An Empirical Study of Different Intrinsic Rewards on a Partially Observable Multiple-Prediction Learning Problem

Animesh Kumar Paul[*]
Department of Computing Science
University of Alberta
Edmonton, AB, Canada
animeshk@ualberta.ca

Abstract

Learning multiple goal-directed value functions can be beneficial to improve exploration and representation learning in a reinforcement learning setting. For learning multiple value functions, we can estimate the intrinsic reward for each independent learner and then maximize that reward using reinforcement learning. As selecting a specific intrinsic reward is difficult for a particular problem setting, Linke et al. [2019] used different intrinsic rewards and made a comparison among those rewards in terms of the agent's learned behavior for each intrinsic reward. In this project, my goal is to investigate and compare different intrinsic rewards for a bandit-like parallel-learning testbed and check whether I could recreate the results of Linke et al. [2019].

1 Introduction

Learning multiple goal-directed value functions in reinforcement learning can generalize to previously unseen goals by generating new policies (Schaul et al. [2015]). Linke et al. [2019] focused on using different independent learners for each value function to learn those value functions accurately when we do not have any external reward as feedback from the environment. To solve this problem, they used intrinsic reward to estimate the system's total learning across all learners. The agent's action selection in each state is adjusted during the learning process by maximizing that intrinsic reward.

Over the past decade, researchers have designed different intrinsic rewards. Authors of Linke et al. [2019] considered 10 intrinsic rewards to check which ones would be best for the learning system on their designed bandit-like parallel learning testbed. They provided a comprehensive empirical comparison of different intrinsic rewards for that testbed. The testbed comprises a single state with multiple actions, and a separate prediction learner is used for estimating the independent target. Here, each target is associated with a different action.

My work aims to recreate the results of Linke et al. [2019] for the Switched Drifter-Distractor problem. From Section 2- 4, I briefly discuss a few ideas: gradient bandit algorithm, two types of learners (non-introspective and Introspective learners), and finally intrinsic rewards. Section 5 presents the experimental results of 10 intrinsic rewards for each type of learner.

2 Gradient Bandit Algorithm for Prediction Learning

Gradient Bandit algorithm is one of the methods to solve multi-armed bandit problems, and it is based on the idea of Stochastic Gradient Ascent (Sutton and Barto [2018]). Its objective function is

[*] Website: https://animesh10kuet.wixsite.com/animeshkumarpaulcse

to maximize the expected reward to learn some preference $H_t(a)$ for each action a over and above the other available actions at that time point. $H_t(a)$ helps the agent select the action with a higher preference value. The action probabilities are computed by using soft-max distribution (refer to eq. 1, here N = the number of available actions).

$$Pr(A_t = a) \doteq \frac{e^{H_t(a)}}{\sum_{b=1}^{N} e^{H_t(b)}} \tag{1}$$

The $H_t(a)$ value is updated using stochastic gradient descent. The equation is given in 2, here $\bar{r}$ is the reward average over t time points, and the initial values of $\bar{r}$ and $H_0(a)$ are both zero - all actions are equally likely to be selected initially. If the reward R_{t+1} is higher than the $\bar{r}$, then the probability of selecting the action A_t in the future is increased, and if the reward R_{t+1} is lower than the $\bar{r}$, then the probability is decreased. The gradient bandit algorithm helps to balance exploration and exploitation (Sutton and Barto [2018]).

$$H_{t+1}(a) \leftarrow \begin{Bmatrix} H_t(a) + \alpha(R_{t+1} - \bar{r})(1 - \pi_t(a)) \quad if \quad A_t = a; \\ H_t(a) - \alpha(R_{t+1} - \bar{r})\pi_t(a) \quad otherwise. \end{Bmatrix} \tag{2}$$

3 Non-introspective Learners vs. Introspective Learners

The authors of Linke et al. [2019] paper considered two types of learners in their experiments: introspective and non-introspective learners. A learner will be introspective if it can adjust its learning rate without external help. In other words, this type of learner should be able to stop its update process if there is no way to make any progress in learning. On the other hand, a non-introspective learner continues its update disregarding the learning progress, and uses a constant step-size parameter to estimate the sample targets on each time point. However, choosing the value of constant step size is a significant problem because if we have an extensive step-size parameter for the high-variance target, then we will see the significant updates by the learner due to the high sample variance, which leads to high prediction error, and if we use the smaller step-size for the tracking target, then the learning will be slower resulting high prediction error.

