Global policy using Gaussian regression

Sameer Kumar[1], Prakyath K[1], and. Myunghee Kim[1]
[1] Department of Mechanical Engineering, University of Illinois at Chicago, Chicago, IL, USA
Email: vbetan3@uic.edu , pkanth3@uic.edu , myheekim@uic.edu

Summary

In this paper, we have generalized control policy for the highly non-linear system using Gaussian Process Regression (GPR). Here we also used model estimation technique based on the Gaussian mixture model (GMM). GPR provides granular control for the tradeoff between bias and variance. We simulated results on the cart-pole problem and analyzed the generalized global policies. We also tested the generalization of new global policies using different local policy optimization such as iterative linear quadratic regulator (iLQR) and differential dynamic programming (DDP). Inferring from the results, we suggest that GPR could be an alternative to the global policy optimizers such as neural network,used in reinforcement learning.

Introduction

Fairly recent advancement in the machine learning and data-driven controls methods have demonstrated successful outcomes in controlling linear and low dimensional models. However, these models and algorithms have faced issues with generalizing to a higher dimension model and dynamic environment. Reinforcement learning (RL), on the other hand, has shown high constancy towards a non-linear system in the dynamic environment. However, RL typically has used the neural networks as global policy optimizers. These policies lack the adaptability to expand when the visual complexity of the real world increases. Especially in policy search methods in RL, policy outputs of deep convolutional neural networks are modeled as the mean of the underlying distribution.In contrast to this, GPR[4] representation uses the learned covariance matrices to describe the distribution of the underlying control signals adequately. The policy output itself can be modeled as a probability distribution as the control action for each observed states can be sampled from the distribution. This motivation leads to modification of global policy optimizer of guided policy search (GPS). GPS is a model-based RL method and which has demonstrated considerable results in manipulation[1]. In GPS, Local trajectory optimization on local policies generates training data which is used to train the global policies. In this paper, we demonstrated the strengths of GPR and how it can be an alternative as a global policy optimizer in RL.

Methods

Here instead of using the actual dynamics of the system, we estimated the dynamics from the observation data, using GMM based prior modeling, we fit the training data of the states, to obtain the posterior dynamics of the system. In the next stage, we used the dynamics obtained in the previous step to generate an optimal trajectory using iLQR, and we also tested with DDP. Researchers used the neural network, taking the states as the input parameters and controller gains (action) as the training labels to produce a non-linear controller [2] to match the local trajectories. In the proposed method of GP [3], we take the inputs as the states and controller gains as the training labels. We are in liberty in choosing the kernel function and also the optimizer to fit the local trajectories; these are the tunable parameters. The advantage of GP is that output of GP itself can be modeled as a probability distribution on the actions and future action can be sampled from it. We performed the local trajectory optimization using iLQR and DDP on the test simulation, results of these cases are illustrated below.

Results and Discussion

We use simulation tools to generate different rollouts for a given local policy. For instance, we applied local policies such that inverted pendulum attached to the cart maintained near vertical position; similarly, we applied different local policies and performed multiple rollouts. Then the dynamics of the system is estimated using this information extracted from the rollouts as mentioned in the following paper [1]. We used the iLQR and DDP as the local trajectory optimizers and compared the results as shown in the figure below. And finally, we used the GPR to train on the data provided by the local trajectory optimizers. To test the robustness of the global policy we generated trajectories using the system dynamics, estimated dynamics, and the GPR.

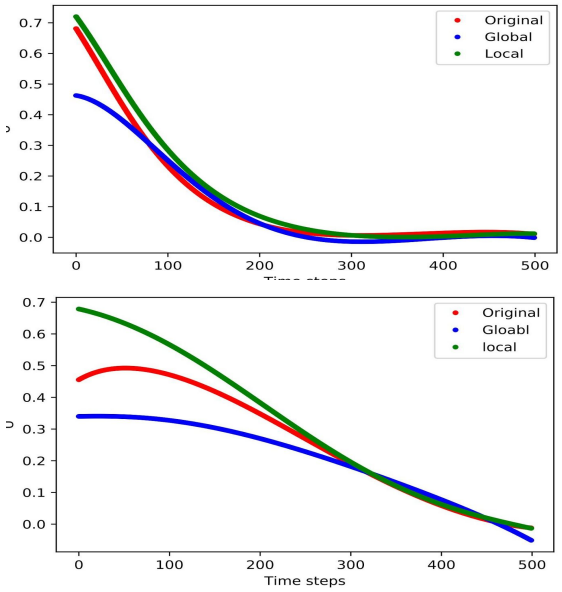

Figure 1: Results of cart-pole problem.

In the above figures we have shown one such trajectory. The red line (original) indicates trajectory generated using system dynamic, blue line (local) indicates trajectory generated using estimated dynamic and green line indicates trajectory generated by GPR (green). Top indicates the control signal applied on the trajectory using DDP. Bottom figure indicates the control signal applied on the trajectory using iLQR. The R2 metric for the GPR for iLQR was 0.83 and GPR for DDP was 0.91 (which is comparable to neural network efficiency [2]).

Future work

These results show that we can use GPR as in place of the neural networks in GPS. In the future, we would extend this work to the hopping robot and gait analysis, where high model uncertainty exists. GPS with GPR could be studied for models which are hard to learn and control such as hopping robot and wearable robots.

References

[1] Sergey et.al. End-to-End Training of Deep Visuomotor Policies. JMLR 2016.
[2] Julian V. et al. Learning a Structured Neural Network Policy for a Hopping Task. *IEEE RAL*, vol. 3, no. 4, pp. 4092-4099, Oct. 2018
[3] Deisenroth et al. "Approximate dynamic programming with Gaussian processes." In 2008 American Control Conference, pp. 4480-4485. IEEE, 2008..
[4] Deisenroth et al "Gaussian processes for data-efficient learning in robotics and control." IEEE transactions on pattern analysis and machine intelligence 37, no. 2 (2015): 408-423.