Snowmass2021 - Letter of Interest

Building Emulators for the Cosmic Frontier

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
Katrin Heitmann (Argonne National Laboratory) heitmann@anl.gov

Authors:
Salman Habib (Argonne National Laboratory), Katrin Heitmann (Argonne National Laboratory), Juliana Kwan (Liverpool John Moores University), Dave Higdon (Virginia Tech), Earl Lawrence (Los Alamos National Laboratory), Zarija Lukić (Lawrence Berkeley National Laboratory), Patricia Larsen (Argonne National Laboratory), Nesar Ramachandra (Argonne National Laboratory)

Abstract: Cosmological surveys continue to explore the nonlinear regime of structure formation. To analyse the rich data these surveys collect, accurate predictions are essential and can often only be obtained from very expensive simulations. It is computationally infeasible to cover the cosmological model space of interest at the required resolution. Therefore, high-precision predictions must be derived from a smaller set of simulations. This requirement has motivated the introduction of emulators in cosmology, fast prediction schemes built from a finite set of simulations spanning the model space of interest. The emulator approach has been proven to deliver very accurate results for a diverse set of measurements. Significant future advances are necessary to fully enable the analysis of cosmological data with emulators. First, the measurements from the simulations have to more closely track observables, as they are actually measured. Second, the underlying simulations must include more physics over time, e.g., going from gravity-only to hydrodynamical simulations. Third, it is essential to develop robust error estimates for emulator results that can be folded into multi-step data analysis.

1 Introduction

Cosmology is inherently an observational science, grounded in data from sky surveys conducted over a wide range of wave-bands. Modern surveys cover large areas of the sky and go to considerable depths, aiming to reach into the early stages of the formation of the Universe. Cosmological analyses of survey observations are essentially statistical in nature and are limited by a number of factors; these include sample size and completeness, observational systematics of various kinds, and the limitations of the theoretical and empirical models underlying the analysis of the data sets. Historically, cosmology was limited by observational reach and statistical errors were a significant concern. Over the last few decades, great strides have been made in instrumentation and, with increased areal coverage and survey depth, this is no longer the case in a number of important situations, e.g., photometric and redshift surveys to medium depths ($z \sim 1$).

The current cosmological Standard Model – ΛCDM – is an excellent fit to the data but has several theoretical shortcomings and is generally perceived, very like the particle physics Standard Model, to possess only a transitory existence. But because ΛCDM is so successful, deviations from it will be subtle and difficult to nail down. Consequently, the next generation of cosmology experiments will be driven not only by the accumulation of statistics but also by the need to understand, mitigate, and control systematic uncertainties. To make substantial headway in the latter task, the ability to create detailed and realistic "virtual universes" on demand is gaining central importance, so much so, that the ultimate scientific success of upcoming sky surveys hinges critically on the success of the modeling and simulation effort.

Scientific inference with sky surveys is a statistical inverse problem, where, given a set of measurement results, one attempts to fit a class of physical models to the data (which include models for the observational process), and to infer the values of the model parameters. Typically, such analyses require many evaluations over a very large number of "virtual universes". The main difficulty lies in the fact that producing each virtual universe requires, in principle, an extremely expensive numerical simulation carried out at high fidelity. Emulators are effectively fast surrogate models that can be used an an alternative route to solving this inverse problem.[1] The objective of this LOI is to outline challenges and opportunities in producing the next generation of emulators, tailoring them to best solve the problems posed by future cosmological surveys.

2 New Challenges and Opportunities

2.1 Emulating the Observable Universe

The importance of providing predictions for cosmological surveys via emulators is now widely recognized; emulation-based predictions have become very popular over the last few years[2–14]. It is now possible to carry out high-quality simulation suites that provide the input to the emulators for a range of cosmological statistics, such as halo mass functions, matter power spectra, and halo bias. (Some of the emulators have gone beyond the Standard Model of Cosmology as well.) However, in order to fully integrate the emulators into the analysis frameworks used by the surveys to extract cosmological parameters, it is very desirable to create emulators connected directly to the survey observables. As a concrete example, the cluster mass function is commonly used to derive cosmological constraints, but it is not a quantity that can be easily extracted from the observations. All measurements, including weak lensing shear do not directly provide mass measurements, but rather approximations or proxies for an idealized "cluster mass". The translation from the observable to the mass function from simulations adds additional uncertainties into derived cosmological parameters and could be avoided if emulators would directly predict the quantity of interest, which is the cluster abundance, measured in a way that is most relevant to how the survey is actually carried out. This level of forward modeling would involve new simulation and analysis efforts and the development of more flexible and sophisticated emulation approaches. Given that a successful implementation can potentially eliminate a major source of uncertainty and bias, this is clearly a worthwhile step. With a broad enough

simulation footprint, such an approach would also enable easier connections across measurements carried out in different wavebands.

