Data Driven Simulation of Cherenkov Detectors using Generative Adversarial Network

Artem Maevskiy
Laboratory of Methods for Big Data Analysis
National Research University Higher School of Economics
11 Pokrovsky boulevard, Moscow 109028, Russia

Denis Derkach
Laboratory of Methods for Big Data Analysis
National Research University Higher School of Economics
11 Pokrovsky boulevard, Moscow 109028, Russia

Nikita Kazeev
Laboratory of Methods for Big Data Analysis
National Research University Higher School of Economics
11 Pokrovsky boulevard, Moscow 109028, Russia

The Yandex School of Data Analysis
11/2 Timura Frunze St. Moscow 119021, Russia

Department of Physics
Sapienza University of Rome,
5 Piazzale Aldo Moro, Rome 00185, Italy

Andrey Ustyuzhanin
Laboratory of Methods for Big Data Analysis
National Research University Higher School of Economics
11 Pokrovsky boulevard, Moscow 109028, Russia

The Yandex School of Data Analysis
11/2 Timura Frunze St. Moscow 119021, Russia

Maksim Artemev
Laboratory of Methods for Big Data Analysis
National Research University Higher School of Economics
11 Pokrovsky boulevard, Moscow 109028, Russia

Lucio Anderlini
Istituto Nazionale di Fisica Nucleare, Sezione di Firenze,
via G. Sansone 1, Sesto Fiorentino 50019, Italy

Abstract

A typical analysis in high-energy physics requires a simulated sample of events that represents the detector response and can be used to judge small effects in the real data sample collected. New generations of collider experiments, in particular the Large Hadron Collider (LHC) upgrade, will require an unprecedented amount

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

of simulated events to be produced to cover the statistics needed. Such large scale productions are extremely demanding in terms of computing resources. Thus new approaches to event generation and the simulation of detector responses are needed. In this paper, we describe a generative approach for obtaining high-level reconstructed observables while bypassing the simulation of the low-level detector interactions. We provide an experimental evaluation of our approach for Cherenkov detectors. Improving the computational efficiency is crucial, as an accurate simulation of Cherenkov detectors is computationally expensive, it takes up to 30% of simulation CPU time for LHCb events. This fast simulation is trained on real data samples collected by the LHCb experiment at the LHC during run 2. We demonstrate that the approach provides high-fidelity results thus physics analyses can benefit from the newly simulated samples.

1 Introduction

Simulation of particle collisions occurring at the Large Hadron Collider (LHC) plays a crucial role in experimental measurements. Often, the amount of the simulated data and the accuracy of description translate to a systematic uncertainty on the experimental result. The demand for simulated events is growing rapidly with the increase of luminosity at the LHC. Given the computational requirements of accurate detector simulation algorithms, they cannot be used to produce all events. Therefore faster approaches to event generation and simulation are needed and were developed in several experiments (see for example, Giammanco [2014]).

The LHCb detector Alves et al. [2008] is one of the four major experiments at the LHC in CERN. It is designed primarily to study particles containing c- and b-quarks. This requires robust particle identification (PID). PID in LHCb is provided by four subsystems: the calorimeter system, the two Ring-imaging Cherenkov (RICH) detectors and the muon stations. Simulating the RICH detectors is particularly computationally expensive due to the need to accurately model the optical photon propagation, as well as low-energy secondary electrons, diffraction, and absorption effects Easo et al. [2005].

The first attempt to apply Generative Adversarial Networks (GANs) Goodfellow et al. [2014] to fast simulation in physics analyses was performed recently in Paganini et al. [2018]. This attempt used physically-motivated generation of the calorimeter response as a training sample with the aim to mimic the low-level detector response.

In this paper, we propose a novel solution to the problem of fast simulation for the RICH detectors at LHCb. It does not rely on physics-based simulation and instead uses a data-driven GAN to directly generate the high-level reconstructed observables.

2 Generative adversarial networks

The key idea behind GANs is simultaneous training of two neural networks. One network, named *generator*, takes samples from a known distribution (typically a standard normal random vector) and transforms them. The goal of the generator training is making its output distributed similarly to data. The other network, *discriminator*, is given both data and generator's output as input and is trained to distinguish between the two. For conditional generation, the typical approach is concatenating the conditions vector to both the generator and discriminator inputs.