To handle the issues of non-introspective learners, Linke et al. [2019] combined AutoStep method with the prediction learner that is called introspective learner. Autostep method performs well in a non-stationary environment, incremental, and online tracking settings. Autostep method has its own one hyper-parameter, named the meta-learning rate, which supervises the rate of changes in the step-size parameter: the value of step-size is increased when the learner's prediction accuracy improves and decreases when the model learning is not improving.

4 Intrinsic Rewards

Without direct feedback from the environment, intrinsic rewards encourage the agent to develop a specific behavior. In this work, I have used 10 different intrinsic rewards that are suggested by Linke et al. [2019]. Prediction error-based intrinsic rewards are Squared Error, Expected Error, and Unexpected Demon Error. Learning progress-based intrinsic rewards are Error Reduction, Error Derivative, Bayesian Surprise, Variance of Prediction, Uncertainty Change, and Weight Change. Along with these, Step-size Change is used for only Introspective Learners. The specific equation and details of each intrinsic reward can be found in Linke et al. [2019].

5 Experiments

5.1 Problem and Experimental Setup

In this experiment, I have used the Switched Drifter-Distractor problem, which was designed by Linke et al. [2019]. This problem has four targets: a) a constant target, b) a drifting target, and c) two high-variance targets. These four targets are each associated with a different action. This problem has two phases: phase 1 contains 50,000 time-steps, and phase 2 contains 100,000 time-steps - phase 2 starts after phase 1. In phase 2, targets and actions associations are interchanged (the information

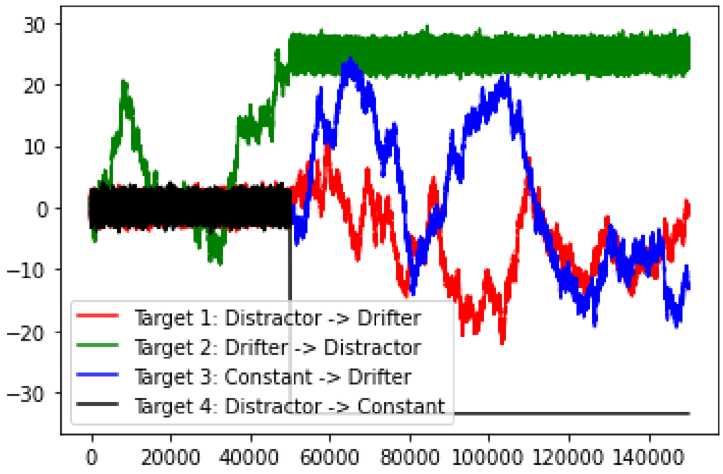

Figure 1: A sample generated targets of Switched Drifter-Distractor problem.

about it is given in Table 1, and a sample generated targets of this problem is given in Figure 1). More detailed information on this problem can be found in Linke et al. [2019].

target	Phase 1	Phase 2
target 1	Distractor	Drifter
target 2	Drifter	Distractor
target 3	Constant	Drifter
target 4	Distractor	Constant

Table 1: Target distributions in Switched Drifter-Distractor problem.

I want to assess the effectiveness of different intrinsic rewards in the partially observable environment. The desired behavior of the agent will be as follows: Initially, the agent should test all four actions, but over time the agent should reduce its prediction error for the constant target, and then it should select the action corresponding to the high-variance target. When the prediction error for the high-variance target goes down to zero, the agent should start choosing the action corresponding to the drifting targets. As the agent would not be able to reduce the error of predicting the drifting targets to zero, it should select the corresponding action of the drifting targets most of its time (Linke et al. [2019]).

The learning system has several hyper-parameters for tuning to obtain the best performance. There are over 50,000 parameter configurations for the learning model with 10 intrinsic rewards (the full list of hyper-parameters can be found in Linke et al. [2019]). Due to the time constraint and computational resource limitation, I have not been able to test all of the hyper-parameter configurations. Among all of my used parameter configurations, I have found the best-performing parameters by minimizing the average of root mean squared error (RMSE) over 200 independent runs. Using those best-performing parameters, I report the best performance for each intrinsic reward by plotting the probability of selecting each action over time according to the agent's policy.

5.2 Results with Non-introspective Learners

Figure 2 shows the behavior of non-introspective learners for different intrinsic rewards. These results are generated by averaging the action selection probability over 200 runs for each intrinsic reward.

