Bisimulation Prioritized Experience Replay

Oscar Guarnizo*[1], Mirco Giacobbe[1], and Leonardo Stella[1]

[1] University of Birminghan, School of Computer Science , United Kingdom

Abstract

Prioritized experience replay has been an effective traditional solution to online deep reinforcement learning challenges, such as data correlation and non-stationarity. However, standard prioritization often overlooks the nuanced, task-specific behaviors of states, leading to a "task-agnostic" sampling problem. This paper introduces a novel approach by incorporating an on-policy bisimulation metric into the experience replay prioritization process. This metric measures behavioral similarities and diversifies the training data, aiming to enhance learning by focusing on behaviorally relevant transitions. The proposed method balances between exploitation and exploration, addressing the limitations of conventional TD-error-based prioritization and enriching the training process with more informative state transitions.

1. Introduction

Incorporating deep learning techniques into reinforcement learning (RL) frameworks has been challenging due disparity in data assumption between deep learning and RL schemes [9]. Traditional deep learning relies on the independence of data samples for effective neural network training, whereas RL is characterized by a temporal sequential process that results in highly correlated states. Moreover, the data distribution in RL is non-stationary; it evolves as the algorithm acquires new behaviors. This dynamism leads to instability in deep learning, which typically assumes a fixed and identical underlying distribution.

Experience replay has been implemented in online RL algorithms, such as DQN [9], DDPG [8], SAC [6] to address both data correlation and non-stationary distributions issues. It facilitates breaking temporal data correlations, leading to approximate independent and identically distributed (iid) data distributions.

While experience replay benefits online RL, significant iterations may still be required for convergence. Schaul et al. [10] note that a DQN algorithm revisits the same experience tuple[†] an average of eight times, not all of which lead to significant improvements. In consequence, they proposed a prioritized experience replay, assigning probabilities to each experience based on the temporal difference (TD) error [11]. The TD-error priority works as an indicator of the expected learning progress; encouraging more frequently replay experiences which lead to higher improvements. Nonetheless, this prioritization can reduce data diversity, an issue alleviated through stochastic prioritization.

Prioritizing purely on TD-error, however, can overlook the task-specific behaviors of states, leading to what we term "task-agnostic" sampling problem, similar to representation learning findings in [12]. This perspective fails to recognize that certain states in a MDP, despite being structural dissimilar, can exhibit similar long-term behaviors under the same policy, resulting in similar expected returns in the long run. Bisimulation metrics [2, 3, 4, 1] provides a means of quantifying this behavioral similarity by considering both the immediate rewards and the expected future rewards (discounted over time), along with how states transition under a given policy.

Leveraging this behavioral concept could prioritize more informative tuples in the experience replay by identifying state pairs with significant behavioral differences as they often correspond to more 'surprising' transitions and improvements. Notice that by prioritizing behavioral dissimilar states, we are encouraging diversity on the sampled data, and consequently more exploration.

In this work, we propose incorporating the on-policy bisimulation metric into the prioritization process of an experience replay to 1) mitigate the loss of diversity caused by TD-error prioritization and 2) emphasize behaviorally relevant transitions, thereby avoiding task-agnostic experience sampling.

2. Background

A finite Markov Decision Process (MDP) is defined as a 5-tuple $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ is a finite set of actions, $\mathcal{P}(s'|s,a)$ is the probability of transitioning from state $s \in S$ to state $s' \in S$, $\mathcal{R} : S \times A \to R$ is the reward function , and $\gamma \in [0, 1)$ is a discount factor.

Initially introduced in the field of concurrency theory, **bisimulation** serves as a form of state abstraction that groups states s_i and s_j that are 'behaviorally equivalent' [7]. Givan et al. [5] later adapted bisimulation relations for MDPs providing a strong form of behavioral equivalence.