The training of the two networks occurs in turns, and typically the loss of the generator is the negative from that of the discriminator. In the classical setup Goodfellow et al. [2014], the metric optimized by the discriminator is cross-entropy, which leads to overall equilibrium achieved when the Jensen-Shannon (JS) divergence between the data and the generated samples is minimized. The GAN training using the JS metric suffers from a number of difficulties. Among them are vanishing gradients for the case of the discriminator significantly outperforming the generator; unstable training in the case any network outperforms the other; mode collapses when the generator learns to cover only a part of the data distribution. Despite that, GANs were a ground-breaking success in the field of data-driven generative models for complex distributions, in particular images.

In order to address the shortcomings of the classical JS GAN, other metrics, such as Wasserstein distance and Cramer distance, were proposed Arjovsky et al. [2017], Gulrajani et al. [2017], Bellemare et al. [2017]. They provide a smooth measure even for disjoint distributions, which helps to prevent mode collapse; they provide non-zero gradients even for a perfectly trained discriminator, allowing for stable training without the need for excessive regularization and hyper-parameters tuning.

It has been shown that a naive implementation of the Wasserstein GAN results in biased gradients estimates in training Bellemare et al. [2017], Derkach et al. [2019]. Since high fidelity is the primary objective of our model, we used the Cramer metric, which combines the stability advantages of Wasserstein distance with unbiased gradient estimates.

3 Setup overview

3.1 LHCb RICH detector

Ring-imaging Cherenkov (RICH) detectors make use of the Cherenkov effect to identify particles. A particle traversing through a transparent medium with speed greater than the speed of light in the medium emits Cherenkov photons at an angle, that is a function of the particle's velocity. Therefore, measuring this angle and momentum allows to reconstruct the mass of the particle and thus provide the necessary PID information. In RICH, the Cherenkov light is focused on pixel hybrid photon detectors Alves et al. [2008], which provide fine spatial resolution and hence allow for the measurement of the Cherenkov angle.

The data from LHCb RICH pixels is processed using the global likelihood approach Forty and Schneider [1998], by finding the optimal particle type hypotheses for each of the tracks. The PID information is then aggregated per charged track in the form of differences between log-likelihood values for a given particle type hypothesis and a pion hypothesis for that track. These differences are named `RichDLL*`, '`*`' standing for `k` (kaon), `p` (proton), `mu` (muon), `e` (electron) and `bt` (below the threshold of emitting Cherenkov light); e. g. `RichDLLp` stands for the log-likelihood difference between a proton and a pion hypothesis for a given track.

3.2 Data

We train our model on data from several real decay samples collected in 2016 Lupton et al. [2016]. These are samples of charged tracks of different particle types that have been selected without the use of information from the PID subsystems response to those tracks. Each particle is a result of a specific decay channel[1].

The `RichDLL*` variables are generated for each track candidate using momentum, pseudorapidity and the number of reconstructed tracks in the event as inputs to the neural network. The design choice of generating `RichDLL*` variables instead of raw RICH pixels is motivated by the performance gain from bypassing the costly discrete likelihood optimization problem. It also improves learning stability due to the reduced dimensionality of the target space.

Due to the fact that the samples are coming from the real data, they contain background noise. The signal `RichDLL*` distributions are extracted from such data using the sPlot technique Pivk and Le Diberder [2005]. This method results in having non-unit sample weights such that weighted `RichDLL*` distributions are those of the signal component. The weights are applied to the loss functions during the training process.

3.3 Network architecture

Both generator and discriminator have 10 fully connected hidden layers with 128 units in each, with rectified linear unit (ReLU) activation functions. The latent space dimensionality for the generator is 64, and the distribution is the standard normal random vector $\{N(0,1), ..., N(0,1)\}$ concatenated with the input parameters, i.e momentum, pseudorapidity and the number of reconstructed tracks in the event.

[1] We use muons from $J/\psi \to \mu^+\mu^-$, kaons from $D^{+*} \to D^0(K^-\pi^+)\pi^+$ and $D_s^+ \to \phi(K^+K^-)\pi^+$, pions from $D^{+*} \to D^0(K^-\pi^+)\pi^+$ and $K_S^0 \to \pi^+\pi^-$, protons from $\Lambda^0 \to p\pi^-$. For each of the processes listed, both the process itself and its charge conjugate are implied

