The non-Gaussian Universe: a challenge in cosmological data analysis

Michele Liguori

Dipartimento di Fisica e Astronomia

Università di Padova

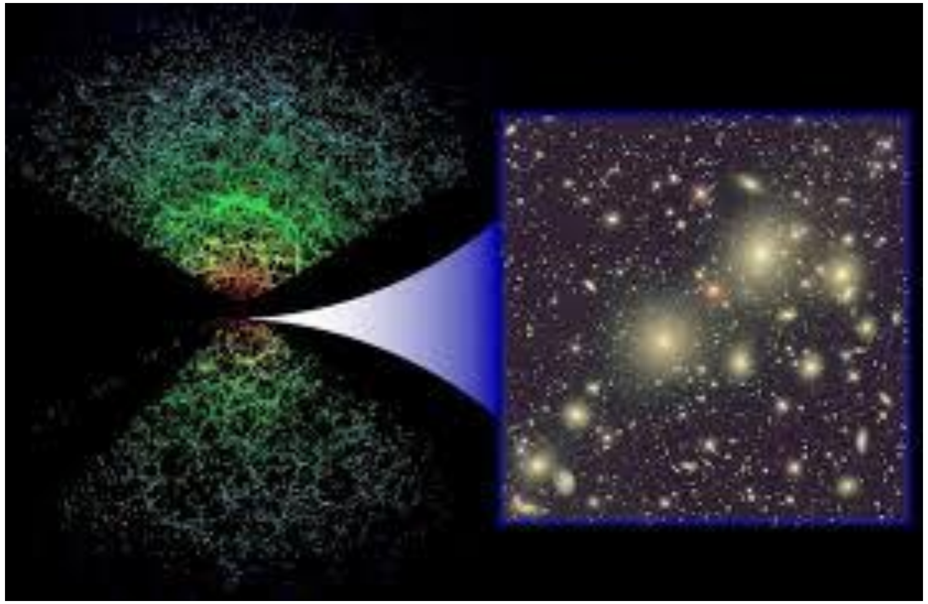
Dipartimento di Fisica e Astronomia Galileo Galilei

Collaborators: M. Baldi, N. Bartolo, W. Coulton, G. Jung, D. Jamieson, D. Karagiannis, S. Matarrese, F. Villaescusa-Navarro, A. Ravenni, H. Shao, M. Shiraishi, L. Verde, B. Wandelt

Main papers:

- ✓ **Fergusson, Liguori, Shellard (2009,2010), https://arxiv.org/abs/0912.5516, https://arxiv.org/abs/1006.1642**
- ✓ **Jung et al. (2022a, 2022b), https://arxiv.org/abs/2211.07565, https://arxiv.org/abs/2206.01624**
- ✓ **Coulton et al. (2022a, 2022b), https://arxiv.org/abs/2206.01619, https://arxiv.org/abs/2206.01619**
- ✓ **Jung et al. (2023) https://arxiv.org/abs/2305.10597**

Cosmological **evolution**



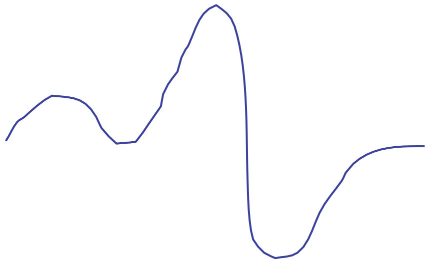
The Universe Large Scale Structure

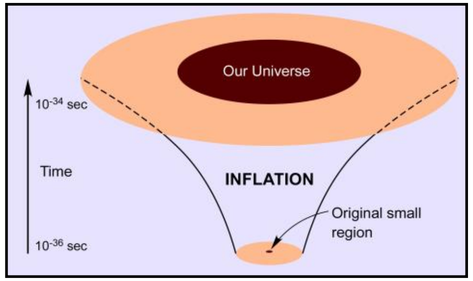
- Cosmological surveys map the distribution of millions of galaxies over very large volumes. These galaxies form groups and cluster in a typical “web-like” pattern.

- If we calculate the average number of galaxies in large boxes of side $\sim 10\ Mpc$, we see that the galaxy density field is nearly homogeneous and isotropic.

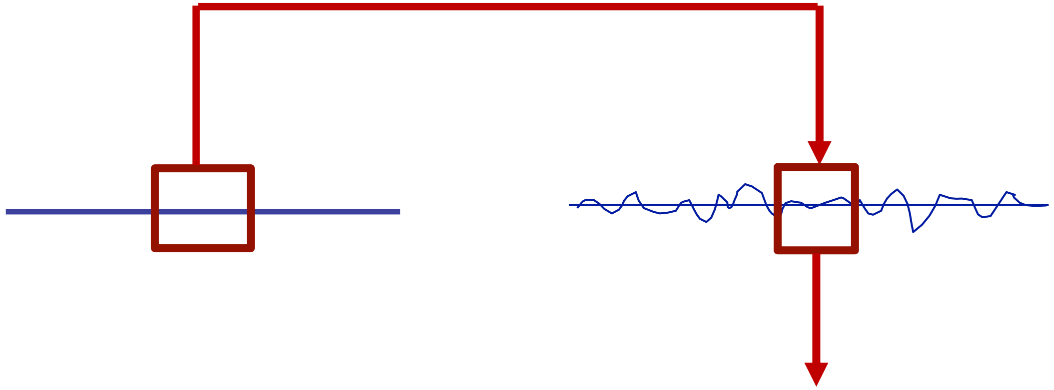
- How do these cosmic structures form and evolve?

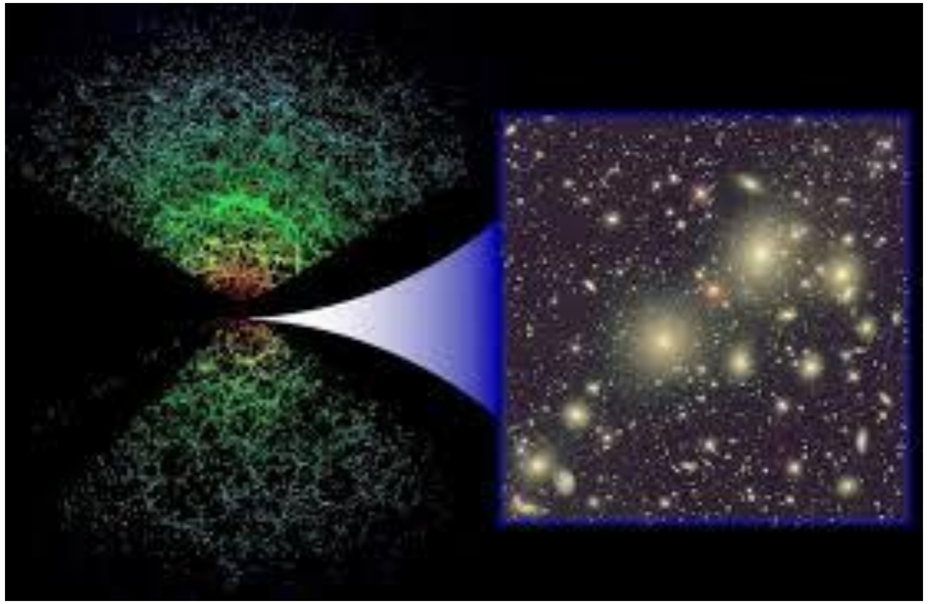
Cosmological evolution

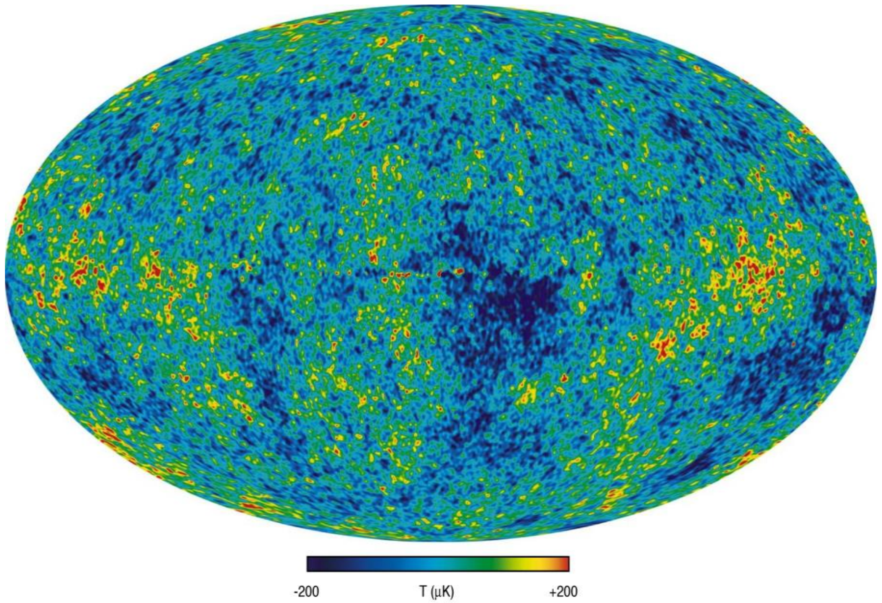




INFLATION







Observational goals

- CMB anisotropies and galaxy clustering originate from a gravitational instability process, starting from primordial **random seeds** (quantum fluctuations) and including the interaction of various particle species (baryons, dark matter, photons, neutrinos)

- CMB (temperature, polarization) anisotropies and observed galaxy clustering are specific realizations of **spatial random processes**.

