Predictions Predicting Predictions

Matthew K. Schlegel*
Department of Computing Science
University of Alberta
mkschleg@ualberta.ca

Martha White
Department of Computing Science
University of Alberta
whitem@ualberta.ca

Abstract

Predicting the sensorimotor stream has consistently been a key component for building general learning agents. Whether through predicting a reward signal to select the best action or learning a predictive world model with auxiliary tasks, prediction making is at the core of reinforcement learning. One of the main research directions in predictive architectures is in the automatic construction of learning objectives and targets. The agent can consider any real-valued signal as a target when deciding what to learn, including the current set of internal predictions. A prediction whose learning target is another prediction is known as a composition. Arbitrarily deep compositions can lead to learning objectives that are unstable or not suitable for function approximators. This manuscript looks to begin uncovering the underlying structure of compositions in an effort to leverage and learn them more effectively in general learning agents. Specifically, we consider the dynamics of compositions both empirically and analytically. We derive the effective schedule of emphasis (or discounts) of future observations with compositions of arbitrary depth, leading to informative observations about the prediction targets. In the empirical simulations, we focus on the unintuitive behavior of compositions, especially in cases that are not easy to analyze. Overall, predictions predicting predictions which predict predictions have interesting properties and can add depth to an agent's predictive understanding of the world.

Keywords: reinforcement learning; prediction; GVFs

Acknowledgements

We would like to thank the Alberta Machine Intelligence Institute, IVADO, NSERC and the Canada CIFAR AI Chairs Program for the funding for this research. We would also like to thank Andrew Patterson for his insightful comments about the connection to digital signal processing and Adam White for sharing the Critterbot dataset.

*mkschleg.github.io

1 Introduction

Reinforcement learning is built on predicting the effect of behavior on future observations and rewards. Many of our algorithms learn predictions of a cumulative sum of (discounted) future rewards, which is used as a bedrock for learning desirable policies. While reward has been the primary predictive target of focus, TD models (Sutton 1995) lay out the use of temporal-difference learning to learn a world model through value function predictions. Temporal-difference networks (Sutton and Tanner 2004) take advantage of this abstraction and build state and representations through predictions. Sutton, Modayil, et al. 2011 and White 2015 further the predictive perspective by developing a predictive approach to building world knowledge through general value functions (GVFs).

GVFs have been pursued broadly in reinforcement learning: Günther et al. 2016 used GVFs to build an open loop laser welder controller, Linke et al. 2020 used predictions and their learning progress to develop an intrinsic reward, Edwards et al. 2016 used GVFs to build controllers for myoelectric prosthetics, using gvfs for auxiliary training tasks to improve representation learning (Jaderberg et al. 2017; Veeriah et al. 2019), to extend a value function's approximation to generalize over goals as well as states (Schaul et al. 2015), and to create a scheduled controller from a set of sub-tasks for sparse reward problems (Riedmiller et al. 2018). Successor representations and features are predictions of the state, learned or given, which have been shown to improve learning performance (Barreto et al. 2018; Dayan 1993; Russek et al. 2017; Sherstan et al. 2018).

Learning predictions of any real-valued signal the agent has access to also opens the possibility of asking compositional predictive questions (White 2015). A compositional question is one whose target is dependent on another prediction internal to the agent. Compositions expand the possible range of predictive questions we can specify as a GVF (Rafols et al. 2006; Schlegel et al. 2021; Sutton and Tanner 2004; White 2015; Zheng et al. 2021). While this may suggest the GVF framework is limited in what questions can be asked, the limitations are necessary so the predictions can be trained *independent of span* (van Hasselt and Sutton 2015). Learning independent of span means the target can be learned using online algorithms regardless of the effective horizon of the prediction. Adding layers of compositional questions have improved the learning in predictive representations (Rafols et al. 2006; Schlegel et al. 2021), and improved the performance of deep reinforcement learning through auxiliary tasks (Zheng et al. 2021). In the automatic specification of learning targets compositions are thought to provide a way for the agent to build complexity (Kearney 2022; Schlegel et al. 2021; Veeriah et al. 2019; Zheng et al. 2021), but often these architectures don't leverage compositions for stability concerns (Schlegel et al. 2021).