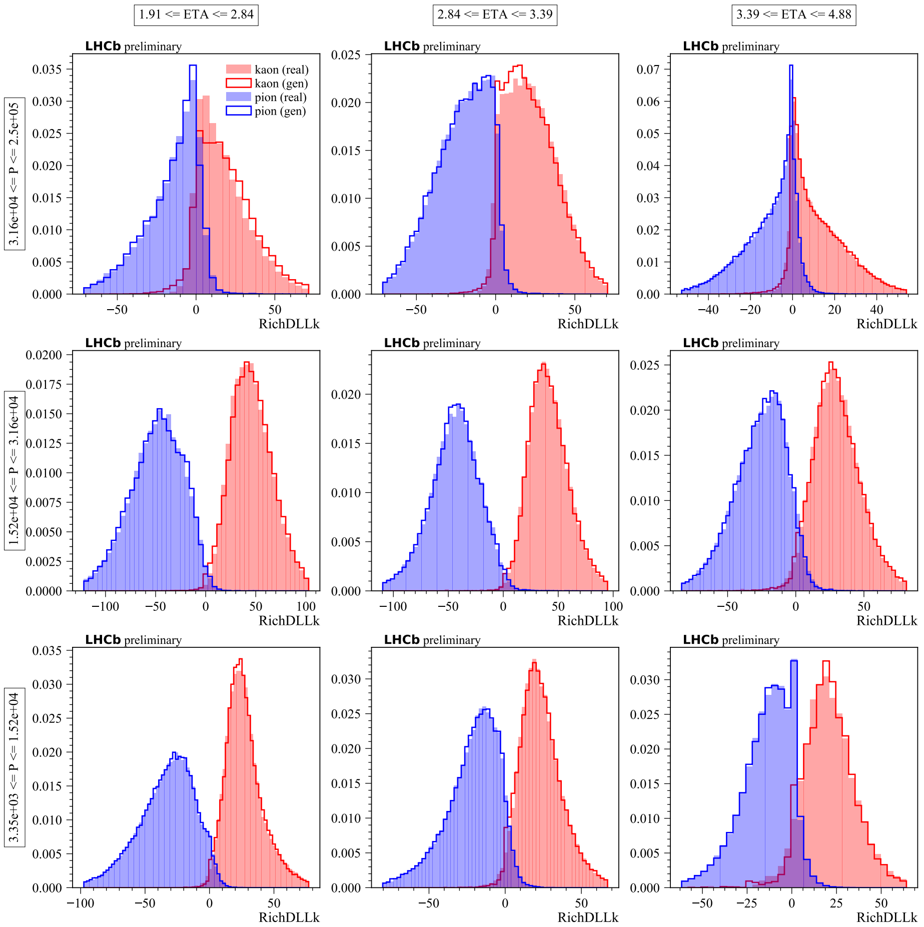


Figure 1: Weighted real data and generated distributions of `RichDLLk` for kaon and pion track candidates in bins of pseudorapidity (ETA) and momentum (P, MeV) over the full phase-space.

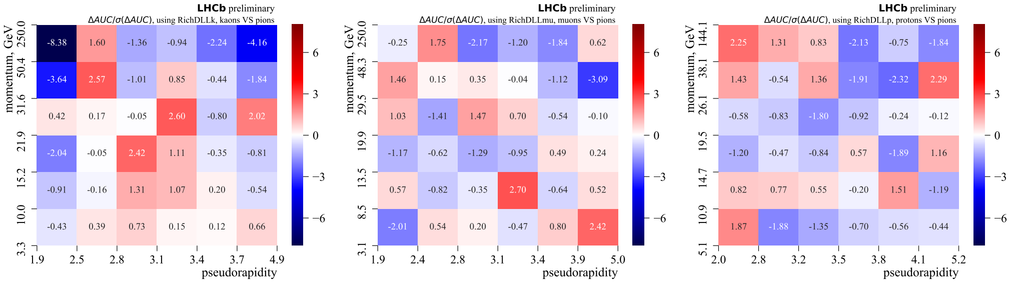
The output dimensionality of the discriminator network is 256 (Cramer metric uses multidimensional output). The output layers of both generator and discriminator do not use activation functions.

We use quantile transformation to transform both features and target variables distributions into standard normal. For our dataset, this results in faster convergence and higher output fidelity, than the commonly used linear scaling. We use exponential learning rate decay.

4 Results

Figure 1 shows a comparison of weighted real data and generated distributions of `RichDLLk` for kaon and pion track candidates, in bins of momentum and pseudorapidity. The binning is only applied when plotting, while the model itself is trained on continuous input.

