Challenges in AI/ML at the Cosmic Frontier

Simone Ferraro
(Lawrence Berkeley National Lab)

SLAC Summer Institute
Aug 8, 2023
Plan for this talk

- **Part I:** What is the Cosmic Frontier and what are the observables?
- **Part II:** Are we being efficient (speed)?
- **Part III:** Are we extracting the whole information?
- **Part IV:** Do we understanding the data?
Part I: What is the Cosmic Frontier and what are the observables?
A simple yet strange Universe

But the model is based on...

- **Dark Matter** (?)
- **Dark Energy** (?)
- **Inflation** (?)
- **Neutrinos and other light particles** (?)

A major goal of the Cosmic Frontier program is to understand these “ingredients”!
A simple yet strange Universe

Fit fully characterized by \textbf{6 numbers}* (with no evidence of needing more):

\[ \{ \Omega_m, \Omega_b, A_s, n_s, \tau, H_0 \} \]

- matter
- “baryons”
- amplitude of primordial fluctuations
- slope of primordial fluctuations
- reionization
- expansion rate today

* and a few assumptions such as flat geometry and minimum mass neutrinos
A brief history of the Universe

94% of photons travel from the CMB to us without scattering*

6% scatter with matter

On small scales, the Cosmic Microwave Background (CMB) contains a “map” of the entire observable universe

*path slightly deflected by gravitational lensing
Cosmic microwave background (CMB)

- Large scales (< 1 deg) □ primordial
- Smaller scales (> 1 deg) □ processed by (known) plasma physics + gravity

Planck Satellite (2018)

“primary fluctuations”

“CMB (angular) power spectrum”
CMB lensing

Paths of CMB photons deflected by matter create statistical anisotropy that can be measured.

Can make maps of the projected matter density (including Dark Matter) to the CMB!

ACT DR6 lensing map
Galaxy (weak) lensing

Rubin Observatory
LSST

"cosmic shear power spectrum"

5 redshift bin tomography
\( \theta_s = 20000 \text{ deg}^2, n_g = 50 \text{ arcmin}^{-2}, \sigma_e = 0.22 \)

\( L(L+1)C_{\ell} / 2\pi \)

large scales \( \ell \) small scales

Simone Ferraro (LBNL)
Large Scale Structure (LSS)

Each dot is a real galaxy!

"galaxy power spectrum"
Neutrino mass effect on LSS

The small-scale matter power spectrum, $k > k_{fs}$, is reduced in presence of massive neutrinos:

- On larger scales vs cluster in the same way as cold dark matter
- Free-streaming vs do not cluster
- The growth rate of CDM and baryon fluctuations is reduced

\[ k_{fs} = 0.018 \left( \frac{M_{\nu}}{1 \text{ eV}} \right)^{1/2} h^{-1/2} \text{ Mpc}^{-1} \]

\[ \frac{\Delta P(k)}{P(k)} = 8f_{\nu} \]

\[ f_{\nu} = \frac{\omega_{\nu}}{\omega_m} = 0.5\% \]

4% suppression minimum!

+ non-linear calculations: additional suppression at large $k$

see Villaescusa-Navarro et al. 2013
Transients

Extragalactic:
- Supernovae/kilonovae
- Fast Radio Bursts, gamma ray bursts
- Tidal disruption events
- Strong lensing time delays
- ...

Galactic:
- Asteroids
- Interacting binaries
- Transiting exoplanets
- Microlensing
- Pulsars
- ...

