Deep Cosmos: Modeling the Universe with Statistical Learning Algorithms

Brian Nord, Associate Scientist
Fermi National Accelerator Laboratory
630 840 8337, nord@fnal.gov
Year Doctorate Awarded: 2010
Number of Times Previously Applied: 0
Topic Area*: Experimental Research at the Cosmic Frontier in High Energy Physics
DOE National Laboratory Announcement Number: **LAB 19-2019**

Abstract

The work described in this proposal will result in an improved understanding of cosmic acceleration and a paradigm shift in computational techniques through the use of statistical learning algorithms. This proposal supports measurements of cosmic acceleration from current and future data-intensive cosmological surveys, like LSST and CMB-S4. To address the growing size and complexity of imaging data from these experiments, we will develop and implement physics-aware deep learning analysis techniques for the extraction of science at multiple analysis levels — from object identification to inference of cosmological parameters.

Motivation: Cosmic science in the era of data-intensive experiments

Modern surveys have great promise to uncover a new understanding of cosmic acceleration, but we lack the modeling tools to take advantage of increasingly rich data sets. New algorithms and modeling methods based on statistical machine learning, but including the power of conventional parametric modeling, will be the key to realizing the potential of future cosmic surveys.

The goal of cosmic survey experiments is to model the origins, evolution, and fate of the universe. Indeed, HEPAP calls out cosmic acceleration as one of the key intertwined science drivers for the cosmic frontier [6]. Late-time acceleration is thought to be driven by dark energy, which is parameterized by the time-varying equation of state, $w(t)$. Early-universe acceleration is theorized to be driven by inflation, whose parameter of interest is the scalar-to-tensor ratio, r. These parameters must be inferred through observations of cosmic probes, which act as tracers of spacetime. The probes are themselves modeled from the raw imaging data acquired through next-generation telescope experiments: LSST in optical wavelengths and CMB-S4 in the microwave regime aim to constrain late- and early-time acceleration, respectively.

Challenges in modeling cosmic probes from imaging data necessarily drive challenges in modeling cosmic acceleration for these surveys. The sensitivity and size of cosmic experiments drive the size and complexity of their data, which conventional algorithms are not prepared to handle. LSST will acquire enormous data sets with billions of objects, seeing more objects than ever before. For example, $\sim 150,000$ strong gravitational lensing systems (two orders of magnitude beyond all current data sets combined) are expected to be discoverable in LSST data, but current analysis methods that rely on human intervention will require too much time. Not only will finding these needles in a haystack be a critical challenge, but analyzing them can take up to a day of human effort to create a model for a single object. The unprecedentedly high-resolution and low-noise CMB-S4 data will have contaminants, like weak gravitational lensing that prohibit new constraints on r. The Quadratic Estimator (QE), a conventionally parameterized model, is the current state of the art for "de-lensing" the CMB signal, but has been shown to be insufficient for future survey data [8].

Conventional algorithms, like those described above rely on physical parameterizations, where the parameters describe and account for the physically interpretable features that humans have identified. However, these types of models can and do miss critical features that have not been explicitly parameterized and identified by humans. On the other hand, deep learning algorithms can learn key features from the data itself, features that are not explicitly parametrized. In recent years, deep learning has made significant strides in applications in society and science, including in astrophysics and cosmology. Strong lens finding and modeling has been accelerated by deep learning algorithms, improving the modeling time by a factor of one million [7]. While this demonstration was carried out on space-based data from the Hubble Space Telescope, our group has developed an algorithm that can work on ground-based data [4]. For the task of removing the contaminating weak lensing signal from CMB data, our team implemented a neural network that outperforms the QE by $50-70\%$ across a wide range of spatial scales [3].

I have been leading teams in analysis of strong lensing, the CMB, and deep learning for three years. In particular, I have been leading the DES Strong Lensing Working Group and have founded the Deep Skies Lab, a collaboration for deep learning in astrophysics. My experience in uniting collaborators from data science and cosmology to attack the key problems in the cosmic frontier makes me uniquely suited to lead this proposal.

Goals and objectives: Understanding cosmic acceleration

The ultimate goal of this proposal is to achieve a new understanding of cosmic acceleration. The objectives that will enable us to achieve this goal are 1) enhanced efficiency and flexibility of modeling algorithms; 2) more effective models of complex imaging data and astrophysical objects; and 3) improved accuracy and precision of cosmological models. These objectives form a short hierarchy, such that one enables the next. In achieving these objectives, we will solve specific critical-path analysis challenges for modern cosmic surveys. The successful completion of these objectives forms a proof of concept that will pave the way for advancements in computational frameworks across cosmic experiments and enable the discovery and construction of new cosmological models.

Deliverables: From software to science

To achieve these objectives, we will deliver new scientific measurements, enhanced data products, and improved software tools for the age of data-intensive cosmic experiments. First, (1) we will create and release refined data products (e.g., catalogs of images and objects) derived from raw imaging data through our deep learning analysis engine. In optical wavelengths, we will create highly complete and pure catalogs of strong lenses, despite their relative scarcity and without the need for intensive human visual inspection. We will also create them in time to take advantage of transient objects, like lensed supernovae, for which follow-up observations will be crucial. At microwave frequencies, we will clean cosmic microwave background images of noise and contamination, like thermal dust signatures and gravitational lensing. Second, (2) we will use the object and imaging catalogs in standard cosmological parameter analysis tools to derive new constraints on cosmic acceleration for the early and late universe. Finally, (3) we will release an open-source software framework built on industry-standard deep learning toolkits. The deep learning algorithms in this framework will be enhanced to solve the current problems facing their application to science data.

