Inference and Machine Learning at Extreme Scales


Simulation Volume

- First-generation surveys
- Single-probe simulations

- Second-generation surveys
- Multi-probe simulations
- Few precision probes
- Intermediate accuracy parameter estimation

- Next-generation surveys
- End-to-end, multi-probe survey-scale simulations
- Multiple cross-calibrated probes
- UQ-enabled cosmic calibration frameworks

Optical (LSST simulated sky)

CMB (Planck)

Computational context for Cosmic Frontier science with cosmological surveys, also showing the growing role of data-intensive computing and the associated development of advanced machine learning and statistical methods

Science with Surveys: Extreme-Scale Computing meets Statistics and Machine Learning

- Modern Precision Cosmology: Use of HPC resources as high-fidelity, large data-volume sources for state-of-the-art data-intensive statistical and machine learning (ML) methods
- ‘Stats at Scale’: Need to speed up methods by many orders of magnitude to enable dealing with datasets in the multi-PB to EB era
- SciDAC-3: Work on emulators is enabling a new era in cosmological analysis
Precision CMB Emulation

- **Science Target:** Precision fast prediction tools via emulators built on a large simulated dataset for South Pole Telescope and future CMB-S4 mission data analysis, speed-up requirement: factor of ~1000

- **Methodology:**
  - Large training/validation data set generated using the CAMB code
  - Dimensional reduction via unsupervised learning
  - High-dimensional non-parametric regression
  - High-accuracy posterior error controls

- **ML/DL method:**
  - Variational autoencoder and PCA-based dimensional reduction methods compared (similar results)
  - Sensitivity analysis via autoencoder-based nonlinear dimension reduction
  - Gaussian Process-based interpolation for both reduction methods

- **Results achieved:**
  - Emulator with factor of ~2000 speed-up compared to CAMB with 1% errors over the desired dynamic range (Top figure; paper in prep.)
Science Target: Search for strong lensing of galactic sources by intervening galaxies (~100K expected in LSST) for precision cosmology measurements; Deep CNN regression for lens properties

Methodology:
- Large synthetic data set based on full ray tracing algorithm with 1) model halo mass distribution as lenses and 2) halos from cosmological simulations, realistic telescope properties (pixelization, noise, etc.); single as well as stacked images
- DL techniques for classification, regression, and other applications (denoising, deblending, —)

ML/DL method:
- Deep CNN classification/regression
- GANs for fast generation of images

Results achieved:
- 80-90% accuracy with very fast classification time (10 microseconds per image)
- Regression testing underway

Image Classification/Regression for Strong Lensing

![Image Classification/Regression for Strong Lensing](image.png)
Science Target: Estimation of galaxy redshift distribution conditioned on photometric information, morphology, and spatial correlations; application to LSST

Methodology:
- Large synthetic data set based on a set of realistic templates
- ML techniques for classification (hidden space variables), use of mixture models; Bayesian learning for posterior PDFs
- Techniques for outlier rejection

ML/DL method:
- Mixture models to follow galaxy sub-populations
- Autoencoders for hidden space variable searches
- Various Gaussian Process-based approaches
- Bayesian Adaptive Regression Tree (BART) methods

Results achieved:
- Multiple synthetic data sets constructed
- Initial analyses with different methods underway
Other Topics; Future Work

- **Emulation Landscape:** 1) Extend work on summary statistics to problems with significantly higher dimensionality, $O(10)$ to $O(100)$; 2) Multi-fidelity emulation; 3) Develop new methods for applications to likelihood-free scenarios (e.g., semi-analytic galaxy modeling); 4) Fast generation of multiple realizations of ‘raw’ sky data (e.g., synthetic catalog/image emulation, prediction of dust maps from 21cm)

- **Image Applications:** Image cross-validation, source de-blending algorithms, application to calibration studies

- **ML/DL Methods on HPC Platforms:** Work on scaling up ML and statistical methods on HPC platforms with GPU acceleration (e.g., Cooley@ALCF, Summit@OLCF)

- **Stats meets ML:** Improve methods by incorporating model information into ‘black box’ techniques; incorporate optimization methods into Bayesian calibration

Catalog-level, emulation-aided, ‘full’ forward modeling approach to cosmological inference