Investigating Curiosity for Multi-Prediction Learning

Cameron Linke
University of Alberta
clinke@ualberta.ca

Nadia M. Ady
University of Alberta
nmady@ualberta.ca

Thomas Degris
DeepMind London
degris@google.com

Martha White
University of Alberta
whitem@ualberta.ca

Adam White
University of Alberta
DeepMind Edmonton
adamwhite@google.com

Abstract

This paper investigates a computational analog of curiosity to drive behavior adaption in learning systems with multiple prediction objectives. The primary goal is to learn multiple independent predictions in parallel from data produced by some decision making policy—learning for the sake of learning. We can frame this as a reinforcement learning problem, where a decision maker's objective is to provide training data for each of the prediction learners, with reward based on each learner's progress. Despite the variety of potential rewards—mainly from the literature on curiosity and intrinsic motivation—there has been little systematic investigation into suitable curiosity rewards in a pure exploration setting. In this paper, we formalize this pure exploration problem as a multi-arm bandit, enabling different learning scenarios to be simulated by different types of targets for each arm and enabling careful study of the large suite of potential curiosity rewards. We test 15 different analogs of well-known curiosity reward schemes, and compare their performance across a wide array of prediction problems. This investigation elucidates issues with several curiosity rewards for this pure exploration setting, and highlights a promising direction using a simple curiosity reward based on the use of step-size adapted learners.

Keywords: Reinforcement Learning, Curiosity, Online Learning

Acknowledgements

We are very grateful to the Alberta Machine Intelligence Institute (Amii), the Reinforcement Learning and Artificial Intelligence lab (RLAI), and DeepMind for their support of our work.

1 Introduction

We consider the case of a lifelong learning agent that receives one stream of experience with many different incoming streams of data (e.g. distance sensors, camera, battery state, etc.). This type of agent is outlined in the Horde architecture (Sutton et al., 2011) where a robotic agent learns about many different sensorimotor streams, off-policy, in parallel, while interacting with its environment. This architecture has been shown to scale up to making thousands of predictions at once, giving the agent a rich prediction-based understanding of its environment. Other architectures such as Universal Value Function Approximators (Schaul et al., 2015) and UNREAL (UNsupervised REinforcement and Auxiliary Learning) (Jaderberg et al., 2016) have shown that in complex environments with deep neural networks as function approximators, additional predictions like these are powerful in improving the ultimate performance of the agent. The extra predictions were used to guide the agent to new areas of the state space, or to add richness to a sparse reward signal. Here we consider how the agent's behaviour changes the quality of its predictions of the signals themselves, in contrast to using the predictions as a target to guide behaviour appropriate to a separate task.

For an agent that is able to learn about many things at once, the challenge remains—how does an agent who gets only one stream of experience decide what to learn, and how long to learn it for? Thought of another way—given a number of target policies, what is the best behaviour policy for the agent to learn to predict and control the streams of data that it is receiving? Prior work on this problem has either used a hand-crafted policy designed to explore the environment in a specific way for the learners (Sutton et al., 2011; Modayil et al., 2014), or had the agent target maximizing one specific reward signal and learning about other signals at the same time (Jaderberg et al., 2016; Schaul et al., 2015; Riedmiller et al., 2018). A number of authors have proposed special reward signals to motivate agent learning; herein we call such signals *curiosity rewards*.

While our ultimate goal is to get to a full Horde setting, it remains a complex environment for testing different types of curiosity reward signals to drive agent behavior. The function approximation used by the agent means that we need to consider representation learning in addition to learning to predict the signal itself. The off-policy nature of the signals being learned also complicates the agent as it brings in choices around the type of learners to use, and further need for sweeps over a larger number of parameters. To specifically study the mechanisms driving the curious agent, we want to be able to remove as many of these other complicating factors as possible and focus on the mechanisms themselves.