Definition 1. (Givan et al. [5]). Given an MDP $\mathcal{M}$, an equivalence relation B between states is a bisimulation relation if, for all states $s_i, s_j \in \mathcal{S}$ that are equivalent under B the following conditions hold:

$$\begin{aligned} \mathcal{R}(s_i, a) &= \mathcal{R}(s_j, a) && \forall a \in \mathcal{A}, \\ \mathcal{P}(G \mid s_i, a) &= \mathcal{P}(G \mid s_j, a) && \forall a \in \mathcal{A}, \quad \forall G \in \mathcal{S}_B, \end{aligned} \tag{1}$$

where $\mathcal{S}_B$ is the partition of $\mathcal{S}$ under the relation B, and $\mathcal{P}(G|s,a) = \sum_{s' \in G} \mathcal{P}(s'|s,a)$.

Two states $s_i, s_j \in S$ are **bisimilar** if there exists a bisimulation relation B such that $(s_i, s_j) \in B$; consequently, their

*__Proposal:__ This work is an early-stage research proposal without results yet, but I have provided sufficient theoretical evidence to support it.

[†]An experience tuple is (state s_t, action a_t, reward R_t, next state s_{t+1})

optimal value functions are equal, $V^*(s_i) = V^*(s_j)$.

The direct use of bisimulation relations is generally problematic because these relations are highly sensitive to infinitesimal variations in the reward function or dynamics, often resulting from data-driven estimations. For this reason, bisimulation metrics [2, 3, 4, 1] have been proposed to provide a smoother notion of similarity than that offered by strict equivalence relations. These metrics are defined within a pseudometric space $(\mathcal{S}, d)$, where a distance function $d : \mathcal{S} \times \mathcal{S} \to \mathbb{R}_{\geq 0}$ quantifies the 'behavioral similarity' between two states.

In this context, Castro [1] proposed an **on-policy bisimulation metric** that accounts for the dynamics induced by the current policy π, rather than those induced by the maximum among each individual action (as explored in Ferns et al. work [2]). This metric is particularly useful in RL schemes, addressing the dynamic nature of the policy, which is iteratively improved as the agent interacts with the environment.

Definition 2. (Castro [1] Theorem 2) π-bisimulation metric.

$$d^\pi(s_i, s_j) = |\mathcal{R}^\pi_{s_i} - \mathcal{R}^\pi_{s_j}| + \gamma \mathcal{W}_d(\mathcal{P}^\pi_{s_i}, \mathcal{P}^\pi_{s_j}) \tag{2}$$

where $\forall G \in S_B, \mathcal{P}^\pi_s(G) = \sum_a \pi(a|s) \sum_{s' \in G} \mathcal{P}(s'|s, a)$, $\mathcal{R}^\pi_s = \sum_a \pi(a|s)\mathcal{R}(s, a)$, and $\mathcal{W}_d$ is the d-Wasserstein metric.

3. Bisimulation Prioritized Experience Replay

Schaul et al. [10] argue that an ideal method for prioritizing experiences in reinforcement learning (RL) is by the **expected learning progress**, which is the amount the RL agent can learn from a transition in its current state. They suggest the TD-error (δ) as a effective surrogate metric, which reflects the 'surprising' nature of a transition. In Deep Q-Network (DQN)*, for an experience tuple $e_t = (s_t, a_t, \mathcal{R}_t, s_{t+1})$, the TD-error is defined as:

$$\delta_t = |\mathcal{R}_t + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^-) - Q(s_t, a_t; \theta)| \tag{3}$$

where Q represents the action-value function.

This work proposes an analogous surrogate through the on-policy bisimulation metric d^π, which stills captures the 'surprising' aspect of a transition, but additionally considers the behavioral similarity with respect to the MDP. Specifically, this metric prioritizes transitions between behaviorally dissimilar states, which can lead to greater expected learning progress. For instance, in Figure 1, while states s and u might exhibit structural difference, they still have similar long-term behaviors, which do not lead to considerable learning improvements. On the contrary, the state s and t exhibit both structural and behavioral differences, which could lead to a higher potential for learning progress.

While the TD-error could potentially exploit some transitions, leading to a loss of diversity [10], the bisimulation metric consistently encourages diversity in the experiences. This trade-off between exploitation and exploration will be regulated by a parameter $\eta \in [0, 1)$, resulting in the following mixed priority:

$$\text{priority}_t = (1 - \eta)\delta_t + \eta d^\pi(s_t, s_{t+1}) \tag{4}$$

*Note that analogous TD-errors could be defined in other RL algorithms.