As well as improving behavior empirically, compositions can provide semantic depth. An excellent example of this can be seen in option-extended temporal difference networks (Rafols et al. 2006), and later explored again in Schlegel et al. 2021. The example is centered in an environment where the agent has a low-powered visual sensor and needs to learn its directionality from the painted walls. Each cardinal direction has a different colored wall. The first layer of predictions the agent makes is to predict what color it will observe if it were to drive straight. The second layer are myopic predictions which ask what the first layer's prediction will be after turning clockwise (or counter-clockwise). The second layer allows the agent to predict which walls are to its sides as well as the wall in the direction the agent is facing. These predictions cannot be specified in the usual GVF framework, but can be easily constructed through compositions. While this may be "repeated information" in a sense, the extra learning objectives makes the learning properties of the predictive representation better as compared to other specifications (Schlegel et al. 2021).

As algorithms for the automatic discovery of complex question networks continue to push the boundaries of what questions are considered by the agent, the properties of compositions should be better studied. When searching for what to learn the questions an agent eventually retains will be dependent on the agent's ability to learn the predictions. While it is clear questions that naturally diverge (say setting the discount $\gamma = 1$) should be avoided, other problems, such as the scale of a target, could be equally as problematic when using function approximation (i.e. end-to-end neural networks). This could mean important predictions are disregarded because the agent is unable to learn the answer without proper strategies to normalize the prediction's magnitude. Better strategies for learning and normalizing predictive targets will come from understanding the effective discount schedule (or emphasis) compositional predictions will have on the targets.

In this report, we consider the effect of compositions on the sequence of discounts, and relegate the effect of off-policy importance weights to future work. We first analyze the sequence of discounts over any number of compositions and constant discounts. We then analyze this sequence to better understand how it emphasizes parts of the data stream. Surprisingly, the effective discount for constant discount compositions have a form which can be described analytically. While this does not include the full spectrum of discount functions, it provides a first step towards understanding compositions. Next we look at simulations using more complex state-dependent discount functions using a simple consistent sequence and two timeseries datasets. In these simulations we focus on the effect of applying the same discount function a large number of times, looking to see if the shape of the returns become regular over the compositions. Finally, several future directions and questions are posed.

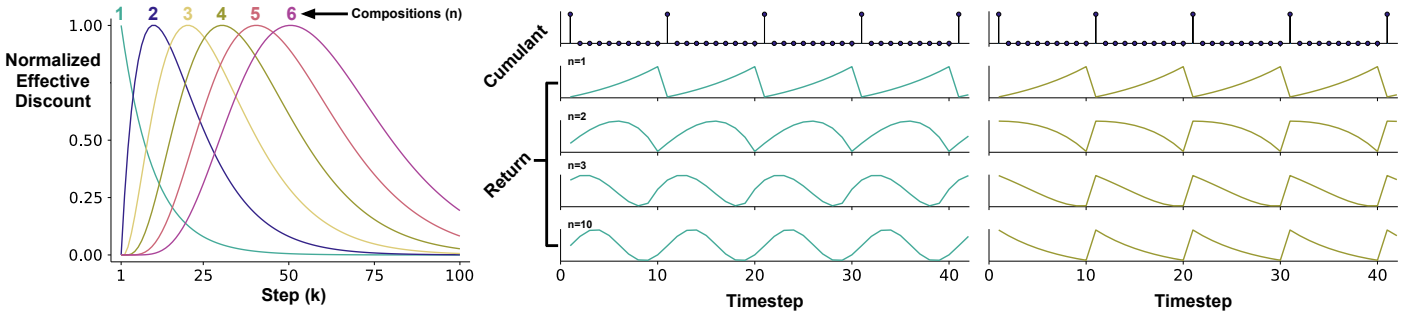

Figure 1: (**left**) The effective discount for n compositions normalized by the maximum value found in section 2. (**middle, right**) The cycle world simulations, with top graph as the cumulant and subsequent plots n compositions with constant and terminating discounts respectively.

2 Analyzing the sequence

In this section, we restrict to the setting where we have an infinite sequence of sensor readings $\mathbf{x} = \{x[0], x[1], \ldots, x[t], \ldots, x[\infty]\}$ where $x[i] \in [x_{\min}, x_{\max}]$ and a constant discount γ. The return of this signal starting at a time step t is $V[t] = \sum_{k=0}^{\infty} x[k]\gamma[k-t]$ where $\gamma[k] = \gamma^{k-1}$ for $k >= 1$ and 0 otherwise. This framing of the return is slightly different from the typical presentation. Specifically, we reinterpret the return as a convolution beteween γ and x [1] and shift the discount sequence over the sensor readings. This implicitly defines an infinite sequence of predictions $V[t]$. In the above equation, if we replace the sequence x with the sequence of predictions V, we get a new set of predictions and for any number of compositions n we have $V^n[t] = \sum_{k=0}^{\infty} V^{n-1}[k]\gamma[k-t]$. Expanding this equation we can define the general sequence of effective discounts for n compositions and the corresponding return as

