Matteo Barbetti • Master Thesis Abstract

matteo.barbetti@stud.unifi.it

Relatore:
Dott. Lucio Anderlini
anderlinil@fi.infn.it

Correlatore:
Prof. Piergiulio Lenzi
piergiulio.lenzi@unifi.it

Techniques for parametric simulation with deep neural networks and implementation for the LHCb experiment at CERN and its future upgrades

The LHCb experiment [1] is one of the four detectors along the accelerator ring of the Large Hadron Collider (LHC) at CERN, and is designed for the study of heavy flavour physics in proton-proton collisions. Its primary goal is to look for indirect evidences of phenomena beyond the Standard Model in CP-violation and in rare decays of b- and c-hadrons.

The upgraded LHCb detector will get back to data taking in 2021 with the LHC Run 3, relying on an increase of its statistical power by at least an order of magnitude thanks to a fully software trigger system [2]. This will allow to reach unprecedented accuracy as long as the Collaboration will be able to provide simulated samples as large. As a direct consequence, the production of such simulated samples will dominate the computing effort of the experiment [3].

Reproducing accurately all the physics processes from the proton-proton collisions to the radiation-matter interactions within the detectors (the *full* simulation approach) is already now incapable to sustain the analysis demands of the various LHCb physics groups, and it is therefore necessary to adopt faster solutions to take full advantage of the upgraded detector [3].

Ultra-fast simulation needs lower computing resources, renouncing to reproduce radiation-matter interactions and parameterizing directly the high-level response of the detector. The LHCb subsystems are based on various physics processes, also very different, that make building high-level parameterizations non-trivial: for example, the *Particle Identification* (PID) system combines information from RICH, calorimeter and muon detectors. This task can be carried out effectively by *Generative Adversarial Networks* (GAN), a powerful class of deep learning algorithms able to reproduce highly faithful and diverse probability distributions thanks to a *generative model* learned directly from data [4].

A large part of this thesis has concerned the development and implementation of *state-of-the-art* GAN algorithms [5] to provide the high-level response of the PID subsystems of LHCb. Following what done in Ref. [6] where GANs are successfully used to reproduce the outputs of the RICH system to traversing particles, I have generalized and formalized such results building generative models based on neural networks able to parameterize faithfully the high-level response of the Global PID system of LHCb.

These neural networks were trained over the calibration samples collected in 2016, in order to provide datasets composed by an unbiased selection of long-lived particles. I have modified the learning procedure to subtract statistically the residual background within the training data, and I have developed an *independent* algorithm capable to measure the quality of the generated samples. This strategy has allowed to build models capable not only to parameterize the high-level response of specific detectors (such as RICH detectors and muon system) to different

particles traversing them, but also to reproduce the distributions of variables resulting from the combination of various detector responses (the Global PID system).

Therefore, given a few basic information such as the particle type, its kinematics and the total number of tracks within the detector, the trained models are able to synthesize accurately a wide range of probability distributions representing the response obtained from a single detector or from their combination.

The second important personal contribution is related to the design and development of the `mambah` framework, a Python package aimed to provide and manage *user friendly* data structures for High Energy Physics applications. All `mambah` objects were designed to take full advantage of a *batch-grained* framework, using the most modern softwares for parallel computing and exploiting efficiently hardware accelerators, such as GPUs or FPGAs.

Within the `mambah` project, I have dealt with the implementation of database management functions and to the design of a simulation framework based on `mambah` and named `mambah.sim`. The `mambah.sim` module allows to selectively generate particles with the kinematics set by the proton-proton collision and to propagate them within the detector defining custom parameterization functions for efficiencies and resolutions: I have developed the parameterization for the PID system of LHCb.

I have proved the correctness of the implemented models, showing the generalization capabilities of GANs in describing decay channels different from the one of training. Lastly, I have shown that the samples produced by `mambah.sim` are competitive with the *full* simulated ones, while ensuring a significant reduction of the computing cost.

References

[1] LHCb Collaboration, A. A. Alves Jr. *et al.*, *The LHCb Detector at the LHC*, JINST **3** (2008) S08005.

[2] LHCb Collaboration, *LHCb Trigger and Online Upgrade Technical Design Report*, Tech. Rep. CERN-LHCC-2014-016. LHCB-TDR-016, CERN, 2014.

[3] LHCb Collaboration, *Upgrade Software and Computing*, Tech. Rep. CERN-LHCC-2018-007. LHCB-TDR-017, CERN, Geneva, 2018.

[4] I. J. Goodfellow *et al.*, *Generative Adversarial Networks*, `arXiv:1406.2661`.

[5] M. G. Bellemare *et al.*, *The Cramer Distance as a Solution to Biased Wasserstein Gradients*, `arXiv:1705.10743`.

[6] LHCb Collaboration, A. Maevskiy *et al.*, *Fast Data-Driven Simulation of Cherenkov Detectors Using Generative Adversarial Networks*, `arXiv:1905.11825`.