Project: Robot Physicist

Team: Manoj Kumar, Phil Yeres, Michele Ceru

Introduction:

Particle physics experiments generally require expensive shared equipment, or time consuming simulations. Therefore, optimizing the experiment simulation procedure to minimize the time required to achieve results is a subject of importance to physicists. The aim of our project is to use data science techniques to reduce resources used in the experimental process.

The experimental model has three components:

1. Experimental configurations such as the energy required to run the particular experiment (phi)
2. Quantities that we would like to infer from the experiment (theta), for instance, we may want to estimate the Weinberg angle (a fundamental parameter of the Standard Model of particle physics).
3. Data generated from the experiment (X)

The aim of our project is to estimate the underlying data generating distribution from simulations which we'll use to predict the settings of the experiment that allow us to obtain theta with high confidence. If we can predict these settings, we can reduce the number of experiments we'd need to run.

Description:

Part 1: Toy data

As a starting point we would estimate theta and phi on toy problems where we know the ground truth of the data generating distribution. For example, X can follow a gaussian distribution with mean theta and variance phi. For this toy data we will take the following steps:

1. For the toy-data generating distribution we would generate a number of samples for a fixed value of theta and phi
2. We would then estimate P(X | theta, phi). For a simple distribution, we could use a histogram to approximate this using evenly-spaced grids.
3. The third step would be to infer P(theta | X, phi) from P(X | theta, phi) using Bayes theorem:

   $$P(\text{theta} \mid X, \text{phi}) = P(X \mid \text{theta}, \text{phi})\ P(\text{theta} \mid \text{phi}) / P(X \mid \text{phi}).$$

   We would also fix the prior P(theta | phi) to some simple distribution and not infer this from data.
4. Run loops from 1-3 for different values of phi such that the information gain in P(theta | X, phi) is maximized. As a possible extension of our project, we may perform bayesian optimization to choose the next value phi intelligently given previous evaluations.

Part 2: Physics data

After we've created the toy data pipeline, we'll extend it for the actual simulated experiments provided by our project advisor (https://github.com/lukasheinrich/higgs-mc-studies). For the actual simulated experiments, the ground truth is not known, so we plan to make use of a liklihood-free inference toolbox (https://github.com/diana-hep/carl) to estimate P(X | theta, phi).

When we move from using the toy data to actual experimental data, and the computational load will increase significantly. Therefore, as discussed with our project advisor, we will look for opportunities to parallelize computations, and we will make use of AWS and docker to deploy the experiment simulator to the cloud.