SIBRE: Self Improvement Based REwards for Adaptive Feedback in Reinforcement Learning

Somjit Nath [1], **Richa Verma**[1], **Abhik Ray**[2], **Harshad Khadilkar**[1]

[1]TCS Research, [2]BITS-Pilani(Goa)

Introduction

We propose a generic reward shaping approach for improving the rate of convergence in reinforcement learning (RL), called **S**elf **I**mprovement **B**ased **RE**wards, or **SIBRE**.

- The reward modification is computationally light (simple average) and can be used to improve the sample efficiency of any RL algorithm.
- SIBRE converges in expectation to the same policy as the original algorithm.
- We empirically observe faster convergence with lower variance on a variety of benchmark environments, with multiple RL algorithms.

SIBRE

In this paper, we propose a modification to the reward function (called **SIBRE**, short for Self Improvement Based REward) that aims to improve the rate of learning in episodic environments and thus addresses the problem of sample efficiency through reward shaping.
SIBRE is a threshold-based reward shaping algorithm for RL, which provides a positive reward when the agent improves on its past performance, and negative reward otherwise. It is done by replacing the reward:

$$r_{k,t}(s_k, a_k, s_{k+1}) = \begin{cases} G_t - \rho_t, & s_{k+1} \in \mathcal{T} \\ R_k, & \text{otherwise} \end{cases}$$

where ρ_t is the performance threshold at episode t, which is updated separately. The update is defined by:

$$\rho_{t+1} = \begin{cases} \rho_t + \beta_t(\sum_{y=t-x+1}^{t} \frac{G_y}{x} - \rho_t) & \text{if updating q-values} \\ \rho_t & \text{otherwise} \end{cases},$$

where $\beta_t \in (0, 1)$ is the step size and is assumed externally defined according to a fixed schedule.

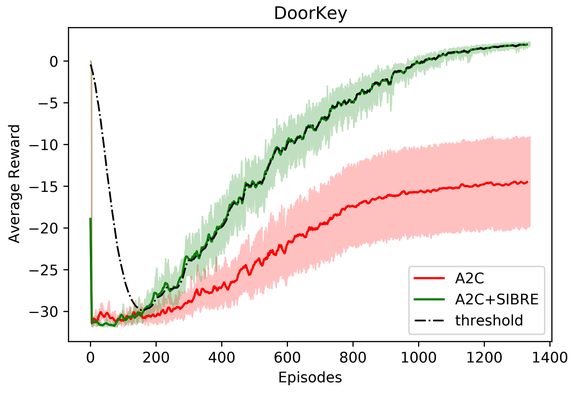
Algorithm

Algorithm 1: Illustration of SIBRE with Q-learning

Algorithm parameters: step size $\alpha \in (0, 1]$, $\epsilon > 0$, $\beta \in (0, 1)$;
Threshold Update after x episodes;
Initialize $Q(s, a)$, for all $s \in \mathcal{S}, a \in \mathcal{A}(s)$, $\rho = 0$;
foreach *episode* **do**
  Initialize S;
  G = 0;
  foreach *step of episode* **do**
    Choose A from S using policy derived from Q (e.g., ε-greedy);
    Take action A, observe R, S';
    $G = G + R$;
    if $S \in$ terminal **then**
      $R = G - \rho$;
      if $episodecount \bmod x = 0$ **then**
        $\rho \leftarrow (1 - \beta)\rho + \beta G$;
      end if
    end if
    $Q(S, A) \leftarrow (1 - \alpha)Q(S, A) + \alpha[R + \gamma \max_a Q(S', a)]$;
    $S \leftarrow S'$;
  end foreach
end foreach

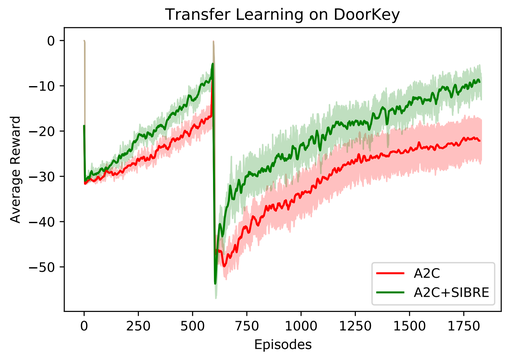
Accelerating Learning

From the learning curve shown on 6x6 DoorKey, we can see how integration of SIBRE can help not only in accelerating learning and but helps in converging to optimal policy.



Transfer Learning

SIBRE learns the value of a threshold which it aims to beat after each episode. We believe that once it has learnt the threshold properly, we get optimal performance. When we use the same model to learn on a bigger state-space with same reward structure, the value of the threshold provides a high initial value to beat and this helps in easy transfer of learning. In the figure below we do see such improvement while transferring from 5x5 to 8x8 grid in Doorkey.



Experiments on a variety of other domains, further hyper-parameter analysis and extension to continuing MDPs along with the exact hyper-parameters for reproduction of such results are presented in [1].

Conclusion

In this work, we showed that an adaptive, self-improvement based modification to the terminal reward (SIBRE) has empirically better performance, both qualitative and quantitative, than the original RL algorithms on a variety of environments. We were able to prove, analytically, that SIBRE converges to the same policy in expectation, as the original algorithms.

References

[1] Somjit Nath et al. *SIBRE: Self Improvement Based REwards for Adaptive Feedback in Reinforcement Learning*. 2020. arXiv: 2004.09846 [cs.LG].