IRIS-HEP Fellowship Proposal:

Developing an automatic pruning utility for statistical models in HistFactory format

Oleksii Kiva
Igor Sikorsky Kyiv Polytechnic Institute

June 2022

Project Details:

The trial of theoretical conjectures about a physical system on actual detector observations is commonly formalized using the framework of statistical models. Commonly, statistical models employed in HEP experiments are large-scale. Thus, the problem of efficient derivation of maximum-likelihood estimates of statistical model parameters, given the observation data, arises when the parameter space of the model in question becomes huge in dimensions.
One way to optimize the process of parameter inference is to reduce the parameter space dimensionality by identifying which parameters of a statistical model are negligible in a certain strictly-defined sense and then redefine the model in a way that will exclude the components of its distributions that are dependent on this subset of parameters.

In pyhf, statistical models are constructed using a modular approach to build a parametrized family of complex probability density functions from more primitive conceptual building blocks, which are combined in a product [1]. What's already available in pyhf is only helpful if one manually decides and specifies exactly what blocks to remove from the model. The respective utility is accessible either via the command line or the Python API.
The goal of this project (suggested by the discussions in [2], [3]) is to devise, implement, document and integrate into the pyhf library framework another tool that will automatically decide which blocks to remove from the statistical model in HistFactory format, given its pyhf-specification in a JSON file and some ‘pruning’ criteria for the blocks such as the error threshold in parameter estimation. The tool should output a JSON serialization of a reduced (‘pruned’) model in the same format.

Multiple algorithms that can determine which pieces of the model should be ‘pruned’ are going to be considered and the most feasible of these will be chosen and implemented. To support the choice, the comparative quantitative analysis will be conducted to estimate the time complexity of these algorithms as well as to understand by how much the results of statistical inference with the model changes due to the specific kind of pruning performed.

This project will be done under the mentorship of Dr. Alexander Held.

Timeline:

- Weeks 1-2: Get familiarized with pyhf & HistFactory.
- Weeks 3-4: Try out pruning by hand & set up a framework to compare results of parameter inference with and without pruning.
- Weeks 5-8: Start implementing pruning algorithms & test them.
- Weeks 9-10: Document implementations & provide tutorial material.
- Weeks 11-12: Prepare a presentation & address any additional topics that will come up during the project.

References:

[1] https://cds.cern.ch/record/1456844
[2] https://github.com/scikit-hep/cabinetry/issues/311
[3] https://github.com/scikit-hep/pyhf/discussions/1735