# Emperical evaluation of policy gradient methods for POMDPs

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Policy gradient algorithms are popular for reinforcement learning in partially observable Markov decision processes (or POMDPs) with finite state controllers. However, policy gradient algorithms are only guaranteed to converge to a locally optimal solution. In this paper, we evaluate three classes of policy gradient algorithms for POMDPs: actor only method with Monte Carlo evaluation, actor-critic method with linear function approximation, and using recurrent policy gradient (RPG) method with policy represented by a recurrent neural network (RNN) with long short term memory (LSTM). These algorithms are evaluated on four simple environments: tiger, voicemail, cheese maze, and 4x4 grid. We find that actor only and recurrent policy gradient based methods converge relatively quickly but almost all sample paths converge to local optima rather than the global optimum.

**Introduction:** Reinforcement learning theory and practice has achieved considerable progress in recent years [1–4]. However, most of this progress is restricted to models where the agent perfectly observes the state of the environment. Learning in environments with partial state observation is not well understood. Partially observable Markov Decision processes (or POMDPs) are perhaps the simplest models with partial state observation. Understanding reinforcement learning for POMDPs is critical to understand learning in more complicated partially observed models such as partially observed semi-Markov models and multi-agent systems.

To understand the difficulty with learning in POMDPs, let's revisit the standard approach to planning in POMDPs [5, 6]. For planning, one constructs the posterior belief of the agent on the state of the system—which is called the belief state—and shows that the belief state is an information state (also called sufficient statistic for control). One then proceeds to write a dynamic program based on the belief state.

The difficulty in extending the planning approach to learning is that, by construction, the belief state depends on the system dynamics and the observation model. So, when an agent is operating in an unknown and partially observed environment, it cannot identify it's current belief state, and hence cannot use value iteration based methods for reinforcement learning. It is for this reason that most of the literature on reinforcement learning for POMDPs focuses on policy search methods, the most popular of which is the policy gradient method [7]. Policy gradient method assumes that the policy is parametrized by a finite number of parameters, estimates the gradient of a parametrized policy from a single sample path using the "log-derivative trick", and then uses stochastic gradient descent [8] to find a locally optimal policy.

Policy gradient methods are particularly attractive for POMDPs because it is known that under mild technical conditions the optimal planning solution for POMDPs is given by a finite state machine [9]. For this reason, most of the literature on policy gradient methods on reinforcement learning for POMDPs assume finite state policies. One of the limitations of policy gradient methods is that they

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

only guarantee convergence to a locally optimal solution. In this paper, we emperically investigate whether different policy gradient algorithms converge to the planning solution.

**Environments:** The following environments are considered— tiger [6], voicemail [10], cheese maze [11], and the 4x4 grid [12]. These environments were chosen because they are low dimensional and is relatively easy to obtain the optimal planning solution.

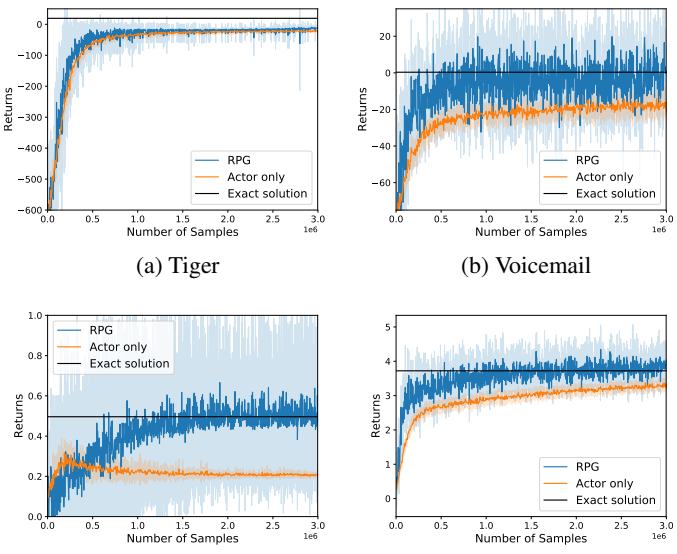
**Policy Parametrization:** The policy assumed to be a finite state controller is given by two functions: the output function $\pi_\theta(\text{action}|\text{obs}, \text{memory})$ and the memory update function $\mu_\phi(\text{next memory}|\text{action}, \text{obs}, \text{memory})$, which are modeled as Gibbs distributions with parameters $\theta$ and $\phi$, respectively.

