Policy Gradient Methods:

Variance Reduction and Stochastic Convergence

Evan Greensmith

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University

March 2005

Typeset in Palatino by TEX and LATEX 2ε.

This thesis is presented in two related parts. The contents of Part I of this thesis appears in a journal paper with Peter L. Bartlett and Jonathan Baxter (Greensmith et al. 2004), as well as in an earlier conference paper (Greensmith et al. 2002); the work was seeded by some ideas of Peter L. Bartlett and Jonathan Baxter. The contents of Part II of this thesis appears in a technical report with Peter L. Bartlett (Greensmith and Bartlett 2005). During the creation process the contents of both parts of this thesis were discussed with Peter L. Bartlett, whereby both technical advice and direction were given. However, except where otherwise indicated, this thesis is primarily my own original work.

Evan Greensmith
23 March 2005

Acknowledgements

I am most grateful for the support of my supervisor Professor Peter Bartlett who supplied direction, comments, and technical advice for the work in this thesis.

I would like to thank Peter for maintaining contact during a period of his life that saw many changes: two continents, three jobs, six homes, and a second child. I would also like to thank Peter, Selena, James and Ella for sharing two of those homes with me on various visits to Berkeley.

I would like to thank Dr. Edward Francis Harrington and Peter for proof reading much this thesis, and Annie Josline for supplying food to go with Ed's criticism. Cheng Soon Ong also proof read portions of this thesis, for which I am grateful.

The group in and around the Computer Science Laboratory have been supportive and helpful. I discussed numerous small problems, some of which still bug me, with Kee Siong Ng and Edward Harrington. I also benefited from more general discussions, both in and out of the ML reading group, with various people, including Cheng Soon Ong, Petra Phillips, Omri Guttman, Tim Sears, Paul Wong, Charles Gretton, Douglas Aberdeen, Olivier Buffet, S.V.N. Vishwanathan. I would also like to thank Michelle Moravec for help with all things administrative and botanical.

I am grateful to my parents—Jan and Don—my brothers and sister, and close family and friends, who I rely upon even when they are not around.

During the term of my PhD I was financially supported by the Australian government through an Australian Postgraduate Award, and by The Australian National University through an ANU Supplementary Scholarship. I also received reimbursement for travel expenses from the University of California, Berkeley.

Abstract

In a reinforcement learning task an agent must learn a policy for performing actions so as to perform well in a given environment. Policy gradient methods consider a parameterized class of policies, and using a policy from the class, and a trajectory through the environment taken by the agent using this policy, estimate the performance of the policy with respect to the parameters. Policy gradient methods avoid some of the problems of value function methods, such as policy degradation, where inaccuracy in the value function leads to the choice of a poor policy. However, the estimates produced by policy gradient methods can have high variance.

In Part I of this thesis we study the estimation variance of policy gradient algorithms, in particular, when augmenting the estimate with a baseline, a common method for reducing estimation variance, and when using actor-critic methods. A baseline adjusts the reward signal supplied by the environment, and can be used to reduce the variance of a policy gradient estimate without adding any bias. We find the baseline that minimizes the variance. We also consider the class of constant baselines, and find the constant baseline that minimizes the variance. We compare this to the common technique of adjusting the rewards by an estimate of the performance measure. Actor-critic methods usually attempt to learn a value function accurate enough to be used in a gradient estimate without adding much bias. In this thesis we propose that in learning the value function we should also consider the variance. We show how considering the variance of the gradient estimate when learning a value function can be beneficial, and we introduce a new optimization criterion for selecting a value function.

In Part II of this thesis we consider online versions of policy gradient algorithms, where we update our policy for selecting actions at each step in time, and study the convergence of the these online algorithms. For such online gradient-based algorithms, convergence results aim to show that the gradient of the performance measure approaches zero. Such a result has been shown for an algorithm which is based on observing trajectories between visits to a special state of the environment. However, the algorithm is not suitable in a partially observable setting, where we are unable to access the full state of the environment, and its variance depends on the time between visits to the special state, which may be large even when only few samples are needed to estimate the gradient. To date, convergence results for algorithms that do not rely on a special state are weaker. We show that, for a certain algorithm that does not rely on a special state, the gradient of the performance measure approaches zero. We show that this continues to hold when using certain baseline algorithms suggested by the results of Part I.

Contents