Cosmological Utility Codes

Prospects on optimisation and transition towards deep learning

Paolo Campeti (Tecnologo @ INFN Sezione di Ferrara)

with Martina Gerbino, Massimiliano Lattanzi

June 14th, 2023 - Spoke 3 General Meeting in Catania

Traditional Cosmological Inference Machinery

It’s likelihood-based

- In both Bayesian (MCMC) and frequentist (profile likelihood) frameworks

Model M with parameters $\theta = \{\theta_1, \theta_2 \dots\}$

→

Theoretically predicted observables (e.g. power spectra)

Data D

→

Observables $O = O_1, O_2 \dots$ (e.g. power spectra computed from maps)

Likelihood $\mathscr{L}(M \,|\, D) \equiv \mathscr{L}(\theta \,|\, O)$ evaluated at each sampling step

Traditional Cosmological Inference Machinery

It’s likelihood-based

- In both Bayesian (MCMC) and frequentist (profile likelihood) frameworks

Model M with parameters $\theta = \{\theta_1, \theta_2 \dots\}$

→

BOTTLENECK

Theoretically predicted observables (e.g. power spectra)

→

Likelihood $\mathscr{L}(M\,|\,D) \equiv \mathscr{L}(\theta\,|\,O)$ evaluated at each sampling step

Data D

→

Observables $O = O_1, O_2 \dots$ (e.g. power spectra computed from maps)

→

Issues with the traditional approach

1. **Curse of dimensionality**:

   - Even for moderate number of nuisance parameters MCMC or profile likelihood require $> \mathcal{O}(10^{5-6})$ likelihood evaluations to converge (~weeks/ months!)

   -Must be addressed for future surveys (LiteBIRD, Euclid…) with many nuisance parameters!

2. **Bayesian model comparison** extremely intensive or even impossible! Fundamental scientific target for future surveys!

3. Can be hard to formulate an **analytical description** of the data in a **likelihood** (e.g. residual non-Gaussian contamination in CMB maps from systematics)

Issues with the traditional approach…

…and possible solutions…

1. **Curse of dimensionality**	→	1.Neural-Net-based **Emulators** (e.g. power spectrum), **GPU porting** of likelihoods and Boltzmann codes
2. Bayesian **model comparison**	→	2. Simulation-based inference with NN
3. Can be hard to formulate an **analytical description** of the data in a **likelihood**	→	3. Simulation-based inference e.g. CNN

WP 3

Emulators for Boltzmann codes

Neural-net-based

- Power spectrum evaluation via Boltzmann codes is the main bottleneck in likelihood computation
- Use **NN** to **map directly cosmological params to power spectra**
- Implemented in e.g. CosmoPower code (*Spurio Mancini+2021*) both matter and CMB angular power spectra
- $\mathcal{O}(10^4)$ **acceleration** compared to Boltzmann codes!!!
- Training can be done in parallel
- Implement also for **higher order correlations** (bispectrum etc.), **interpretability**

