Active Learning for Probabilistic Machine Learning based modeling of Dynamical Systems

Tamil Arasan*, Selva Kumar *, Murugesan*, Meiyazhagan *, Gopinath * and Malaikannan *

* *Saama AI lab, Chennai, India*

Abstract. Machine learning models have shown a significant impact in modeling physical simulations. While machine learning has many benefits, one major drawback is the need for a large amount of data. We explore active learning as a remedy for addressing this issue. In our work, we studied the dynamics of two different dynamical systems, modeled their behavior using the Active Learning (AL)-enabled Gaussian Process (GP), and compared it with the vanilla GP. We demonstrate that AL-enabled GP shows superior performance with a lesser number of data points. AL can query for salient samples given a larger dataset to achieve orders of magnitude better Mean Square Error (MSE). In particular, for some instances, we were able to cut down the required samples by $\frac{1}{3}^{rd}$ and reduce the error rate by more than 10 orders of magnitude.

Introduction

Scientific Computing has begun adopting data-driven techniques such as Probabilistic Machine learning, Deep Learning, and more to accurately model physical phenomena using dispersed, noisy observations from coarse-grained grid-based simulations. In particular, Machine Learning models based on neural networks are data-hungry, and their performance is directly affected by the quality and quantity of the data. Moreover, the supervised machine learning models demand labeled data, which for scientific experiments, is expensive to gather in large quantities. This issue can be tackled using Active Learning, a special case of supervised learning. In this work, we used Active Learning to train the Gaussian Process to model different Nonlinear Dynamical systems, namely, Nonlinear Schrödinger (NLS) equation and Gross-Pitaevskii equation (GPE).

Method

Active Learning is selecting data in an iterative fashion to improve model performance by maximizing information acquisition with limited training samples. AL adds a certain cleverness on which samples to choose to improve the model's accuracy. In this method, the model is initially trained with a small subset of the data, and then a query strategy is used to acquire more useful samples from the dataset. Our work aims to sample more points in the steep regions and fewer points in the smooth regions i.e to bring good fitting in the entire parameter space of the wave function. The query strategy searches for the samples using their confidence of the model for the samples, i.e samples with the highest variance are selected for training the Gaussian Process [1] in the next iteration.

Following our earlier work [2], we have experimented with modeling the ground state wave function of One and Two Component GPE for using an active learning framework. We train Gaussian Process (GP) models with and without Active learning method enabled. We use modAL python package for running the experiments. We compare GP without Active learning (GP) and GP with Active Learning enabled (GP_AL) based two metrics, **(i) Number of samples required to achieve same error rate and (ii) Error rate for same number of samples.** For GP, we use 500 samples for training, whereas for GP_AL, we start the training with 50 samples and let the query strategy figure out the next samples. We let the training proceed until the error rate of GP_AL reaches up to GP for a fair comparison. We also compare GP_AL and GP by letting GP_AL consume the same number of samples as GP and compare the error rate of both. We observe a multi-fold decrease in the error rate of GP with active learning enabled for the same number of samples. To pronounce the versatility of our work, we experimented with the same setting for the case of NLSE, and it reduced the data requirement by 10% of the original data.

Conclusion

In this work, we propose a novel approach to exploit the Active Learning framework for training machine learning models like GP for modeling dynamical systems. We compared models trained with and without Active Learning. We observed that the Active Learning method could successfully query for salient samples to achieve the same error rate with a far smaller subset of the data. When feature space is complex, Active Learning achieves orders of magnitude better error rate for the same amount of data. Future direction will explore how to incorporate the Active Learning method into the data generation and gathering process for Physical Simulations.

References

[1] MacKay, D. The humble gaussian distribution (2006).https://www.seas.harvard.edu/courses/cs281/papers/mackay-2006.pdf

[2] Bakthavatchalam, T. A., Ramamoorthy, S., Sankarasubbu, M., Ramaswamy, R., & Sethuraman, V. (2021). Bayesian Optimization of Bose-Einstein Condensates. Scientific Reports, 11(1), 1-9. https://doi.org/10.1038/s41598-021-84336-0