Bayesian Trust Region Policy Optimization

Akella Ravi Tej[*] **Kamyar Azizzadenesheli**[†] **Mohammad Ghavamzadeh**[‡]

Animashree Anandkumar[†]

Abstract

Trust region policy optimization (TRPO) methods have gained a distinct appeal among other policy gradient (PG) algorithms because of their data efficiency and guarantees for monotonic policy improvements. However, these approaches do not address the issue of high variance in gradient estimates or utilize the uncertainty in their predictions for computing the policy updates. In this work, we propose Bayesian Trust Region Policy Optimization (BTRPO), an actor-critic algorithm that employs a Bayesian class of parametric critics modeled using Gaussian processes (GP). The posterior distribution over the action-value function returned by the GP-based critic is then used to obtain a low variance estimate of the policy gradient and a refined estimate of the trust region. Our approach highlights a shortcoming in the Monte-Carlo estimation of policy gradient and has interpretable connections to Natural Policy Gradient (NPG) and TRPO. We demonstrate a superior sample complexity with the proposed approach relative to other online policy gradient algorithms over versatile robotic locomotion tasks.

1 Background

Monte-Carlo (MC) methods for policy gradient estimation (Williams, 1992) violate the *likelihood principle* (Berger & Wolpert, 1988) and suffer from a high variance in gradient estimates. However, the simplicity and computational efficiency of MC approaches makes for its ubiquitous application in modern policy gradient algorithms. On the other hand, Bayesian quadrature (BQ) (O'Hagan, 1991) provides a Bayesian alternative for policy gradient estimation using samples drawn from a policy distribution. Unlike MC methods which return a point estimate, BQ returns the complete posterior distribution over the policy gradient. Moreover, BQ approaches comply with the *likelihood principle* and exhibit a lower variance in gradient estimates compared to MC (Ghavamzadeh & Engel, 2007). However, the exact computation of BQ approaches has a prohibitive $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ storage requirement.

2 Main Contributions

1. We present an approximate Bayesian actor-critic algorithm for optimizing large non-linear policies in challenging continuous domain environments.
2. In addition, we formulate a novel constrained optimization problem that uses the uncertainty in the policy gradient posterior to compute robust updates with monotonic policy improvements.
3. The resulting algorithm which we call Bayesian Trust Region Policy Optimization has two variants, *viz.* BTRPO and A-BTRPO, that share close connections to prior Monte-Carlo

[*]Indian Institute of Technology Roorkee, ravitej.akella@gmail.com

[†]Caltech, {kazizzad,anima}@caltech.edu

[‡]Facebook AI Research (FAIR), mgh@fb.com

A short preview of our work. Full manuscript is under preparation.

policy gradient (MCPG) algorithms, namely NPG (Kakade, 2001) and TRPO (Schulman et al., 2015).

4. We also show that the kernel families of the Bayesian equivalents of MCPG have a non-existent posterior distribution over the state-value function. This analysis provides fresh insights about the high variance problem in MCPG and the kernel families that yield a better sample efficiency.
5. The practical implementations of BTRPO and A-BTRPO algorithm clearly outperforms other online policy gradient methods while remaining comparable to MCPG in computational complexity.

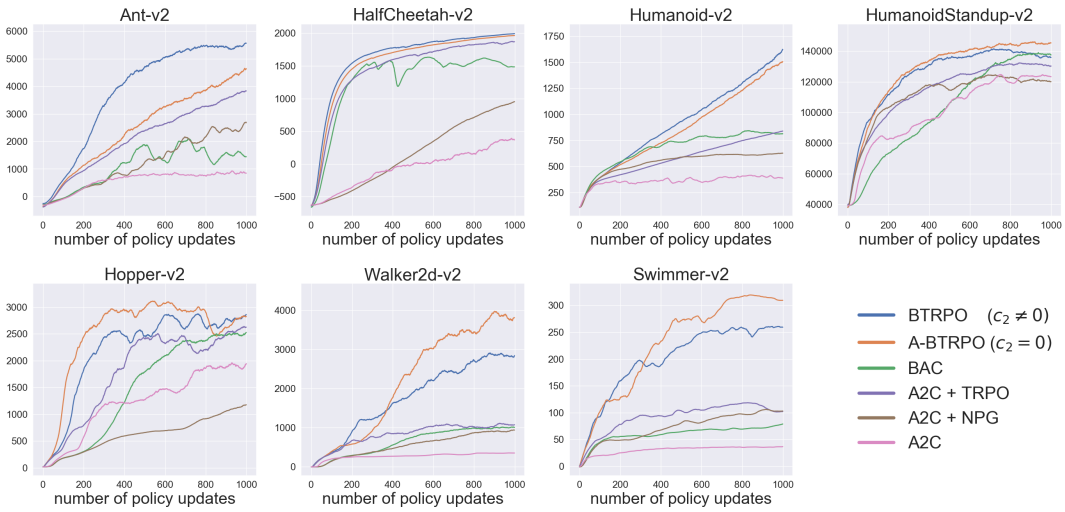

Figure 1: Comparison of the proposed approach with other online policy gradient algorithms on several MuJoCo environments. The performance of an algorithm ($y - axis$) is measured using the average episodic reward over the last 100 episodes across 3 random seeds.

References

James O. Berger and Robert L. Wolpert. *Chapter 3: The Likelihood Principle and Generalizations*, volume Volume 6 of *Lecture Notes–Monograph Series*, pp. 19–64. Institute of Mathematical Statistics, Hayward, CA, 1988. doi: 10.1214/lnms/1215466214. URL `https://doi.org/10.1214/lnms/1215466214`.

Mohammad Ghavamzadeh and Yaakov Engel. Bayesian actor-critic algorithms. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pp. 297–304, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273534. URL `http://doi.acm.org/10.1145/1273496.1273534`.

Sham Kakade. A natural policy gradient. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*, NIPS'01, pp. 1531–1538, Cambridge, MA, USA, 2001. MIT Press. URL `http://dl.acm.org/citation.cfm?id=2980539.2980738`.

A. O'Hagan. Bayes-hermite quadrature. *Journal of Statistical Planning and Inference*, 29 (3):245–260, November 1991. URL `http://www.sciencedirect.com/science/article/B6V0M-45F5GDM-53/1/6e05220bfd4a6174e890f60bb391107c`.

John Schulman, Sergey Levine, Philipp Moritz, Michael I. Jordan, and Pieter Abbeel. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3-4):229–256, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992696. URL `https://doi.org/10.1007/BF00992696`.