Master's Thesis Project - Executive Summary

Letizia Palmas

Word count: 982

The aim of this project is to reproduce the methods outlined in the paper Unbinned multivariate observables for global SMEFT analyses from machine learning (2023). In our study, we will explore the Standard Model Effective Field Theory (SMEFT) framework by analysing high-energy particle interactions and inferring the presence of New Physics within our data. SMEFT represents an extension to the Standard Model of particle physics (SM) that addresses its flaws and parametrises the effects of New Physics (NP) while preserving the fields and symmetries of the SM. Despite its success, the SM is thought to represent a low-energy approximation of a more fundamental theory. SMEFT addresses this argument by providing a strategy to perform indirect searches of NP and investigate any deviations of the data from the SM. These strategies are especially beneficial when NP occurs at energy scales beyond the reach of current colliders.

In order to perform an extensive SMEFT study, it is necessary to provide a precise analysis of the events detected by the collider. The kinematic information of the features in the dataset can be employed to perform parameter inference and compare the results with the EFT predictions.

The datasets we will study belong to three categories:

- The observed data;
- The Monte Carlo generated samples at the parton level;
- The Monte Carlo samples with the full set of features.

The observed data represents a toy dataset reproducing the events measured at high-energy particle accelerators such as the Large Hadron Collider (LHC). We consider the process of top-quark pair production as a result of proton-proton collisions: $pp \to \tau\tau \to b\ell^+\nu_\ell \bar{b}\ell^-\bar{\nu}_\ell$, where $\ell^\pm$ are the positive and negative leptons, ν_ℓ are the lepton (anti)neutrinos, and $b\bar{b}$ is the b-jet pair. Our analysis will detect the presence of NP within this data and aims to accurately infer the coefficients that parameterise the SMEFT model.

The Monte Carlo-generated samples at the parton level will be used in a primordial implementation of deep learning strategies as a reference for a more complete analysis. These samples are assigned to 5 separate datasets, one representing the SM, and 4 representing the linear and quadratic definitions of the 2 Wilson coefficients that parametrise the EFT hypothesis: $c_{\mathrm{dt}}^{(8)}$ and $c_{\mathrm{qt}}^{(8)}$. The parton level data represents a simplified scenario for which the analytical form of the likelihood is known.

The complete samples are structured in the same manner as the parton level data but instead are described by the set of features of the observed events. These features represent the kinematic variables of the final-state particles and allow us to provide a more exhaustive analysis of the SMEFT.

The initial section of our study involves the construction of binned observables to perform a classical statistical inference of the two Wilson coefficients. The binned approach allows us to represent the likelihood of the data without resorting to numerical simulations. The drawback of this method is, however, the inherent loss of information that results from averaging the data points that fall within the same bins. We can devise an optimal binning strategy for each of the observables in the dataset, and construct the log-likelihood of the data based on the χ^2, assuming Gaussian distributed events:

$$-2\log\mathcal{L} = \sum_{i=1}^{N} \frac{[X_i - E_i(\mathbf{c})]^2}{\sigma_i^2} \equiv \chi^2 \tag{1}$$

where E_i represents our model, encapsulating both the SM and the EFT effects, and X_i represents the measured data. In this case, the variables correspond to the transverse momentum of the positive lepton, $p_{t\ell}^{+}$, but the formula applies equally to all observables, provided a pertinent binning. We employ this likelihood function and a uniform prior to analyse the posterior distribution of the parameter space through Nested Sampling. Based on the algorithm's outcome, we identify the points at which the likelihood function reaches its maximum value and provide confidence bounds for both Wilson coefficients.

The second component of the project implements supervised Machine Learning (ML) classification techniques to train four distinct Neural Networks (NN) and extract the distribution of the data at the parton level. The Multi-Layer-Perceptron (MLP) architecture is kept consistent for each dataset, whereas hyperparameters are tuned individually for each NN. Each model is trained to minimise the Binary Cross-Entropy Loss function and optimised using the Adam optimiser with decoupled weight decay (AdamW). The output of each model is a decision boundary function, $g(\mathbf{x}, \mathbf{c})$. This function allows us to derive the cross-section ratios that describe the likelihood of the data. To validate the models, we compare their outputs with the analytical cross-sections and observe that these align well with the ML parametrisation of the likelihood. This confirms the efficacy of ML in this setting and validates its potential application within the full framework.

In the final part of our analysis, we integrate the previously outlined techniques to produce a ML parametrisation of the unbinned likelihood, which allows us to construct the posterior distribution of the observed data using Nested Sampling (NS). We train our models on the full MC samples, employing the learned decision boundaries to characterise the complete unbinned log-likelihood that will be evaluated in the NS algorithm. From this, we are able to infer the posterior probability distribution in the EFT parameter space. Following the traditional approach, we then determine the confidence level intervals for the unbinned observables. We eventually observe that the bounds associated with the 95% confidence level are significantly narrower than those obtained from the classical analysis, indicating a higher level of precision in the results. To illustrate our findings, we include a table comparing the maximum likelihood estimation and the 95% confidence intervals for the Wilson coefficients in both the traditional approach and the ML framework:

	Coefficient	**Traditional Analysis**	**ML Parametrisation**
MLE	$\boldsymbol{c}_{\mathbf{dt}}^{(\mathbf{8})}$	-0.669	0.453
	$\boldsymbol{c}_{\mathbf{qt}}^{(\mathbf{8})}$	-0.856	-0.653
95% CI	$\boldsymbol{c}_{\mathbf{dt}}^{(\mathbf{8})}$	$[-1.256, 0.324]$	$[0.271, 0.602]$
	$\boldsymbol{c}_{\mathbf{qt}}^{(\mathbf{8})}$	$[-0.952, 0.021]$	$[-0.753, -0.536]$

Table 1: Comparison of the Nested Sampling results following the posterior evaluation of the parameter space using binned likelihoods and the Machine Learning approach.

These results further underline the increased potential of our strategy to accurately constrain the values of the Wilson coefficients.

Our study shows that the application of Machine Learning to the study of particle physics events holds significant promise for advancing our understanding of the Standard Model Effective Field Theory. We have reproduced a robust framework for parameter estimation that significantly enhances our ability to detect and infer the presence of New Physics.