Toward Machine Learning Optimization of Experimental Design

MODE Collaboration:

Atılım Güneş Baydin[1], Kyle D. Cranmer[2], Pablo de Castro Manzano[3], Christophe Delaere[4], Denis Derkach[5], Julien Donini[6], Tommaso Dorigo[7], Andrea Giammanco[4], Jan Kieseler[8], Lukas Layer[7,9], Gilles Louppe[10], Fedor Ratnikov[5], Giles C. Strong[7,11], Mia Tosi[7,11], Andrey Ustyuzhanin[5,12], Pietro Vischia[4], and Hevjin Yarar[7,11]

[1]HSE University, [1]University of Oxford, [2]New York University, [3]Treelogic, [4]Université catholique de Louvain, [5]HSE University, [6]Université Clermont Auvergne, LPC, CNRS/IN2P3, [7]INFN, Sezione di Padova, [8]CERN, [9]Università di Napoli "Federico II", [10]University of Liège, [11]University of Padova, [12]National University of Science and Technology MISIS

February 5, 2021

Abstract

The effective design of instruments that rely on the interaction of radiation with matter for their operation is a complex task. A full optimization of the many parameters involved may still be sought by leveraging recent progress in computer science. Key to such a goal is the definition of a utility function that models the true goals of the instrument. Such a function must account for the interplay between physical processes that are intrinsically stochastic in nature and the vast space of possible choices for the physical characteristics of the instrument. The construction of a differentiable model of all the ingredients of the information-extraction procedures, including data collection, detector response, pattern recognition, and all existing constraints, then allows the automatic exploration of the vast space of design choices and the search for their best combination.

In this document we succinctly describe the research program of the MODE Collaboration (an acronym for Machine-learning Optimized Design of Experiments), which aims at developing tools based on deep learning techniques to achieve end-to-end optimization of the design of instruments via a fully differentiable pipeline capable of exploring the Pareto-optimal frontier of the utility function. The goal of MODE is to demonstrate those techniques on small-scale applications such as muon tomography or hadron therapy, to then gradually adapt them to the more ambitious task of exploring innovative solutions to the design of detectors for future particle collider experiments.

1 Introduction

The design of instruments that rely on the interaction of radiation with matter for their operation is a quite complex task if our goal is to achieve near-optimality on some well-defined utility function $\mathcal{U}$, such as the expected precision of a set of planned measurements achievable with a given amount of collected data. This complexity stems from the interplay between physical processes that are intrinsically stochastic in nature—the quantum phenomena that take place at the subnuclear level—and the vast space of possible choices for the physical characteristics of the instrument and its detection elements, as defined in its design space. The precision of pattern recognition of detected signals and the power of information-extraction procedures that directly affect the value of $\mathcal{U}$ both depend on these characteristics. In the majority of realistic cases, $\mathcal{U}$ may be represented as a combination of performance and cost considerations that should be balanced within reasonable limitations.

Neural networks are naturally suitable for the task mentioned above [1] They can also be effectively used as surrogates for simulators, to enable gradient-based optimization in cases where a simulator is non-differentiable. In addition, automatic differentiation (AD) techniques developed in the 1980s [2] and now commonly available in the most popular machine learning (ML) frameworks [3, 4] make it possible to rely on efficient implementations of the back-propagation algorithm. The MODE Collaboration [5] (an acronym for Machine-learning Optimized Design of Experiments) aims at developing tools based on deep neural networks and modern AD techniques to implement a full modelling of all the elements of experimental design, achieving end-to-end optimization of the design of instruments via a fully differentiable pipeline capable of exploring the Pareto-optimal frontier of $\mathcal{U}$. Exploratory studies have shown that very large gains in performance are potentially achievable even for very simple apparata [6, 7]. MODE has the goal of showing how these techniques may be adapted to the complexity of modern and future particle detectors and experiments, while remaining adaptable to a number of applications outside of that domain. Below we succinctly describe the research program of the MODE Collaboration.

2 The MODE program

2.1 Architecture development

A generic optimization pipeline for a complex system can be constructed by assembling modules that take on different modeling tasks. The modules interact by receiving input data and processing them to provide an output that satisfies specified external constraints dependent on the value of the parameters under study; the output of each module is fed to the next one, until an objective function can be computed. The computation of each module is differentiable, so that the composition of such modules is also differentiable through the chain rule of differentiable

calculus, enabling the gradient of $\mathcal{U}$ to be computed and used in the search for extrema of the objective function [8]; the search may be performed in steps, by freezing some modules while updating others, to simplify the parameter space scan.

