

**Efficient Approximate Policy Iteration Methods  
for Decision Making in Reinforcement Learning**

Reinforcement learning is a promising learning paradigm in which an agent learns how to make good decisions by interacting with an (unknown) environment. This learning framework can be extended along two dimensions: the number of decision makers (single- or multi- agent) and the nature of interaction (collaborative or competitive). This characterization leads to the four decision making situations that are considered in this thesis and are modeled as Markov decision process (MDPs), Collaborative Markov Decision Processes, Zero-Sum Markov Games, and Team Zero-Sum Markov Games.

Existing reinforcement learning algorithms have not been applied widely on real-world problems, mainly because the required resources grow fast as a function of the size of the problem. Exact, but impractical, solutions are commonly abandoned in favor of approximate, but practical, solutions. Unfortunately, research on efficient and stable approximate methods has focused mainly on the prediction problem, where an agent tries to learn the outcome of a fixed decision making policy. This thesis contributes two efficient and stable algorithms for decision making problems, based on the general framework of approximate policy iteration.

Least-Squares Policy Iteration (LSPI) is an algorithm for learning good policies based on a least-squares fixed-point approximation of the value function. LSPI makes efficient use of sample experience and therefore is most appropriate for domains where training data are expensive or a simulator of the process is not available. Rollout Policy Iteration (RLPI) on the other hand learns good policies based on the use of rollout outcomes to train a classifier that represents an approximate policy. For that reason, RLPI is most appropriate for domains where experience comes at no cost or where a simulator is available. LSPI and RLPI are complementary in the sense that one can be applied in domains where the other cannot and vice-versa. They also exhibit several nice theoretical properties, rarely found in approximate RL algorithms, and they can benefit by advances in feature selection / subspace projections and supervised learning for classification, respectively.

The proposed algorithms are demonstrated on a variety of learning tasks: inverted pendulum balancing, bicycle balancing and riding, the game of Tetris, multiagent system administration, distributed power grid control, server-router flow control, two-player soccer game, and multiagent server-router flow control. These results demonstrate clearly that the efficiency and applicability of the new algorithms achieve fast learning even for large control problems.