2.2 Extending the Physics Content of Simulations

Most emulator efforts so far have relied upon gravity-only simulations. These are an order of magnitude less expensive than hydrodynamic simulations, and yet carrying out a high-quality suite of gravity-only simulations has only become possible in the last few years due to increases in available computing resources. For hydrodynamics simulations such campaigns are still out of reach because of the much larger time to solution per simulation. Additionally, hydrodynamics simulations have many modeling (nuisance) parameters and uncertainties, increasing the design space for the emulation and in turn increasing the number of simulations to be carried out. (For gravity-only simulations, after having established criteria for precision simulations, only the fundamental cosmological parameters have to be varied, keeping the simulation campaign size manageable.) With the advent of exascale supercomputing resources (and beyond) in the coming years, and employing strategies such as multi-fidelity simulations, this problem can be significantly reduced, assuming the hydrodynamics codes can take full advantage of the new generation of architectures. If this turns out to be the case, many opportunities open up: Emulators can be built to investigate and optimize subgrid model parameters, be deployed to gain a better understanding of the interplay of subgrid model and cosmology parameters, and to directly predict observational quantities for different surveys accessing different wavelength regimes.

2.3 Robust Error Estimation and Parameter Exploration

Error estimation with emulators is a potentially very powerful avenue of research but remains to be properly realized in many of the current generation of emulators. Partly this is because error estimation is inherently difficult and partly because the methods used have been too informal, and insufficiently sharp. Because the formal statistical uncertainties in the observations are reducing with time, the onus is on modeling systematic errors, including the errors in the emulators. This area is relatively little-studied in the statistics literature although there are some useful investigations in discrepancy modeling[15]; the power of the results obtained, however, is relatively limited.

Another problem is that the dynamic range in cosmology is vast and it is simply impossible to model all the relevant processes via a first principles approach. Consequently some of the inputs in the subgrid models must be empirical, based on known results from observations. This adds another layer of complexity to error estimation because the proper treatment of such evidence in a cosmological analysis potentially requires a separate set of investigations for each empirical input. However, we note that continuous inclusion of observational data in emulator construction will be helpful in reducing the volume of parameter space that needs to be explored. (As mentioned previously, multi-fidelity simulations are also useful here in minimizing the amount of computational work.) Adaptive sampling methods are very useful in time-domain applications and they can be easily transplanted to cosmology, provided error analyses can continue to be undertaken in a robust manner.

3 Summary

Emulators have been proven to be very useful in delivering accurate predictions for a range of cosmological statistics and have already been used by surveys for their cosmological analyses. Their continuous development into the future is an important topic that needs to be taken up in the Snowmass process. In this LOI we have outlined three promising venues for extending this line of research. These should only be viewed as a subset of several possible interesting directions. Given that we will soon enter the exascale era in high-performance computing many more opportunities will exist in the coming decade to further develop these

approaches and to fully integrate them into the analysis frameworks of the major surveys.

References

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

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

[3] E. Lawrence, E., K. Heitmann, M. White, et al., Astrophys. J.713, 1322 (2010)

[4] K. Heitmann, E. Lawrence, J. Kwan, S. Habib, and D. Higdon, D., Astrophys. J. 780, 111 (2014)

[5] E. Lawrence, K. Heitmann, J. Kwan, J., et al., Astrophys. 847, 50 (2017)

[6] S. Bocquet, K. Heitmann, K., S. Habib, E. Lawrence, T. Uram, N. Frontiere, A. Pope and H. Finkel, arXiv:2003.12116 [astro-ph.CO], accepted for publication in ApJ (2020)

[7] T. McClintock, E. Rozo, A. Banerjee, M. R. Becker, J. DeRose, S. McLaughlin, J. L. Tinker, R. H. Wechsler and Z. Zhai, [arXiv:1907.13167 [astro-ph.CO]].

[8] Z. Zhai, J. L. Tinker, M. R. Becker, J. DeRose, Y. Y. Mao, T. McClintock, S. McLaughlin, E. Rozo and R. H. Wechsler, Astrophys. J. 874, 95 (2019)

[9] T. McClintock, E. Rozo, M. R. Becker, J. DeRose, Y. Y. Mao, S. McLaughlin, J. L. Tinker, R. H. Wechsler and Z. Zhai, Astrophys. J. **872**, no.1, 53 (2019)

[10] T. Nishimichi, M. Takada, R. Takahashi, K. Osato, M. Shirasaki, T. Oogi, H. Miyatake, M. Oguri, R. Murata, Y. Kobayashi and N. Yoshida, Astrophys. J. 884, 29 (2019)

[11] Y. Kobayashi, T. Nishimichi, M. Takada, R. Takahashi and K. Osato, arXiv:2005.06122 [astro-ph.CO].

[12] S. Heydenreich, B. Brück and J. Harnois-Déraps, arXiv:2007.13724 [astro-ph.CO].

[13] A. Mootoovaloo, A. F. Heavens, A. H. Jaffe and F. Leclercq, arXiv:2005.06551 [astro-ph.CO].

[14] M. Knabenhans *et al.* [Euclid], Mon. Not. Roy. Astron. Soc. 484, 5509-5529 (2019)

[15] J. Brynjarsdottir and A. O'Hagan, Inverse Problems **30**, 114007 (2014)