$$\gamma^n[k] = \begin{cases} 0 & \text{if } k < n \\ \frac{\prod_{i=1}^{n-1}(k-i)}{(n-1)!}\gamma^{k-n} \end{cases} \qquad V^n[t] = \sum_{k=0}^{\infty} x[k]\gamma^n[k-t]$$

where $\gamma^1[k] = \gamma[k]$ defined above and $V^n[t]$ is the target of the nth composition at timestep t. For any value n there are two sequences multiplied together. The original discounting shifted by the number of applications $\gamma^1[k-n]$ and a diverging series

$$Q^n[k] = \frac{\prod_{i=1}^{n-1}(k-i)}{(n-1)!} = \frac{\Gamma(k)}{\Gamma(k-n+1)\Gamma(n)}$$

where $\Gamma(k) = (k-1)!$ for $k \in \mathbb{Z}$ is known as the Gamma function, and can be used to analyze the function with $k \in \mathbb{R}$.

We know for any particular application of the convolution γ on a series with known domain $[x_{\min}, x_{\max}]$ the value function can take values bounded by $V^1[t] \in [\frac{x_{\min}}{1-\gamma}, \frac{x_{\max}}{1-\gamma}]$. This extends to n compositions in a straightforward way where the range of the value function becomes $V^n[t] \in [\frac{x_{\min}}{(1-\gamma)^n}, \frac{x_{\max}}{(1-\gamma)^n}]$. While normalizing the value function to take values within in the range $[0, 1]$ has been used in various settings (Schlegel et al. 2021), as we add more compositions we see the effective range of values shrinking considerably.

Given the effective discounting sequence above, we can begin to piece together the which observations are emphasized in the predictions. The first 100 steps of the effective discount function for several values of n can be seen in figure 1. These sequences are normalized to be in the range $[0, 1]$ for a visual comparison. The emphasis becomes increasingly spread as n increases, with the peak of this function moving further to the future at a consistent rate.

To find the maximum value we take the derivative of the log of the sequence with respect to k getting

$$\frac{\delta}{\delta k} \ln \gamma^n[k] = \psi(k) - \psi(k-n+1) + \ln \gamma$$

where $\psi(z+1) = H_z - C$ is the digamma function, $H_z = \sum_{i=1}^{z} \frac{1}{i} \leq \int_1^z \frac{1}{x} dx = ln(z)$ is the Euler harmonic number, and C is the Euler-Mascheroni constant. Using the approximation above, we can find where we should expect the maximal value is (to an approximation) $k = hn - (h-1) = h(n-1) + 1$, where $h = \frac{1}{1-\gamma}$ is sometimes known as the horizon of discount γ. Of course this is an approximation from above and the real value falls in $k \in [h(n-1), h(n-1)+1)]$.

[1] In digital signal processing (Oppenheim and Schafer 2010) often the convolution, in this case γ, is mirrored across t and the inifinte sequence of sensor readings is $\mathbf{x} = \{x[-\infty], \ldots, x[t], \ldots, x[\infty]\}$. The corresponding convolution would be $V[t] = \sum_{k=-\infty}^{\infty} x[k]\gamma[t-k]$ which would change how we define the sequence of γ. To be consistent with the reinforcement learning literature, we don't follow this here and instead implicitly define γ as the mirrored version and only consider the sequence starting at $k = 0$.

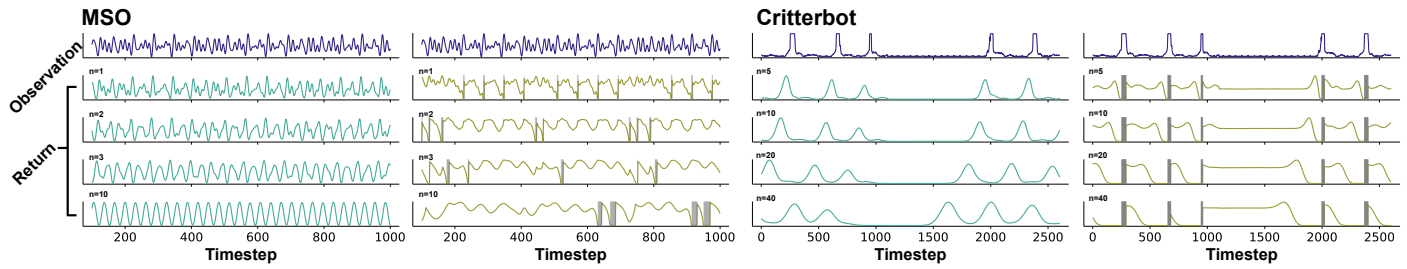

Figure 2: (**left two**) Returns of the multiple sinusoidal oscillator (MSO) synthetic data set with constant and terminating discount respectively. The gray vertical lines are where the return terminates. (**right two**) Returns of Critterbot data set over the light3 sensor with constant and terminating discount respectively.