As a specific example we may consider the optimization of the layout of a muon radiography [9] apparatus for material identification within a volume of interest (one of the use cases described below). A random generation of cosmic rays, in the form of incoming particle four-vectors, is fed to a fast simulation of detection apparatus and scanned volume. The simulation of multiple scattering, particle propagation, and resulting electronic signals in the detector may be directly produced by a differentiable program. Alternatively, the simulation output may inform a differentiable module based on deep generative models such as variational autoencoders (VAE) [10], generative adversarial networks (GAN) [11], or flow models [12], or through the use of local generative surrogates of the gradients [7]; a generation/validation loop must be available to keep the model appropriate as the layout parameters are modified during the optimization task. The output of the particle detection module is fed to a reconstruction module, which produces incoming and outgoing track measurements through a fit to the detected signals; these are again a function of the detector parameters. Downstream, an information-extraction module accumulates information on the density of material in the container. Its output may be used to compute a loss function that describes as closely as possible the real goal of the system. In a simplified setup this function could be defined as the type-II error rate on the detection of a given amount of a particular material within the volume of interest, as modeled by the simulation. A sketch of the described pipeline is shown in Figure 1.

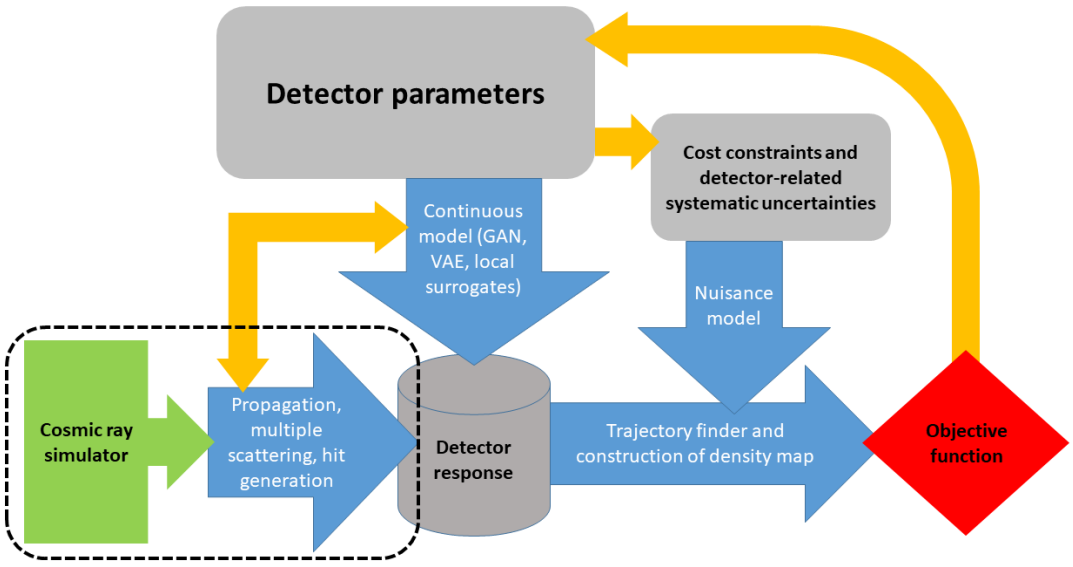

Figure 1: *Conceptual layout of an optimization pipeline for a muon radiography apparatus. Modules within the dashed black box inform the validation of a continuous model and are not part of the optimization flow.*

2.2 Use cases

Given its considerable complexity, the development of a pipeline for the optimization of experiment designs should start with the study of simpler use cases, and proceed incrementally by adding complexity. Below we succinctly exemplify a set of use cases that might be considered in series in the development of our research plan.

2.2.1 The MUonE detector

MUonE is a detector proposed to precisely measure the q^2-differential cross section of elastic muon-electron scattering, to reduce dominant systematic uncertainties in the g-2 experiment [13]. Given the simplicity of investigated physics process and baseline detector layout, MUonE was taken as an example for geometry optimization studies that did not employ AD techniques [6]. A reanalysis within a full feedback loop which considers all geometry parameters together with reconstruction-driven systematic uncertainties, cost, and a more precise definition of the utility function is relatively straightforward to produce, and may thus constitute a valid initial benchmark for comparison of automatic optimization searches and discrete scans.

