Benchmarking context- and gradient-based Meta-Reinforcement Learning

Patrik Okanovic
pokanovic@ethz.ch

Rafael Sterzinger
rsterzinger@ethz.ch

Fatjon Zogaj
fzogaj@ethz.ch

1 Problem Description

In order to tackle the problem of needing large amounts of data to learn a particular task, we aim at utilizing meta-learning to find a common representation of a group of similar tasks. We analyze, benchmark and compare a variety of meta-reinforcement learning algorithms on multiple tasks, initially trained on some underlying, common task distribution.

2 Challenges

Potential challenges mainly include the overall computational requirements for an exhaustive evaluation as well as gathering training samples for the offline meta-reinforcement learning setting.

As the tasks/environments will also highly differ in their dimensionality, it could be that some algorithms might need more fine-tuning than others or are less robust in certain settings. This further adds to the possible computational burden and the complexity of comparing the algorithms.

Finally, as mentioned in [1], we assume that offline algorithms will perform worse than online variants due to the lack of exploration during testing. For this, we might have to consider different metrics to compare these two types of algorithms fairly such as performance in relation to sample efficiency.

3 Methods/Algorithms/Implementations

We will implement the following list of algorithms for meta-reinforcement learning:

- R^2 [2]
- MAML [3]
- TMAML [4]
- MACAW [1]

(Optional, if time permits)

- PEARL [5]
- VariBAD [6]
- BOReL [7]

The R^2 algorithm is encoded in the weights of an RNN, which are learned through a general-purpose "slow" RL algorithm. The RNN receives all information a typical RL algorithm would and retains its state across episodes. The activations of the RNN store the state of the “fast” RL algorithm on the current previously unseen MDP [2]. A prior work of Wang et al. describes a similar context-based algorithm which we will also take a look into.

MAML is a model-agnostic meta-learning algorithm compatible with any model trained using gradient descent. It is applicable to a variety of different learning problems, including classification,

regression, and reinforcement learning [3]. Building upon that, TMAML proposes a surrogate objective function that adds control variates into the gradient estimation via automatic differentiation [4].

Lastly, to also include an offline-variant, we take a look at MACAW which uses Meta-Actor Critic with Advantage Weighting. It imitates the general approach of pre-training on an existing dataset and then fine-tuning with little data to a new task [1].

4 Evaluation

To evaluate the suggested algorithms exhaustively, we will make use of a variety of existing benchmark environments. For this, we propose testing the implemented algorithms on simple environments such as Multi-Armed Bandits/Tabular MDPs [2] as well as more sophisticated tasks from the OpenAI Gym environment which includes HalfCheetah and Ant [3]. Within these, we will mainly compare the gathered cumulative reward and the number of iterations.

References

[1] Eric Mitchell, Rafael Rafailov, Xue Bin Peng, Sergey Levine, and Chelsea Finn. Offline Meta-Reinforcement Learning with Advantage Weighting. page 12, 2021.

[2] Yan Duan, John Schulman, Xi Chen, Peter L. Bartlett, Ilya Sutskever, and Pieter Abbeel. RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning. November 2016. URL http://arxiv.org/abs/1611.02779. arXiv: 1611.02779.

[3] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1126–1135. PMLR, July 2017. URL https://proceedings.mlr.press/v70/finn17a.html. ISSN: 2640-3498.

[4] Hao Liu, Richard Socher, and Caiming Xiong. Taming MAML: Efficient unbiased meta-reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4061–4071. PMLR, May 2019. URL https://proceedings.mlr.press/v97/liu19g.html. ISSN: 2640-3498.

[5] Kate Rakelly, Aurick Zhou, Deirdre Quillen, Chelsea Finn, and Sergey Levine. Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables. page 10, 2019.

[6] Luisa Zintgraf, Kyriacos Shiarlis, Maximilian Igl, Sebastian Schulze, Yarin Gal, Katja Hofmann, and Shimon Whiteson. VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning. February 2020. URL http://arxiv.org/abs/1910.08348. arXiv: 1910.08348.

[7] Ron Dorfman, Idan Shenfeld, and Aviv Tamar. Offline Meta Learning of Exploration. February 2021. URL http://arxiv.org/abs/2008.02598. arXiv: 2008.02598.

[8] Jane X. Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. January 2017. URL http://arxiv.org/abs/1611.05763. arXiv: 1611.05763.