- Goal: using observations, study the **statistical properties** of the galaxy density and CMB anisotropy field, in order to:

  - Measure the abundance of different components (e.g. Ω_b, Ω_c, Ω_Λ…)
  - Study gravity on cosmological scales
  - Test and constrain inflation

Observational goals

- Most inflationary models predict primordial cosmological fluctuations to be **Gaussian** distributed $\Rightarrow$ CMB and galaxy density fluctuations on large scales (> 10 Mpc) are Gaussian random fields (with zero average).

- To characterize a zero average Gaussian random field, all we need is its covariance.

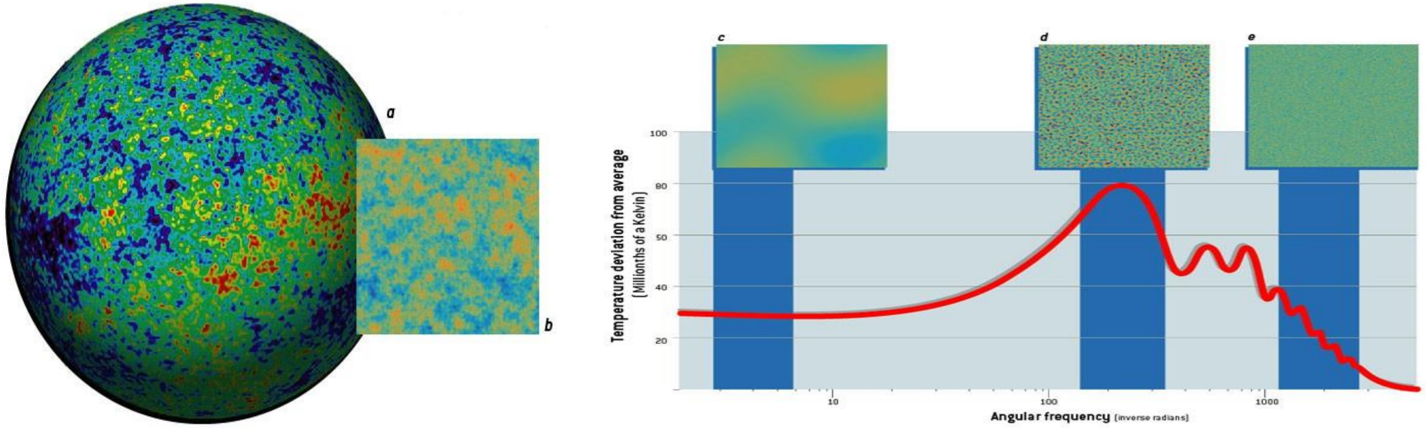
- Homogeneity and isotropy $\Rightarrow$ all the information is in the variance of Fourier modes of the field (**power spectrum**)

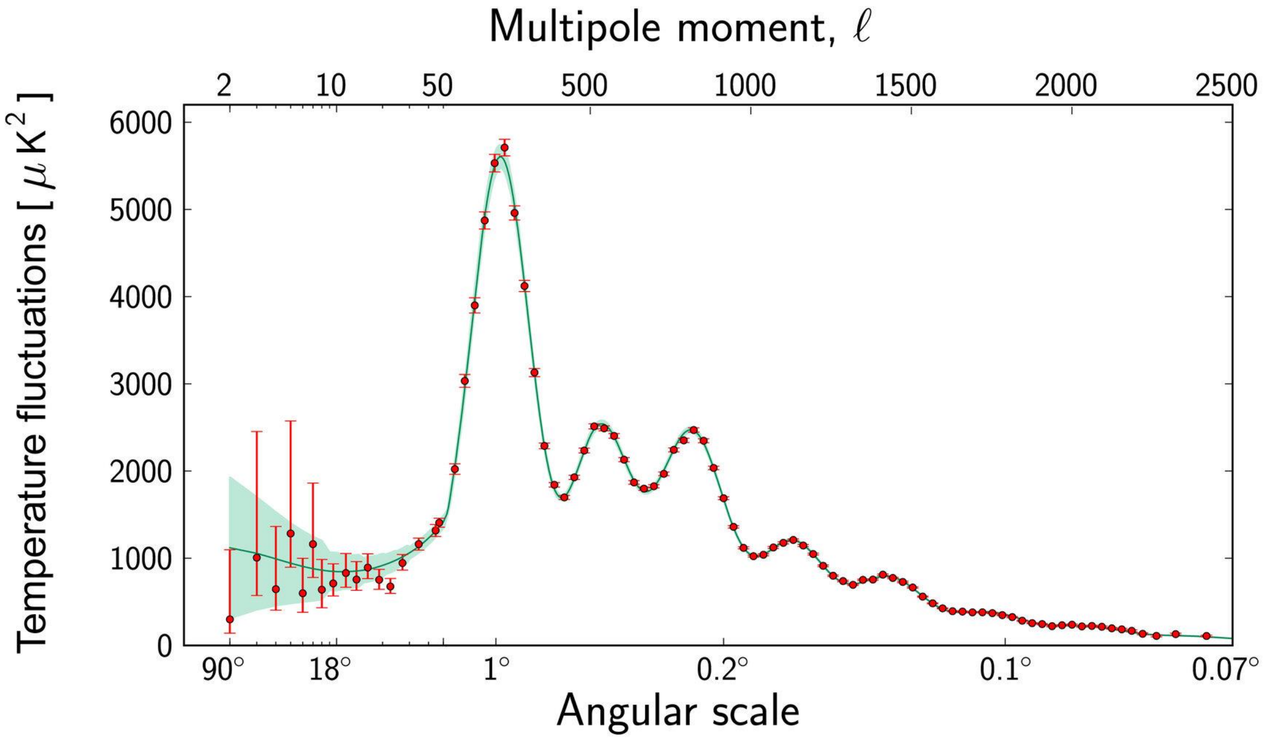
- A significant part of observational Cosmology therefore deals with the problem of predicting and measuring power spectra of CMB and LSS observables