2.2.2 Muon Radiography

The abundant natural flux of atmospheric muons and their large penetration power have been exploited for the imaging of a large variety of objects spanning in size from $\mathcal{O}(1\text{ m})$ to $\mathcal{O}(1\text{ km})$, with applications including archaeology, volcanology, border control, nuclear safety, and industrial process control [9]. In some applications, the volume of interest can be sandwiched between two trackers and one can measure the scattering of the muons through the target volume, which is correlated with the atomic number Z of the material. When the target volume is very large (e.g., a mountain or an entire building), a single tracker is placed downstream to measure the absorption of the muon flux through the target, from which a density map can be derived. By optimizing the layout of the detectors, large gains in the resolution and material identification potential of a muon tracking system are achievable. A recent project [14] aims at the development of compact, autonomous, portable, and modular muon radiography setups based on small-area resistive plate chambers (RPC), a technology chosen because of its good trade-off between cost, ease of construction, and position and time resolution. The goal is to allow a high degree of modularity for the geometry of the complete setup: ideally, the already mounted individual RPC planes would be produced in large numbers and deployed in situ in the arrangement that best fits the specific use case while respecting the local constraints (e.g., the optimal location may be in a narrow tunnel). The same RPC layers may be arranged to form one or two trackers depending on the relative importance of absorption and scattering on the final discrimination power; for a single tracker sometimes it is not obvious a priori whether it is more convenient to have a few layers with large areas to collect more data, or to maximize the num-

ber of layers crossed by the muons to improve tracking resolution. An automatic optimization algorithm would be able to provide a quick redesign of the geometry for new measurements of different targets.

2.2.3 Proton Therapy

Effective irradiation of non-operable tumors with intense proton or light-hadron beams could be achieved if rapid imaging techniques are used to create 3D maps of the target and surrounding tissue. The imaging resolution depends on the possibility of acquiring sufficient data within seconds, avoiding target movements. A fast calorimeter has been developed by the iMPACT collaboration [15] for this effort. The optimization of the layout of the detection elements, and the optimal addition of a magnetic field to the setup, are important aspects well suitable to an investigation with AD means. We plan to collaborate with iMPACT to investigate the space of detector solutions, with the goal of maximizing the benefit of the imaging tool produced.

2.2.4 A hybrid calorimeter for a future collider

So far, the guiding principle when building high-energy physics detectors has been strongly governed by the idealized requirements of classic reconstruction algorithms. As a consequence, general-purpose detectors follow the principle of tracking charged particle trajectories within a magnetic field in a low-material-budget tracker, where nuclear interactions and multiple scatterings are kept to a minimum to provide good conditions for the track helix fit; only in a second step both neutral and charged particles are brought to a stop to measure their energy in a dense calorimeter. With recent advances in machine-learning-based particle reconstruction from raw detector signals [16], it is possible to break this paradigm and optimally combine position and energy measurements. We foresee the exploitation of these advancements with the study of feasibility and design of an optimized hybrid calorimeter, with material density increasing with distance from the interaction point. Such a device would ideally allow one to exploit the distinct nuclear interactions of different particle species with the material together with the probabilistic information from a detailed tracking of the evolving shower through the detector. However, this approach would require an optimization of the hybrid reconstruction before its inclusion into the global optimization pipeline [17]. Creating a precise, fully differentiable model for the optimization of such a system is a terrific challenge, with possible enormous gains.

2.3 Computing requirements and infrastructure

The basic pipeline should be generic and customizable for different detector optimization problems, and have a well-defined structure of encapsulated functional blocks. This enables running blocks separately for the initial decomposition of the

full problem, as well as validation of individual blocks. The pipeline optimization loop should use containerization technologies and be executable on common computing infrastructure. Various blocks use gradient optimization underneath, thus access to accelerated tensor computing hardware is essential. The infrastructure should provide an interface for the communication of input/output values for every block. The optimization target should be flexible to support a variety of constraint metrics, such as physics performance, detector performance, and cost; in cases when multiple criteria are specified, a Pareto-optimal selection of possible configurations should be returned. Besides optimized detector parameters, the pipeline should produce trained surrogate models and reconstruction algorithms suitable for interactive analysis of possible trade-offs between alternative options. The infrastructure should also allow users to substitute any ML-powered block with a reference baseline implementation of the same functionality. This provides a direct way for collecting reference data to train the surrogate models ML implementations, as well as for validating and evaluating the corresponding blocks. The framework should also support tuning of ML models using a combination of real and simulated data.

3 Concluding remarks

Recent advances in computer science make it possible to give a truer meaning to the word "optimization" when discussing the design of instruments that operate via the interaction of radiation with matter. The MODE collaboration aims at developing a versatile, modular, and scalable software architecture, which can be customized to different optimization problems and provide a full exploration of the space of their design choices and information extraction procedures. We believe that such a tool may offer enormous potential gains to a wide range of research and industry applications.