Simone Ferraro (LBNL)
DESI

Dark Energy Spectroscopic Instrument: Massively multiplexed spectroscopic survey with 5000 robotic fibers, over ~14,000 sq. deg

Five target classes
35 million redshifts (SDSS x20)

2.4 million QSOs
Lyα z > 2.1
Tracers 1.0 < z < 2.1

17 million ELGs
0.6 < z < 1.6

6 million LRGs
0.4 < z < 1.0

10 million Brightest galaxies
0.0 < z < 0.4
The Vera C. Rubin Observatory

**Location:** El Peñón, Cerro Pachon, Chile (median seeing 0.67 arcsec)

**Specs:** 8.4m mirror, 3.2 Gigapixels camera
- 9.6 sq. deg. field of view (~40 full moons),
- 6 broadband filters (ugrizy)

It will perform the 10 year LSST survey of the sky
The CMB landscape – mid 2020s

- 10 Countries
- 40+ institutions

Large Aperture Telescope
one 6 meter in diameter

Small Aperture Telescopes
42 cm refractors

Large frequency coverage (30 – 270 GHz)

Fully funded
6-year program
First light in 2024!
• **CMB S4**: next generation ground based experiment
  • Factor of ~10 increase in sensitivity
  • ETA ~late in this decade

• Multi-agency effort (DOE & NSF)
Looking ahead: the “explosion” of surveys

Simone Ferraro (LBNL)

SPIDER-2

ADV-ACT

SIMONS ARRAY

SIMONS OBS

BICEP ARRAY

SPT-3G

BICEP3

KECK ARRAY

KIDS

eBOSS

DESI

HETDEX

HSC

LiteBIRD

PICO

CMB-HD

CMB-S4

SPHEREx

PUMA?

EUCLID

VERA RUBIN OBS (LSST)

NANCY GRACE ROMAN (WFIRST)

MSE/MegaMapper/SpecTel

credit: D. Kirkby
Dark Matter direct and indirect detection

IF Dark Matter interacts (weakly) with the Standard Model, can look for scattering/recoil (**direct detection**).
Several targets: Xenon, Germanium, etc

Also: “**indirect detection**” in astrophysical systems (Eg. Fermi gamma ray satellite)

Maria Elena Monzani’s lecture on anomaly detection!
Part II: are we being efficient (speed)?
Challenge 1: theoretical model

- **Complex & non-linear** dependence of theory on cosmological parameters even for power spectrum (2 point function). Often **no analytical form**, and prediction relies on expensive numerical simulations.

Calculating

\[ \text{theory}(\Omega_m, \Omega_{DE}, A_s, n_s, \tau, \ldots, \{\text{nuisance parameters}\}) \]

Can take minutes to hours (or more). Often too slow for parameter inference!

**SOLUTION:** **Build emulators!**
Reduce theory calculation to O(ms) per call

- [Joe DeRose’s lecture on emulators](#)
Challenge 2: parameter inference

- High dimensional problem: typically > 100 parameters (dimensions) inference. Slow or impossible.

  - The probability distribution \( P(\Theta|d) \) (posterior) for model parameters \( \Theta \) given data \( d \) can be related to the probability \( P(d|\Theta) \) (likelihood) of an experiment giving data \( d \) for model parameters \( \Theta \) using the Bayes’ theorem:

    \[
    P(\Theta|d) = \frac{P(d|\Theta)P(\Theta)}{P(d)} 
    \]  

    where \( P(\Theta) \) is called the prior and \( P(d) = \sum P(d|\Theta)P(\Theta) \) is used for the normalization purpose.

Typically you would sample the posterior by Monte-Carlo. In high-dimension (>100 parameters), algorithms involving gradients are more efficient (eg. Hamiltonian Monte Carlo).

SOLUTION: Build a differentiable likelihood + differential emulators!
Mock data generation

Lensing data

Both cases based on (conditional) Variational Auto-Encoder (VAE)

Similar applications with Generative Adversarial Networks (GANs)

"Gas pasting" on Dark-Matter only simulations

Diffusion models?

See lectures on generative models applications
Mock data generation

Can you tell which one is generated by an (expensive) hydrodynamical simulation and which is generated by CVAE?
Part III: are we extracting the whole information?
Beyond the power spectrum

- Power spectrum (2pt function) contains the whole information only for Gaussian fields
- Can perturbatively consider 3pt function and higher, but (in general), limited information available.
- But in general, no guidance on what's the most informative statistic...