While the two kinds of derived data products may appear highly disparate due to their different spatial scales, a key insight is that deep learning models handle these data structures with equivalent efficiency and accuracy, regardless of spatial scale. The computational framework will be constructed to take advantage of this feature of deep learning algorithms.

A new approach: statistical deep learning algorithms

To produce the deliverables, we propose to develop analysis techniques based on deep learning algorithms and to demonstrate their efficacy on key problems for cosmic surveys.

Algorithm development begins with well-tested deep neural network architectures, using supervised learning for optimization. For classification and regression of individual objects, we start with Residual and Inception architectures, which have exhibited the greatest efficiency and accuracy to date. For image analysis on large scales — like noise removal — we start with self-supervised networks, like autoencoders and u-nets. Finally, generative adversarial networks have proven highly versatile both for image segmentation and noise-removal, so these will be explored as well.

We will develop and improve the algorithms and models through an iterative process of testing on perfectly understood simulated data and real-sky data. We first train and test the model on simulated data, then predict on real-sky data in regimes where we already know the correct answer from conventional techniques and analyses. We will then augment the simulated training samples to include features required for accurate modeling of the real-sky data. This iterative cycle is standard for the development of models with deep learning algorithms.

We propose to implement a number of key innovations to overcome challenges in the application of neural networks to scientific data. We will develop an active learning mechanism that automates the iterative model-tuning cycle by using the differences between the simulated and real-sky data to automatically produce new simulations for training. Another key innovation will be the incorporation of physical parameters into the deep learning model. This performs a kind of regularization that will allow the algorithm to optimize the model more quickly and more accurately. Finally, we will implement a mechanism to propagate uncertainties through the deep learning model to obtain physically interpretable error bars on the final measurement. Bayesian neural networks [9] and the method of Concrete Dropout [5] offer the best avenues for implementing uncertainty measurement.

Deep learning algorithms require a large and diverse set of examples for effective training. For these training sets, we will create simulated data sets with currently available software, and then further develop the simulation software to accommodate the needs of the learning algorithm. For strong gravitational lens simulations, we will start with Lenstronomy [2], and for CMB simulations, we start with Monte Carlo simulations of the primary CMB and apply gravitational lensing with Quicklens [1] and other foregrounds with PySM [10].

The success of our algorithms for creating new models will be assessed through a) the speed of the inference process when applied to data; b) the accuracy and generalizability of the models in representing complex real-sky data; and c) the factors of improvement on figures of merit for cosmic acceleration parameters. Our three key innovations in the development of deep learning algorithms will be required to achieve these objectives.

Potential for impact: A new paradigm of cosmic data analysis

The desired results of this proposal are to enable discovery science next-generation cosmic surveys and to revolutionize analysis techniques for data-intensive cosmic survey experiments. If successful, it would not only provide a new understanding of cosmic acceleration, but significant time and cost decreases in our analyses of imaging data. The long-term impact also includes the solution of key challenges for the application of deep learning to science problems — incorporating existing physics knowledge and statistical measures of uncertainties into statistical deep learning models.

References

[1] quicklens. `https://github.com/dhanson/quicklens`.

[2] S. Birrer et al. *Physics of the Dark Universe*, 22:189–201, Dec. 2018.

[3] J. Caldeira et al. *arXiv e-prints*, page arXiv:1810.01483, Oct. 2018.

[4] C. de Bom, J. Poh, and B. Nord. in prep. 2019.

[5] Y. Gal et al. *arXiv e-prints*, page arXiv:1705.07832, May 2017.

[6] HEPAP Panel. Building for Discovery. https://science.energy.gov/hep/hepap/reports/.

[7] Y. D. Hezaveh et al. *ArXiv e-prints*, page arXiv:1708.08842, Aug. 2017.

[8] M. Millea et al. *ArXiv e-prints*, page arxiv:1708.06753, Aug. 2017.

[9] V. Mullachery et al. *arXiv e-prints*, page arXiv:1801.07710, Jan. 2018.

[10] B. Thorne et al. MNRAS, 469:2821–2833, Aug. 2017.

Collaborators and Other Affiliations

A. Agnello, ESA; A. Amara, ETH Zurich; C. Avestruz, KICP; B. Bassett, SAAO; S. Birrer, UCLA; M.R. Becker, Argonne; C. Chang, UChicago; T. Collett, U. Portsmouth; F. Courbin, EPFL; A. Dey, NOAO; H. Dominguez Sanchez, Penn; C. de Bom, CFETC; A. Farahi, CMU; A. Galda, UChicgo; K. Glazebrook, Swinburne; R. Hlozek, Toronto; T. Humble, ORNL; C. Jacobs, Swinburne; A. Manzotti, IAP; A. McCaskey, ORNL; C. Miller, UMichigan; A. Nicola, Princeton; J. Peek, STScI; J. Forero-Romero, U. de los Andes; A. Refregier, ETH Zurich; E. Rozo, U. Arizona; E.S. Rykoff, SLAC; I. Sevilla-Noarbe , UIUC; P. L. Schechter, MIT; L. Shaw, Yale; J. Simon, Carnegie; G. Snyder, STScI; T. Treu, UCLA; S. Trivedi, MIT; R.H. Wechsler, Stanford; K. Wei, UChicago; W. L. K. Wu, KICP; S. Young, ORNL;

Graduate and Postdoctoral Advisors:
A.E. Evrard, UMichigan; T.A. McKay, UMichigan; G. Tarlé, UMichigan; J. McMahon, UMichigan;

Thesis Advisor and Postgraduate-Scholar Sponsor:
Graduate students advised: Shubhendu Trivedi, PhD, MIT, Brown;
Postdoctoral scholars advised: J. Caldeira, PhD, FNAL;