Snowmass2021 - Letter of Interest

Scientific AI Approaches in Computational Cosmology

Thematic Areas: (check all that apply □/■)
□ (CompF1) Experimental Algorithm Parallelization
■ (CompF2) Theoretical Calculations and Simulation
■ (CompF3) Machine Learning
□ (CompF4) Storage and processing resource access (Facility and Infrastructure R&D)
□ (CompF5) End user analysis
□ (CompF6) Quantum computing
□ (CompF7) Reinterpretation and long-term preservation of data and code

Contact Information:
Salman Habib (Argonne National Laboratory) habib@anl.gov
Nesar Ramachandra (Argonne National Laboratory) nramachandra@anl.gov

Authors:
Salman Habib (Argonne National Laboratory), Nesar Ramachandra (Argonne National Laboratory), Xiaofeng Dong (University of Chicago), Sandeep Madireddy (Argonne National Laboratory), Katrin Heitmann (Argonne National Laboratory), Jonas Chaves-Montero (Argonne National Laboratory)

Abstract: The use of Machine Learning algorithms has been successfully demonstrated in numerous scientific applications – including particle physics and cosmology. Many of these applications conform to a broad theme of building parametric and nonparametric statistical surrogates for classification, regression or surrogate modeling. We identify key avenues for developments in the field of Artificial Intelligence, that not only comply with current uncertainty quantification requirements, but also provide opportunities for future explorations of physical laws in a scalable manner.

1 Introduction

Cosmic Frontier experiments provide a rich area for AI applications for several reasons. Cosmology is based on large observational datasets rather than on isolated experiments. The observational nature of the field makes it oftentimes impossible to extract the full information content from the data without the use of optimized learning algorithms. There are many examples of AI-based analysis approaches that have have been developed actively and have had initial impacts in cosmology already, including algorithms to disentangle images of galaxies (deblending), approaches for the analysis of photometric data for redshift estimation, and feature extraction to identify, e.g., strong lenses.

The AI methodologies employed cover as broad a range as the problems to be solved; they include deep learning and active learning methods, random forest classifications and also more traditional machine learning approaches such as Gaussian process modeling. An important feature is the close connection with statistics, in particular, sampling theory and Bayesian methods. There is a significant focus on topics such as detailed verification and validation, typically not considered in non-scientific applications (e.g., market predictions) where one does not demand rigorously controlled results.

2 Challenges and Avenues for Development

2.1 Effective physical model building

Generative models for cosmological functions (typically summary statistics)[1;2] employ data-driven emulators as a replacement for expensive numerical simulations. In addition to training data, one may also employ partial or complete information of physical conservation laws and other underlying symmetries in optimizing the fitting to the statistical models. Such *regularized* data-driven models that penalize divergence from first principles as well as empirical disagreement from training data could generalize better, with far fewer training data sets than purely data-driven models that are currently employed for classification and regression tasks in the physical sciences.

Another approach to physical modeling is the development of physics abstractions for the equations with appropriate parameterizations for modeling the decision space. Techniques of Reinforcement Learning may be employed to refine broad statistical models into a physically meaningful one via incrementally constraining physics-informed reward mechanisms.

2.2 Interpretability and explainability of AI systems

Due to the presence of a large number of trainable parameters, deep learning approaches like Convolutional Neural Networks lack robust frameworks for interpretation. Often, visualizations of the internally-generated convolution filters[3] or the latent space distribution[4] provide hints of relevant features of the data. However, many scientific problems require interpretations in terms of domain-specific summary statistics or semantic information (the analog of "scene understanding" in computer vision), rather than the raw features in the data (like individual pixels in astronomical images). Identification of feature importance measures[5–7] and providing intuitive explanations would be key in improving the explainability of machine learning models in the context of scientific research.

2.3 Uncertainty Quantification in Deep Learning

Current state-of-the-art deep learning algorithms originating from non-scientific domains are primarily deterministic, hence are poor estimators of predictive uncertainty, relative to principled, probabilistic approaches such as Gaussian Processes. In combination with a tendency to over-fit the data, the absence of robust uncertainties results in a difficult adaptation to scientific problems with sub-percent accuracy requirements, such as inference of cosmological parameters.