The CMB power spectrum

$$\frac{\Delta T}{T}(\hat{n}) = \sum_{\ell m} a_{\ell \mathrm{m}} Y_{\ell m}(\hat{n})$$

$$C_\ell = \langle |a_{\ell m}|^2 \rangle$$





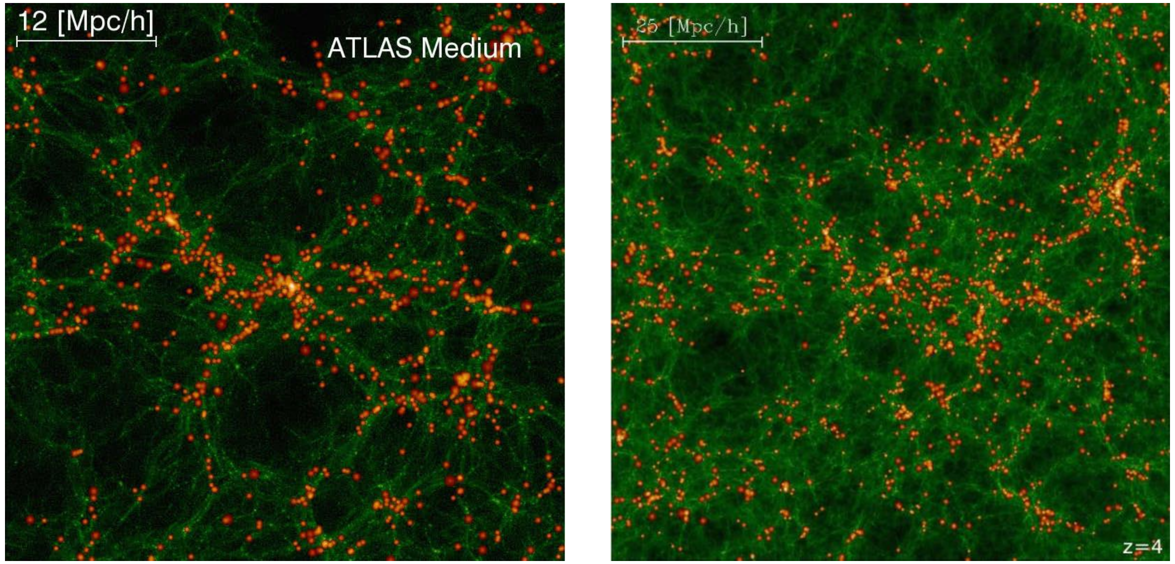
- Isotropy: the power spectrum does not depend on m:
- Homogeneity: different ℓ (angular frequency) are uncorrelated

The matter power spectrum

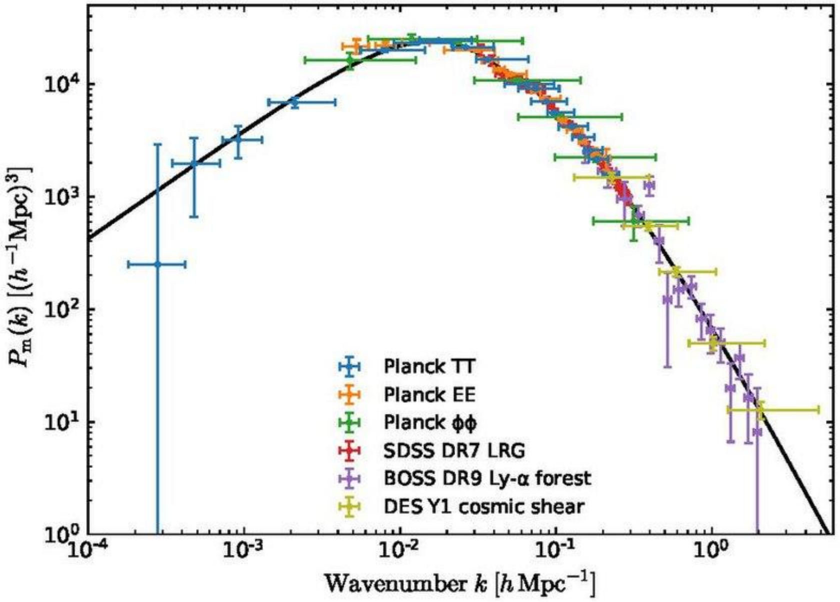
$$\delta_{\mathrm{g}} = b_g(z)\,\delta_{\mathrm{m}}$$

$$P(k) = \langle|\delta(\vec{k})|^2\rangle$$

- Isotropy: the power spectrum does not depend on orientation of k:
- Homogeneity: different k are uncorrelated



Wang et al. (ATLAS collaboration) arXiv:1802.01539



Planck collaboration 2018

Beyond the power spectrum: non-Gaussianity

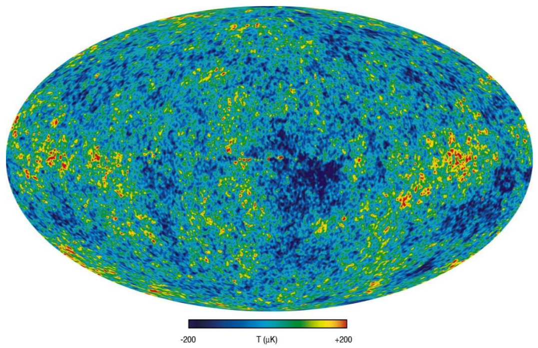
Does the power spectrum contain all relevant information?
(i.e., are the CMB and LSS fluctuation fields always Gaussian, at all scales?)

No, it does not

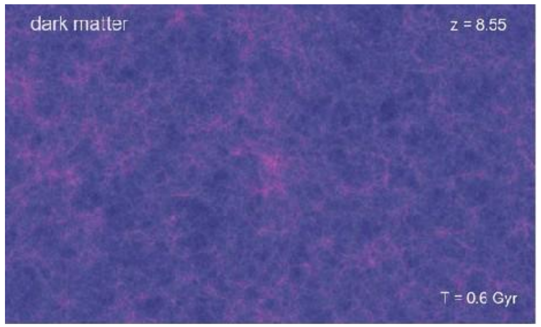
Inflation can produce small, model dependent non-Gaussianity (NG) In the primordial density field, in presence of deviations from standard single-field, slow-roll, e.g., multi-field, non-standard kinetic terms, features in the inflaton potential and more

Gravitational instability introduces non-linearity in the perturbation evolution process. When $\delta > 1$, the matter fluctuation field becomes non-Gaussian

$\frac{\Delta T}{T} \ll 1,$
linear transfer



$\delta \ll 1,$
linear transfer

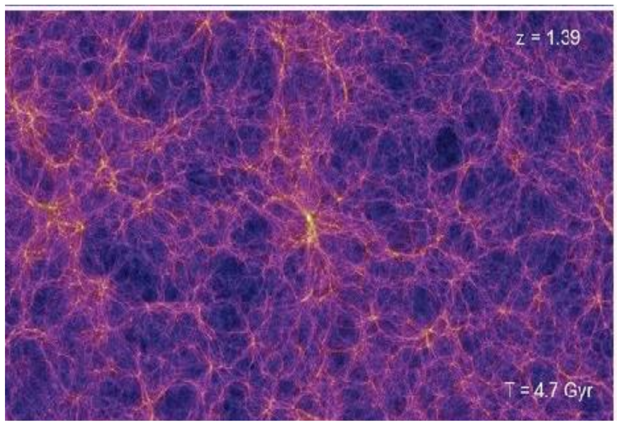


Primordial perturbations: mildly NG if non-standard inflation

CMB anisotropies: at linear transfer level, mildly NG only if also initial conditions are.

High redshift (early times) matter field: "same as CMB"

$\delta > 1,$
non-linear transfer



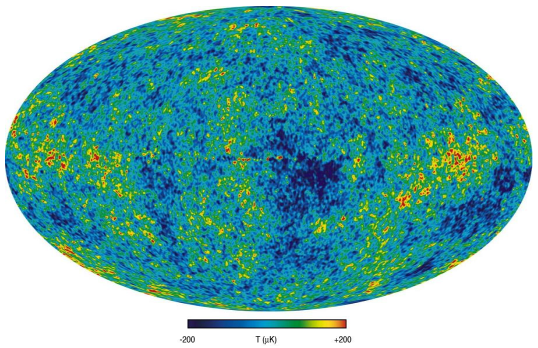
Low-redshift (late times) matter field
Strong NG from gravitational evolution of structures

Hierarchical structure formation (small scales collapse first)

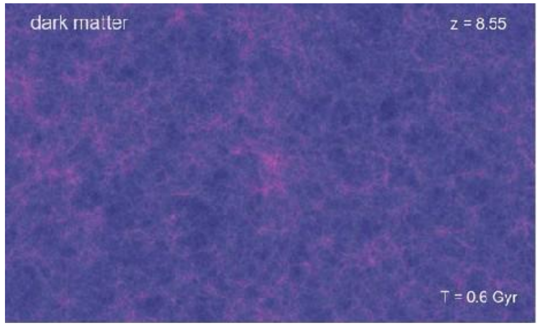
Large scales (k < 0.1 h/Mpc), low-pass filter => Gaussian (or small NG from inflation)

Small scales (k > 0.1 h/Mpc), high-pass filter => non-Gaussian

$\frac{\Delta T}{T} \ll 1$,
linear transfer



$\delta \ll 1$,
linear transfer

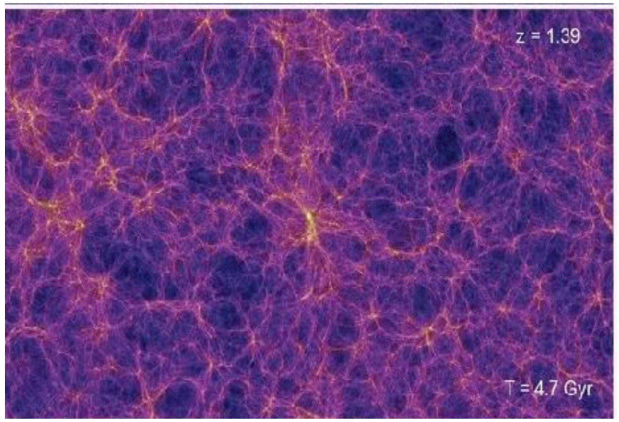


Primordial perturbations: mildly NG if non-standard inflation

CMB anisotropies: at linear transfer level, mildly NG only if also initial conditions are.