3 Empirical observations

While we can describe the effective discount for composing constant discount predictions, the same techniques are difficult to apply to a non time-invariant discount (i.e. state-dependent discounts (Sutton, Modayil, et al. 2011; White 2015)). Instead, in this section we look at the ideal returns of various signals using constant discounting and a terminating discounting functions. We use three datasets moving from highly synthetic to real-world robot sensori-motor data. The goal of this section is to show the non-intuitive behavior of compositions to motivate further analysis and exploration. All code can be found at `https://github.com/mkschleg/CompGVFs.jl`. Below $\gamma = 0.9$ unless otherwise stated.

The first series is based off the cycle world, where the agent observes a sequence of a single active bit followed by 9 inactive bits, where the length of the sequence is $m = 10$. The cumulant is the observation itself, and in this report we learn using TD($\lambda = 0.9$) with learning rate $\alpha = 0.1$ and an underlying tabular representation where each component is the place in the sequence. We learn two chains of compositions. The first is that of the continuous discounting described above, and the second is a series of discounts which terminate (i.e. $\gamma[t] = 0$) when the observation is active. The predictions of a single run can be seen in figure 1. For the constant discount, as the number of compositions increases we see the prediction sequence converge to what looks to be a sinusoid with frequency of 10, and amplitude driven by the analysis above. We expect this to be the case following from the central limit theorem. For the terminating discount, the wave form is more interesting. The first layer of predictions look very similar to the constant discount with amplitude shifted by $\frac{\gamma^m}{1-\gamma^m}$. But as there are more compositions the effect seems to be the prediction is at its height farther away from the active bit. As the agent gets closer to the observation, the sequence of summed values is shorter leading to smaller values. Given the sequence we use it is easy to mistake this as the agent creating a trace of the cumulant, but we must remember the prediction is about future cumulants.

Next we use a subset of the Critterbot dataset (Modayil et al. 2014; White 2015), focusing on light sensor 3. This gives a sequence of spikes similar to the cycle world sequence and a long pause in-between consistent saturations of the light sensor. We are able to see with the current setting the predictions look more like shifted and spread spikes. But with many more compositions, the return reverts to a similar form as before. The terminating discounts (with termination at sensor saturation $x[t+1] > 0.99$) provides a nice demonstration of how the returns are predicting the signal, just with a decaying prediction instead of the usual growing prediction. The results are similar in the multiple sinusoidal oscillator (Jaeger and Haas 2004). We use a slightly different terminating discount where the return terminates when the previous normalized prediction is $y^{n-1}[t+1] > 0.9$ rather than when the observation is saturated. While there are decays as the MSO sequence peaks, as we increase the depth of the composition, these periods are less frequent. Deep compositions may indicate parts of the sequence where there are fewer saturations in the original sequence.

4 Future Directions

This work suggests a number of interesting research directions and questions. While we mostly analyzed the sequence on discrete steps and applications of the filter, the general form does lend itself to continuous and complex values of n and k. In a similar vein, we focused on real valued exponential discounting while several discounting schemes exist which could be applied to our formulation. We are particularly interested in complex discounting (De Asis et al. 2018) and hyperbolic discounting (Fedus et al. 2019). Applying a diverse set of discounting schemes in compositions provide an interesting way to extend the power of value functions while maintaining learnability through efficient algorithms like temporal-difference learning.

The approach used in this paper is unable to analyze state-dependent discount functions. One way around this might be in analyzing truncated sequences and taking an expectation over a distribution of sequence lengths. This might lead to a expected effective discounting sequence, but how this will interact with an underlying Markov process is unclear. This

is an important next step for understanding the effects of compositions in general value functions, and could also help in analyzing off-policy compositions.