Due to a large number of network parameters and enormous amounts of training data, fully Bayesian approaches (that use Monte Carlo methods) are often substituted by relatively inexpensive loss-optimization training schemes. However proxies for the posterior distribution function may be employed depending on the specific problem of interest, such as Gaussian mixture modeling for galaxy photometric redshift estimation. In the recent literature, advances in variational approximation methods[8] have offered a middle ground for scalable probabilistic modeling. These approaches have enabled handling flexible models (with millions of parameters) efficiently on modest hardware and obtain state-of-the-art predictive accuracy. In spite of this promise, the uncertainty quantification using these approaches is limited compared to the fully Bayesian approaches due to the heuristic nature of the approximate posteriors[9]. This is an active research area with research largely focused around improving prior assumptions, and variational distribution families that can handle multi-modality and thus provide tighter uncertainty bounds.

In the context of cosmological parameter estimation, where systematic effects may continue to be the dominant source of uncertainty, applications of Probabilistic Neural Networks still remains nontrivial. Hence, uncertainty quantification appears to be a crucial missing step in scientific machine learning.

2.4 Scalability

Leveraging the next generation of High Performance Computing (HPC) resources and specialized AI-engines is another avenue of performing at-scale AI-based analyses of future data sets. Most of the current scaling efforts are focused on: 1) Data-parallelism techniques to train using large data-sets, 2) Model-parallelism for dividing the training of complex networks across multiple computing nodes, 3) Scalable hyperparameter optimization and neural architecture searches, such as[10], to avoid ad-hoc model choices.

In addition to the above, cosmological analyses would also benefit from *in-situ* training and deployment of AI systems concurrently with numerical simulations. For instance, resource-intensive sub-grid modeling could be replaced by AI-surrogates in numerical simulations in order to provide feedback during the execution. Finally, posterior approximation methods like Hamiltonian Monte Carlo (HMC) allow for efficient explorations of target distributions. In conjunction with automatic differentiable frameworks such as `JAX`[11], and embedding deep neural networks to improve expressiveness in high dimensions[12], HMC can efficiently exploit future HPC systems for estimating posterior distributions of thousands of parameters.

3 Summary

While a significant amount of applications on astrophysical data have hinted at the remarkable versatility of AI algorithms, they have also pointed to the vast potential in the era of exascale computation and data-driven modeling. We list some of the key topics of interest at the intersection of computational cosmology, Bayesian inference and scientific machine learning. While this list is by no means exhaustive, we have identified a few areas that will be instrumental in future integration of AI within rigorous cosmological studies.

References

[1] https://www.hep.anl.gov/cosmology/CosmicEmu/emu.html

[2] K. Heitmann, D. Higdon, C. Nakhleh, and S. Habib, Astrophys. J. 646, L1, 2006

[3] Ribli, D., Pataki, B. Á., & Csabai, I. 2019, Nature Astronomy, 3, 93

[4] Madireddy, S., Li, N., Ramachandra, N., et al. 2019, arXiv:1911.03867

[5] Singh, C., Ha, W., Lanusse, F., et al. 2020, arXiv:2003.01926

[6] Lundberg, S. & Lee, S.-I. 2017, arXiv:1705.07874

[7] Tulio Ribeiro, M., Singh, S., & Guestrin, C. 2016, arXiv:1602.04938

[8] Zhang, C., Butepage, Judith., Kjellstrom, Hedvig., & Mandt, Stephan. 2017, arXiv:1711.05597

[9] Charnock, T., Perreault-Levasseur, L., & Lanusse, F. 2020, arXiv:2006.01490

[10] Balaprakash, P., Egele, R., Salim, M., et al. 2019,. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC19).

[11] https://github.com/google/jax

[12] Levy, D., Hoffman, M. D., & Sohl-Dickstein, J. 2017, arXiv:1711.09268.