On a Connection between Importance Sampling and the Likelihood Ratio Policy Gradient

Jie Tang
jietang@eecs.berkeley.edu

Pieter Abbeel
pabbeel@cs.berkeley.edu

EECS Department, University of California, Berkeley

Introduction

- Likelihood ratio policy gradient methods are state of the art techniques for reinforcement learning in continuous state spaces.
  - Learning to hit balls with a bat [7]
  - Learning legged robot gaits [9]
- Model-free learning with strong convergence guarantees
- In this work:
  - Show how policy gradient methods can be derived from an importance sampling perspective
  - Show more general form of optimal baselines.
  - Present a new policy search method which leverages these insights to outperform standard likelihood ratio PG methods.

Background

Preliminaries:

- States $s_t \in S = \mathbf{R}^n$, actions $a_t \in A = \mathbf{R}^m$.
- Reward function $R(s_t, a_t) \in \mathbf{R}$.
- Consider a class of stochastic policies parameterized by θ.
- Let $\pi_\theta : S \times A \to [0,1]$ denote a policy in this class.
- Sample state-action sequences $(s_0^i, a_0^i, s_1^i, a_1^i, ..., s_H^i, a_H^i)$ using policy π_θ.

Policy Gradient Methods:

- Directly optimize expected reward U as a function of policy parameters θ.

$$U(\theta) = \mathrm{E}\left[\sum_{t=1}^{H} R(s_t, a_t) \;\middle|\; \pi_\theta\right]$$

- Can compute gradient of U from sample state-action sequences:

$$\nabla_\theta U(\theta) \approx \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{H} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \sum_{t=1}^{H} R(s_t, a_t)$$

- Can add zero-mean baseline term to reduce variance. [6]

$$\nabla_\theta U(\theta) \approx \frac{1}{m}\sum_{i=1}^{m}\sum_{t=1}^{H} \nabla_\theta \log \pi_\theta(a_t \mid s_t) \sum_{t=1}^{H} (R(s_t, a_t) - b_t)$$

Importance Sampling:

- Importance sampling reweights samples gathered under old policy parameters to form an (unbiased) estimate of the expected return of novel, arbitrary policy parameters.
- Given sample state-action sequences $\{(s_t, a_t)\}_i \sim \theta_2$, estimate expected return function $\widehat{U}^{IS}(\theta_1)$:

$$\widehat{U}^{IS}(\theta_1) \approx \frac{1}{m}\sum_{i=1}^{m}\prod_{t=1}^{H} \frac{\pi_{\theta_1}(a_t \mid s_t)}{\pi_{\theta_2}(a_t \mid s_t)} \sum_{t=1}^{H} R(s_t, a_t)$$

Main Result: IS and Policy Gradients

Proposition:

The sample estimate of the gradient of $\widehat{U}^{IS}(\theta)$ evaluated using only sample trajectories drawn under π_θ is equal to the likelihood ratio based sample estimate of the gradient of $U(\theta)$.

- Suggests LRPG does not make full use of data.

Main Result: Generalized Baselines

Proposition:

For any distribution $P_\theta(X)$, any scalar valued function $f(X)$, and any fixed vector b:

$$E_{P_\theta(X)}[f(X)] = E_{P_\theta(X)}\left[f(X) - b^T \nabla_\theta \log P_\theta(X)\right]$$

- Set b to minimize variance of the estimator
- Directly applies to estimating $\widehat{U}^{IS}$

Algorithm Summary

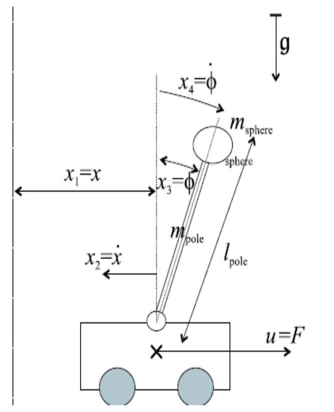
Input: domain of policy parameters Θ, initial policy $\pi_{\widehat{\theta}_0^{MEM}}$

