State-Wise Adaptive Discounting from Experience (SADE): A Novel Discounting Scheme for Reinforcement Learning (Student Abstract)

Milan Zinzuvadiya, Vahid Behzadan
Secured and Assured Intelligent Learning Lab
University of New Haven
West Haven, CT 06516
mzinz1@unh.newhaven.edu, vbehzadan@newhaven.edu
www.sail-lab.org

Abstract

In Markov Decision Process (MDP) models of sequential decision-making, it is common practice to account for temporal discounting by incorporating a constant discount factor. While the effectiveness of fixed-rate discounting in various Reinforcement Learning (RL) settings is well-established, the efficiency of this scheme has been questioned in recent studies. Another notable shortcoming of fixed-rate discounting stems from abstracting away the experiential information of the agent, which is shown to be a significant component of delay discounting in human cognition. To address this issue, we propose State-wise Adaptive Discounting from Experience (SADE) as a novel adaptive discounting scheme for RL agents. SADE leverages the experiential observations of state values in episodic trajectories to iteratively adjust state-specific discount rates. We report experimental evaluations of SADE in Q-learning agents, which demonstrate significant enhancement of sample complexity and convergence rate compared to fixed-rate discounting.

Introduction

In Markov Decision Processes (MDPs), the effect of delayed rewards on utilities is often captured by exponential discounting with a fixed discount rate γ. However, recent studies question the efficacy of this approach. For instance, ((Naik et al. 2019)) establishes that in Reinforcement Learning environments with continuous tasks, the stationary formulation of discounted MDPs is not an optimization problem. Inspired by the evidence supporting the adaptive dynamics of temporal discounting in human cognition (e.g., (Kurth-Nelson, Bickel, and Redish 2012)), we propose a novel scheme for State-Wise Adaptive Discounting from Experience (SADE), that leverages experience to iteratively adjust state-specific discount rates. Furthermore, we experimentally investigate the performance of SADE in comparison to commonly used fixed-rate approaches, namely: exponential discounting and hyperbolic discounting ((Fedus et al. 2019)).

Preliminaries

Accounting for delayed consequences is one of the defining features of sequential decision-making problems modeled as MDPs, since the objective of such problems is to maximize Return G, defined as the sum of all rewards received during an episode. MDPs are defined by the tuple $MDP = (S, A, R, P, \gamma)$, where S is the set of reachable states, A is the set of permissible actions, R is the mapping of transitions to immediate numeric rewards, P represents the transition probabilities (i.e., dynamics), and $\gamma \in [0, 1]$ is the discount factor, which is traditionally assumed to be a fixed value for all states and timesteps. In such formulations of MDPs, the return at timestep t_i, denoted by G_{t_i}, is the expected discounted sum of rewards. Formally,

$$G_t = \mathbb{E}[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}] \quad (1)$$

Where k is the number of timesteps from t, R_t is the reward at timestep t and γ is the discount factor. The discount rate measures the present utility of future rewards: the reward received k timesteps in the future is adjusted by a factor of γ^{k-1} to account for preference over sooner rewards.

State-wise Adaptive Discounting from Experience (SADE)

Considering the settings of episodic RL, SADE replaces the classical discount factor γ with a discount function $\lambda : S \rightarrow [0, 1]$, which provides a mapping of states to a corresponding discount rate. SADE hypothesizes that in the evolution of trajectories in consecutive episodes, the expected return increases for states that are more likely to be on or close to optimal trajectories. Accordingly, SADE increases the discount rate of states with increasing estimate of returns, and decreases the discount rate for vice versa.

In SADE, each state s is mapped to a specific discount factor, denoted by λ_s. Assume that at timestep t, expected return is G_t^{SADE} and the expected return of the immediate next state is G_{t+1}^{SADE} then, $G_t^{SADE} = R_t + \lambda_s * G_{t+1}^{SADE}$. The proposed discount function incorporates the past experiences of agent by adjusting the discount rate of each state based on the time-steps needed to achieve it. Accordingly, the SADE-discounted return of an agent within the finite horizon k is given by:

$$G^{SADE}(s_t) = R_t + \sum_{k=1}^{\infty} (\prod_{1}^{k} \lambda_k(s)) R_{t+k+1} \quad (2)$$

Where R_t is the reward at timestep t, and k is number of steps in the horizon.

Initially every λ_s are assigned a value $\in (0, 1)$. After each episode during training, values of λ_s for all $s \in$ trajectory will be adjusted with a predefined adjustment rate according to SADE hypothesis. To prevent $\lambda(s)$ from becoming zero during the adjustment, we also define the values $0 < \lambda_{min} < \lambda_{max} < 1$ where λ_{min} and λ_{max} are the lowest and highest permissible values of $\lambda(s)$ for all states, respectively. After passing initial value of λ_{in} ($\lambda_{min} \leq \lambda_{in} \leq \lambda_{max}$), The adjustment procedure is presented in Algorithm 1.

