Towards Bounded Optimal Meta-Level Control: A Case Study

Daishi Harada
UC Berkeley
daishi@cs.berkeley.edu

The Idea

In this thesis abstract, due to space limitations, we omit references and assume that the reader is familiar with the basic ideas and terminology from the fields of reinforcement learning, search, and rationality. In this first section, we motivate our idea using MDPs.

Let $\langle X, U, T, R\rangle$ be a MDP of ⟨states, controls, transitions, rewards⟩. Assume that we have an estimate of the value function $\hat{V}$. The standard method of obtaining a policy π from this estimate is to define $\pi_{\hat{V}}(x) = \mathrm{argmax}_{u \in U}\, E\left[R(x,u) + \gamma \hat{V}(Y)\right]$. It is clear that this may be interpreted as the controller performing a depth 1 search/lookahead.

Now suppose we instead allow the controller to perform arbitrary search, and to base its control on the backed up information. To do this, we need to make decisions about the following: the order in which search nodes are expanded, and when to stop searching and actually "commit" to a control. The approach that we take is to view these decisions as the meta-level control problem. With some care in the formulation, it can be seen that a solution to this meta-level control problem will provide us with a bounded optimal controller. We would like to solve this problem by using algorithms from reinforcement learning.

Some Discussion

Let us now consider the issues and problems presented by this approach. First, we would certainly like a formalism which allows us to express the joint system of the domain and the controller as an asynchronous, "actively" interacting pair of systems. This is difficult because the standard formulation of MDPs does not include a notion of time. Although SMDPs address the issue of time, it does not easily express the idea of strongly coupled subsystems. Currently the most promising approach seems to be that of merging the formalisms considered by the two communities of reinforcement learning and concurrent systems.

Given a formalism, the theoretical results which we would like to show concern whether it is possible to find the optimal policy. It is fairly clear that given the "more sophisticated" formalism posited above it will be necessary to similarly extend the canonical "learning algorithms" from reinforcement learning. In particular, it will be necessary to integrate the techniques explored by the hierarchical/modular architecture subcommunity with that of the function approximation subcommunity. That the latter is relevant follows directly from the assumption that we have an estimate of the value function. The former also follows since we have a system with subcomponents (the domain and the controller); indeed, even stronger connections follow from the fact that we are considering a controller based on search.

Setting theoretical concerns aside, it still seems plausible that intuition can be used to guide us towards building a system that is at least useful in practice. Unfortunately, this is also non-trivial. Returning to the idea presented in the previous section, note that since we have assumed that having the controller search ahead to improve its performance is reasonable, the problem which remains is to construct an architecture for the meta-level controller which guides the search.

The search community has been exploring this issue soley within the context of search for quite some time. In general, the underlying intuition has been to formulate the "value" of searching further along a particular path in terms of the volatility of the search nodes. There is not, however, a foundation to motivate the idea of "value". Our perspective is an attempt to address this. In principle, the problem of real time control provides a framework within which one can formulate a coherent notion of the value of computation.

Based on the practical successes of the intuitions from the search community, it seems reasonable to use some measure of volatility as a relevant feature for the meta-level controller. Using this information to obtain a meta-level decision efficiently returns to the hierarchical framework alluded to above. In particular, the sequence of decisions from the root concerning which branch to expand next has a nice interpretation within this framework.