

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

© Evan Greensmith

Typeset in Palatino by T_EX and L^AT_EX 2_ε.

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 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

Acknowledgements	v
Abstract	vii
1 Introduction	1
1.1 Performance	2
1.2 Model	3
1.2.1 A Note on Performance	7
1.3 Value Function Techniques	8
1.4 Policy Gradient Techniques	11
1.5 Actor-Critic Techniques	15
1.6 Policy Search Techniques	17
2 The Policy Gradient Approach	21
2.1 Gradient Calculation	22
2.2 Gradient Estimation	23
2.3 Approximate Gradient Estimation	30
2.4 Policy Gradient Algorithms	32
2.4.1 On the GPOMDP Estimate	37
2.5 Convergence of Policy Gradient Algorithms	38
2.6 Thesis Contribution	41
I Reducing Estimation Variance	47
3 Variance of Policy Gradient Algorithms	49
3.1 Variance of Sample Averages on Markov Chains	50
3.1.1 Covariance of Markov Samples	54
3.1.2 Proof of Theorem 3.4	62
3.2 Variance of the Baseline Estimate	65
3.2.1 Proof of Theorem 3.12	68
4 Selecting Baselines for Policy Gradient Algorithms	73
4.1 Optimal Baseline	75
4.2 Optimal Constant Baseline	77
4.3 Algorithms for Learning Baselines	80

5	Selecting Value Functions for Actor-Critic Algorithms	85
5.1	The Error of the Δ_T^V Gradient Estimate	87
5.2	On the Bias and the Variance	89
5.2.1	Zero Variance, Zero Bias Example	90
5.2.2	Minimum Variance Example	91
5.3	Algorithms for Learning Value Functions	93
5.3.1	Minimizing Bias Error	101
5.3.2	Minimizing Sample Error	101
6	Experiments	103
6.1	Three State MDP	103
6.2	Online Training	105
6.3	Locating a Target	107
II	Online Optimization	111
7	Convergence of Stochastic Gradient Algorithms	113
7.1	The Stochastic Error	114
7.2	Convergence Result	120
7.3	Proof of Theorem 7.1	123
7.3.1	The Error	127
7.3.2	Full Intervals	138
7.3.3	Almost Surely There	143
8	Convergence of Online Policy Gradient Algorithms	145
8.1	The COLMDP Algorithm	149
8.2	Adding a Baseline	155
8.3	Partially Observable Setting	159
8.4	Smooth Gradient Results	160
8.5	Proof of Convergence of the COLMDP Algorithm	170
8.5.1	Relating the COLMDP Algorithm to Theorem 7.1	170
8.5.2	Constructing Auxiliary Sequences	173
8.5.3	Satisfying the Conditions of Theorem 7.1	176
8.6	Proof of Convergence when Learning the Baseline	186
9	Conclusion	197
A	On Irreducibility and Aperiodicity	199
B	A Law of Large Numbers	203