Surprise and Curiosity for Big Data Robotics

Adam White, Joseph Modayil, Richard S. Sutton
Reinforcement Learning and Artificial Intelligence Laboratory
University of Alberta, Edmonton, Alberta, Canada

Abstract

This paper introduces a new perspective on curiosity and intrinsic motivation, viewed as the problem of generating behavior data for parallel off-policy learning. We provide 1) the first measure of surprise based on off-policy general value function learning progress, 2) the first investigation of reactive behavior control with parallel gradient temporal difference learning and function approximation, and 3) the first demonstration of using curiosity driven control to react to a non-stationary learning task—all on a mobile robot. Our approach improves scalability over previous off-policy, robot learning systems, essential for making progress on the ultimate big-data decision making problem—life-long robot learning.

Off-policy, life-long, robot learning is an immense big-data decision making problem. In life-long learning the agent's task is to learn from an effectively infinite stream of interaction. For example, a robot updating at 100 times a second, running 8 hours a day, with a few dozen sensors can produce gigabytes of raw observation data every year of its life. Beyond the temporal scale of the problem, off-policy life-long learning enables additional scaling in the number of things that can be learned in parallel, as demonstrated by recent predictive, learning systems (see Modayil et al 2012, White et al 2013). A special challenge in off-policy, life-long learning is to select actions in way that provides effective training data for potentially thousands or millions of prediction learners with diverse needs, which is the subject of this study.

Surprise and curiosity play an important role in any learning system. These ideas have been explored in the context of option learning (Singh et al 2005, Simsek and Barto 2006, Schembri et al 2007), developmental robot exploration (Schmidhuber 1991, Oudeyer et al, 2007), and exploration and exploitation in reinforcement learning (see Baldassarre and Mirolli 2013 for an overview). Informally, surprise is an unexpected prediction error. For example, a robot might be surprised about its current draw as it drives across sand for the first time. An agent might be surprised if its reward function suddenly changed sign, producing large unexpected negative rewards. An agent should, however, be unsurprised if its prediction of future sensory events falls within the error induced by sensor noise. Equipped with a measure of surprise, an agent can react—change how it is behaving—to unexpected situations to encourage relearning. This reactive adaptation we call *curious* behavior. In this paper we study how surprise and curiosity can be used to adjust a robot's behavior in the face of changing world.

In particular, we focus on the situation where a robot has already learned two off-policy predictions about two distinct policies. The robot then experiences a physical change that significantly impacts the predictive accuracy of a single prediction. The robot observes its own inability to accurately predict future battery current draw when it executes a rotation command, exciting its internal surprise measure. The robot's behavior responds by selecting actions to speed relearning of the incorrect prediction—spinning in place until the robot is no longer surprised—then returning to normal operation.

This paper provides the first empirical demonstration of surprise and curiosity based on off-policy learning progress on a mobile robot. Our specific instantiation of surprise is based on the instantaneous temporal difference error, rather than novelty, salience, or predicted error (all explored in previous work). Our measure is unique because 1) it balances knowledge and competence-based learning and 2) it uses error generated by off-policy reinforcement learning algorithms on real robot data. Our experiment uses commodity off-the-shelf iRobot Create and simple camera resulting in real-time adaptive control with visual features. We focus on the particular case of responding to a dramatic increase in surprise due to a change in the world—rather than initial learning. The approach described in this paper scales naturally to massive temporal streams, high dimensional features, and many independent off-policy learners common in life-long robot learning.

Background

We model an agent's interaction with the world (including the robot's body) as a discrete time dynamical system. On each time step t, the agent observes a feature vector $\mathbf{x}_t \in \mathcal{X} \subset \mathbb{R}^n$, that only partially characterizes the environmental state $s_t \in \mathcal{S}$. We assume s_t is the current state of some unobservable Markov Decision Process (MDP), and thus $\mathbf{x}_t$ is computed from any information available to the

Copyright © 2014, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

agent at time t. On each step the agent takes an action $a_t \in \mathcal{A}$, resulting in an transition in the underlying MDP, and the agent observes a new feature vector $\mathbf{x}_{t+1}$.

