Sparse Distributed Memories in a Bounded Metric State Space: Some Theoretical and Empirical Results.

Alexandre Bouchard-Côté

August 23, 2004

Abstract

Sparse Distributed Memories (SDM) is a linear, local function approximation architecture that can be used to represent cost-to-go or state-action value functions of reinforcement learning (RL) problems. It offers a possibility to reconcile the convergence guarantees of linear approximators and the potential to scale to higher dimensionality typically found only in nonlinear architectures. In this presentation, I will expose the results of our investigations in both avenues to see if SDM can fulfill its promises in the context of a RL problem with a bounded metric state space. On the theoretical side, algorithms from the two main categories of techniques to solve RL problems (value and policy iteration) were considered. Surprisingly, while a version of Q-learning can be proven to converge with SDM (approximate value iteration algorithms are usually expected to have no such guarantees, even with linear approximation architectures), an important result on the non-divergence of SARSA, an optimistic approximate policy iteration algorithm, failed to apply in our case. On the other hand, one of the most important convergence result in reinforcement learning, namely the convergence of TD(λ), was successfully translated into the language of measure theory to cover the case of Markov processes with an invariant measure taken from a general probability space. This forms the foundation for an eventual proof of convergence of a "continuous" version of an approximate policy iteration algorithm. On the empirical side, the first step was to design and implement a specialized data structure to store and retrieve hard locations (basis points). The specifications of this hash-based data structure will be exposed, as well as statistics on the performances of SDM equipped with this data structure on RL tasks. A disturbing observation suggested by those statistics is that the SARSA algorithm had a "better behavior" during convergence than the Q-learning algorithm, an apparent contradiction to our theoretical results. We propose an explanation to this discrepancy and conclude that policy iteration algorithms should be preferred over value iteration algorithms in our framework.