Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

ANALYSIS OF MEASURE-VALUED DERIVATIVES IN A REINFORCEMENT LEARNING ACTOR-CRITIC FRAMEWORK

Kim van den Houten

Delft University of Technology
Van Mourik Broekmanweg 6
Delft, 2628XE, THE NETHERLANDS

Emile van Krieken
Bernd Heidergott

Vrije Universiteit Amsterdam
De Boelelaan 1105
Amsterdam, 1081HV, THE NETHERLANDS

ABSTRACT

Policy gradient methods are successful for a wide range of reinforcement learning tasks. Traditionally, such methods utilize the score function as stochastic gradient estimator. We investigate the effect of replacing the score function with a measure-valued derivative within an on-policy actor-critic algorithm. The hypothesis is that measure-valued derivatives reduce the need for score function variance reduction techniques that are common in policy gradient algorithms. We adapt the actor-critic to measure-valued derivatives and develop a novel algorithm. This method keeps the computational complexity of the measure-valued derivative within bounds by using a parameterized state-value function approximation. We show empirically that measure-valued derivatives have comparable performance to score functions on the environments Pendulum and MountainCar. The empirical results of this study suggest that measure-valued derivatives can serve as low-variance alternative to score functions in on-policy actor-critic and indeed reduce the need for variance reduction techniques.

1 INTRODUCTION

Reinforcement learning (RL) methods are increasingly successful in areas such as robotics (Carvalho et al. 2021), self-driving cars (Kiran et al. 2022), and energy systems (Perera and Kamalaruban 2021). After the break-trough of the REINFORCE algorithm (Williams 1992), and the policy gradient theorem (Sutton et al. 1999), a variety of successful policy gradient methods have been developed, such as A2C (Mnih et al. 2016), PPO (Schulman et al. 2017), and DDPG (Lillicrap et al. 2016). Policy gradient methods are techniques from RL that optimize a policy with respect to the expected cumulative reward by applying gradient ascent. Most policy gradient algorithms utilize the score function (SF) as gradient estimator for the stochastic objective function. Unfortunately, the policy gradient algorithms developed in the past decades suffer from the excessive variance of the SF gradient estimates (Sutton and Barto 2018). Using SF, several variance reduction techniques are necessary for convergence to the optimal policy, such as implementation of a baseline (Greensmith et al. 2002; Li 2018; Mohamed et al. 2020).

Stochastic optimization seeks gradient estimators with both low variance, and bias. This motivates the study of the measure-valued derivative (MVD), known for having significantly lower variance than SFs. The SF is not widely used in the RL community. We hypothesise that the low variance of the MVD reduces the need for variance reduction techniques. There have been few controlled studies that compare differences in performance between SF, and MVD for RL. However, the promising results from Bhatt et al. (2019), and Carvalho et al. (2021) motivate further research in using MVD for RL purposes.

In this paper, we present a novel algorithm (AC-MVD), in which the MVD is implemented in an on-policy actor-critic algorithm with parameterized state-value functions. We compare the results of this algorithm with the SF variant (AC-SF). The compared algorithms are equivalent in structure, besides the

978-1-6654-7661-4/22/$31.00 ©2022 IEEE

Haarnoja, T., A. Zhou, P. Abbeel, and S. Levine. 2018. "Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement". In *Proceedings of the 35th International Conference on Machine Learning*. July 10th-15th, Stockholm, Sweden, 1861-1870.

Heidergott, B., and H. Leahu. 2010. "Weak Differentiability of Product Measures". *Mathematics of Operations Research* 35(1):27–51.

Heidergott, B., F. J. Vázquez-Abad, and W. Volk-Makarewicz. 2008. "Sensitivity Estimation for Gaussian Systems". *European Journal of Operational Research* 187(1):193–207.

Heidergott, B., and W. Volk-Makarewicz. 2016. "A Measure-Valued Differentiation Approach to Sensitivities of Quantiles". *Mathematics of Operations Research* 41(1):293–317.

Kiran, B., I. Sobh, V. Talpaert, P. Mannion, A. Sallab, S. Yogamani, and P. Perez. 2022. "Deep Reinforcement Learning for Autonomous Driving: A Survey". *IEEE Transactions on Intelligent Transportation Systems* 23(6):4909–4926.