In conventional reinforcement learning, the agent's objective is to predict the total discounted future reward on every time step. The reward is a special scalar signal $r_{t+1} = r(\mathbf{x}_{t+1}) \in \mathbb{R}$, that is emitted by the environment on each step. To predict reward the agent learns a value function $v : \mathcal{S} \to \mathbb{R}$. The time scale of the prediction is controlled by a discount factor $\gamma \in [0, 1)$. The precise quantity to be predicted is the *return* $g_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$, and the value function is the expected value of the return,

$$v(s) = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s \right],$$

where the expectation is conditional on the actions (after t) selected according to a particular policy $\pi : \mathcal{X} \times \mathcal{A} \to [0, 1]$ as denoted by the subscript on the expectation operator. As is common in reinforcement learning, we estimate v with a linear approximation, $v_w(s_t) = \mathbf{w}^\top \mathbf{x}_t \approx v(s_t)$, where $\mathbf{w} \in \mathbb{R}^n$.

We use generalized notions of reward and termination to enable learning a broader class of predictions than is typically considered in conventional reinforcement learning. First notice we can define the reward to be any bounded function of $\mathbf{x}_t$, such as the instantaneous value of an IR sensors, which we call pseudo reward. Second, γ need not be constant, but can also be defined as a function the features $\gamma : \mathcal{X} \to [0, 1]$. These changes require the definition of a general value function (GVF)

$$v(s) = \mathbb{E} \left[\sum_{k=0}^{\infty} \left(\Pi_{j=1}^{k} \gamma(\mathbf{X}_{t+j}) \right) r(\mathbf{X}_{t+k+1}) \right],$$

but no special algorithmic modifications are required to learn GVFs. See Modayil et al (2014) for a more detailed explanation of GVF prediction.

In order to learn many value functions in parallel, each conditioned on a different policy, we require *off-policy learning*. Off-policy reinforcement learning allows the agent to learn predictions about one policy while following another. In this setting, the policy that conditions the value function—the target policy π—is different from the policy used to select actions and control the robot, called the behavior policy $\mu : \mathcal{X} \times \mathcal{A} \to [0, 1]$. Separating the data generation policy from the policy to be learned allows learning about many value functions in parallel from a single stream of experience. Each prediction $V_t^{(i)} = v_w^{(i)}(s_t)$ is about future pseudo reward $r^{(i)}(\mathbf{x}_t)$ observed if the agent follows $\pi^{(i)}$ with pseudo termination according to $\gamma^{(i)}(\mathbf{x}_t)$. These policy-contingent predictions can each be learned by independent instances of off-policy reinforcement learning algorithms.

One such off-policy algorithm is GTD(λ) which preforms stochastic gradient descent on the Mean Squared Projected Bellman Error (MSPBE), with linear time and space complexity. The update equations for GTD(λ),

$$\begin{aligned}
\mathbf{e}_t &\leftarrow \frac{\pi(\mathbf{x}_t, a_t)}{\mu(\mathbf{x}_t, a_t)} (\mathbf{x}_t + \gamma(\mathbf{x}_t)\lambda \mathbf{e}_{t-1}) \\
\mathbf{w}_{t+1} &\leftarrow \mathbf{w}_t + \alpha(\delta_t \mathbf{e}_t - \gamma(\mathbf{x}_{t+1})(1-\lambda)(\mathbf{e}_t^\top \mathbf{h}_t)\mathbf{x}_{t+1}) \\
\mathbf{h}_{t+1} &\leftarrow \mathbf{h}_t + \beta(\delta_t \mathbf{e}_t - (\mathbf{x}_t^\top \mathbf{h}_t)\mathbf{x}_t),
\end{aligned}$$