- Several options available:
  - Field-level inference
  - Compression in a “small” number of summary statistics
  - In both cases, likelihood may not be known analytically (likelihood-free/simulation-based inference)
The initial conditions are very close to Gaussian: they contain the whole information. Can we reconstruct them?
Field-level inference

"solving the inverse problem by optimization"

D. Bartlett
Field-level inference

Optimization converges in $O(20 \text{ steps})$, even though $N_{\text{dim}} = N_{\text{pix}} \sim 10^6$ or more!

But: want to marginalize over the initial conditions to extract cosmological parameters. Active area of research and questions remain!

Can use Laplace approximation or MUSE (see arXiv:2112.09354).

Or… full HMC sampling (eg. BORG https://www.aquila-consortium.org/)

See “Introduction to Differentiable Programming in Jax” (F. Lanusse)

More generally: the full problem

Often we need more freedom than a traditional likelihood approach:

- We may not know what the likelihood is (Gaussian approximation is often a bad one!)
- We may summarize, cut, mask the data any way we want
- Observational or instrumental effects are hard to treat analytically but easy to simulate.

Simulating data is often much easier than deriving an accurate likelihood

**Simulation-Based Inference** (SBI)

SBI = Inference “engine” when explicit likelihood is intractable or unknown, but simulation is possible.
Simulation based inference introduction

Simplest implementation of "Approximate Bayesian Computation" (ABC)

Suffers from severe "Curse of dimensionality"
Compression beyond the power spectrum

- Beyond the power spectrum, we have little guidance on what’s the most informative statistic…

- **Score compression** and **Information Maximizing Neural Networks (IMNN)**: \( t(d) : \mathbb{R}^N \rightarrow \mathbb{R}^p \) \((p < N)\) produce a small number of summary statistic that maximize the retained Fisher information.

- See arXiv:1802.03537 for more info.
Likelihood-free inference (LFI)

Likelihood-Free Inference (LFI) with summaries

The statistical model is learned from simulations.

Alsing, Wandelt & Feeney
arXiv:1801.01497

Alsing, Charnock, Wandelt & Feeney
arXiv:1903.00007

Slide from B. Wandelt
Discovering new relations: symbolic regression

- Search the space of analytic equations to fit some data
- Often done by hand but efficient algorithms exist!
- Concise
- Interpretable

arXiv:2201.01305
arXiv:2305.01582
https://github.com/MilesCranmer/PySR
Part IV: Do we understand the data?
Classification

- Source classification (stars vs galaxies vs quasars etc)
- Transient classification
- Classification of the cosmic web (voids, sheets, filaments etc)
- Photometric redshifts
- ...

Both supervised and unsupervised methods. Need to allow for the existence of unexpected patterns!
Anomaly detection

Several goals:
• Eliminate the influence of outliers or contaminants
• Find new signals when the signal is rare
  • known unknowns: supernovae, strong lenses, transients, …
  • unknown unknowns: eg. discovery of pulsars

credit: M. Lochner
https://github.com/MichelleLochner/astronomaly
CAMELS: an “MNIST” dataset for ML in Cosmology

- A suite of 4,233 simulations
- 2,049 N-body; Gadget-III
- 2,184 state-of-the-art (magneto-)hydrodynamic sims
- AREPO/IllustrisTNG + GIZMO/SIMBA
- 6 parameters: \( \Omega_m, \sigma_8, A_{SN1}, A_{SN2}, A_{AGN1}, A_{AGN2} \)
- More than 100 billion resolution elements over combined volume of \( (400 \text{ Mpc/h})^3 \)
- More than 2,000 cosmologies & astrophysics models; more than 140,000 snapshots
- Designed for machine learning applications
Simons Collaboration on “Learning the Universe”

https://www.learning-the-universe.org/
Conclusions

- ML will help us solve cosmological problems that are intractable today.

- With great power, come great responsibility! Astrophysical systems are complex and often not fully understood. Model misspecification can lead to issues and great care needs to be taken.

- Finally, for a comprehensive list of ML applications to cosmology, see https://github.com/georgestein/ml-in-cosmology

Thanks!