- **For phase 1 (up to 50,000 time-steps)**: From this experiment, I found that the Gradient Bandit agent with intrinsic reward based on violated expectations, including Expected Error, UDE, and Squared Error, selects primarily the action corresponding to the drifting target. Moreover, we can also see that during phase 1, the agent's behavior is to focus on the drift action quickly -this is the desired behavior from our agent- when using Error

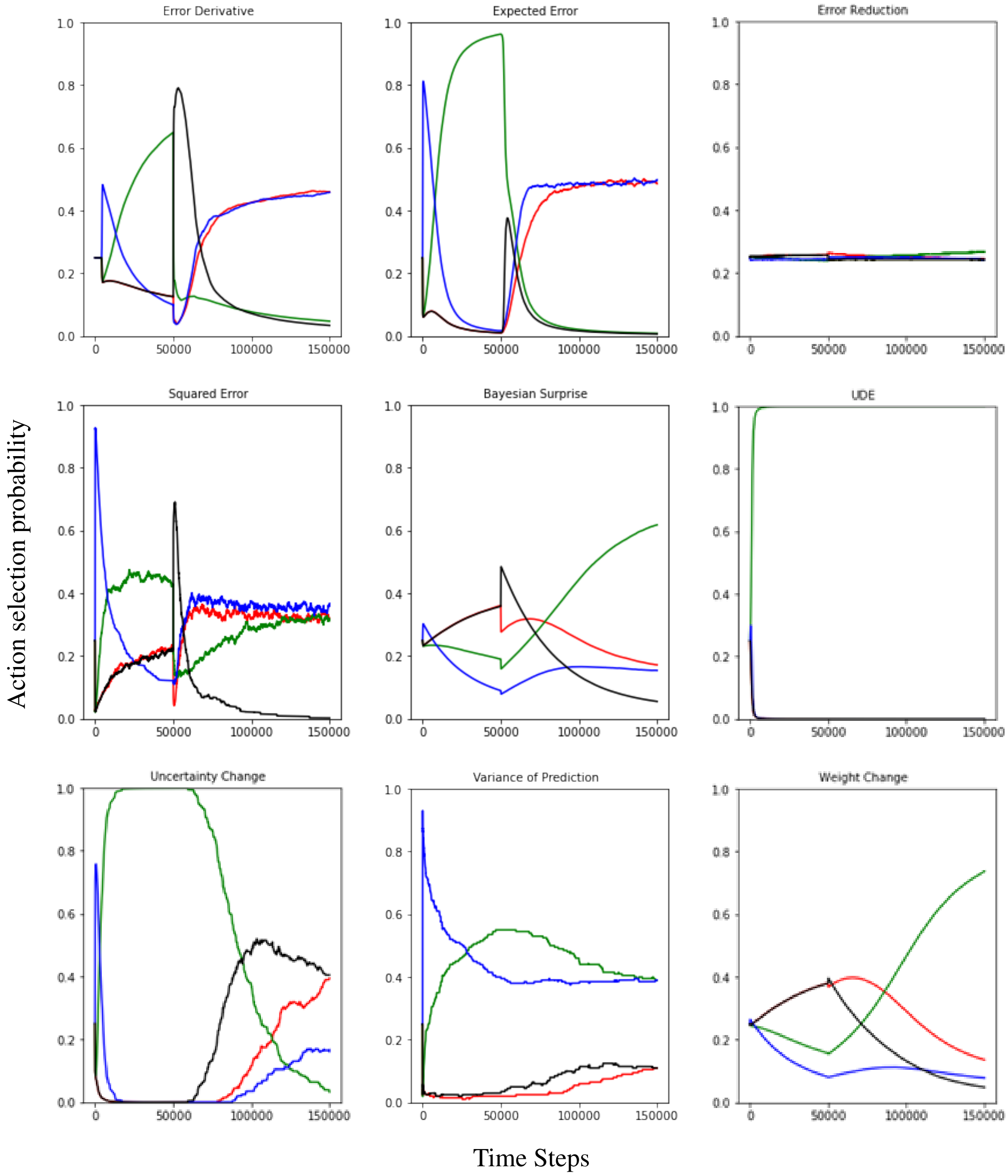

Figure 2: Resultant behavior of Non-introspective Learners.

Derivative or Uncertainty Change reward. Unfortunately, Bayesian Surprise and Weight Change rewards-based agents are stuck in selecting the action for the high-variance target, which is a sub-optimal behavior. But, the Error Reduction and Variance of Prediction rewards create unseemly behavior.