This paper has three contributions. First we propose a new benchmark for evaluating curiosity-driven learning agents. We propose using a bandit setting to focus on the single learning decision that the agent needs to take at each time step—what action can I take that will maximize learning? This is different from the traditional exploration/exploitation trade-off given by the bandit literature. We are instead focused on maximizing the long-term learning of the agent, not the long-term reward. The setup of our bandit is to have multiple arms each with their own Least Mean Square (LMS) learner predicting the value of the arm. Each arm's learner can be thought of as a single prediction about a sensorimotor stream, and the agent can take an action that will teach it more about that stream of data. This setup models the situation that is faced by an agent in the Horde setting—at each step it is learning about a set of predictions, and a curious agent will want to take an action that allows it to learn the most. The bandit is set up to model the types of challenges a curious agent will face—how does an agent learn to best improve its predictions while avoiding getting stuck on unlearnable signals? Mechanisms that are effective at driving curious behavior in this setting should be effective in the larger setting. Alternatively, mechanisms that are not able to solve this distilled setting will not be effective in the larger Horde setting.

Second, we propose a comprehensive empirical comparison of current approaches to curiosity on our proposed benchmark. We compare many of the well-known methods from reinforcement learning and active learning. Much recent work has been in large-scale domains where it is difficult to tease out which of the many moving parts is driving the benefit of the method. In this comparison we distill the methods down to the specific mechanism driving curious behavior and compare them in our focused domain. This allows us to both compare how the mechanisms perform against each other, as well as evaluate the specific properties of how each mechanism deals with the problems each environment tests. While we did not compare every curiosity method that exists, we did make sure to have a representative method from each of the different classes we identified as grouping similar underlying mechanisms.

Third, we show the effectiveness of capable learners in directing curious behavior. Capable learners, such as step-size adaptation methods, are able to quickly adapt to unlearnable targets, such as those with high variance. This allows the agent to avoid unlearnable signals that may seem interesting due to their high error, but ultimately are not worth the agent continuing to spend time trying to learn. In this work, we show one particular curiosity reward—Weight Change—becomes one of the best curiosity rewards in this setting when paired with capable learners. While we examine capable learners in a simplified setting with LMS learners, nothing precludes these results from being moved back to the Horde setting. The more general case of off-policy learners with function approximation also have step-size adaptation methods (Jacobsen et al., 2019; Kearney et al., 2018) that allow capable learners to be applied.

2 Related Work

Many learning systems draw inspiration from the exploratory behavior of young humans and animals, uncertainty reduction in active learning, and information theory—and the resulting techniques could all be packed into the suitcase of curiosity and intrinsic motivation. In an attempt to distill the key ideas and perform a meaningful yet inclusive empirical study, we consider only methods applicable to our bandit formulation of multi-prediction learning. As such, we ignored work related to curriculum learning (Graves et al., 2017), methods that rely on predictions about state (e.g., (Pathak et al., 2017)), or traditional bandit exploration methods.

Several curiosity rewards are based on **violated expectations**, or surprise. This notion can be formalized using the error in some prediction (of a signal or multiple signals) to compute the instantaneous *Absolute Error* or *Squared Error*. We can obtain a less noisy measure of violated expectations with a windowed average of the error, which we call *Expected Error*. Regardless of the specific form, if the error increases, then the curiosity reward increases, encouraging further sampling for that target. Such errors can be normalized, such as was done for *Unexpected Demon Error* (White et al., 2014), to mitigate the impact of the differing magnitudes of and noise in the targets.

Another category of methods focus on **learning progress**, and assume that the learning system is capable of continually improving its policy or predictions. For approaches designed for tabular stationary problems, this is trivially true: (Chentanez et al., 2005; Still and Precup, 2012; Little and Sommer, 2013; Meuleau and Bourgine, 1999; Barto and Şimşek, 2005; Szita and Lőrincz, 2008; Lopes et al., 2012). The most well-known approaches for integrating intrinsic motivation make use of rewards based on improvements in (model) error: including *Error Reduction* (Schmidhuber, 1991, 2008), and Oudeyer's model *Error Derivative Change* approach (Oudeyer et al., 2007). Improvement in the value function can as be used to construct rewards, and can be computed from the *Positive Error Part* (Schembri et al., 2007), or by tracking improvement in the value function over all states (Barto and Şimşek, 2005).