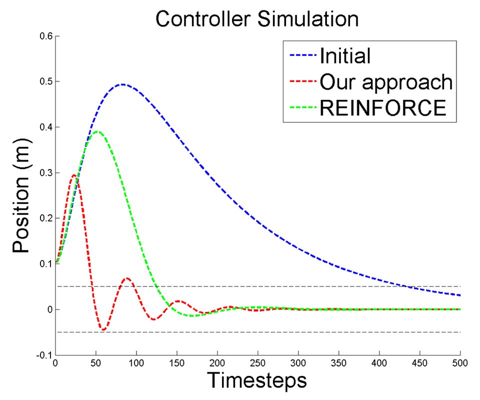
for $i = 0$ to ... **do**
  1. Run M trials under policy $\pi_{\widehat{\theta}_i^{MEM}}$
  2. Search within ESS region
  for $j = 1 : i$ **do**
    $\theta_j \leftarrow \widehat{\theta}_j^{MEM}$
    while $\widehat{U}(\theta_j)$ is improving **do**
      $g_j \leftarrow$ step direction$(\widehat{U}(\theta_j))$
      $\alpha_j \leftarrow$ ESS line search$(\widehat{U}(\theta_j), g_j)$
      $\theta_j \leftarrow \theta_j + \alpha_j g_j$
    end while
  end for
  3. Update policy: $\widehat{\theta}_{i+1}^{MEM} = \arg\max_{\theta_j} \widehat{U}(\theta_j)$
end for

- Our approach: find local optima of $\widehat{U}^{IS}$ through **memory-based optimization**
- Use general **minimum variance baseline** for estimating $\widehat{U}^{IS}$.
  - Minimum variance b is also an expectation: in principle, can reapply baseline trick recursively.
  - Introducing baselines increases model complexity. Requires more samples (or IS).
- Use **effective sample size (ESS)** to limit search areas of θ space with many samples [4]
- Do **optimal line search** (Armijo rule) [2]

Cartpole

- $s = (x_1, x_2, x_3, x_4) \in \mathbf{R}^4$, $a = u \in \mathbf{R}$.
- $\theta = (K, \eta) \in \mathbf{R}^5$, $\eta \in \mathbf{R}$
- Policy $\pi_\theta(a|s) = N\left(Ks, 0.1 + \frac{1}{1+e^\eta}\right)$
- Reward is 0 inside target region, -2 on failure, -1 o.w.
- Right: cartpole system under initial policy, policy learned with our approach, and policy learned with REINFORCE. Black lines show the target region where cost is 0.

