The Steering Approach for Multi-Criteria Reinforcement Learning

Shie Mannor and Nahum Shimkin
Department of Electrical Engineering
Technion, Haifa 32000, Israel
{shie,shimkin}@{tx,ee}.technion.ac.il

Abstract

We consider the problem of learning to attain multiple goals in a dynamic environment, which is initially unknown. In addition, the environment may contain arbitrarily varying elements related to actions of other agents or to non-stationary moves of Nature. This problem is modelled as a stochastic (Markov) game between the learning agent and an arbitrary player, with a vector-valued reward function. The objective of the learning agent is to have its long-term average reward vector belong to a given target set. We devise an algorithm for achieving this task, which is based on the theory of approachability for stochastic games. This algorithm combines, in an appropriate way, a finite set of standard, scalar-reward learning algorithms. Sufficient conditions are given for the convergence of the learning algorithm to a general target set. The specialization of these results to the single-controller Markov decision problem are discussed as well.

1 Introduction

This paper considers an on-line learning problem for Markov decision processes with vector-valued rewards. Each entry of the reward vector represents a scalar reward (or cost) function which is of interest. Focusing on the long-term average reward, we assume that the desired performance is specified through a given target set, to which the average reward vector should eventually belong. Accordingly, the specified goal of the decision maker is to ensure that the average reward vector will converge to the target set. Following terminology from game theory, we refer to such convergence of the reward vector as approaching the target set.

A distinctive feature of our problem formulation is the possible incorporation of arbitrarily varying elements of the environment, which may account for the influence of other agents or non-stationary moves of Nature. These are collectively modelled as a second agent, whose actions may affect both the state transition and the obtained rewards. This agent is free to choose its actions according to any control policy, and no prior assumptions are made regarding its policy.

This problem formulation is derived from the so-called theory of approachability that was introduced in [3] in the context of repeated matrix games with vector payoffs. Using a geometric viewpoint, it characterizes the sets in the reward space that a player can guarantee for himself for any possible policy of the other player, and provides appropriate policies for approaching these sets. Approachability theory has been extended to stochastic (Markov) games in [14], and the relevant results are briefly reviewed in Section 2. In this paper we add the learning aspect, and consider the problem of learning such approaching policies on-line, using Reinforcement Learning (RL) or similar algorithms.

Approaching policies are generally required to be non-stationary. Their construction relies on a geometric viewpoint, whereby the average reward vector is “steered” in the direction of the target set by the use of direction-dependent (and possibly stationary) control policies. To motivate the steering viewpoint, consider the following one dimensional example of an automatic temperature