In order to quantify the quality of the model in various regions of the phase space, area under the ROC curve (AUC) values were calculated in momentum-pseudorapidity bins for binary classification cases using both real data and generated variables. Figure 2 shows differences between AUCs divided by uncertainties for real and generated samples for discriminating kaons, muons and protons form pions, classifying with the `RichDLLk`, `RichDLLmu`, and `RichDLLp` variables, respectively, in bins of momentum and pseudorapidity. The uncertainty of the differences between AUC values was estimated using bootstrap Bertail et al. [2009]. Most of the differences are not greater than a few standard deviations, with no obviously biased regions, possibly with the exception of marginal bins that lack training statistics.



(a) kaons vs pions

(b) muons vs pions

(c) protons vs pions

Figure 2: Differences between the real- and generated-sample areas under ROC-curves divided by uncertainties for discriminating kaons, muons and protons from pions, classifying with the `RichDLLk`, `RichDLLmu` and `RichDLLp` variables, respectively, in bins of momentum and pseudorapidity.

5 Conclusion

High-quality fast simulation of the RICH detectors at LHCb can be achieved using generative models. In particular, GANs have the potential to be a good candidate for such an approach. As training can be done on real data directly, there is no need for later tuning and corrections of the model, compared to the way regular accurate detector simulation algorithms are used.

The proposed model shows a good approximation of the real data distributions with some imperfections in low-statistics regions. The systematic effects due to the usage of this approach of fast simulation fast simulation are thus expected to be small, which gives good prospects to the the future measurement.

Acknowledgements

The research leading to these results has received funding from Russian Science Foundation under grant agreement n° 19-71-30020.

References

A. A. Alves, Jr. et al. The LHCb Detector at the LHC. *JINST*, 3:S08005, 2008. doi: 10.1088/1748-0221/3/08/S08005.

M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv e-prints*, art. arXiv:1701.07875, Jan 2017.

M. G. Bellemare, I. Danihelka, W. Dabney, S. Mohamed, B. Lakshminarayanan, S. Hoyer, and R. Munos. The Cramer Distance as a Solution to Biased Wasserstein Gradients. *arXiv e-prints*, art. arXiv:1705.10743, May 2017.

P. Bertail, S. J. Clémençcon, and N. Vayatis. On bootstrapping the roc curve. In *Advances in Neural Information Processing Systems*, pages 137–144, 2009.

D. Derkach, N. Kazeev, F. Ratnikov, A. Ustyuzhanin, and A. Volokhova. Cherenkov Detectors Fast Simulation Using Neural Networks. In *10th International Workshop on Ring Imaging Cherenkov Detectors (RICH 2018) Moscow, Russia, July 29-August 4, 2018*, 2019. doi: 10.1016/j.nima.2019.01.031.

S. Easo, I. Belyaev, G. Corti, C. Jones, A. Papanestis, W. Pokorski, F. Ranjard, and P. Robbe. Simulation of lhcb rich detectors using geant4. *IEEE Transactions on Nuclear Science*, 52(5):1665–1668, Oct 2005. doi: 10.1109/TNS.2005.856766.

R. W. Forty and O. Schneider. RICH pattern recognition. Technical Report LHCb-98-040, CERN, Geneva, Apr 1998. URL `http://cds.cern.ch/record/684714`.

A. Giammanco. The fast simulation of the CMS experiment. *Journal of Physics: Conference Series*, 513(2):022012, jun 2014. doi: 10.1088/1742-6596/513/2/022012. URL `https://doi.org/10.1088%2F1742-6596%2F513%2F2%2F022012`.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative Adversarial Networks. *arXiv e-prints*, art. arXiv:1406.2661, Jun 2014.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5767–5777. Curran Associates, Inc., 2017. URL `http://papers.nips.cc/paper/7159-improved-training-of-wasserstein-gans.pdf`.

O. Lupton, L. Anderlini, B. Sciascia, and V. Gligorov. Calibration samples for particle identification at LHCb in Run 2. Technical Report LHCb-PUB-2016-005. CERN-LHCb-PUB-2016-005, CERN, Geneva, Mar 2016. URL `https://cds.cern.ch/record/2134057`.

M. Paganini, L. de Oliveira, and B. Nachman. Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multilayer Calorimeters. *Phys. Rev. Lett.*, 120(4):042003, 2018. doi: 10.1103/PhysRevLett.120.042003.

M. Pivk and F. R. Le Diberder. SPlot: A Statistical tool to unfold data distributions. *Nucl. Instrum. Meth.*, A555:356–369, 2005. doi: 10.1016/j.nima.2005.08.106.