Acknowledgements

A. Giammanco's work was partially supported by the EU Horizon 2020 Research and Innovation Programme under the Marie Sklodowska-Curie Grant Agreement No. 822185, and by the Fonds de la Recherche Scientifique (FNRS) under Grant No. T.0099.19. P. Vischia's work was supported by the FNRS under the Grant No. 40000963. H. Yarar and L. Layer are supported by the European Union's Horizon 2020 research and innovation programme under Grant Agreement No.765710. D. Derkach, F. Ratnikov and A. Ustyuzhanin are supported by Russian Science Foundation under Grant Agreement No. 19-71-30020. A.G. Baydin is supported by EPSRC/MURI grant EP/N019474/1 and by Lawrence Berkeley National Lab. T. Dorigo would like to thank Piero Giubilato for insight in proton therapy applications.

References

[1] Giuseppe Carleo et al. "Machine learning and the physical sciences". In: *Rev. Mod. Phys.* 91 (4 Dec. 2019), p. 045002. DOI: 10.1103/RevModPhys.91.045002. URL: https://link.aps.org/doi/10.1103/RevModPhys.91.045002.

[2] Atilim Gunes Baydin et al. "Automatic Differentiation in Machine Learning: a Survey". In: *Journal of Machine Learning Research* 18.153 (2018), pp. 1–43. URL: http://jmlr.org/papers/v18/17-468.html.

[3] Martín Abadi *et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[4] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[5] The MODE Collaboration. *MODE: Machine-learning Optimized Design of Experiments*. Website of the collaboration. 2020. URL: https://mode-collaboration.github.io/.

[6] Tommaso Dorigo. "Geometry optimization of a muon-electron scattering detector". In: *Physics Open* 4 (2020), p. 100022. ISSN: 2666-0326. DOI: https://doi.org/10.1016/j.physo.2020.100022. URL: http://www.sciencedirect.com/science/article/pii/S2666032620300090.

[7] Sergey Shirobokov et al. "Black-Box Optimization with Local Generative Surrogates". In: (June 2020). arXiv: 2002.04632 [cs.LG].

[8] Mitar Milutinovic, Atılım Güneş Baydin, Robert Zinkov, William Harvey, Dawn Song, Frank Wood, Wade Shen. "End-to-End Training of Differentiable Pipelines Across Machine Learning Frameworks". In: *NeurIPS* (2017). URL: https://openreview.net/forum?id=ryh7qqGRZ.

[9] L. Bonechi, R. D'Alessandro, and A. Giammanco. "Atmospheric muons as an imaging tool". In: *Rev. Phys.* 5 (2020), p. 100038. DOI: 10.1016/j.revip.2020.100038. arXiv: 1906.03934 [physics.ins-det].

[10] Diederik P Kingma and Max Welling. *Auto-Encoding Variational Bayes*. 2013. arXiv: 1312.6114 [stat.ML].

[11] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML].

[12] George Papamakarios et al. *Normalizing Flows for Probabilistic Modeling and Inference*. 2019. arXiv: 1912.02762 [stat.ML].

[13] G Abbiendi. *Letter of Intent: the MUonE project*. Tech. rep. CERN-SPSC-2019-026. SPSC-I-252. The collaboration has not yet a structure, therefore the names above are for the moment an indication of contacts. Geneva: CERN, June 2019. URL: https://cds.cern.ch/record/2677471.

[14] S. Basnet et al. "Towards portable muography with small-area, gas-tight glass Resistive Plate Chambers". In: *JINST* 15.10 (2020), p. C10032. DOI: 10.1088/1748-0221/15/10/C10032. arXiv: 2005.09589 [physics.ins-det].

[15] S. Mattiazzo et al. "The iMPACT project tracker and calorimeter". In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 845 (2017). Proceedings of the Vienna Conference on Instrumentation 2016, pp. 664–667. ISSN: 0168-9002. DOI: https://doi.org/10.1016/j.nima.2016.04.105. URL: http://www.sciencedirect.com/science/article/pii/S0168900216303412.

[16] S. Farrel et al. "The HEP.TrkX Project: deep neural networks for HL-LHC online and offline tracking". In: *EPJ Web Conf.* 150 (2017), p. 00003. DOI: 10.1051/epjconf/201715000003. URL: https://doi.org/10.1051/epjconf/201715000003.

[17] Fedor Ratnikov et al. "Using machine learning to speed up new and upgrade detector studies: a calorimeter case". In: *24th International Conference on Computing in High Energy and Nuclear Physics*. Mar. 2020. arXiv: 2003.05118 [physics.ins-det].