Konidaris, G., and S. Osentoski. 2011. "Value Function Approximation in Reinforcement Learning Using the Fourier Basis.". In *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, edited by D. Leake, R. Morris, M. Wellman, and S. Ludvik, 380–385. Menlo Park, California, United States: the Association for the Advancement of Artificial Intelligence Press.

Li, Y. 2018. "Deep Reinforcement Learning". *arXiv e-prints*. http://arxiv.org/abs/1810.06339.

Lillicrap, T. P., J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. 2016. "Continuous Control with Deep Reinforcement learning". In *Proceedings of the 4th International Conference on Learning Representations*. May 2nd - 4th, San Juan, Puerto Rico.

Mnih, V., A. P. Badia, M. Mirza, A. Graves, T. Harley, T. P. Lillicrap, D. Silver, and K. Kavukcuoglu. 2016. "Asynchronous Methods for Deep Reinforcement Learning". In *Proceedings of the 33rd International Conference on Machine Learning*. June 19th-24th New York City, New York, United States, 1928–1937.

Mohamed, S., M. Rosca, M. Figurnov, and A. Mnih. 2020. "Monte Carlo Gradient Estimation in Machine Learning". *Journal of Machine Learning Research* 21(1):5183–5244.

Pan, H. 2020. *AC-SF*. https://github.com/workofart/openai-gym-baselines/blob/master/Pendulum-v0/actor_critic_baseline.py.

Perera, A., and P. Kamalaruban. 2021. "Applications of Reinforcement Learning in Energy Systems". *Renewable and Sustainable Energy Reviews* 137:110618.

Pflug, G. 1989. "Sampling Derivatives of Probabilities". *Computing* 42:315–328.

Schulman, J., S. Levine, P. Abbeel, M. Jordan, and P. Moritz. 2015. "Trust Region Policy Optimization". In *Proceedings of the 32nd International Conference on Machine Learning*. July 7th-9th, Lille, France, 1889-1897.

Schulman, J., F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. 2017. "Proximal Policy Optimization Algorithms". *arXiv e-prints*. http://arxiv.org/abs/1707.06347.

Sutton, R., and A. Barto. 2018. *Reinforcement Learning: An Introduction*. 2nd ed. Cambridge, Massachusetts, United States: The MIT Press.

Sutton, R. S., D. McAllester, S. Singh, and Y. Mansour. 1999. "Policy Gradient Methods for Reinforcement Learning with Function Approximation". In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, edited by S. Solla, T. Leen, and K. Müller, 1057—1063. Cambridge, Massachusetts, United States: MIT Press.

van den Houten, K. 2022. *AC-MVD*. https://github.com/kimvandenhouten/AC-MVD.

Williams, R. J. 1992. "Simple Statistical Gradient-following Algorithms for Connectionist Reinforcement Learning". *Machine Learning* 8:229–256.

AUTHOR BIOGRAPHIES

KIM VAN DEN HOUTEN is a PhD candidate in Algorithmics at the Technische Universiteit Delft, the Netherlands. She received her master's degree in Operations Research from the Vrije Universiteit, Amsterdam. Her research interests are on the intersection of optimization and machine learning, and simulation. Her email adress is k.c.vandenhouten@tudelft.nl.

EMILE VAN KRIEKEN is a PhD candidate in Artificial Intelligence at the Vrije Universiteit Amsterdam, the Netherlands. He holds a master's degree in Artificial Intelligence from the University of Amsterdam. His research interests include neuro-symbolic AI, probabilistic deep learning and optimization. His email adress is e.van.krieken@vu.nl.

BERND HEIDERGOTT is the professor of Stochastic Optimization at the Department of Operations Analytics at the Vrije Universiteit Amsterdam, the Netherlands. He received his PhD degree from the University of Hamburg, Germany, in 1996, and held postdoc positions at various universities before joining the Vrije Universiteit. Bernd is research fellow of the Tinbergen Institute and board member of the Amsterdam Business Research Institute. His research interests are optimization and control of discrete event systems, perturbation analysis, Markov chains, max-plus algebra, and social networks. His email adress is b.f.heidergott@vu.nl