High redshift (early times) matter field: “same as CMB”

$\delta > 1$,
non-linear transfer



Low-redshift (late times) matter field
Strong NG from gravitational evolution of structures

Hierarchical structure formation (small scales collapse first)

Large scales (k < 0.1 h/Mpc), low-pass filter => Gaussian (or small NG from inflation)

Small scales (k > 0.1 h/Mpc), high-pass filter => non-Gaussian

Primordial NG and the Cosmic Microwave Background

A simple model ("local NG", from multifield scenarios): $\Phi = \Phi_G + f_{NL}\,(\Phi_G^2 - \langle \phi_G^2 \rangle) + \ldots$

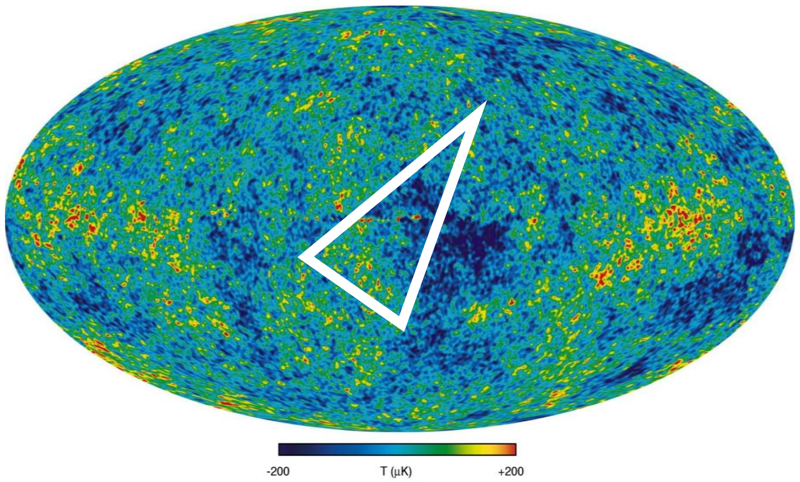
Primordial potential

G part
$\Phi_G \sim 10^{-5}$

NG amplitude
$f_{NL} < 10$

NG part

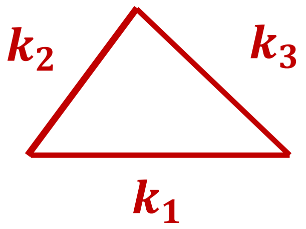
In this perturbative regime, most information is stored in the 3-point function of the primordial potential (bispectrum, in Fourier space). This can be tested by measuring the 3-point correlation function of the CMB



The CMB angular bispectrum

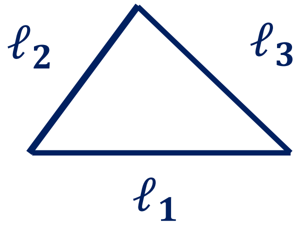
Due to homogeneity and isotropy, it is always convenient to work in Fourier space. Primordial bispectrum:

$$\langle \Phi(k_1)\Phi(k_2)\Phi(k_3)\rangle = B(k_1, k_2, k_3)\delta^D(\mathbf{k_1} + \mathbf{k_2} + \mathbf{k_3})$$

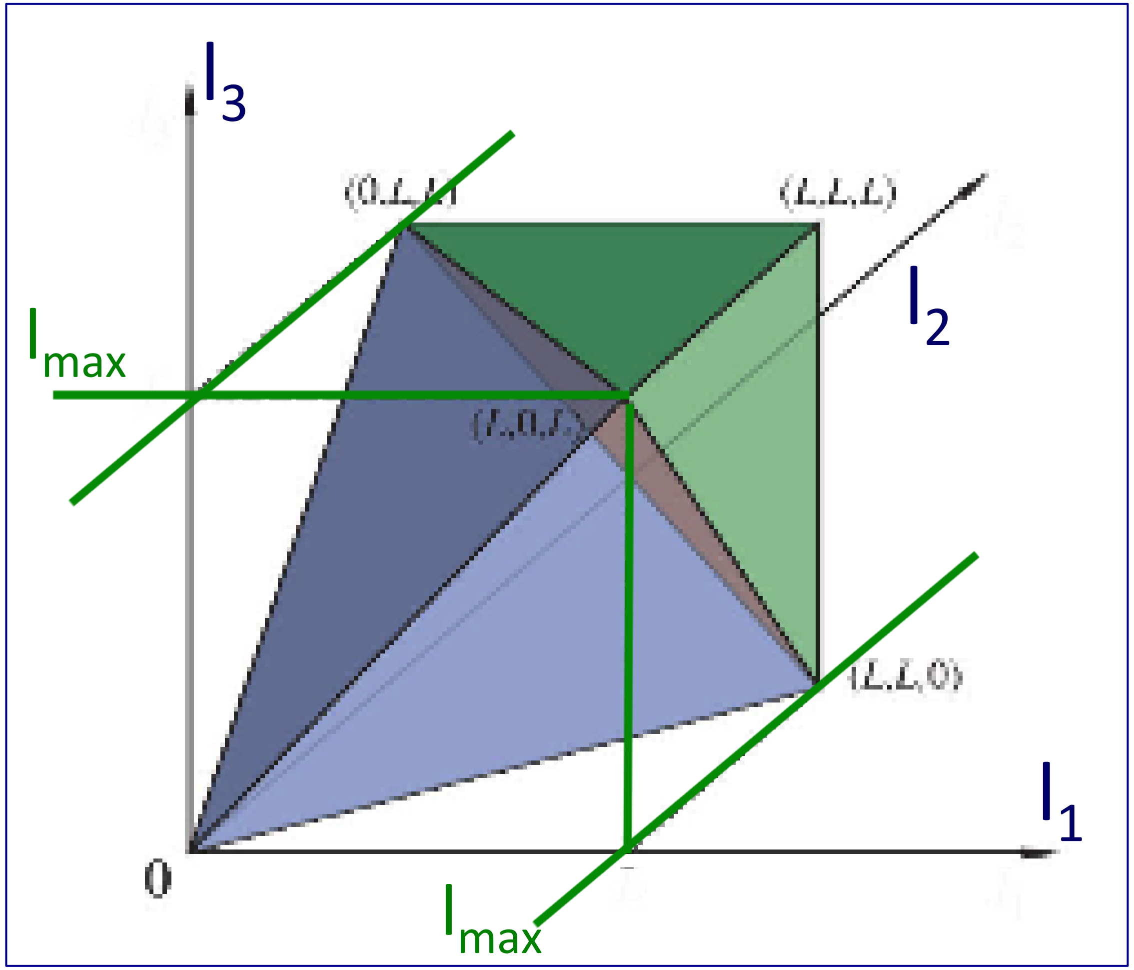


We work in harmonic space and compute the multipole 3-point correlation function:

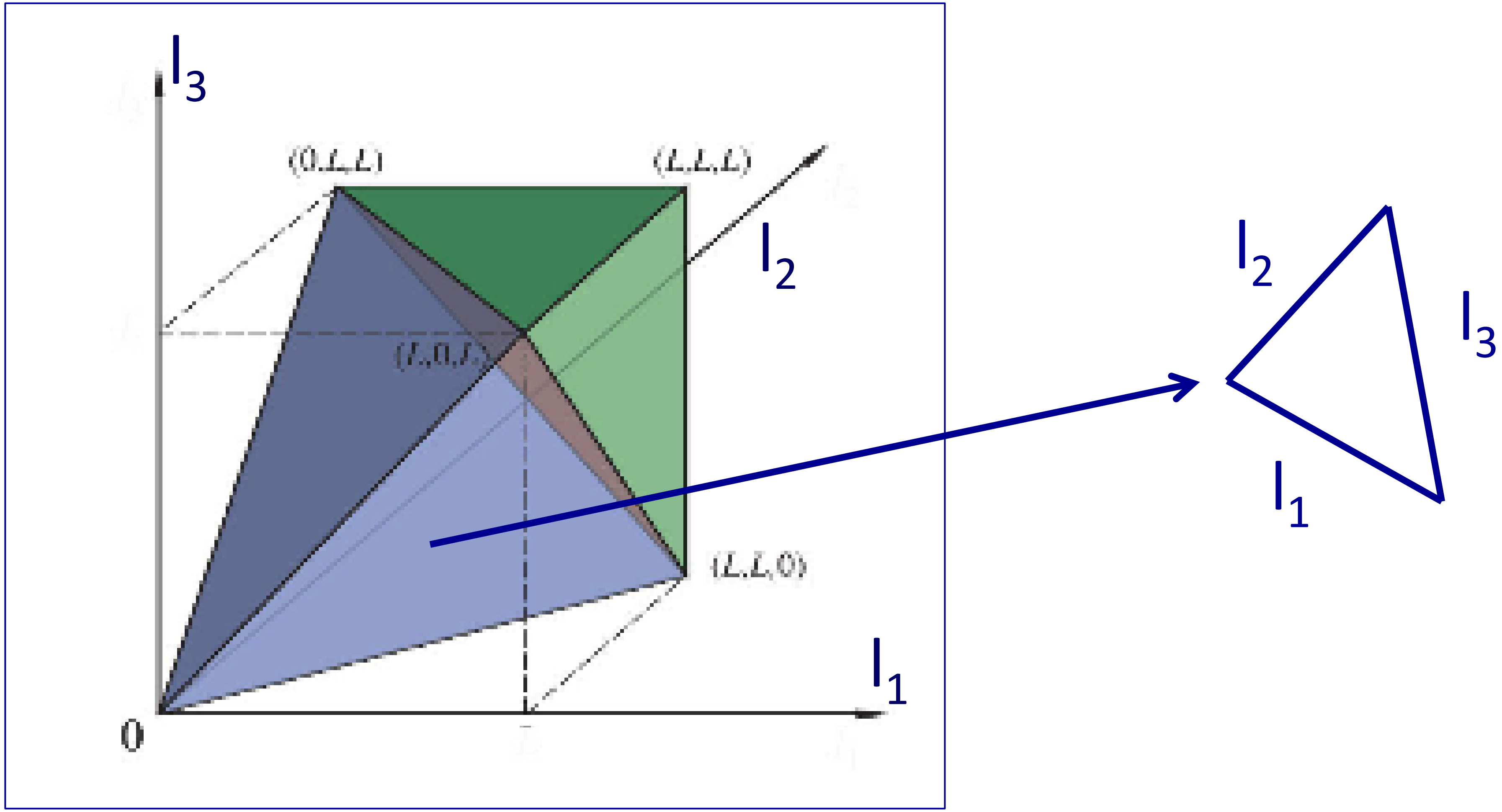
$$\left\langle a_{\ell_1 m_1} a_{\ell_2 m_2} a_{\ell_3 m_3} \right\rangle = \mathrm{b}_{\ell_1 \ell_2 \ell_3} \times (\textbf{Gaunt factor})$$



