HSF-INDIA PROPOSAL
April 2023

Sourabh Dube
Arnab Laha

In the last decade, usage of machine learning (ML) algorithms has exploded in particle physics[1]; uses range from obvious binary classifiers for identifying BSM particles over standard model backgrounds to unfolding detector effects across several kinematic variables. Significantly, ML algorithms are used in 'everyday' analysis tasks by a small team or a single graduate student.

The usage patterns in Run1 and Run2 of the LHC show that most actual R&D of novel techniques happens typically on smaller university clusters or individual workstations, with larger farms used for either generation of simulation or the downsizing of datasets to smaller sizes. This downsizing is achieved by aggressive skimming and slimming of datasets, and by the use of modern tools such as multidimensional histograms which can be projected, and software such as Uproot which allows powerful data analysis libraries of the python ecosystem to be used seamlessly.

The production of faster simulation using ML algorithms, specifically generative adversarial networks (GANs) has also appeared in the literature. However, so far it has not been implemented in any LHC analysis. One possible bottleneck is that implementing a GAN that gives usable output requires large complicated networks and high amount of computing.

In our project, we propose to combine dimensionality reduction algorithms with GANs to construct an example pipeline for faster ML-based simulation to be implemented on a workstation.
(a) Algorithms such as PCA or UMAP will be used to obtain a latent dimensional (n) representation of the original kinematic variables (k). (Of course $n < k$).
(b) Algorithms such as a GAN or a variational autoencoder will be used to generate the n variables with a high degree of accuracy.
(c) The generated n variables will then be used to obtain the k "generated" kinematic variables.

Generating a smaller subset of numbers accurately (n instead of k), will be advantageous computationally, as well as in terms of the "loss-landscape" for the GAN. As an example, we shall test the generation by training a classifier which uses the kinematic variables as inputs. An auxiliary study of using the latent dimensional representation to perform the classification study will also be done. This project will be conducted on a workstation with an Intel XEON E5 processor, with 128 GB of memory. We hope that the success of this will encourage more users to consider implementing such faster ML-based simulation algorithms as 'everyday' tasks.

The task requires a student reasonably well-versed with python and preferably the basics of experimental particle physics analyses. We have local expertise within our group with ML algorithms, GANs, VAEs, as well as various dimensionality reduction algorithms. In addition, the training resources of the IRIS-HEP project are an important resource. The modular nature of the tasks leads to well-defined goals at each step. We anticipate reasonable results within a timeframe of approximately 5 to 6 months, given the appropriate trainee.

[1] https://github.com/iml-wg/HEPML-LivingReview