ASCR Workshop Position Paper:
Challenges and Opportunities in High Energy Physics

Nick Smith, Oz Amram, Prasanth Shyamsundar, Stephen Mrenna, Aleksandra Ciprijanovic
Fermi National Accelerator Laboratory, Batavia, IL
{ncsmith,oamram,prasanth,mrenna,aleksand}@fnal.gov

FERMILAB-FN-1271-CSAID-PPD-SQMS

Topics: Inverse problems using incomplete, noisy, or multi-modal data; Uncertainty-aware hybrid modeling for solving inverse problems; Scalable algorithms for inverse problems

Challenge

High energy particle physics and cosmology concern themselves with estimating fundamental parameters of nature, such as the masses and interactions of fundamental particles like the Higgs boson and the rate of expansion of the universe. In doing so, they analyze exabyte-scale datasets, some of the largest in all of science, and face many challenges in subsequent data analysis. These challenges are shared between the two disciplines, but we focus on particle physics to highlight one specific domain. In particle physics, the standard method for estimating parameters involves performing Monte Carlo (MC) integration as a function of both parameters of interest and nuisance parameters using an expensive simulator, counting the number of observed collision events (i.i.d. samples) from an experiment in the corresponding integration domains, and forming a Poisson likelihood function. This likelihood function is then used in a Frequentist manner to construct a maximum likelihood point estimate (MLE) and confidence set for the parameters. To sufficiently populate the high-dimensional integration domains, simulators consume billions of CPU-hours annually and produce hundreds of petabytes of intermediate output data. Several techniques have been developed to: optimize definitions of the integration domains so as to be maximally sensitive to a particular subset of parameters, efficiently estimate the integrals, and build robust surrogate models by interpolating between integral evaluations at different parameter points. One can view this whole endeavor as *classical* Simulation-Based Inference (SBI).

This approach has several limitations. Given the continual increase in collected experimental data, the current usage of simulators as well as inference methodologies face scaling challenges. The larger dataset size allows for improved parameter estimation only if accompanied by a corresponding decrease in the statistical uncertainties of the MC integration. The curse of dimensionality and usage of simulators requires inference to be performed only in a low-dimensional subspace of the true data dimension, reducing the amount of extracted information. Additionally, the simulators are known to be imperfect models of the true data. The resulting *domain shift*, between the simulated and real data distributions, means all inference tasks must take into account uncertainties that are often difficult to quantify. A more precise description of the simulator uncertainty will require improved methods of quantifying and parameterizing its deficiencies.

This manuscript has been authored by FermiForward Discovery Group, LLC under Contract No. 89243024CSC000002 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

Opportunity

Recent developments suggest that Machine Learning (ML) methods may allow for more efficient and more sensitive construction of the likelihood function. ML SBI techniques can exploit the higher-dimensional information contained in the simulator much more efficiently than traditional methods. The predominant Frequentist interpretation of the likelihood allows for a convenient simplification: these methods regress an approximation of the integrand but need not estimate the associated Jacobian, as it does not affect the determination of the MLE or the likelihood ratio used to construct confidence sets. As the observation space is i.i.d. samples, one only needs to estimate the likelihood function for single samples. Furthermore, if the simulator is differentiable, local information about the shape of the likelihood can be extracted to improve the convergence of the approximation. Established techniques for interpolation in classical SBI may be incorporated as inductive bias in the ML approximation as well. Generative models trained on simulations and/or auxiliary data can be used as computationally efficient replacements for the simulator and mitigate the impact of domain shift.

Innovation

As the total number of collision events numbers in the billions, the fidelity requirements of a regressed likelihood function are stringent. Therefore, the computational cost of training the likelihood regression is likely to exceed the simulation costs of classical SBI. New techniques are necessary to keep this cost under control. Generative AI has great potential to meet these scaling challenges, properly account for domain shift, and provide improved parameter estimation. However, the usage of ML models, for either likelihood construction or for generative alternatives to the classical simulator, requires appropriate uncertainty quantification on their outputs. Best practices need to be established to minimize the impact of the simulation-reality domain shift and corresponding uncertainties. This challenge is particularly striking for generative models, which produce output in high-dimensional space, making traditional methods for quantifying these deficiencies infeasible. Inclusion of additional domain adaptation methods, which force the ML-SBI to learn only dataset-invariant features, would enhance out-of-distribution robustness and enhance our ability to perform accurate inference on real data.

The deployment of these new techniques would revolutionize data analysis in particle physics, vastly improving the precision on fundamental physics parameters extracted from subatomic particle collisions. This improved precision could allow the observation of new physical phenomena at subatomic scales, revolutionizing our understanding of nature's most fundamental constituents.