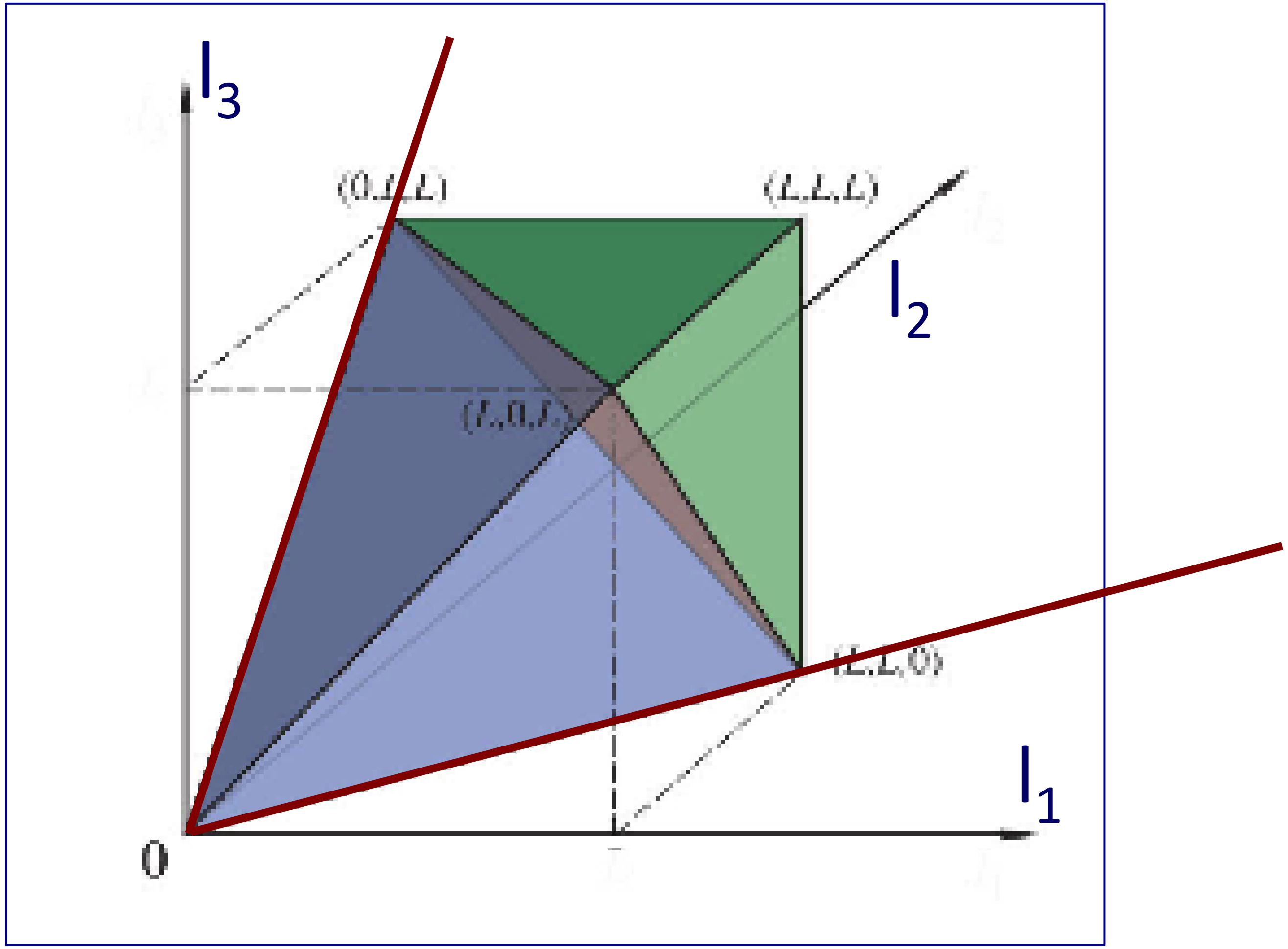
The CMB angular bispectrum

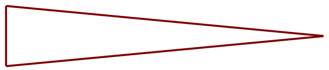


The CMB angular bispectrum

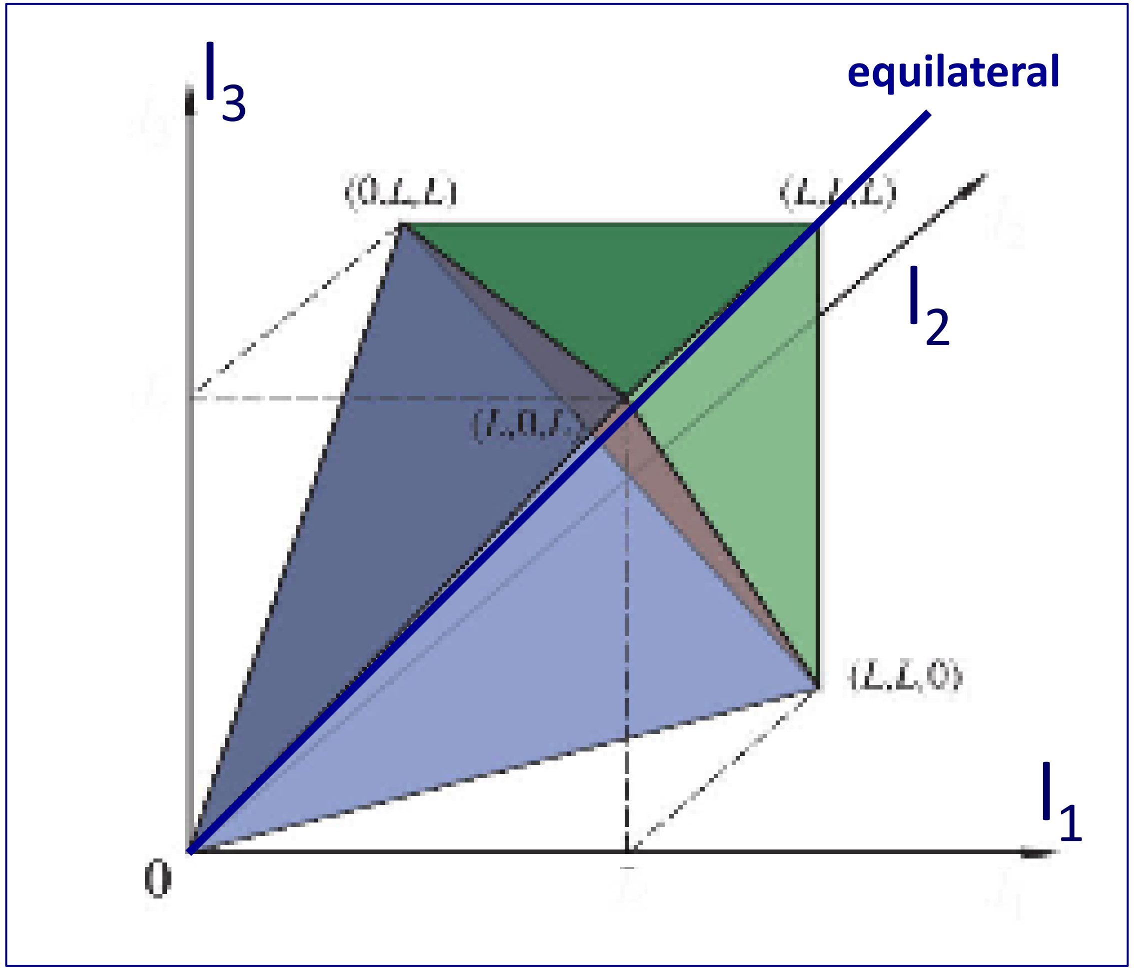


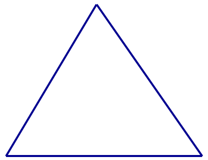
The CMB angular bispectrum





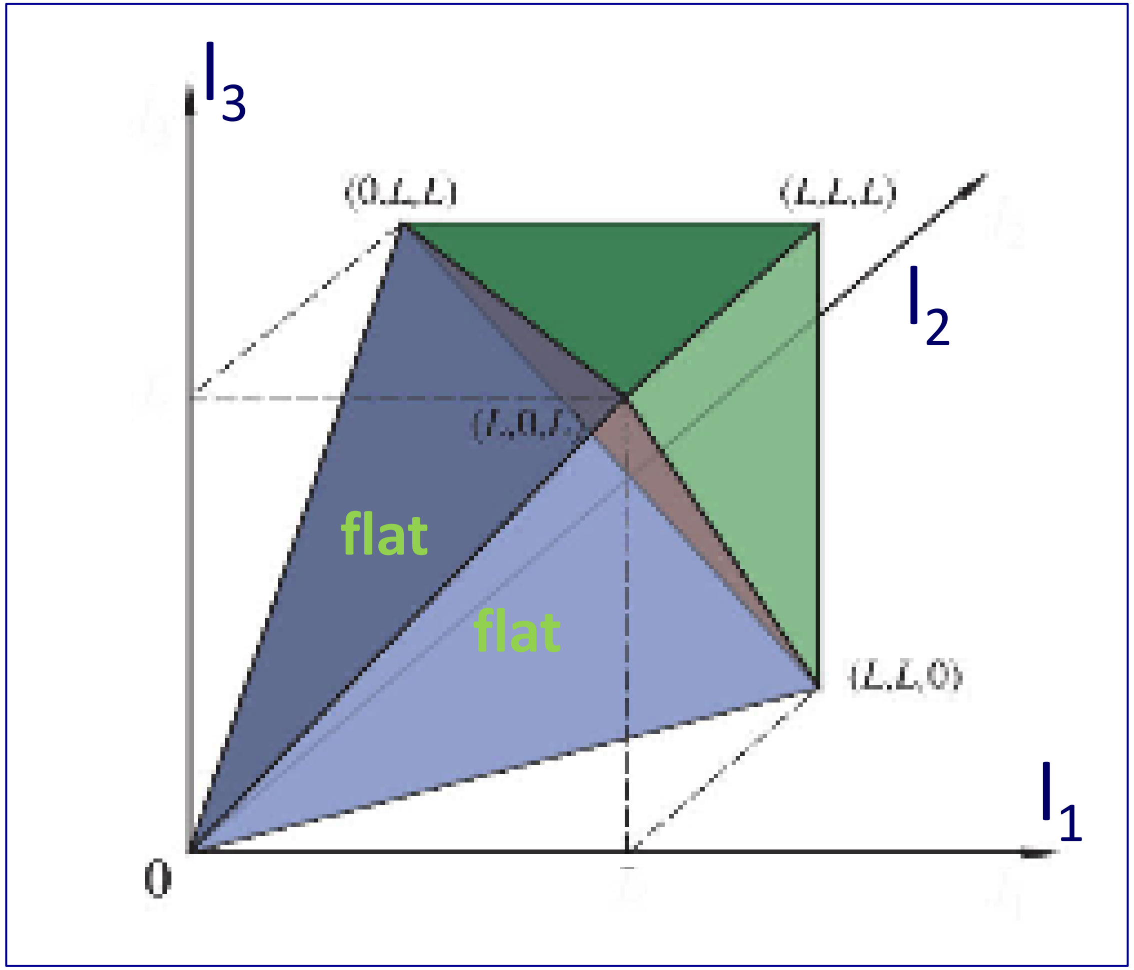
The CMB angular bispectrum





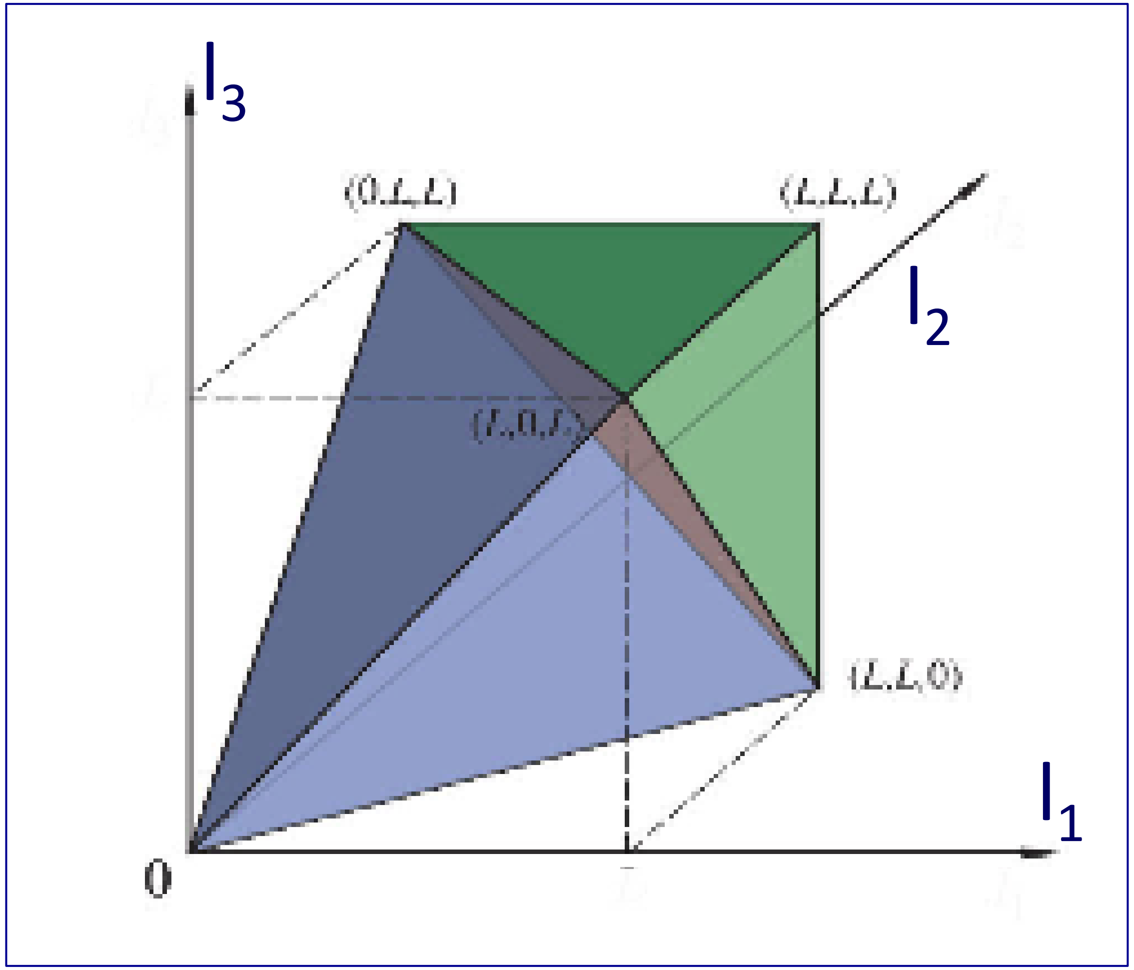
$l_1 \sim l_2 \sim l_3$

The CMB angular bispectrum

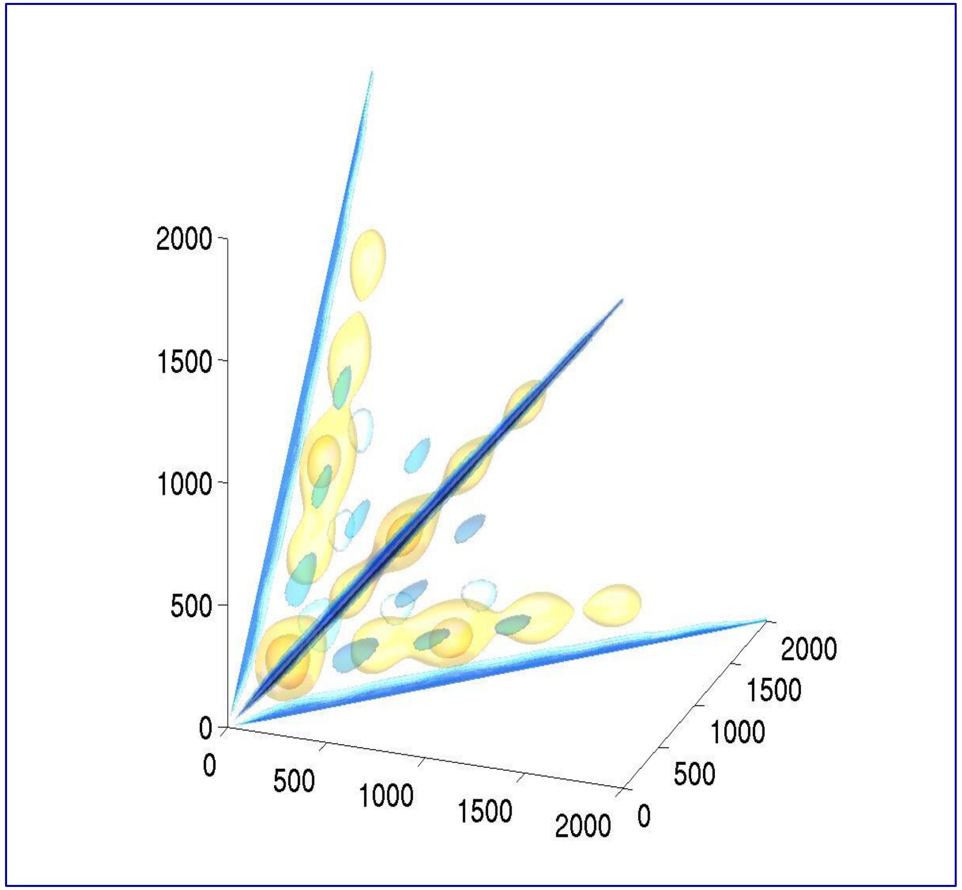


