Week 12 Paper Review

Niharika Shrivastava
School of Computing
National University of Singapore
Singapore, 119077
niharika@comp.nus.edu.sg

Abstract

This is a brief review of [1], [2], [3].

1 Guided Policy Search

In this paper, the authors introduce a Guided Policy Search algorithm that directs policy learning by incorporating guided samples in the trajectories to avoid local optima, scale to high-dimensional systems, and generalize to complex tasks.

These guided samples are generated using Differential Dynamic Programming (DDP) that assists policy search by exploring high-reward regions. Furthermore, they are combined with samples generated using an importance sampling variant of the maximum likelihood ratio estimator which alleviates the need for off-policy samples or a well-chosen learning rate. Moreover, to mitigate the issue of samples with zero weights in high-dimensional domains, they introduce a regularizer that acts as a soft maximum over the logarithms of the weights which helps to control the divergence of the policy from the samples. These guiding samples can be adaptive and are constructed at each iteration of the policy search starting from the previous DDP solution. This helps in generalizing to unseen tasks where a sole DDP policy would fail due to un-modelled/incorrect physics.

The experiments use diverse tasks and demonstrate the importance of guiding samples along with the regularizer to direct policy optimization for swimmer, hopper, and walker. It also showed comparable results to TBDP in terms of orientation generalization while outperforming it on new terrains.

Overall, the paper presents a sustainable approach of minimizing the divergence of policy learning from its representative samples by incorporating multiple DDP solutions in its training. However, it still relies on a model-based DDP algorithm to succeed.

2 SAM-RL: Sensing-Aware Model-Based Reinforcement Learning via Differentiable Physics-Based Simulation and Rendering

The authors propose the SAM-RL (sensing-aware model-based reinforcement learning) framework that leverages a differentiable physics simulator to learn effective world dynamics. SAM-RL automatically updates the model by comparing the raw observations between the simulation and the real world and produces the policy efficiently. It also allows robots to select an informative viewpoint to better monitor the task process, thereby not requiring a sequence of camera poses at each step anymore. The simulated policies are learnt via DAgger and the sensing-aware Q function is learnt via supervised learning. Moreover, simulated actions are modified using an actor network (residual policy) before directly applying them in the real world to lessen the gap between the simulated and real world.

The experiments involve 3 tasks: peg-insertion, flipping a pancake, and needle-threading. They demonstrate that SAM-RL learns informative camera viewpoints for effective robot manipulation, is

Preprint. Under review.

around 80% effective compared to MFRL approaches, and the effectiveness of model-updation and a residual policy via ablation studies. Even though the ideas presented in the paper seem novel, in my opinion, it is poorly structured and needs thorough experiments with existing MBRL approaches for a fair comparison.

3 Differentiable Physics Simulations with Contacts: Do They Have Correct Gradients w.r.t. Position, Velocity and Control?

In this paper, the authors compare 4 gradient computation techniques using existing differentiable physics simulation tools namely LCP, convex optimization models, compliant models, and PBD, on three tasks and find that not all the computed gradients are correct.

All the tasks involve simple frictionless collision and differentiable cases. It is shown that in the task of simple collision, only the implementations using Time of Impact (TOI) compute accurate gradients. Whereas, the gradients wrt to velocity and control have the same direction but are inaccurate. However, even then they end up with reasonable solutions. In the task of reaching a specific target, the solutions by each implementation follow a different trajectory which demonstrates inconsistency from an optimization perspective. Furthermore, for task 3, none of the gradients performs nearly at par with the analytical gradients.

Through this paper, the authors encourage further research in investigating why inaccurate gradients can still result in successful optimizations. However, even though the authors manage to pinpoint a flaw in the current systems, they do not provide any novel approaches to mitigate the issue in this paper.

References

[1] Levine, Sergey and Koltun, Vladlen. (2013). Guided Policy Search. Proceedings of the 30th International Conference on Machine Learning. https://proceedings.mlr.press/v28/levine13.html

[2] Lv, J., Feng, Y., Zhang, C., Zhao, S., Shao, L., & Lu, C. (2022). SAM-RL: Sensing-Aware Model-Based Reinforcement Learning via Differentiable Physics-Based Simulation and Rendering. ArXiv. /abs/2210.15185

[3] Zhong, Y. D., Han, J., & Brikis, G. O. (2022). Differentiable Physics Simulations with Contacts: Do They Have Correct Gradients w.r.t. Position, Velocity and Control? ArXiv. /abs/2207.05060