- **For phase 2 (from 50,001 to 150,000 time-steps)**: Error Derivative, Expected Error, and Squared Error rewards-based agents show the desired behavior - selecting the drift actions most of the time. Bayesian Surprise and Weight Change rewards cause the agent to choose the action corresponding to the high variance targets, which is similar to the results of phase 1. During phase 2, an agent with Uncertainty Change reward reduces the action selection of the drifting target and provides more focus on selecting the actions of high variance targets. UDE, Error Reduction, and Variance of Prediction generate inappropriate behavior.

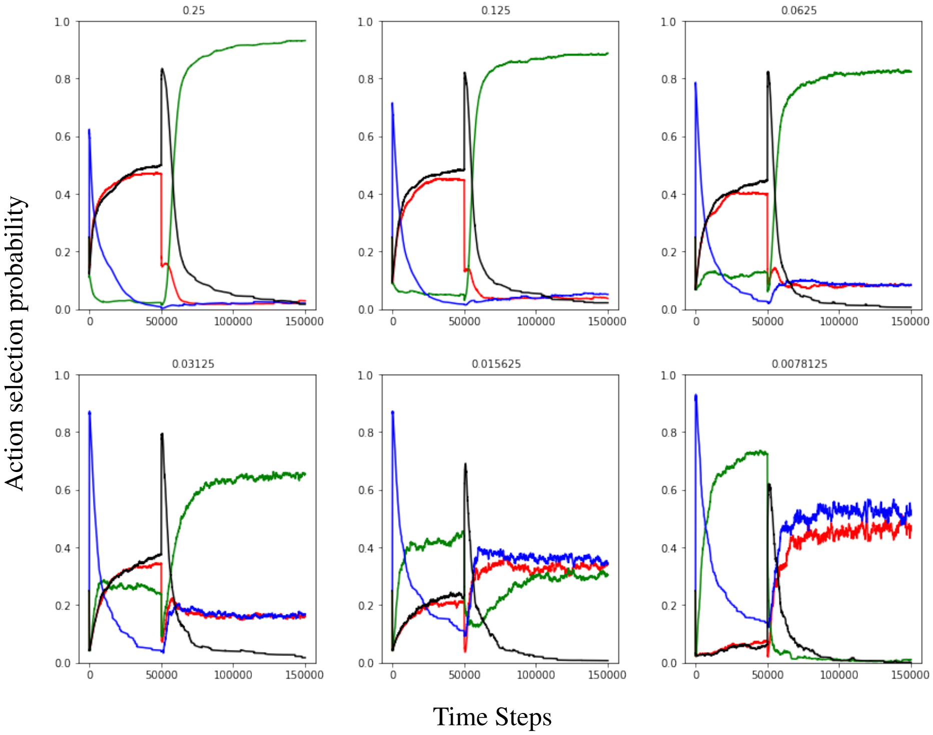

Figure 3: The impact of varying the step-size parameter with Squared-Error reward.