require an eligibility trace vector $\mathbf{e} \in \mathbb{R}^n$, scalar learning rate parameters α and β, and the usual temporal difference error $\delta_t = r(\mathbf{x}_{t+1}) + \gamma(\mathbf{x}_{t+1})\mathbf{w_t}^\top \mathbf{x}_{t+1} - \mathbf{w_t}^\top \mathbf{x}_t$. The MSPBE is a convex objective that can be directly estimated from data, $\texttt{MSPBE}(\mathbf{w}) = \mathbb{E}_\mu[\delta \mathbf{e}]^T \mathbb{E}_\mu[\mathbf{x}\mathbf{x}^T]^{-1} \mathbb{E}_\mu[\delta \mathbf{e}]$, useful for tracking the learning progress of off-policy predictions (White et al 2013).

Experiment

The objective of this paper is to investigate one potential benefit of using surprise to influence decision making. Specifically we seek a concrete demonstration of adapting the behavior policy automatically, in response to a perturbation to the agent's sensorimotor stream. The change is unexpected and not modelled by the agent. The change is designed to influence the accuracy of only one GVF prediction. Consider a measure of surprise based on the temporal difference error of each GVF

$$Z_t^{(i)} = \frac{\overline{\delta^{(i)}}}{\sqrt{\texttt{var}[\delta^{(i)}]}}, \tag{1}$$

where $\bar{\cdot}$ denotes an exponentially weighted average. This measure of surprise increases when instantaneous errors fall considerably outside the mean error.

A *curious behavior* is any policy that uses a measure of surprise from several GVFs to influence action selection. Here we consider a rule-based curious behavior that uses surprise to determine whether the behavior should continue selecting actions according to target policy $\pi^{(i)}$ or switch to another target $\pi^{(j)}$. Assuming the behavior μ had selected actions according to $\pi^{(i)}$ for k consecutive steps, the agent decides to continue following $\pi^{(i)}$ for k more steps, or switch to a new target policy:

$$\begin{aligned}
&\texttt{if } Z_t^{(i)} < \tau \texttt{ then} \\
&\qquad j = \text{argmax}_{j \neq i} Z_t^{(j)} \\
&\qquad \texttt{if } Z_t^{(j)} < \tau \\
&\qquad\qquad \texttt{pick } j \texttt{ randomly} \\
&\qquad \mu = \pi^{(j)} \\
&\texttt{follow } \mu \texttt{ for } k \texttt{ consecutive steps}
\end{aligned} \tag{2}$$

In our experiment reported here k was set to 120 steps or approximately four seconds, τ was set to 0.2, and the decay rate of the exponential average in Equation 1 was 0.01.

Intuitively we expect a curious behavior to select actions that encourage or facilitate re-learning of an inaccurate GVF. In the case of a perturbation that affects a single GVF $v^{(i)}$,

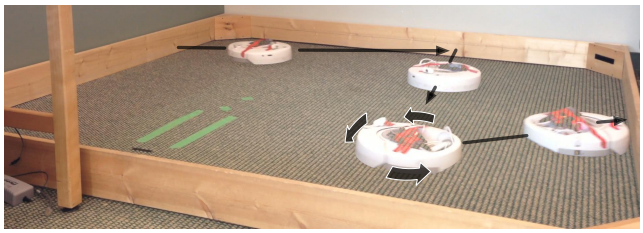

Figure 1: Non-reactive control of the Create in its pen.

we expect the curious behavior (2) to continually select actions according to the corresponding target policy $\pi^{(i)}$ until the surprise has subsided, and thus the new situation has been learned.

