Policy Gradient Methods for Reinforcement Learning with Function Approximation

Richard S. Sutton, David McAllester, Satinder Singh, Yishay Mansour

You may optionally include all authors on the cover page.

Keywords: Policy Gradient, Compatible Function Approximation

Summary

In this example the summary is the same as the paper abstract.

Function approximation is essential to reinforcement learning, but the standard approach of approximating a value function and determining a policy from it has so far proven theoretically intractable. In this paper we explore an alternative approach in which the policy is explicitly represented by its own function approximator, independent of the value function, and is updated according to the gradient of expected reward with respect to the policy parameters. Williams's REINFORCE methods and actor-critic methods are examples of this approach. Our main new result is to show that the gradient can be written in a form suitable for estimation from experience aided by an approximate action-value or advantage function. Using this result, we prove for the first time that a version of policy iteration with arbitrary differentiable function approximation is convergent to a locally optimal policy.

Contribution(s)

lthough not a requirement, e recommend ordering the ntributions by significance.

1. This paper presents conditions under which substituting an approximation of the action-value function in place of the true action-value function within an expression for the policy gradient does not introduce error. This result applies to both the discounted and episodic setting and the average-reward and continuing setting.
   Context: The conditions require a parametric approximation of the action-value function using optimal parameters—parameters that are not known in practice, but which can be approximated.
2. This paper presents the first proof that a form of policy iteration with function approximation is convergent to a locally optimal policy.
   Context: Like the standard policy iteration algorithm, implementing this approximate form of policy iteration still requires knowledge of the transition function and reward function of the underlying Markov decision process.

   The context could correct common expected misconceptions about the meaning of a contribution, even if the careful reader should already be able to infer this context.
3. This paper presents an expression for the policy gradient in terms of the action-value function and partial derivative of the parametric policy with respect to its parameters. This single expression applies to both the discounted and episodic setting and the average reward and continuing setting.
   Context: Expressions for the policy gradient have been derived previously for the episodic setting only (Williams, 1992) and the continuing setting only (Marbach & Tsitsiklis, 1998), although neither of these were written in terms of the action-value function. Other policy gradient algorithms that apply to both the episodic and continuing settings also exist (Baxter & Bartlett, 1999).

The cover page uses the me bibliography as the rest ne paper—it does not have its own list of references. Although there may be ferences that only occur on the cover page, this should be uncommon.