Solving Simulator-Defined MDPs for Natural Resource Management

Thomas G. Dietterich, Majid Alkaee Taleghan,
Mark Crowley, Sean McGregor
Oregon State University, Corvallis, Oregon, USA

Natural resource management problems, such as forestry, fisheries, and water resources, can be formulated as Markov decision processes. However, solving them is difficult for two reasons. First, the dynamics of the system are typically available only in the form of a complex and expensive simulator. This means that MDP planning algorithms are needed that minimize the number of calls to the simulator. Second, the systems are spatial. A natural way to formulate the MDP is to divide up the region into cells, where each cell is modeled with a small number of state variables. Actions typically operate at the level of individual cells, but the spatial dynamics couple the states of spatially-adjacent cells. The resulting state and action spaces of these MDPs are immense.

We have been working on two natural resource MDPs. The first involves the spread of tamarisk in river networks. A native of the Middle East, tamarisk has become an invasive plant in the dryland rivers and streams of the western US. Given a tamarisk invasion, a land manager must decide how and where to fight the invasion (e.g., eradicate tamarisk plants? plant native plants? upstream? downstream?). Although approximate or heuristic solutions to this problem would be useful, our collaborating economists tell us that our policy recommendations will carry more weight if they are provably optimal with high probability.

A large problem instance involves 4.7×10^6 states with 2187 actions in each state. On a modern 64-bit machine, the action-value function for this problem can fit into main memory. However, computing the full transition function to sufficient accuracy to support standard value iteration requires on the order of 3×10^{20} simulator calls.

The second problem concerns the management of wildfire in Eastern Oregon. In this region, prior to European settlement, the native ponderosa pine forests were adapted to frequent, low-intensity fires. These fires allow the

ponderosa pine trees (which are well-adapted to survive fire) to grow very tall while preventing the accumulation of fuel at ground level. These trees provide habitat for many animal species and are also very valuable for timber. However, beginning in the early 1900s, all fires were suppressed in this landscape, which has led to the build up of huge amounts of fuel. The result has been large, catastrophic fires that kill even the ponderosa trees and that are exceptionally expensive to control. The goal of fire management is to return the landscape to a state where frequent, low-intensity fires are again the normal behavior. There are two concrete fire management problems: LetBurn (decide which fires to suppress) and FuelTreatment (decide in which cells to perform fuel reduction treatments).

Note that in these problems, the system begins in an unusual, non-equilibrium state, and the goal is to return the system to a desired steady state distribution. Hence, these problems are not problems of reinforcement learning, but rather problems of MDP planning for a specific start state. Many of the assumptions in RL papers, such as ergocity of all policies, are not appropriate for this setting. Note also that it is highly desirable to produce a concrete policy (as opposed to just producing near-optimal behavior via receding horizon control). A concrete policy can be inspected by stakeholders to identify missing constraints, state variables, and components of the reward function.

To solve these problems, we are exploring two lines of research. For tamarisk, we have been building on recent work in PAC-RL algorithms (e.g., MBIE , UCRL, UCRL2, FRTDP, OP) to develop PAC-MDP planning algorithms. We are pursuing two innovations. First, we have developed an exploration heuristic based on an upper bound on the discounted state occupancy probability. Second, we are developing tighter confidence intervals in order to terminate the search earlier. These are based on combining Good-Turing estimates of missing mass (i.e., for unseen outcomes) with sequential confidence intervals for multinomial distributions. These reduce the degree to which we must rely on the union bound, and hence give us tighter convergence.

For the LetBurn wildfire problem, we are exploring approximate policy iteration methods. For FuelTreatment, we are extending Crowley's Equilibrium Policy Gradient methods. These define a local policy function that stochastically chooses the action for cell i based on the actions already chosen for the cells in the surrounding neighborhood. A Gibbs-sampling-style MCMC method repeatedly samples from these local policies until a global equilibrium is reached. This equilibrium defines the global policy. At equilibrium, gradient estimates can be computed and applied to improve the policy.