An alternative to learning progress is to reward **amount of learning**. Such rewards do not penalize errors becoming worse, and instead only measure that estimates are changing: the prediction learner is still adjusting its estimates and so is still learning. *Bayesian Surprise* (Itti and Baldi, 2006) formalizes the idea of amount of learning. For a Bayesian learner, which maintains a distribution over the weights, Bayesian Surprise corresponds to the KL-divergence between this distribution over parameters before and after the update. Other measures based on information gain have been explored (Still and Precup, 2012; Little and Sommer, 2013; Achiam and Sastry, 2017; de Abril and Kanai, 2018; Still and Precup, 2012), though they have been found to perform similarly to Bayesian Surprise (Little and Sommer, 2013). Note that several learning progress measures, discussed in the previous paragraph, can be modified to reflect amount of learning by taking the absolute value, and so removing the focus on increase rather than change. We can additionally consider non-Bayesian strategies for measuring amount of learning, including those based on how much the variance in the prediction changes—*Variance of Prediction*.

3 Experimental Design

To simulate a curious learning problem that an agent may face we introduce the drifter-distractor environment. The environment is modelled in the bandit setting and has four arms, two which respond with random noise, one which slowly drifts, and one which returns a constant. This environment simulates a common situation for learning agent: having distracting signals that will produce large prediction errors, but are ultimately unlearnable.

The ideal behaviour of a curious learning agent in this environment is to first test out all of the arms, getting a sense of what each of the signals is. After it has learned a bit about each signal it should begin to hone in on the signals that it is poor at predicting. This will initially lead it to pulling the noisy arms a fair amount. The agent should then fairly rapidly move away from pulling the noisy arms, instead focusing on the drifter arm as it has a signal that is learnable, but needs to continually be tracked to keep its predictions accurate. To measure the effectiveness of each agent we can look at the prediction error of each of its learners.

Capable learners are able to adapt to the the learnability of a signal. Previous work has not investigated the impact of capable learners in a multi-prediction setting. Here step-size adaptation methods achieve this learnability adaptation by adjusting the step size up or down. We can see the benefits on the drifter-distractor problem. Capable learners allow us to use the step size to make the noisy signals less interesting to our agent. We use Autostep (Mahmood et al., 2012) to update the agent's step size online where the reward bonus of Weight Change is given by:

$$\|w_t - w_{t+1}\|_1 = \alpha_{t,i}\|\hat{y}_{t,i} - \hat{y}_{t-1,i}\|_1 = \alpha_{t,i}|\delta_{t,i}| \tag{1}$$

Where $w_t \in \mathbb{R}$ is the weight at time t, $\alpha_{t,i} \in \mathbb{R}$ is the step size of arm i at time t, $\hat{y}_{t,i} \in \mathbb{R}$ is the prediction of arm i at time t, and $\delta_{t,i}$ is the prediction error for the agent's prediction of arm i at time t. We can see clearly that a step-size adaptation method allows us to temper the error signal by driving down the step size in unlearnable situations. This property permits us to give a curiosity reward that allows the agent to make learning progress based on the error, because it only affects the reward if the prediction is learnable.

