IRIS-HEP Fellowship Proposal

Advancing an Active Learning Algorithm for
Efficient Searches of New Physics at the LHC

Author: Jason Hipkins — University of Maryland — hipkinsj@terpmail.umd.edu
Mentors: Lukas Heinreich — CERN Staff Scientist — lukas.heinrich@gmail.com
Kyle Cranmer — New York University — kyle.cranmer@nyu.edu
Irina Morales — New York University — iem244@nyu.edu

June 2, 2021

1 Background

Current methods for computing excursion sets of black-box functions (equivalently finding iso-hypersurfaces of n-dimensional scalar multivariate functions) are embarrassingly parallel and computationally expensive [1]. An active learning algorithm appropriately named 'excursion' has reduced the compute resources necessary to find excursion sets so that researchers can quickly classify BSM theories.

The 'excursion' package uses an iterative algorithm that solves a Bayesian optimization problem. When given a gaussian probability distribution of a function value an acquisition method will sample points from the parameter space which maximally reduce the uncertainty of the estimated excursion set. The algorithm can densely sample points near the iso-hypersurface by exploiting mutual information between previously selected best points and the next point to learn parameters of signal Monte Carlo that samples the points. Previous approaches to finding excursion sets are much slower since they test many more points in the parameter space. The benefits of the active learning approach can also be scaled in functions defined over higher dimension parameter spaces. Ph.D student Irina Espejo Morales implemented GPyTorch [2] which computes gaussian processes faster using the parallel processing architecture of GPUs. It has been able to find excursion sets of functions defined over 5-dimensional parameter spaces within a day. This is an exciting prospect in the search for new physics as it significantly reduces the time necessary to test BSM theories.

Participating in the development of the 'excursion' python package will advance my knowledge of machine learning and computational physics. It will allow me to explore in more detail concepts related to Monte Carlo experiments. I am excited for the opportunity to participate in the IRIS-HEP community full-time this summer. I will be able to greatly expand and apply my skill set with the mentorship of Kyle Cranmer, Lukas Heinrich, and Irina Espejo Morales. I will work to improve 'excursion' to efficiently solve problems with up to 11 dimensions or more. Providing this code to physicists will be a significant benefit in e.g. testing theories for SUSY and many other searches for New Physics.

2 Proposed Project

The current 'excursion' package is only able to run a few included black-box functions. Thus the goals of the project are twofold, implementing a python package that allows scaling the speed up of gaussian processes for higher dimensions and deploying the 'excursion' package so that physicists may input their own black-box functions. In a prior meeting to plan this proposal, Irina informed us that an author of BoTorch [3] suggested their library can allow scaling up to 19 dimensions or more. Implementing this code will be one portion of the project. Afterwards the 'excursion' package will be integrated into an application pipeline that can take a general black-box function from pMSSM. Work will also be done to add further parallelization to the algorithm. Improvements made to the 'excursion' package will be documented for presentation.

2.1

Timeline by 0.5 FTE Months	
Deliverable 1: Merge, refactor code, and implement BoTorch	
Jun 14 - Jun 25	Merge Irina's fork with Lukas's code. Get code running on local machine and refactor to reduce code complexity while adding comments and documentation. Select a git development method from my software engineering course materials and prepare code for containerization. Add code to measure run times and data structures to record those for future analysis.
Jun 28 - Jul 09	Begin research on implementing BoTorch and testing it. Ensure results are valid. Implement parallelization of for loops for gp runs. Measure changes in resource utilization along time and space dimensions. Establish performance improvements from parallelization and Irina's GPyTorch implementation.
Jul 12 - Jul 23	Finish refactoring code with BoTorch. Measure changed resource utilization along time and space dimensions. Analyze improvements from BoTorch when scaling up dimensionality.
Deliverable 2: Prepare documentation and implement general black-box function	
Jul 26 - Aug 06	Finish containerization and implement application on physics black box function (pMSSM observable).
Aug 09 - Aug 20	Extra time to buffer any delays or adding extra features. Bug checks for edge cases. Begin writing formal documentation of work done and code improvements.
Deliverable 3: Finish documentation and presentation	
Aug 23 - Sep 03	Finalize documentation and its results and prepare work for presentation.

References

[1] L. Heinrich, G. Louppe, and K. Cranmer, in Excursion set estimation using sequential entropy reduction for efficient searches for new physics at the lhc, 19th ACAT Workshop, Oral Presentation with slides (ACAT, Mar. 2019).

[2] I. M. Espejo, *Irinaespejo/excursion*, `https://github.com/irinaespejo/excursion`, 2020.

[3] M. Balandat, B. Karrer, D. Jiang, S. Daulton, B. Letham, A. G. Wilson, and E. Bakshy, in Advances in neural information processing systems, Vol. 33, edited by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (2020), pp. 21524–21538.