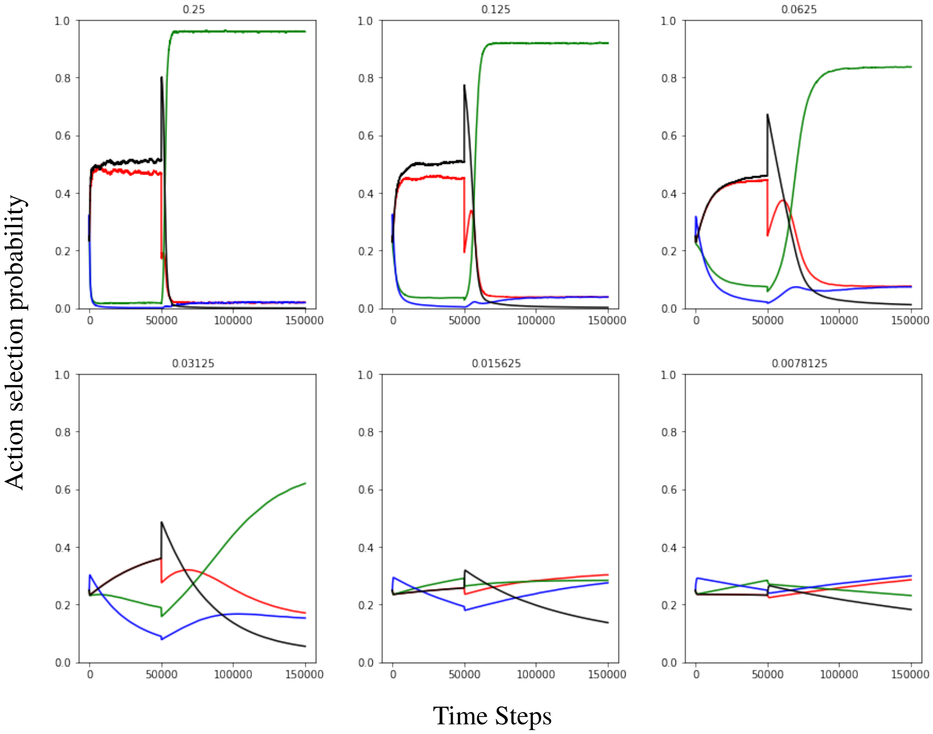

Figure 4: The impact of varying the step-size parameter with Bayesian Surprise reward.

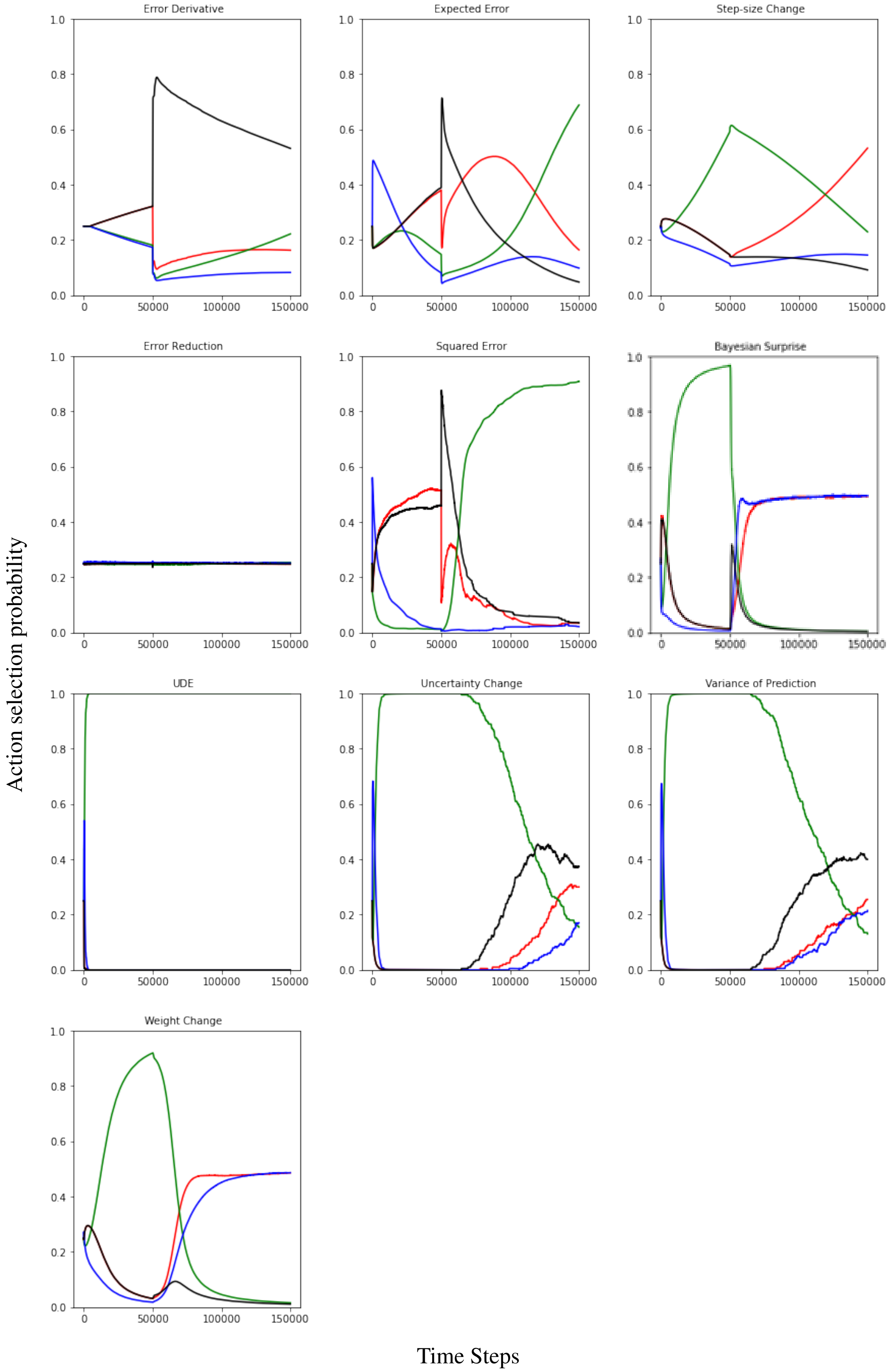

Figure 5: Resultant behavior of Introspective Learners.

