# Samuele Grossi

*Second Year PhD Report*

*samgrs1997@gmail.com*

## Tutors

Riccardo Torre and Simone Marzani

## Research Activity

Over the past year, my research focused on two distinct projects. The first was a continuation and conclusion of my work from the first year of my PhD, the second was centered on the statistical method of two-sample testing: we analyzed various "simple" metrics to establish a robust benchmark for testing generators.
In the following I will briefly discuss the former, given that this year we refined some technical aspects, while the general information and procedure can be found in the First Year Report and I will focus more on the latter, as it represents a completely new line of research. I will refer to the two projects as "SMEFT Drell-Yan" and "Generative models evaluation" respectively.

**SMEFT Drell-Yan:**
I have analysed how dimension-six SMEFT operators impact the Drell-Yan processes

$$\mathrm{q\bar{q}} \to \ell^+\ell^- \,, \qquad \mathrm{u\bar{d}} \longrightarrow \ell^+\nu_\ell \,, \qquad \mathrm{d\bar{u}} \longrightarrow \ell^-\bar{\nu}_\ell \,,$$

in order to constrain the Wilson coefficients. This is useful to investigate new physics scenarios in an indirect way. This year we focused our attention on determining whether a single binned analysis with fine binning could be as competitive as a multi-differential analysis. We found that this is the case for certain operators, particularly the ones less sensitive to angular variables. We submitted our findings to the EPJ C journal which published our work.

**Generative models evaluation:**
The idea for this work originates from a current problem in high-energy physics. In this field, simulations to reproduce experimental results rely on complex model-based Monte Carlo generators like, for example, Powheg, which provides highly accurate outcomes but at a significant computational cost. This could become problematic in the near future, when High Luminosity LHC will start working, because it will generate a larger volume of data to simulate with respect to now. As a consequence, the community has started exploring alternative approaches to generate synthetic results, considering in particular machine learning algorithms which learn from the robust generators. The advantage of such algorithms is their speed, as they can deliver results much faster than Monte Carlo generators, while the drawback is their accuracy, as they learn from finite datasets without achieving 100% accuracy (to avoid overfitting). The accuracy issue, i.e. if it is true that synthetic data generated with a machine learning algorithm are as reliable as the ones produced with a Monte Carlo generator, is of paramount importance in the context of high energy physics, given the stringent precision of experimental measurements. For this reason, recently, some effort has been put to compare the dataset produced by machine learning algorithms versus the one produced by the traditional Monte Carlo generators. Formally this type of problem can be seen as a two sample test.
A two sample test aims to understand if two independent data samples are drawn from the same probability density function (PDF). This is achieved by employing a so-called *test statistic* or *metric*, which is a function that quantifies the "distance" between the two samples. Intuitively, the closer they are, the higher the probability they are drawn from the same PDF is. For simple hypothesis, the most powerful test statistic is the (log-)likelihood ratio (LLR) test, as stated by the Neyman-Pearson lemma. However, the analytic form of the PDFs generating the two samples is usually unknown, making the likelihood ratio (LR) test impractical. Nonetheless some machine learning algorithms, such as classifiers, can be trained to approximate the likelihood ratio and can be used to perform a two-sample test between different datasets. This approach is currently being explored by a part of the research community.
What we have done in the second part of this year is useful to set a benchmark in this context. While classifiers can be effective, they require long training and their "black-box" nature often makes their results challenging to interpret. Additionally, current machine learning algorithms do not mimic Monte Carlo generators with the required precision, making classifiers excessively powerful for a tasks that could be addressed with simpler methods. Therefore we

selected some *test statistic* which are based on one dimensional Integral Probability Measures (IPM): the mean Kolmogorov-Smirnov (MKS), the sliced Kolmogorov-Smirnov (SKS) and the sliced Wasserstein distance (SWD). These *metrics* have straightforward analytical form and can be computed quickly. Additionally, we considered two recent proposals by Kansal et al.: an estimate of the Fréchet gaussian distance (FGD) at infinite sample size and the Maximum Mean discrepancy (MMD).
We considered toy models, namely multi-dimensional gaussians and mixture of multi-dimensional gaussians, and physics inspired datasets, namely the gluon initiated jets that can be found in the JetNet dataset. We deformed them in various ways and we extracted two samples: one from the *reference* model and one from the *deformed* model. Each deformation is described by a parameter $\epsilon$. Applying the selected metrics we can establish, for each deformation, the smallest value of $\epsilon$ we are sensitive to. Additionally we estimated the errors on the $\epsilon$ values found, to ensure the robustness of our test-statistics. We considered also varying space-dimensionalities and sample sizes, that along with the different deformation types ensure that our metrics can be used in a general framework. When considering the toy models we also performed the two sample test using the LR test, to obtain the best reach and compare it with the results of the selected metrics.
We found that:
- the LR test, when available, is an order of magnitude more sensitive then each metric tested.
- our metrics based on IPMs has performances comparable or even superior to the FGD and MMD while being usually faster and easier to implement.

Additionally, given the generality of our approach and the fact that our results have a clear interpretation, we believe we have established a clear and robust benchmark for comparison when using classifiers trained to approximate the LR test. Given the current state of generators we think that, in cases where training a classifier would be too complex or resource-intensive, the proposed test statistics can be an attractive option thanks to their simplicity and low computational cost.

## Courses

Attended:
- Advanced Statistics for Data Analysis (F. Badaracco, F. Di Bello, F. Parodi, 3CFU).
- Introduction to the Foundations of Quantum Mechanics and Applications (P. Solinas, N. Zanghì, 3 CFU),

Exam passed:
- Non-Abelian Gauge Theory (N. Maggiore, 3CFU),
- Fisica Teorica (G. Ridolfi, 3 CFU),
- QCD and Collider physics (S. Marzani, 3 CFU).

## Publications

- *More variables or more bins? Impact on the EFT interpretation of Drell–Yan measurements*, published the 19/07/2024 in: The European Physics Jornal C 84, 713 (2024).
- (In progress) *Comparison of the performances of non-parametric two-sample tests for high dimensional samples*. Soon to be available on arxiv.

## Schools, Seminars and Conferences

- PhD School: "Theory meets experiments, The high intensity frontier of particle physics", at GGI, in Florence, Italy, from 20 to 24 November 2023.
- Seminars: Weekly seminars organized by INFN in Genoa through the whole year.
- Conference: "European AI for fundamental physics conference, EuCAIFCon24", in Amsterdam, from 30 April to 3 May 2024. I presented a poster for the work *Comparison of the performances of non-parametric two-sample tests for high dimensional samples*.
- Conference: "BOOST 2024, 16th International Workshop on boosted object phenomenology, reconstruction, measurements, and searches ar colliders", in Genova, from 29 July to 2 August 2024.
- (Soon to be attended) Conference: "PHYSTAT Workshop on "Statistics meets ML", in London, from 9 to 12 September 2024 - Poster.
- (Soon to be attended) Conference: "Fourth MODE Workshop on Differentiable Programming for Experiment Design", in Valencia, from 23 to 25 September 2024 - Poster.