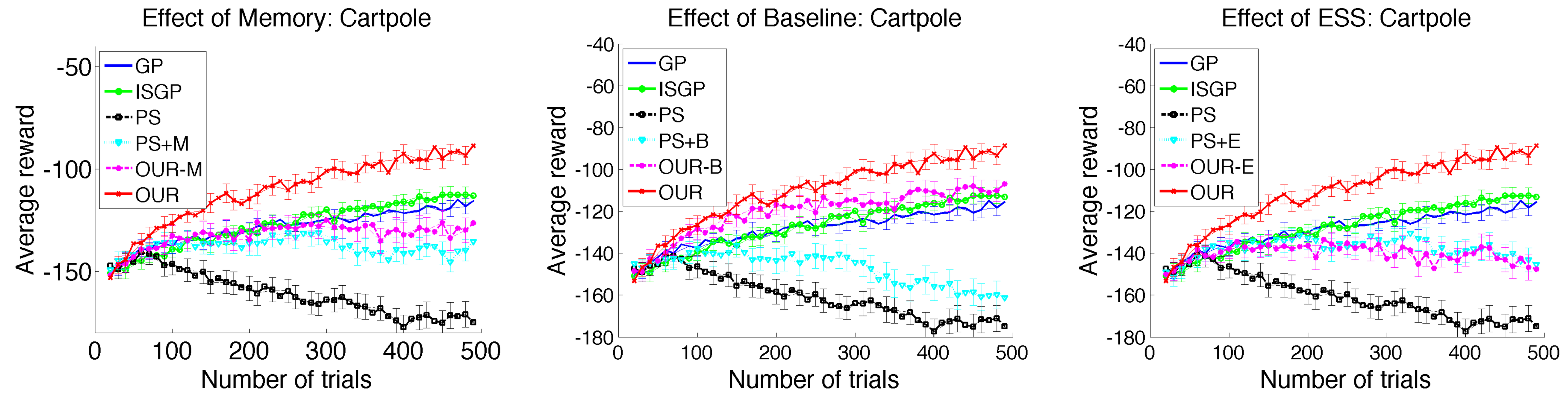




References

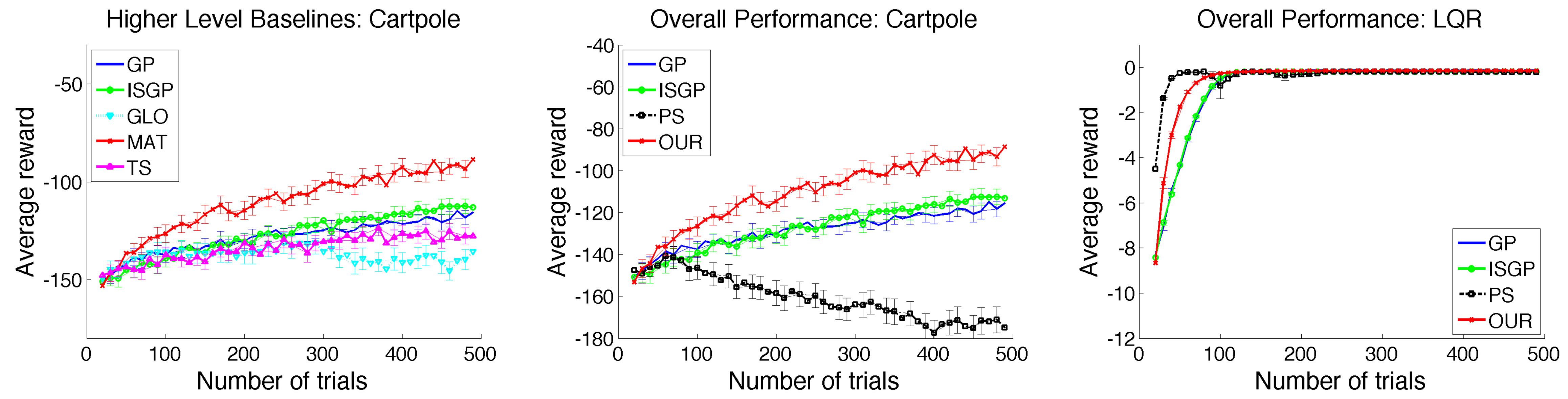
[1] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10, 1998.
[2] D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2004.
[3] S. Kakade. A natural policy gradient. In *Advances in Neural Information Processing Systems*, volume 14, 2001.
[4] J. S. Liu. *Monte Carlo Strategies in Scientific Computing*. Springer Publishing Company, Incorporated, 2008.
[5] L. Peshkin and C. R. Shelton. Learning from scarce experience. In *Proceedings of the Nineteenth International Conference on Machine Learning*, 2002.
[6] J. Peters and S. Schaal. Policy gradient methods for robotics. In *Proceedings of the IEEE International Conference on Intelligent Robotics Systems*, 2006.
[7] J. Peters, S. Vijayakumar, and S. Schaal. Natural actor-critic. In *Proceedings of the European Machine Learning Conference (ECML)*, 2005.
[8] M. Riedmiller, J. Peters, and S. Schaal. Evaluation of policy gradient methods and variants on the cart-pole benchmark. In *IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, 2007.
[9] R. Tedrake, T. W. Zhang, and H. Seung. Learning to walk in 20 minutes. In *Proceedings of the Fourteenth Yale Workshop on Adaptive and Learning Systems*, 2005.

Experiments



(a) (b) (c)

This figure demonstrates the effect of (a) memory based search (b) optimal baselines, and (c) ESS search region on cartpole performance. In each figure, we show the performance of Peshkin and Shelton's approach (PS) and our approach (OUR). In addition, we show the performance with memory only (PS+M), baselines only (PS+B), and ESS only (PS+E), and our approach with memory (OUR-M), baselines (OUR-B), and ESS (OUR-E) removed. GPOMDP (GP) and IS GPOMDP (ISGP) are also plotted for reference purposes.



(a) (b) (c)

(a) Performance of various choices for the higher level baselines in our approach. We have a matrix baseline (MAT), and a recursive baseline (REC). For reference, we also plot our approach without an optimal baseline (GLO), GPOMDP (GP), and IS GPOMDP (ISGP). (b), (c) Performance evaluation on Cartpole and LQR (another common benchmark). The algorithms considered are the GPOMDP likelihood ratio policy gradient method (GP), GPOMDP with importance sampling (ISGP), Peshkin and Shelton's algorithm (PS), and our approach (OUR).

Future Work

- Generalizing GPOMDP-like temporal decomposition to the IS setting by starting from the following expression for $U(\theta)$.

$$U(\theta) = \sum_{t=1}^{H} \mathrm{E}_{(s_0, a_0, ..., s_t, a_t) \sim \theta}\left[R(s_t, a_t) \mid \pi_\theta\right]$$

- Optimistic bandit-style exploration.
- Investigating dependence between samples.
- Validation on more complex domains.

Conclusions

- Policy gradient methods are a special case of gradient descent over $\widehat{U}^{IS}$.
- Baselines used in policy gradients are a special case of generalized baselines.
  - Minimum variance unbiased estimators can be computed for estimating $\widehat{U}^{IS}$.
  - Optimal baselines are themselves expectations, which can be given generalized baselines recursively.
- Our experiments show fewer trials are needed to learn good controllers.