Time Hopping Technique for Reinforcement Learning and its Application to Robot Control

Petar Kormushev
Department of Computational Intelligence and Systems Science
Tokyo Institute of Technology, Japan
Email: petar@hrt.dis.titech.ac.jp

ABSTRACT

To speed up the convergence of reinforcement learning (RL) algorithms by more efficient use of computer simulations, three algorithmic techniques are proposed: Time Manipulation, Time Hopping, and Eligibility Propagation. They are evaluated on various robot control tasks.

The proposed *Time Manipulation* [1] is a concept of manipulating the time inside a simulation and using it as a tool to speed up the learning process. It is applicable to a subset of RL problems whose goal is to learn a control policy to avoid failure events. Time Manipulation works by turning back the time of the simulation on failure events, thus avoiding redundant state transitions and exploring deeper the state space. This is impossible to be done in the real world, but it can easily be done in a simulation. In order to evaluate the proposed algorithm, experiments on a classical control benchmark problem are conducted: an inverted pendulum balancing robot task. The aim of the RL algorithm is to find a control policy which can prevent the pendulum from falling by moving the robot left or right, without hitting the edges of the given track. The experimental results show that Time Manipulation speeds up the learning process by 260%. It also improves the state space exploration by 12%, because it allows the RL algorithm to explore better the state space in proximity of failure states.

The proposed *Time Hopping* [2] is a generalization of Time Manipulation, able to make arbitrary "hops" between states and this way traverse rapidly throughout the entire state space. Time Hopping extends the applicability of time manipulations to include not only failure-avoidance problems, but also continuous optimization problems, by creating new mechanisms to trigger the time manipulation events, to make prediction about the possible future rewards, and to select promising time hopping targets. The proposed implementation of the Time Hopping technique consists of 3 components: Hopping trigger (decides when the hopping starts), Target selection (decides where does it hop to), and Hopping (performs the actual hopping). For the implementation of the Hopping trigger component, a Gamma pruning technique is proposed, which detects and prunes unpromising exploratory paths. For the Target selection component, a Best Lasso Target Selection technique is proposed, which selects a target among the proximity of the current best policy.

The evaluation of Time Hopping is performed on a biped crawling robot task. The crawling robot has 2 limbs, each with 2 segments, for a total of 4 degrees of freedom (DOF), 80 possible actions at each time step, and 13689 possible robot states. The goal of the learning process is to find a crawling motion with the maximum speed. The reward function for this task is defined as the horizontal displacement of the robot after every action. The experimental results show that Time Hopping accelerates the learning process more than 7 times. A very strong point of Time Hopping is that it is completely transparent for the RL algorithm, which offers various opportunities for combining Time Hopping with other approaches for speeding up the learning process. In addition, it can also be used as a tool for re-shaping the state probability distribution as desired [3]. This is achieved by changing the target selection strategy appropriately.

The proposed *Eligibility Propagation* [4] is a mechanism to further speed up Time Hopping. It provides similar abilities to what eligibility traces provide for conventional RL. Eligibility traces are one of the basic mechanisms for temporal credit assignment in reinforcement learning. An eligibility trace is a temporary record of the occurrence of an event, such as the visiting of a state or the taking of an action. Eligibility Propagation uses the transitions graph to obtain all predecessor states of an updated state. Regardless of the actual order in which Time Hopping visits the states, this oriented graph contains a record of the correct chronological sequence of state transitions. Once this oriented graph is available, it is used to propagate state value updates in the opposite direction of the state transition edges, thus making the propagation flow logically backwards in time.

The evaluation of Eligibility Propagation is performed on the same biped crawling robot task as for Time Hopping. The results show that Time Hopping with Eligibility Propagation achieves 99% of the maximum possible speed almost 3 times faster than Time Hopping alone, and more than 4 times faster than conventional Q-learning. This significant speed-up of the learning process is achieved despite the additional computational overhead of maintaining the transitions graph. The reason for this is the improved Gamma-pruning based on more precise future reward predictions.

All experiments are conducted on a custom developed Java-based software application system, and a custom developed 2D robot physics simulation engine. The significant speed-ups achieved by the proposed algorithms make them very suitable for a wide range of robot control problems, where reducing the computational cost is important [5].

ACKNOWLEDGMENTS

This work was supported in part by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT).

REFERENCES

[1] P. Kormushev, K. Nomoto, F. Dong, and K. Hirota, "Time manipulation technique for speeding up reinforcement learning in simulations," *International Journal of Cybernetics and Information Technologies*, vol. 8, No. 1, pp. 12–24, 2008.

[2] P. Kormushev, K. Nomoto, F. Dong, and K. Hirota, "Time hopping technique for faster reinforcement learning in simulations," *International Journal of Cybernetics and Information Technologies*, vol. 11, no. 3, pp. 42–59, 2011.

[3] P. Kormushev, F. Dong, and K. Hirota, "Probability redistribution using time hopping for reinforcement learning," in *Proc. 10th International Symposium on Advanced Intelligent Systems, ISIS 2009*, Busan, Korea, 2009.

[4] P. Kormushev, K. Nomoto, F. Dong, and K. Hirota, "Eligibility propagation to speed up time hopping for reinforcement learning," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 13, No. 6, 2009.

[5] P. Kormushev, "Time hopping technique for reinforcement learning and its application to robot control," PhD thesis, Tokyo Institute of Technology (TiTech), Japan, 2009.