To investigate this hypothesis, we conducted an experiment in which two GVFs were learned from off-policy experience. The first GVF predicted future discounted battery current draw while counter-clockwise spinning, encoded as $\pi^{(0)}(\mathbf{x}, \texttt{rotateCCW}) = 1.0$, $\gamma^{(0)}(\mathbf{x}) = 0.8 \ \forall \ \mathbf{x} \in \mathcal{X}$, and $r_t^{(0)} = \texttt{battery_current}_t$. We used a simple discrete action set of {`forward`, `reverse`, `stop`, `rotate cw`, and `rotate ccw`}. The second GVF predicted the expected number of time steps until bump if the robot drove forward, encoded as $\pi^{(1)}(\mathbf{x}, \texttt{forward}) = 1.0 \ \forall \ \mathbf{x} \in \mathcal{X}$, $\gamma^{(1)}(\mathbf{x}) = 0.0$ on bump and 0.95 otherwise (a 6.6 second prediction horizon), and $r_t^{(1)} = 1.0$.

Each GVF had a unique feature representation. The rotation GVF's binary feature vector was produced by a single tiling of a decaying trace of the observed battery current, with a tile width of 1/16th. The forward GVF used a feature vector constructed from 120×160 web-camera images sampled at 30 frames per second. At the start of the experiment, 100 pixels were selected at random, and from these pixels either the luminance or color channel was selected at random, and these selected values were used to construct features from the most recent image. Each value (between 0 and 255) was independently tiled into 16 non-overlapping tiles of width 16, producing a binary feature vector $\mathbf{x}_t$, with 16000 components, of which 100 were active on each time step (see Sutton and Barto (1998) for a review of tile coding). The camera was mounted directly on top of the robot facing forward.

Both GVFs were learned using a separate instances of GTD($\lambda = 0.9$) with $\alpha = 0.05$ and $\beta = 0.0005$. All learning was done directly on a raspberry pi directly connected to an iRobot Create with an update cycle of 30 ms. The perturbation involved putting a heavy load in the back of the robot, which changes the current draw and directly affects the rotation GVF. The drive speed and thus the forward GVF prediction will be unaffected.

Our experiment involved two phases. During the first phase (roughly ten mins) the robot followed a hand-coded *non-reactive behavior* policy that alternated between driving forward until bumping, rotating counter clockwise in free-space (not against the wall), and rotating away from the wall after bump. Figure 1 shows a visualization of the non-

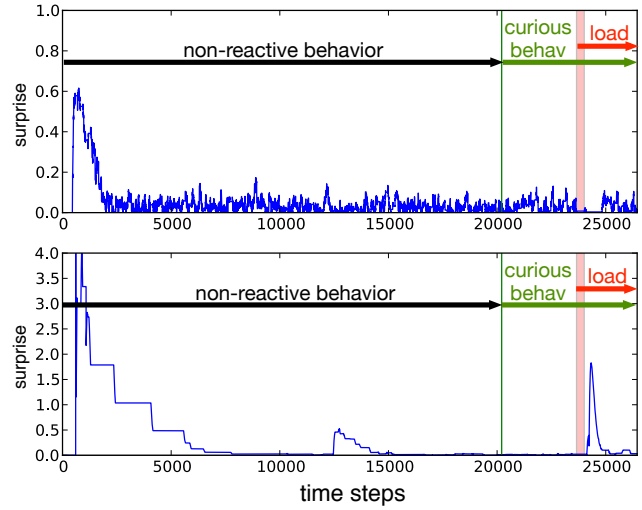

Figure 2: A time-series plot of the surprise measure for the Forward (top) and Rotation (bottom) GVFs. Both graphs show the three phases of the experiment: 1) initial learning under the non-reactive behavior, 2) learning under the curious behavior, and 3) learning after load was added to the robot. Notice that during phase 1, a spike in rotation surprise is present. This was the robot's first rotation just after bumping, which generates a novel current profile. A second spike in surprise occurs after the load is added in phase 3. The later spike is larger, but reduces faster because the robot effectively modifies its behavior.

reactive behavior during the first phase of learning. Phase one was interrupted after each GVF had learned to an acceptable level of accuracy, after which time the behavior was switched to a *curious behavior* policy (described in 2). After about two minutes, a 5 pound load was placed in the cargo bay of the Create. The load had a significant effect on the battery current draw, but was not heavy enough to affect the robot's ability to achieve the requested wheel velocity for the drive forward actions.