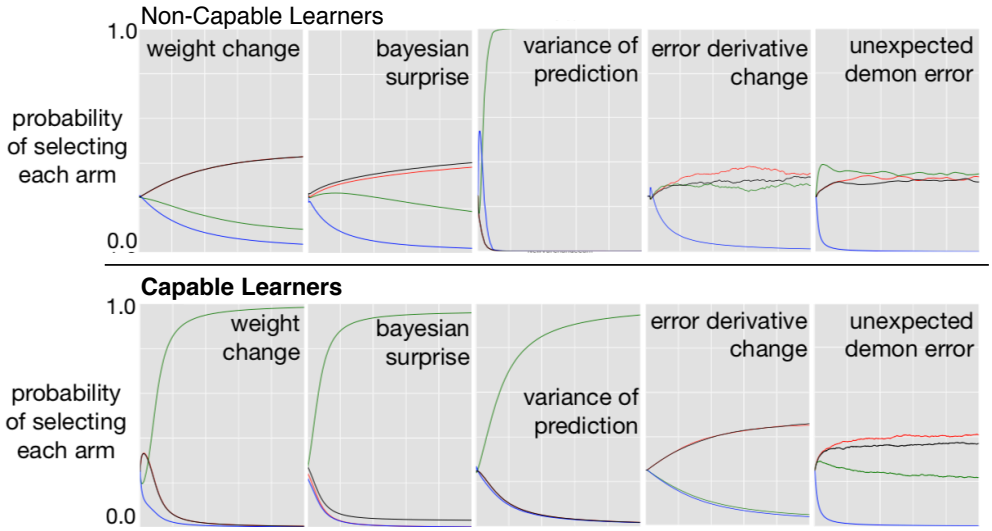

Figure 1: **The Drifter-Distractor Problem on both non-capable and capable learners.** Each subplot summarizes the learning of the control agent, over 50,000 time steps, using different curiosity rewards. Each plot shows the agent's preference for each arm over time. The drifting arm is green. We can see that the inclusion of Autostep (our capable learners) allows the simple reward function based on Weight Change and Bayesian Surprise to efficiently solve the problem. Variance of Prediction was able to solve the problem even with a weak learner, however this was due to extensive parameter sweeps and long running agents—shorter runs caused an incorrect preference for the noisy arms. The rewards based on variance result in faster preference for the drifting arm when combined with Autostep, but prefer the drifting arm less in the long run because the prediction varies less when using step-size adaptation.

4 Results

We conducted two experiments in the drifter-distractor environment, one with non-capable learners and one with capable learners. In Figure 1 we show the results for five of the methods tested in this domain. For each experiment, an extensive parameter search was conducted over the the parameters of control agent (Gradient Bandit), the prediction learners, and the reward functions.

We can see from Figure 1 that the inclusion of a capable learner allowed both the Weight Change agent and the Bayesian Surprise agent to effectively track the learnable arms while ignoring the noisy arm. Without a capable learner these agents were dominated by the error of the noisy arms and were not effective. Variance of Prediction was able to solve the problem without a capable learner—this, however, was an artifact of the parameter sweep, which chose very long running averages, enabling the larger variance to be detected. Shorter averaging windows for the variance computations caused an incorrect preference for the noisy arm.

There are two key conclusions from this experiment. First, capable learners were critical for curiosity rewards based on amount of learning, particularly Weight Change and Bayesian Surprise. Without Autostep, both Weight Change and Bayesian Surprise incorrectly cause the agent to prefer the two high-variance arms because their targets continually generate changes to the prediction. With Autostep, however, the weights converge for the constant and high-variance arms, and both agents correctly prefer the drifting arm. Second, measures based on violated expectations—Unexpected Demon Error and Error Derivative Change—either induce uniform selection or focus on noisy arms, with or without Autostep. Full results with all of the tested methods in further domains can be found in the full paper online.

5 Conclusions

The goal of this work was to investigate curiosity rewards in the multi-prediction setting. This paper has three main contributions: (1) Introduce a new benchmark for curiosity-driven exploration. (2) Survey existing ideas for curiosity driven exploration and test them on our proposed benchmark. (3) Show how capable learners can allow curiosity mechanisms, specifically the step-size adapted Weight Change, to perform strongly on the proposed task.

The problem introduced in this work formalizes the multi-prediction setting as a non-stationary multi-armed bandit. This formalism allowed us to draw clear conclusions about the efficacy of existing curiosity rewards for the task of choosing what to learn about if we hope to minimize error for multiple predictions. The focus of future work is on scaling up to a larger setting with a actor-critic or SARSA behavior policy (Sutton and Barto, 2018), with a Horde (Sutton et al., 2011) of prediction and control learners rather than the LMS learners used here.

