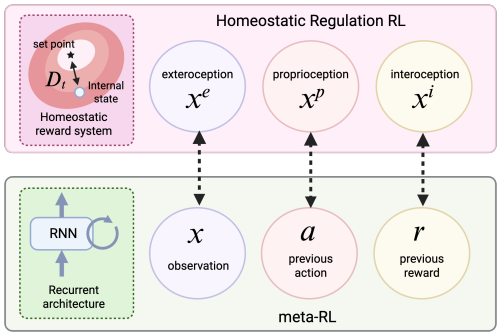

Figure 1: Correspondence between HRRL and meta-RL.

Minimal Architecture for meta-RL in HRRL

Wang et al. proposed a form of meta-reinforcement learning (meta-RL) in the context of deep reinforcement learning (deep RL), in which specific neural architectures induce advanced cognitive processes—such as exploration and one-shot learning—within the network, enabling the agent to retain adaptive capabilities to the environment even after training (Wang et al., 2016, 2018). In this work, we focus on the potential correspondence between meta-RL and HRRL (Figure 1). Specifically, meta-learning capabilities in meta-RL require access to external observations x_t, the most recent action selection a_{t-1}, the most recent reward r_{t-1}, and a recurrent connection within the architecture. These multimodal observations are thought to correspond, respectively, to exteroception x^e, proprioception x^p, and interoception x^i in HRRL (Yoshida, Daikoku, Nagai, & Kuniyoshi, 2024). Because the reward is defined as a function of the interoceptive sequence, x^i implicitly encodes information about the reward signal. Therefore, incorporating a recurrent layer into the model architecture of an HRRL agent may be sufficient to induce meta-learning capabilities by implicitly acquiring the mapping from interoception to the most recent reward.

Experiments

To examine our hypothesis, we conducted computational experiments using a simple non-stationary bandit task designed to maintain the agent's homeostasis (Figure 2). In this experiment, the environment requires the agent to sustain home-

ostasis in a one-dimensional energy state by consuming food from a multi-armed bandit with three non-stationary arms. If the agent selects the arm containing food, it receives the food with a probability of 0.9; otherwise, no food is provided. The location of the food is assumed to change with a small probability ($p = 0.01$). The agent monitors its own energy state through interoception x^i. The energy state is updated according to the following simple metabolic dynamics: $x^i_{t+1} = x^i_t - \alpha + \beta I_t$.

Here, $\alpha = 0.05$ represents the energy change due to the agent's metabolism. $\beta = 0.12$ is a constant representing the energy inflow when the agent consumes food. I_t is an event function that takes the value of 1 when food is consumed and 0 otherwise. The drive function was defined as $D(x^i) = (x^i - x^i_*)^2$, with the set point as $x^i_* = 0$. We used $k = 100$ to define the homeostatic reward. Furthermore, to facilitate the visualization of behavior, a small constant penalty was applied to all actions except the resting action. The maximum survival step in the environment was set to 2,000 steps. If the energy level exceeded the range $|x^i| < 1$, both the environment and the agent were reset (death). If the agent reached the maximum number of time steps, the environment was also reset. For comparison, the network's weight parameters were fixed at regular learning step intervals, and the survival time (episode length) was measured during test trials.

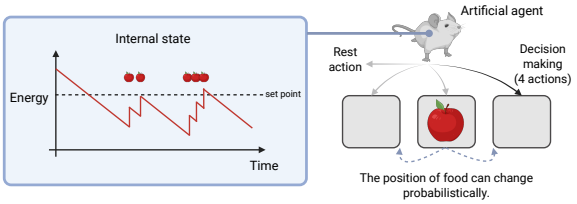

Figure 2: Overview of the experiment.

Agent architecture The proposed model architecture (meta-HRRL) consisted of a fully connected layer with a ReLU activation function (Krizhevsky, Sutskever, & Hinton, 2012) for input embedding, followed by a recurrent layer implemented with LSTM (Hochreiter & Schmidhuber, 1997). The LSTM's hidden state was mapped to produce the policy π and a value prediction V_π. The agent was trained using PPO algorithm (Schulman, Wolski, Dhariwal, Radford, & Klimov, 2017). As a variation of the network structure, we compared performance using a sequence of interoceptive inputs, $\bar{x}^i \triangleq [x^i_{t-K+1}, \ldots, x^i_t]$, with $K = 1$ or 4, resulting in the final observation defined as obs $\triangleq [\bar{x}^i_t, x^p_t]$, where proprioception is the previous action $x^p_t = a_{t-1}$[1]. As a baseline for HRRL without meta-RL capability, we also evaluated a structure that excluded proprioception, i.e., obs $\triangleq [x^i_t]$.