Finally, the return can be re-interpreted as a convolution over the infinite sequence of observations. While this interpretation was only used to better the notation in this manuscript, further connections to convolutions and digital signal processing should be explored. Better filter designs might inspire different discounting schedules to squeeze more information from the data stream. We also have only analyzed these convolutions in the time domain. The frequency domain might give us more insight into how consistent signals like the cycle world dataset will be effected by compositions.

References

Barreto, Andre et al. (2018). "Transfer in Deep Reinforcement Learning Using Successor Features and Generalised Policy Improvement". In: *International Conference on Machine Learning*. PMLR.

Dayan, Peter (1993). "Improving Generalization for Temporal Difference Learning: The Successor Representation". In: *Neural Computation*.

De Asis, Kristopher, Brendan Bennett, and Richard Sutton (2018). "Predicting Periodicity with Temporal Difference Learning". In: *arXiv preprint arXiv:1809.07435*.

Edwards, Ann L et al. (2016). "Application of real-time machine learning to myoelectric prosthesis control: A case series in adaptive switching". In: *Prosthetics and orthotics international* 40.5, pp. 573–581.

Fedus, William et al. (2019). "Hyperbolic discounting and learning over multiple horizons". In: *arXiv preprint arXiv:1902.06865*.

Günther, Johannes et al. (2016). "Intelligent laser welding through representation, prediction, and control learning: An architecture with deep neural networks and reinforcement learning". In: *Mechatronics* 34, pp. 1–11.

Jaderberg, Max et al. (2017). "REINFORCEMENT LEARNING WITH UNSUPERVISED AUXILIARY TASKS". In: *International Conference on Representation Learning*.

Jaeger, Herbert and Harald Haas (2004). "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication". In: *science*.

Kearney, Alex (2022). "What Should I Know? Using Meta-Gradient Descent for Predictive Feature Discovery in a Single Stream of Experience". In: *Unpublished*.

Linke, Cam et al. (2020). "Adapting Behavior via Intrinsic Reward: A Survey and Empirical Study". In: *Journal of Artificial Intelligence Research* 69, pp. 1287–1332.

Modayil, Joseph, Adam White, and Richard Sutton (2014). "Multi-Timescale Nexting in a Reinforcement Learning Robot". In: *Adaptive Behavior*.

Oppenheim, Alan V and Ronald W Schafer (2010). *Discrete-time Signal Processing*. Pearson Higher Education.

Rafols, Eddie, Anna Koop, and Richard Sutton (2006). "Temporal Abstraction in Temporal-difference Networks". In: *Advances in Neural Information Processing Systems 18*. Ed. by Y. Weiss, B. Schölkopf, and J. C. Platt. MIT Press.

Riedmiller, Martin et al. (2018). "Learning by playing solving sparse reward tasks from scratch". In: *International conference on machine learning*. PMLR, pp. 4344–4353.

Russek, Evan M. et al. (2017). "Predictive Representations Can Link Model-Based Reinforcement Learning to Model-Free Mechanisms". In: *PLOS Computational Biology*.

Schaul, Tom et al. (2015). "Universal Value Function Approximators." In: *International Conference on Machine Learning*.

Schlegel, Matthew et al. (2021). "General value function networks". In: *Journal of Artificial Intelligence Research* 70, pp. 497–543.

Sherstan, Craig, Marlos C. Machado, and Patrick M. Pilarski (2018). "Accelerating Learning in Constructive Predictive Frameworks with the Successor Representation". In: *arXiv:1803.09001 [cs, stat]*. arXiv: 1803.09001 [cs, stat].

Sutton, Richard (1995). "TD models: Modeling the world at a mixture of time scales". In: *Machine Learning Proceedings 1995*. Elsevier, pp. 531–539.

Sutton, Richard, Joseph Modayil, et al. (2011). "Horde: A Scalable Real-time Architecture for Learning Knowledge from Unsupervised Sensorimotor Interaction". In: *The 10th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*. AAMAS '11. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Sutton, Richard and Brian Tanner (2004). "Temporal-Difference Networks". In: *Advances in Neural Information Processing Systems*.

van Hasselt, Hado and Richard Sutton (2015). "Learning to predict independent of span". In: *arXiv:1508.04582*.

Veeriah, Vivek et al. (2019). "Discovery of useful questions as auxiliary tasks". In: *Advances in Neural Information Processing Systems*, pp. 9306–9317.

White, Adam (2015). "Developing a Predictive Approach to Knowledge". In: *University of Alberta*.

Zheng, Zeyu et al. (2021). "Learning State Representations from Random Deep Action-conditional Predictions". In: *Advances in Neural Information Processing Systems* 34.