We demonstrated how this formalism can represent a variety of types of targets we expect to see in a multi-prediction setting, including those that are noisy, drifting or easy-to-predict, and investigated performance of these curiosity re-

wards across several such settings, with both weak learners and capable learners. We surveyed and studied 15 different curiosity rewards, suitable for our pure exploration setting. We reach a surprisingly clear conclusion, particularly considering the number of approaches surveyed. Simple curiosity rewards based on learner parameters, such as change in weights, can be highly effective when paired with a capable learner.

References

Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *arXiv:1703.01732*, 2017.

Andrew G Barto and Ozgür Şimşek. Intrinsic motivation for reinforcement learning systems. In *Yale Workshop on Adaptive and Learning Systems*, 2005.

Nuttapong Chentanez, Andrew G Barto, and Satinder P Singh. Intrinsically motivated reinforcement learning. In *Adv. in Neural Inform. Process. Sys.*, pages 1281–1288, 2005.

I Magrans de Abril and Ryota Kanai. Curiosity-driven reinforcement learning with homeostatic regulation. *arXiv:1801.07440*, 2018.

Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. *arXiv:1704.03003*, 2017.

Laurent Itti and Pierre F Baldi. Bayesian surprise attracts human attention. In *Adv. in Neural Inform. Process. Systems*, 2006.

Andrew Jacobsen, Matthew Schlegel, Cameron Linke, Thomas Degris, Adam White, and Martha White. Meta-descent for online, continual prediction. 2019.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. *arXiv:1611.05397*, 2016.

Alex Kearney, Vivek Veeriah, Jaden B Travnik, Richard S Sutton, and Patrick M Pilarski. TIDBD: Adapting temporal-difference step-sizes through stochastic meta-descent. *arXiv:1804.03334*, 2018.

Daniel Ying-Jeh Little and Friedrich Tobias Sommer. Learning and exploration in action-perception loops. *Front. in Neural Circuits*, 2013.

Manuel Lopes, Tobias Lang, Marc Toussaint, and Pierre-Yves Oudeyer. Exploration in model-based reinforcement learning by empirically estimating learning progress. In *Adv. in Neural Inform. Process. Sys.*, pages 206–214, 2012.

A. Rupam Mahmood, Richard S Sutton, Thomas Degris, and Patrick M Pilarski. Tuning-free step-size adaptation. In *ICASSP*, 2012.

Nicolas Meuleau and Paul Bourgine. Exploration of multi-state environments: Local measures and back-propagation of uncertainty. *Machine Learning*, 35(2):117–154, 1999.

Joseph Modayil, Adam White, and Richard S Sutton. Multi-timescale nexting in a reinforcement learning robot. *Adaptive Behavior*, 22 (2):146–160, 2014.

Pierre-Yves Oudeyer, Frédéric Kaplan, and Verena Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 2007.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, 2017.

Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Van de Wiele, Volodymyr Mnih, Nicolas Heess, and Jost Tobias Springenberg. Learning by playing-solving sparse reward tasks from scratch. *arXiv:1802.10567*, 2018.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *International Conference on Machine Learning*, pages 1312–1320, 2015.

Massimiliano Schembri, Marco Mirolli, and Gianluca Baldassarre. Evolving childhood's length and learning parameters in an intrinsically motivated reinforcement learning robot. In *International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems*, 2007.

Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, 1991.

Jürgen Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Workshop on Anticipatory Behavior in Adaptive Learning Systems*, 2008.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction, 2nd Edition*. MIT press, 2018.

Richard S Sutton, Joseph Modayil, Michael Delp, Thomas Degris, Patrick M Pilarski, Adam White, and Doina Precup. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *International Conference on Autonomous Agents and Multiagent Systems*, 2011.

István Szita and András Lőrincz. The many faces of optimism: a unifying approach. In *International Conference on Machine Learning*, pages 1048–1055, 2008.

Adam White, Joseph Modayil, and Richard S Sutton. Surprise and curiosity for big data robotics. In *AAAI Workshop on Sequential Decision-Making with Big Data*, 2014.