
A Connection Between Importance Sampling and Likelihood Ratio Policy Gradients

Jie Tang, Pieter Abbeel *

1 Introduction

Likelihood ratio policy gradient methods have been some of the most successful reinforcement learning algorithms, especially for learning on physical systems. There exists a rich literature on different policy gradient techniques, from simple gradient estimators like REINFORCE [3], Baxter and Bartlett’s GPOMDP [2], natural gradient techniques [1] which leverage the curvature of the space, and more sophisticated value function approximators like Peters’ episodic Natural Actor Critic [5]. Policy gradient methods have been widely applied to a variety of complex real-world reinforcement learning problems, e.g. hitting a baseball with an articulated arm robot [5], learning gaits for a legged robot quickly [6]. In these settings the most time consuming factor in the learning process is the number of real-world trials.¹²

We describe a novel connection between likelihood ratio based policy gradient methods and importance sampling. The likelihood ratio policy gradient estimate is equivalent to taking the derivative of a particular importance sampled estimate of the value function. This particular importance sampled estimate of the value function only leverages data from the current policy in the search. This indicates that likelihood ratio policy gradients are quite naive in terms of data use.

Likelihood ratio policy gradient methods perform a (stochastic) gradient ascent over the policy parameter space Θ to find a local optimum of $U(\theta)$. It is well known (see, e.g., [3]) that the gradient $\nabla_{\theta}U(\theta)$ can be re-expressed as follows:

$$\nabla_{\theta}U(\theta) = \sum_{\tau} P(\tau; \theta) \nabla_{\theta} \log P(\tau; \theta) R(\tau)$$

This gives us the following expression for a Monte Carlo estimate of the policy gradient from m sample paths under policy π_{θ} :

$$\nabla_{\theta} \log P(\tau^{(i)}; \theta) = \sum_{t=0}^H \underbrace{\nabla_{\theta} \log \pi_{\theta}(u_t^{(i)} | s_t^{(i)})}_{\text{no dynamics model required!!}}$$

Hence even without access to a dynamics model, the likelihood policy gradient method is able to provide an unbiased estimate of the policy gradient.

Importance sampling can be readily leveraged to evaluate stochastic policies [4] as follows:

$$U(\theta_1) \approx \frac{1}{m} \sum_{i=1}^m \frac{P(\tau^{(i)}; \theta_1)}{P(\tau^{(i)}; \theta_2)} R(\tau^{(i)}), \quad \tau^{(i)} \sim P(\tau; \theta_2)$$
$$\frac{P(\tau^{(i)}; \theta_1)}{P(\tau^{(i)}; \theta_2)} = \frac{\prod_{t=0}^{H-1} \pi_{\theta_1}(u_t | s_t)}{\prod_{t=0}^{H-1} \pi_{\theta_2}(u_t | s_t)}$$

*The authors are with the Department of Electrical Engineering and Computer Sciences, UC Berkeley, CA 94720, U.S.A. Email: jietang@eecs.berkeley.edu, pabbeel@cs.berkeley.edu.

¹Topic: learning algorithms

²Preference: Poster

Hence we can estimate the utility of a policy π_{θ_1} from sample paths obtained with a policy π_{θ_2} .

Let e_i denote a vector with all entries equal to zero, except for the i 'th entry being equal to one. Then, by using the importance sampling based estimate for U , we obtain the following expression for the gradient of the utility function evaluated at a point θ :

$$\begin{aligned}
 \frac{\partial U}{\partial \theta_i}(\theta) &= \lim_{\epsilon \rightarrow 0} \frac{U(\theta + \epsilon e_i) - U(\theta - \epsilon e_i)}{2\epsilon} \\
 &= \lim_{\epsilon \rightarrow 0} \frac{1}{m} \sum_{i=1}^m \frac{R(\tau^{(i)})}{P(\tau^{(i)}; \theta)} \frac{P(\tau^{(i)}; \theta + \epsilon e_i) - P(\tau^{(i)}; \theta - \epsilon e_i)}{2\epsilon} \\
 &= \frac{1}{m} \sum_{i=1}^m \frac{R(\tau^{(i)})}{P(\tau^{(i)}; \theta)} \frac{\partial P(\tau^{(i)}; \theta)}{\partial \theta_i} \\
 &= \frac{1}{m} \sum_{i=1}^m \frac{\partial \log P(\tau^{(i)}; \theta)}{\partial \theta_i} R(\tau^{(i)}) \\
 &= \text{likelihood ratio based gradient}
 \end{aligned}$$

This shows that an importance sampled estimate drawn at a single instance of θ corresponds to the standard likelihood ratio based policy gradient. However, a more sophisticated approach uses all available data points (samples drawn from many different θ 's) to build a lower variance importance sampled estimate of the full utility function $U(\theta)$ ³

Specifically, using our estimate of the complete value function, we can find a local optimum policy for the estimated value function and use it to sample from our actual system. However, a naive gradient ascent type algorithm fails to account for the variance of importance sampled estimates. If our policy parameters become substantially different from parameters we have explored in the past, the data we have gathered becomes less useful for accurately estimating the value function. Our learning methods address this by (i) applying hard thresholds to the variance we are willing to tolerate during a gradient ascent step and (ii) adding a penalty term for areas of policy parameter space with high variance.

By leveraging the observed connection between importance sampling and likelihood ratio policy gradient estimates we obtain algorithms which outperform policy gradient methods on toy problems when a small number of samples are available. Efficient use of samples is desirable for real world applications, where gathering samples is often expensive and/or time consuming. The benefits stem from several factors: (i) Because our methods provide a full approximation of the true value function, we can apply line search and gradient ascent methods which do not require tuning a stepsize parameter for every new problem domain. (ii) The importance sampling based approaches leverage past experience in a theoretically sound way; policy gradient methods generally forget data from past policies entirely or forget data in ad-hoc ways. Our algorithms show promise in standard example settings often used in the evaluation of policy gradient methods (LQR and cartpole).

- [1] Amari, S. Natural gradient works efficiently in learning. *Neural Computation*, 10, 1998.
- [2] Baxter, J. and Bartlett, P. Direct gradient-based reinforcement learning, 1999. URL citeseer.nj.nec.com/baxter99direct.html.
- [3] Glynn, P. "Likelihood ratio gradient estimation: An Overview." In *Proceedings of the 1987 Winter Simulation Conference*, Atlanta, GA, 1987.
- [4] Leonid Peshkin, Christian R. Shelton. Learning from scarce experience. In *Proceedings of the Nineteenth International Conference on Machine Learning*, 2002.
- [5] Peters, J. Vijayakumar, S., and Schaal, S. Natural actor-critic. In *Proceedings of the European Machine Learning Conference (ECML)*, 2005.
- [6] R. Tedrake, T.W. Zhang, H.S. Seung. Learning to walk in 20 minutes. In *Proceedings of the Fourteenth Yale Workshop on Adaptive and Learning Systems*, 2005.

³This analysis provides justification for existing importance-sampling based analogues to popular policy gradient techniques (REINFORCE, GPOMDP, natural gradient, with and without optimal minimum variance baselines).