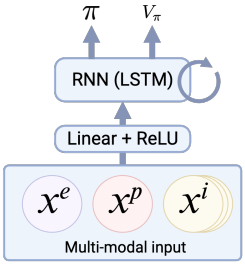

Figure 3: Proposed architecture

[1] In the case of HRRL, exteroception is not present in this task.

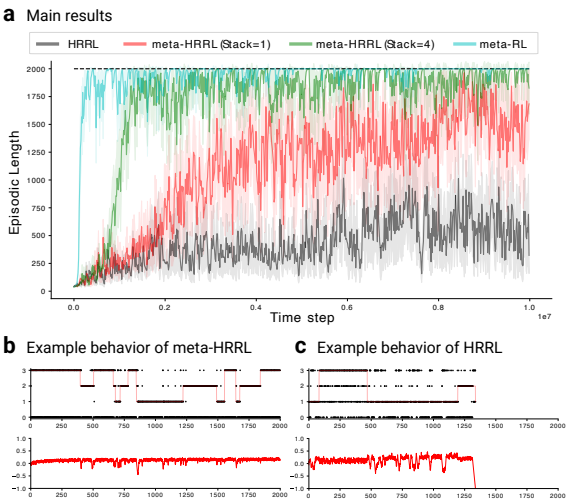

Figure 4: Overview of the results. Results in the panel **a** are the average of 20 trials and 95% confidence interval.

Additionally, we compared the performance of a conventional meta-RL agent, which received the observation obs $\triangleq [x^i_t, a_{t-1}, r_{t-1}]$, as a performance upper bound.

Results

Figure 4a illustrates the test performance in terms of episode length. As expected, the simple HRRL agent failed to achieve long-term survival in the environment. In contrast, the meta-RL architecture quickly reached the maximum episode length, indicating the acquisition of efficient exploratory behaviors. The results of our proposed architecture demonstrated superior performance compared to HRRL, regardless of whether the stack size was set to $K = 1$ or 4. This supports our hypothesis that the agent successfully acquired meta-RL capabilities. Additionally, a larger stack size yielded better results. This suggests that increasing the stack size facilitates the acquisition of information equivalent to the reward signal, as homeostatic rewards are inherently evaluated based on interoception over two time steps. Figures 4b and 4c show examples of the behavior of meta-HRRL ($K = 4$) and HRRL agents, respectively. As demonstrated in the top panel, the meta-HRRL agent adapts to the changing food location (pale solid line) and adjusts its action selection (black dots), successfully controlling its internal state even in a dynamic environment (bottom panel). In contrast, the simple HRRL agent is able to survive only when food happens to appear consistently in a specific arm.

Concluding Remark and Discussion

This study experimentally demonstrated that, in HRRL, agents with multimodal inputs and recurrent structures can acquire meta-RL capabilities. The results also indicated that the conventional meta-RL architecture achieves higher learning efficiency. Humans are sometimes capable of perceiving rewarding stimuli as conscious experiences (Berridge & Kringelbach, 2008). This ability to perceive rewarding stimuli as part of a multimodal observation may contribute to enhanced learning efficiency in situations where advanced cognitive behaviors, such as exploration, must be acquired through learning.

Acknowledgments

This research is supported by Japan Society for the Promotion of Science KAKENHI grant 24K23892. Figure 1-3 were created using BioRender.com.

References

Ashby, W. R. (1952). *Design for a brain.* Wiley.

Berridge, K. C., & Kringelbach, M. L. (2008). Affective neuroscience of pleasure: reward in humans and animals. *Psychopharmacology*, *199*, 457–480.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735–1780.

Hull, C. L. (1943). *Principles of behavior: An introduction to behavior theory*. New York: Appleton-Century-Crofts.

Juechems, K., & Summerfield, C. (2019). Where does value come from? *Trends in cognitive sciences*, *23*(10), 836–850.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial intelligence*, *101*(1-2), 99–134.

Keramati, M., & Gutkin, B. S. (2011). A reinforcement learning theory for homeostatic regulation. In *Advances in neural information processing systems* (pp. 82–90).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097–1105.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., ... Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, *21*(6), 860–868.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., ... Botvinick, M. (2016). Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*.

Yee, D. M. (2024). Neural and computational mechanisms of motivation and decision-making. *Journal of Cognitive Neuroscience*, *36*(12), 2822–2830.

Yoshida, N., Daikoku, T., Nagai, Y., & Kuniyoshi, Y. (2024). Emergence of integrated behaviors through direct optimization for homeostasis. *Neural Networks*, *177*, 106379.