$l_2 \sim l_3$

Local primordial NG

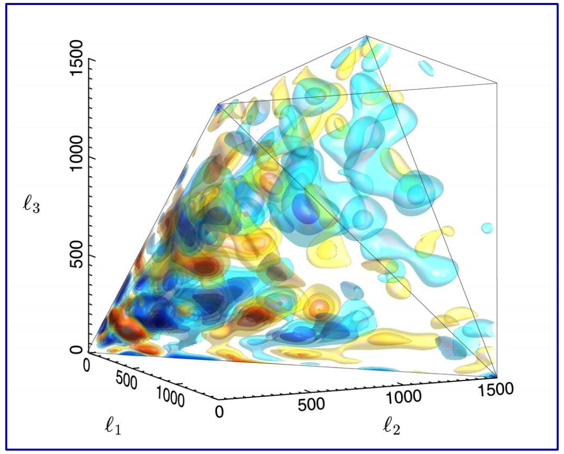


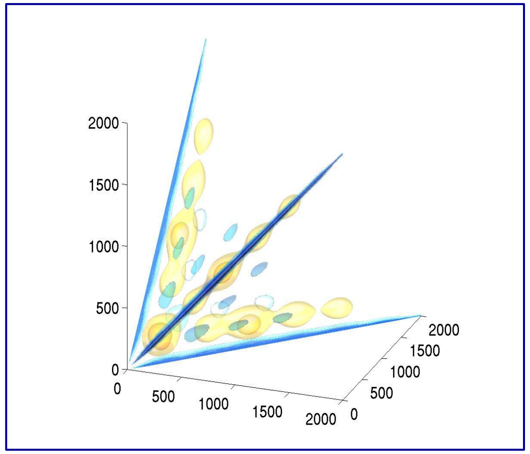
$$\Phi = \Phi_G + f_{NL}\,(\,\Phi_G^2\ - \langle \phi_G^2 \rangle) + \ldots$$

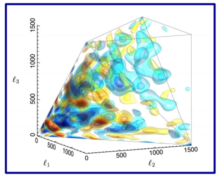
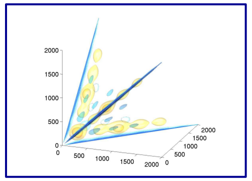


Testing local primordial NG

How do we measure local f_{NL} using CMB data? Schematically: 1. Extract the data bispectrum, 2. Compute the relevant theoretical bispectrum template and fit it.