Algorithm 1 SADE Algorithm

```
input : adjustment rate $a\%$, upper and lower bounds
        $\lambda_{min}$,$\lambda_{max}$ and Initial $\lambda_{in}$
$\lambda_s \leftarrow \lambda_{in} \forall s \in S$
 After Each Episode for $\forall s \in Trajectory$ do
    if $G^{SADE}(s_{t-1}) < G^{SADE}(s_t)$ then
        $\lambda_{s_t} \leftarrow \min(\lambda_{max}, a\% - increased - \lambda_{s_t})$
    end
    if $G^{SADE}(s_{t-1}) > G^{SADE}(s_t)$ then
        $\lambda_{s_t} \leftarrow \max(\lambda_{min}, a\% - decreased - \lambda_{s_t})$
    end
end
```

Experimental Analysis

Grid World is a rectangular grid setting with $m \times n$ cells as states, where an RL agent starting with (x,y) state aims to traverse through to reach goal state (i, j) while avoiding blocked grids (i.e., walls). where $x, i < m$ and $y, j < n$. For comparison of SADE with fixed discounting, we implemented the standard Q-learning algorithms in Grid World using both schemes. In these experiments, we took $\lambda_{min} = 0.3$, $\lambda_{max} = 0.7$ and $\lambda_{in} = 0.5$. Our experiments were performed in 6 different dimensions of the Grid World , each trained for up to 1400 iterations or until it convergence. As illustrated in Fig. 1, in each setting, the agent is trained with exponential discounting, hyperbolic discounting ((Fedus et al. 2019)) and SADE. To measure the performance of each scheme, we measure the training efficiency defined as:

$$\text{Efficiency} = \frac{\sum R}{\text{No. Of Visited States In Training}} \tag{3}$$

We also experimented with 14 different adjustment ratesin each of the settings, and recorded the count of winners (i.e., agents with higher total scores), as demonstrated in Fig. 2.

It is observed that SADE outperforms both fixed-discounting schemes (i.e., exponential and hyperbolic) if the appropriate range of λ_s and adjustment rates are selected.

Discussion and Conclusion

As the experimental results demonstrate, SADE outperforms fixed exponential and hyperbolic discounting by at least a factor of 2 not only in terms of speed of convergence, but also in reward to episode-length ratio. As all 3 versions of Q-learning agents have identical settings except for their discounting schemes, our experiments strongly support the advantages of adaptive discounting over the classical fixed discounting.

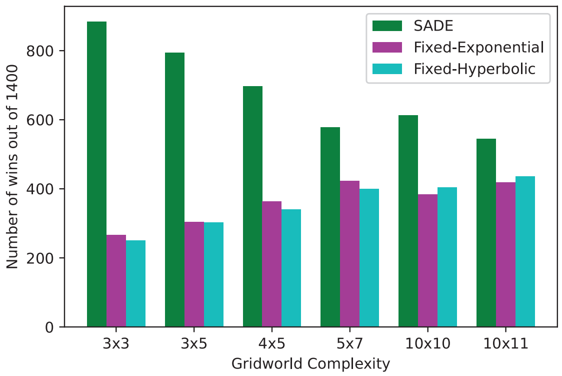

Figure 1: Comparison of Efficiency between Exponential, Hyperbolic and SADE Discounting

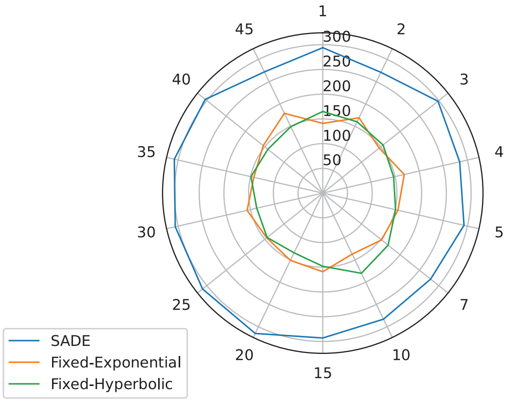

Figure 2: Comparison of performance between SADE with various adjustment rates and Exponential and Hyperbolic discounting

References

Fedus, W.; Gelada, C.; Bengio, Y.; Bellemare, M. G.; and Larochelle, H. 2019. Hyperbolic discounting and learning over multiple horizons. *arXiv preprint arXiv:1902.06865* .

Kurth-Nelson, Z.; Bickel, W.; and Redish, A. D. 2012. A theoretical account of cognitive effects in delay discounting. *European Journal of Neuroscience* 35(7): 1052–1064.

Naik, A.; Shariff, R.; Yasui, N.; and Sutton, R. S. 2019. Discounted reinforcement learning is not an optimization problem. *arXiv preprint arXiv:1910.02140* .