Figure 3 shows the effect of step-size parameter α_p for Squared Error reward. We can see that if the α_p is small, then the agent's behavior is more in line with our expectations, but when $\alpha_p = 0.0078125$, the RMSE is 1.3245, compared to a RMSE of 1.006 when $\alpha_p = 0.25$. Figure 4 shows similar characteristics for Bayesian Surprise reward- when $\alpha_p = 0.0078125$, the RMSE is 1.4472, compared to a RMSE of 1.0701 when $\alpha_p = 0.25$.

5.3 Results with Introspective Learners

Figure 5 presents the behavior of the Introspective agent, which uses the Autostep method to adjust its own learning step-size parameter. Here, to highlight the agent's behavior, I plot the average of action selection probabilities over 200 independent runs.

- **For phase 1 (up to 50,000 time-steps)**: Agents with UDE and Uncertainty Change with or without Autostep show similar behavior of mainly selecting the actions with the drifting targets (refer to Figure 2 and 5). Although the non-introspective agents with Bayesian Surprise, Weight Change, and Variance of Prediction rewards cannot choose the preferred action, the introspective agents can select drift action more often than other actions. Additionally, the Step-size Change reward also chooses the desired action. But, Error reduction produces uniform action selection.
- **For phase 2 (from 50,001 to 150,000 time-steps)**: Introspective agents with Weight Change and Bayesian Surprise rewards select our expected action compared to the non-introspective agents. Step-size Change, Uncertainty Change, and Variance of Prediction rewards encourage the agent to reduce the frequency of selecting the constant target as well as increase the rate of action selection for the drifting targets. But, Error Derivative, Expected Error, and Squared Error do not show the preferred behavior. UDE provides unfavorable results with or without Autostep.

6 Conclusion

In this work, I have tried to recreate the Linke et al. [2019] results for the Switched Drifter-Distractor problem. The experimental results showed that an introspective learner could produce effective behavior with the amount of learning measure based intrinsic reward, such as Weight Change, as this learner can adjust its own learning rate. Although for most of the intrinsic rewards, the behavior of the non-introspective or introspective agents is consistent with the given results in Linke et al. [2019], some intrinsic rewards, such as Error Derivative and UDE, show different behaviors in phase 2 of the experiment.

I acknowledge that this work is part of the course project on Reinforcement learning at the University of Alberta, and I have tried to contact the first two leading authors of this paper to get their best performing hyper-parameters for reproducing their results, but unfortunately, my attempts were not successful. And, due to the time constraint and computational resource limitation, I have not been able to check every possible hyper-parameters combination for the detailed investigation of the 10 different intrinsic rewards' effectiveness in the given environmental settings. One of the reasons behind these few inconsistent results between this work and Linke et al. [2019] might be the lack of considering all of the combinations of hyper-parameters.

References

Cam Linke, Nadia M. Ady, Martha White, Thomas Degris, and Adam White. Adapting Behaviour via Intrinsic Reward: A Survey and Empirical Study. *Journal of Artificial Intelligence Research*, 69:1287–1332, jun 2019. ISSN 10769757. doi: 10.1613/JAIR.1.12087. URL `https://arxiv.org/abs/1906.07865v4`.

Tom Schaul, Dan Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *32nd International Conference on Machine Learning, ICML 2015*, volume 2, 2015.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An Introduction (2nd edition 2018)*, volume 3. 2018. ISBN 0262193981. URL `https://books.google.com/books?id=CAFR6IBF4xYC&pgis=1%5Cnhttp://incompleteideas.net/sutton/book/the-book.html%5Cnhttps://www.dropbox.com/s/f4tnuhipchpkgoj/book2012.pdf`.

A Appendix

I have provided a python notebook of this experiment: `https://github.com/animeshkumarpaul/An-Empirical-Study-of-Different-Intrinsic-Rewards-on-a-Partially-Observable-Multiple-Prediction-Lear`