Figure 2 shows the effect of the load on each GVF's prediction via the surprise measures. The forward GVF's prediction is largely unaffected by the load; the robot was unsurprised. The rotation GVF's pseudo reward was based on current draw, and was therefore significantly affected as seen in a substantial increase in surprise. The extra load generated a clear and obvious change in the robot's behavior. The first time the curious behavior selected a counter-clockwise rotation (after the load was added) a large increase in surprise was produced. The increased surprise caused an extended counter-clockwise rotation. When neither surprise measure exceeds the threshold, the curious behavior will drives forward, only occasionally executing rotations in free space. The effect of the load produced a very distinctive and extended change in behavior. Figure 3 provides two quantitative measures of how the actions were selected before the load was added, during the surprise period, and after relearning. After several seconds of rotation the surprise for rotating subsided below τ, constant rotation ceased, and the behavior returned to normal operation.

The experiment was repeated several times varying the

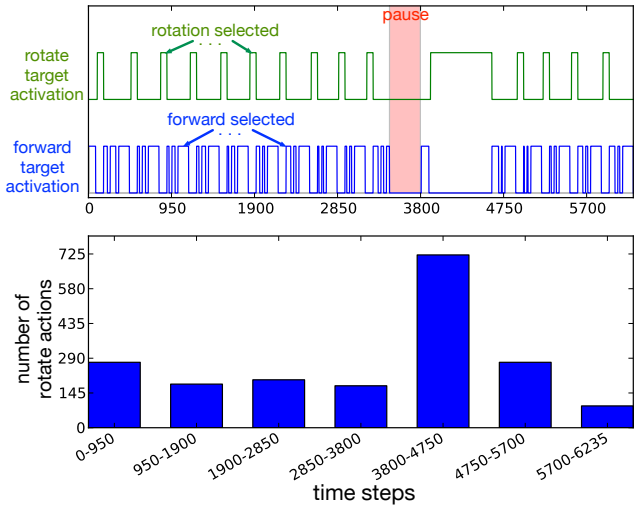

Figure 3: The top figure shows which GVF policy is active while the robot was under curious behavior control. The bottom histogram shows the number of time steps on which the rotation action was selected over the same time period. The upper figure clearly shows that the robot continually selected actions according to the rotate target policy after the load was added (marked by the red bar), and then returned to normal operation. The lower figure illustrates the same phenomenon, and also shows the variability in the number of rotations before the load was added to the robot.

lighting conditions, wall-clock duration of each phase, and camera orientation. The increase in surprise and resultant behavior modification was reliably demonstrated each time.

Discussion

Our experiment was designed to highlight one way in which the off-policy learning progress of multiple GVFs can influence control. We crafted the GVFs and introduced a change that would be immediately detected, and would produce a distinctive change in behavior (constant rotation). Although limited, our demonstration is novel, and demonstrates a fundamental idea in a scalable way.

Surprisingly this work is the first demonstration of adaptive control for the behavior policy of an off-policy learning system on a robot. All previous GVF learning systems used non-adaptive behaviors (see Sutton et al 2010, Modayil et al 2012, White et al 2013). Typically intrinsically motived systems are segregated into two camps: 1) knowledge-driven sequential task learning (Schmidhuber 1991, Oudeyer et al 2007), and 2) competence-driven option or subtask training (Singh et al 2005, Simsek and Barto 2006, Schembri et al 2007). Our approach unifies these two approaches. A GVF can be used to represent a prediction conditioned on a fixed target policy (as used here) and encode an state-action-value function used to learn a policy (learned with greedy-GQ(λ)). Therefore adapting behavior based-on GVF error drives both predictive knowledge acquisition and improving competence. Finally, our work is the first to highlight and demonstrate the role of surprise and curiosity in a non-stationary setting.