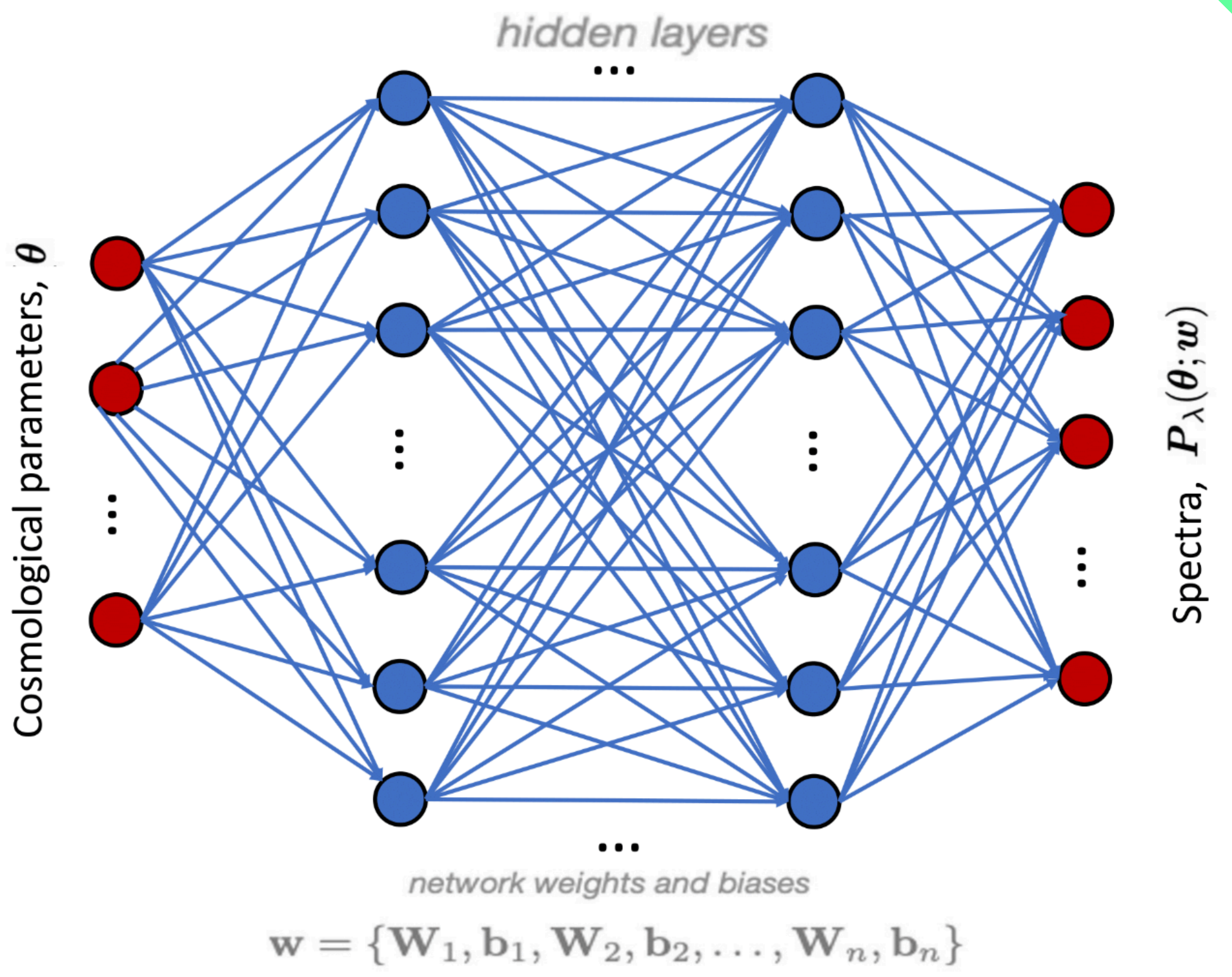

Figure from *Spurio Mancini+2021*

GPU porting of likelihoods and Boltzmann codes

WP 1 & 2

Likelihood codes

- Theory power spectrum computation isn't the only bottleneck

- Other numerical operations and loops inside the likelihood (e.g. bin cross-correlations) can be expensive

➡ Rewrite likelihoods in TENSORFLOW to exploit **GPU parallelisation**

➡ In general, **optimize simulation/ inference pipelines** (e.g. for LiteBIRD)

Boltzmann codes

- NN emulators require re-training for each cosmological model being compared to data (e.g. beyond ΛCDM models)

- Re-training can be expensive

➡ **Porting Boltzmann codes** like CAMB or CLASS **to GPU**

Comparing cosmological models

requires computation the Bayesian Model Evidence:

WP 3

$$p(\theta | d, M) = \frac{p(d | \theta, M) p(\theta | M)}{p(d | M)} \qquad z = p(d | M) = \int p(d | \theta, M) p(\theta | M) d\theta$$

Bayesian
Evidence

- Extremely expensive multi-dimensional integral even in moderate parameter space: traditional sampling is typically unfeasible

- Solution: **simulation-based inference**, in particular **neural density estimation** (e.g. *Delaunoy+2022*, *Cole+2022*, *Spurio Mancini+2022*, *Vasist+2023*)

- Ideal case of use of **deep neural nets**: train estimator q_ϕ parametrized by weights ϕ to approximate target prob density (e.g. posterior, likelihood…) from a training set of N pairs of prior samples and simulations $\{\theta_i, d_i\}_{i=1}^N$, accurate for $N \to \infty$

Another application of Simulation-Based Inference

useful when modelling analytically data in a likelihood is hard

WP 3

- E.g. residual non-Gaussian contamination in Planck polarization maps affecting cosmological parameters determination

➡ estimate cosmo parameters directly at map level using Convolutional NN (e.g. *Wolz, Krachmalnicoff & Pagano 2023*)

- SBI also useful when it is difficult to define the optimal summary statistic

➡ Reduction of data to summary statistic loses information

Our program

- Develop new and better tools and/or extend scope of existing tools for efficient **cosmological inference**, such as:
    - Faster and more efficient power spectrum computation:
        - Via **NN-based emulators** **WP 3**
        - Via **GPU porting** of **Boltzmann codes** **WP 1 & 2**
    - Faster and more efficient **likelihood codes** via **GPU porting** **WP 1 & 2**
    - **NN-based** tools for **model comparison** and **parameter estimation** **WP 3**
- Develop tools and/or optimize **pipelines** for INFN-funded experiments (e.g. LiteBIRD, Euclid) **WP 1 & 2**
- **Synergic experiments** starting now (e.g. Simons observatory) will also benefit from our **timely** studies and provide **ideal & immediate test-bed** for development and tuning of our tools!