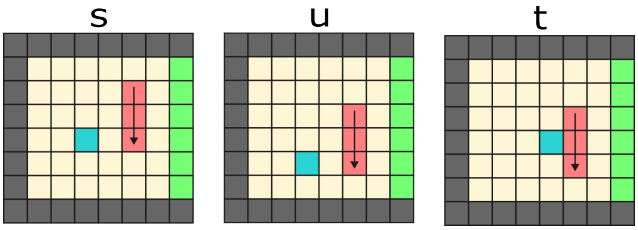

Figure 1. A simple toy example: The goal of the sky-blue agent is to find the shortest path to the green area while avoiding a moving obstacle that shifts one cell at a time. Both u and t are possible next states from state s. The states s and u are behavioral similar, while states s and t are not.

with sampling probability for the experience e_t given by

$$P(t) = \frac{\text{priority}_t^\alpha}{\sum_k \text{priority}_k^\alpha} \tag{5}$$

where α controls the degree of prioritization.

This probability is assigned to each incoming experience transition and utilized to sample data from the experience replay during the training phase. Similar to the standard prioritized experience replay method [10], certain implementation considerations must be addressed, especially when dealing with large experience replay databases. These considerations include potential biases and sampling challenges, which are mitigated through the use of importance sampling and efficient sampling techniques.

4. Proposed Experiments

While the bisimulation metric offers a robust theoretical framework for behavioral similarity, calculating it, especially the Wasserstein term $\mathcal{W}_1$, remains difficult in high-dimensional or continuous state spaces. Based on Castro's work [1], which indicates computing the Wasserstein metric is no longer necessary under a system with deterministic transitions, our experimentation will concentrate on deterministic MDP environments. In these settings, each state-action pair leads to a unique subsequent state, thus ensuring certainty of the next state. Additionally, Castro proposes a learnable approximation of the on-policy bisimulation metric for large (or continuous) state spaces, which we will adopt and train within the main RL loop.

The proposed methods will be tested in two different setups focusing on deterministic MDPs: 1) Grid Worlds, similar to those shown in Figure 1, where calculating the bisimulation metrics is relatively straightforward; and 2) Atari 2600 benchmark suite, with modified background distractors akin to those described in [12], to explore how the algorithm prioritizes behavioral dissimilarity over structural dissimilarity..

5. Conclusion

This work pioneers the integration of the on-policy bisimulation metric into experience replay prioritization, offering a nuanced alternative to traditional methods. The proposed mechanism, blending TD-error and bisimulation metrics, demonstrates a promising direction for handling the inherent complexities of RL environments. By emphasizing behavioral dissimilarity, the approach ensures a diverse and informative training set, potentially accelerating learning and enhancing model performance.

References

[1] P. S. Castro. Scalable methods for computing state similarity in deterministic markov decision processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10069–10076, 2020. 1, 2

[2] N. Ferns, P. Panangaden, and D. Precup. Metrics for finite markov decision processes. In *UAI*, volume 4, pages 162–169, 2004. 1, 2

[3] N. Ferns, P. Panangaden, and D. Precup. Bisimulation metrics for continuous markov decision processes. *SIAM Journal on Computing*, 40(6):1662–1714, 2011. 1, 2

[4] N. Ferns and D. Precup. Bisimulation metrics are optimal value functions. In *UAI*, pages 210–219, 2014. 1, 2

[5] R. Givan, T. Dean, and M. Greig. Equivalence notions and model minimization in markov decision processes. *Artificial Intelligence*, 147(1-2):163–223, 2003. 1

[6] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018. 1

[7] L. Li, T. J. Walsh, and M. L. Littman. Towards a unified theory of state abstraction for mdps. *AI&M*, 1(2):3, 2006. 1

[8] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015. 1

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013. 1

[10] T. Schaul, J. Quan, I. Antonoglou, and D. Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015. 1, 2

[11] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018. 1

[12] A. Zhang, R. McAllister, R. Calandra, Y. Gal, and S. Levine. Learning invariant representations for reinforcement learning without reconstruction. *arXiv preprint arXiv:2006.10742*, 2020. 1, 2