The approach described here is small, but surprisingly scalable. Consider an alternative approach to reacting to change like following each target policy in sequence. The robot might be learning thousands of different GVFs conditioned on thousands of target polices, as in previous work (White et al 2013). It would be infeasible to rely on sequencing target policies to efficiently detect and react to changes when the number of GVFs is large.

Assuming any change to the robot's world only affects a subset of the GVFs, our approach to curiosity will provide data to GVF's that are capable of more learning. A more general approach would be to learn the curious behavior with reinforcement learning and a surprise-based reward. If the behavior were generated via average reward actor critic, then the robot could balance the needs of many GVFs in its action selection without the restriction of following any target policy.

The approach explored here is the first step in exploring the natural synergy between parallel off-policy learning and curious behavior. Although surprise is useful for adapting to non-stationarity, it can be usefully deployed for a wide range of settings. Imagine a setting where new GVFs are continually created over time. A curious behavior, in a similar way to adapting to a perturbation, can adjust action selection to provide relevant data for new GVFs. We focused here on using curiosity after initial learning—each GVF had been learned to high accuracy. Curiosity can also be used during initial learning to avoid the inherent limitations of hand-coded behaviors. Finally, what does an robot do when its no-longer surprised or bored? Could a curious behavior select actions in such a way to ready itself to react efficiently to new perturbations or new GVFs? These questions are left to future work.

Conclusion

This paper provides 1) the first measure of surprise based on off-policy GVF learning progress, 2) the first investigation of reactive behavior control with parallel gradient TD learning and function approximation, and 3) the first demonstration of using curiosity driven control to react to non-stationarity—all on a mobile robot. The ability to determine which off-policy predictions are substantially inaccurate, and modifying robot behavior online to improve learning efficiency is particularly important in large-scale, parallel, off-policy learning systems.

Acknowledgements

This work was supported by grants from Alberta Innovates Technology Futures and the National Science and Engineering Research Council of Canada.

References

Baldassarre, G., Mirolli, M. (Eds.). (2013). *Intrinsically motivated learning in natural and artificial systems*. Berlin: Springer.

Maei, H. R. (2011). Gradient Temporal-Difference Learning Algorithms. *PhD thesis*, University of Alberta.

Modayil, J., White, A., Sutton, R. S. (2012). Multi-timescale nexting in a reinforcement learning robot. In *From Animals to Animats 12*, 299–309.

Oudeyer, P. Y., Kaplan, F., Hafner, V. (2007). Intrinsic Motivation Systems for Autonomous Mental Development. In *IEEE Transactions on Evolutionary Computation 11*, 265–286

Schembri, M., Mirolli, M., Baldassarre, G. (2007). Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot. In *Development and Learning*, 282–287.

Schmidhuber J. (1991). A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proceedings of the 1st International Conference on Simulation of Adaptive Behavior*, 222–227.

Simsek, O., Barto, A. G. (2006). An intrinsic reward mechanism for efficient exploration. In *Proceedings of the 23rd international conference on Machine learning*, 833–840.

Singh S., Barto, A. G., Chentanez, N. (2005). Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems 17*, 1281–1288.

Sutton, R. S., Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Sutton, R. S., Maei, H. R., Precup, D., Bhatnagar, S., Silver, D., Szepesvári, Cs., Wiewiora, E. (2009). Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning*.

Sutton, R. S., Modayil, J., Delp, M., Degris, T., Pilarski, P. M., White, A., Precup, D. (2011). Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proceedings of the10th International Conference on Autonomous Agents and Multiagent Systems*.

White, A., Modayil, J., Sutton, R. S. (2012). Scaling life-long off-policy learning. In *Development and Learning and Epigenetic Robotics*, 1–6.