Reinforcement Learning Targeting Arbitrary Expected Rewards

Bachelor's Thesis by Antonia Halbig
Advisors: Thomas Gabor, Thomy Phan
Supervisor: Prof. Dr. Claudia Linnhoff-Popien

In this thesis we introduce a variant of reinforcement Learning (RL) similar to UDRL proposed by Schmiduber in 2019.[1] Instead of teaching an agent to maximize the accumulated reward as in traditional RL, our algorithm aims to **accumulate an arbitrary target reward**. For this we utilize a Deep Q-Network (DQN).

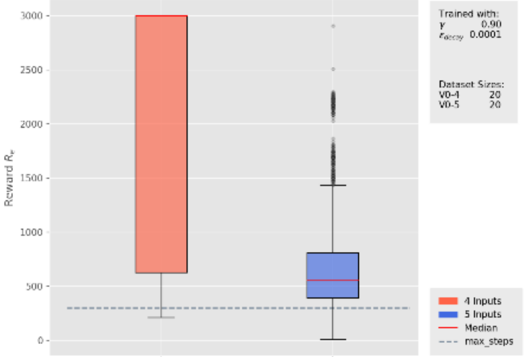

Fig. 1 Evaluation of V0

As a target reward we select a random number T_R each episode of each training. This new parameter can be integrated into the code in two ways. The **first version (V1)** adds the target reward during action selection by choosing the action with the expected accumulated reward closer to the target. The **second version (V2)** inputs the difference d of the received reward R_e to the target reward directly into the neural net as an additional part of the observation. With reward manipulation the net changes its output to the negative future difference between the expected reward R and T_R. In addition to reward shaping we explore curriculum learning (CL) to further improve performance in V2. For comparison we implemented a **base version (V0)** with traditional Q Learning.

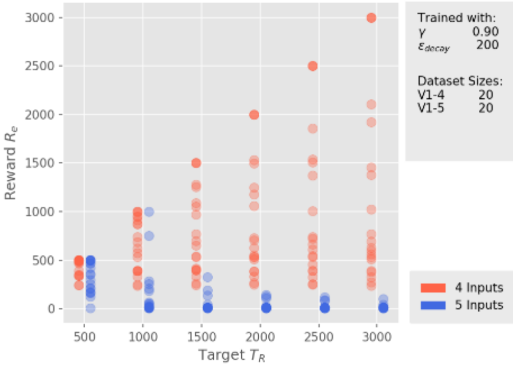

Fig. 2 Evaluation of V1

The approach is tested and developed in the CartPole domain, which is based on the inverted pendulum problem. The utilized neural networks are small feed forward neural networks. They have the same structure, only one is given 4 and the other 5 inputs. The first 4 inputs are set by the environment, the 5th input is the difference d. V0 and V1 are trained with both the 4- and 5-input network. We choose $T_R \in [50, 250]$.

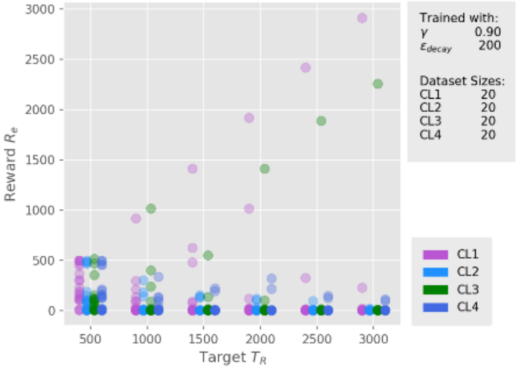

Fig. 3 Evaluation of V2 with CL

The results of V0 show that, while the 4- and 5-input network perform very similar in training, **the 5-input network deteriorates in the evaluation** with an average reward just above 500, as shown in Fig. 1. 4-input networks are capable of reaching rewards of 3000 and more. When examining the 5-input network more closely, we observe that weights of the additional input are not set to zero and therefore effect the agent's choice. **With V1 as well as V2 we trained agents able to precisely target arbitrary rewards**. In V1 both agents are successful at accumulating T_R with an average difference of 0.2. Aiming for values above the trained interval, the agent with 4 inputs surpasses the agent with 5. It can target values up to 3000 and higher, as seen in Fig. 2. In V2 the best reward shaping method reaches an average loss of 1.5 while also being able to target values up to 1500. By combining it with CL it can successfully aim for values up to 3000 as well as further improving its accuracy to 0.5, which is shown in Fig. 3.

[1] Schmidhuber, Juergen: Reinforcement Learning Upside Down: Don't Predict Rewards – Just Map Them to Actions. arXiv preprint arXiv:1912.02875, 2019.