**RL algorithms:** We consider the following basic RL algorithms for POMDPs. We use actor only methods where the policy evaluation is done using Monte Carlo simulation [13], actor critic methods with finite state controllers using linear function approximator to estimate a critic [14], and recurrent policy gradients (RPG) where the policy is represented using recurrent neural network (RNN) with long short term memory (LSTM) [15].

**Results:** We evaluate all three RL algorithms on the four environments. For all four environments, the optimal planning policy is a finite state controller. For the actor only and actor critic methods, we consider finite state controllers with size of the memory same as that of the optimal planning policy. For RPG, the policy is represented by a two layer network with 10 layers each and the value is represented by a single layer network of 20 neurons. We repeat each experiment 15 times and plot the mean and standard deviation across the runs in the plot below. Both actor only method with Monte Carlo evaluation and Recurrent Policy Gradient methods converge relatively quickly. In our experiments, the actor critic method with linear function approximation did not converge for any environment; for clarity, we haven't included it in the plots below.

The actor only method converges relatively quickly for all environments. For the cheese maze environment, it converges to a bad local mimina along all sample paths; for the other environments, it converges to a policy whose performance is close to the optimal performance. For all environments, there is very little variation across different runs.

The RPG method converges relatively quickly for all environments and in all cases it converges to a policy that performs close to the optimal policy. However, there is a relatively large variance across different runs. Note that we are plotting the networks own evaluation of its performance and that is the reason that the along some runs the estimated performance is higher than the optimal performance.




(a) Tiger (b) Voicemail

(c) Cheese maze (d) 4x4 grid

Figure 1: Comparison of policy gradient based RL algorithms for different POMDP environments

As mentioned earlier, in our experiments, the actor critic method failed to converge. A possible explanation of this could be that the linear features computed for the joint MDP represented by the controller states, observations and the actions were not sufficient enough to calculate the conditional mean of the true value function of the POMDP.

# References


[1] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

[2] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[3] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[4] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In *International Conference on Machine Learning*, pages 1889–1897, 2015.

[5] Karl J Astrom. Optimal control of markov decision processes with incomplete state estimation. *Journal of mathematical analysis and applications*, 10:174–205, 1965.

[6] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[7] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

[8] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

[9] Huizhen Yu and Dimitri P Bertsekas. On near optimality of the set of finite-state controllers for average cost POMDP. *Mathematics of Operations Research*, 33(1):1–11, 2008.

[10] Jason D Williams and Steve Young. Partially observable markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422, 2007.

[11] R Andrew McCallum. Overcoming incomplete perception with utile distinction memory. In *Proceedings of the Tenth International Conference on Machine Learning*, pages 190–196, 1993.

[12] Anthony R Cassandra, Leslie Pack Kaelbling, and Michael L Littman. Acting optimally in partially observable stochastic domains. In *AAAI*, volume 94, pages 1023–1028, 1994.

[13] Jonathan Baxter and Peter L Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

[14] Huizhen Yu. A function approximation approach to estimation of policy gradient for POMDP with structured policies. In *UAI '05, Proceedings of the 21st Conference in Uncertainty in Artificial Intelligence, Edinburgh, Scotland, July 26-29, 2005*, pages 642–657, 2005.

[15] Daan Wierstra, Alexander Förster, Jan Peters, and Jürgen Schmidhuber. Recurrent policy gradients. *Logic Journal of the IGPL*, 18(5):620–634, 2010.