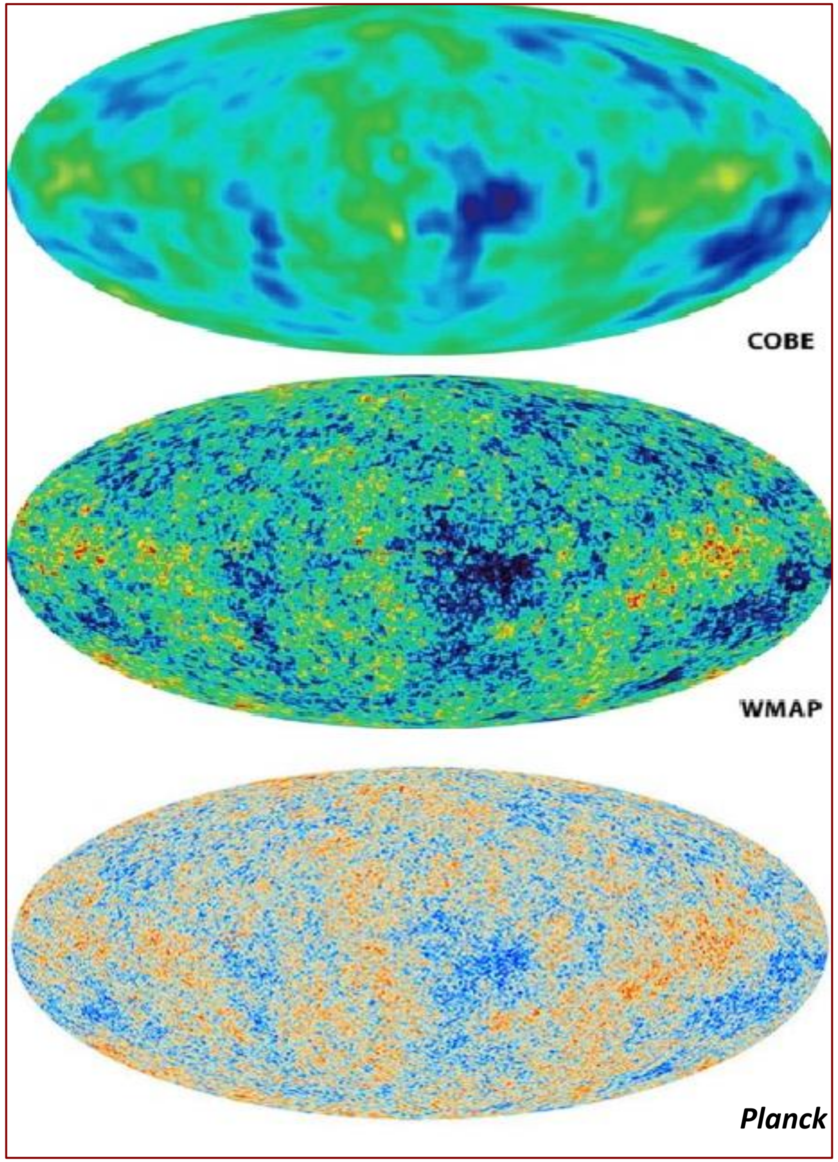


$\chi^2(f_{NL}) = \sum_{\Delta}$ ( - )^2 / σ_{Δ}^2

1. Complex observational issues (e.g. foregrounds, non-stationary noise, sky masking)
2. Lots of triangles!

COBE, WMAP, Planck.
Computational requirements

- For a COBE-like surveys (early 90s), ~ 5000 triangles. A brute force computation is possible (Komatsu et al. 2000), but low S/N.
- For WMAP and Planck, $\sim 10^8$, $\sim 10^9$ triangles respectively. A brute force computation is unfeasible
- Local NG is just one example. We typically want to fit hundreds of inflationary motivated template
- Need some form of <u>data compression</u> and/or a methodology to speed up bispectrum template fitting



$N_{pix} \sim 10^4$

FWHM $\sim 7^o$

$l_{max} \sim 30$

$N_{pix} \sim 3\times10^6$

FWHM $\sim$ 12 arcmin

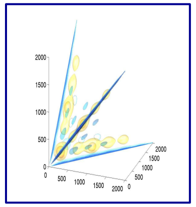
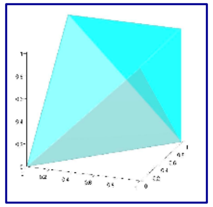
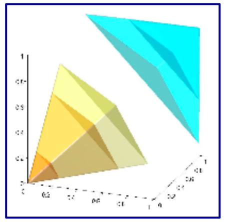
$l_{max} \sim 1000$

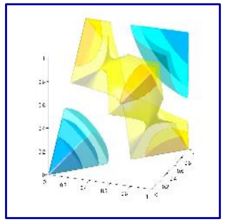
$N_{pix} \sim 5\times10^7$

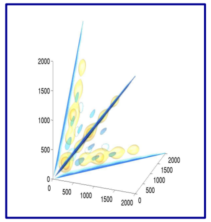
FWHM $\sim$ 5 arcmin

$l_{max} \sim 3000$

Modal expansion

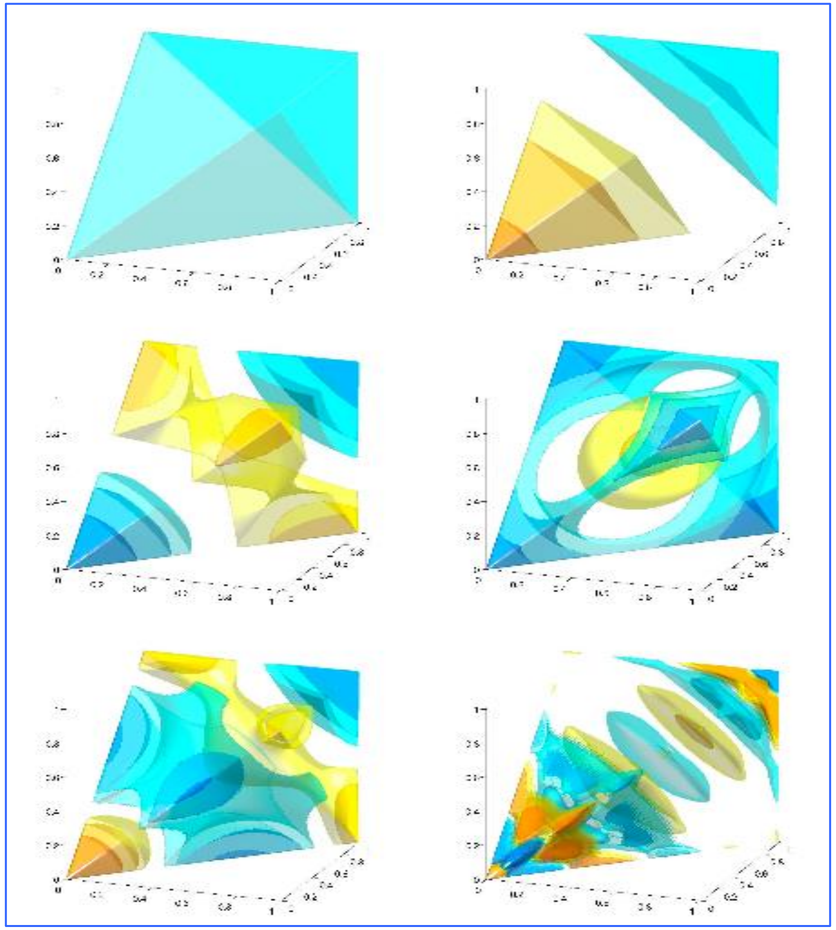
 $= \alpha_0$  $+ \alpha_1$ 

$+ \alpha_2$  $+ \ldots$

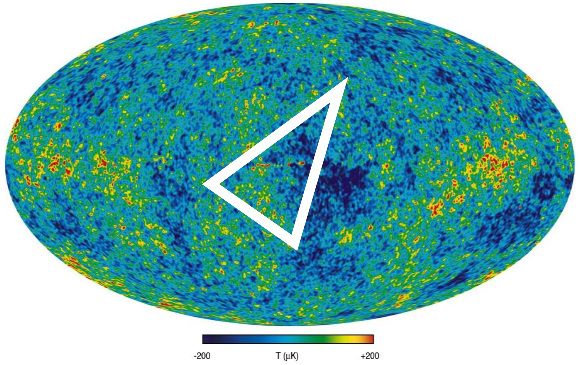
 $\Rightarrow (\alpha_1, \alpha_2, \alpha_3, \ldots, \alpha_n)$

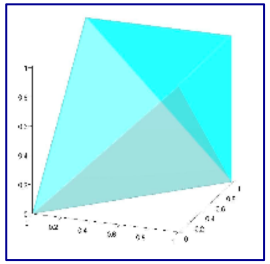
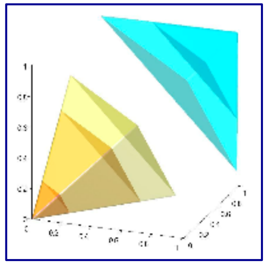
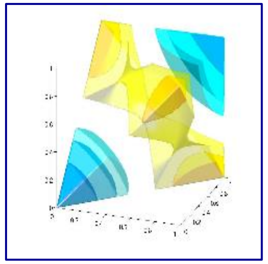
At Planck resolution, we can expand all models with ~ 1000 modes, by picking an efficient basis

Basis templates ("bispectrum modes")



Modal expansion



$= \beta_0$  $+ \beta_1$  $+ \beta_2$  $+ \ldots$

- Fit all basis modes to the data to estimate β amplitudes . *The modes can be constructed with suitable mathematical properties to make the computation very fast.*

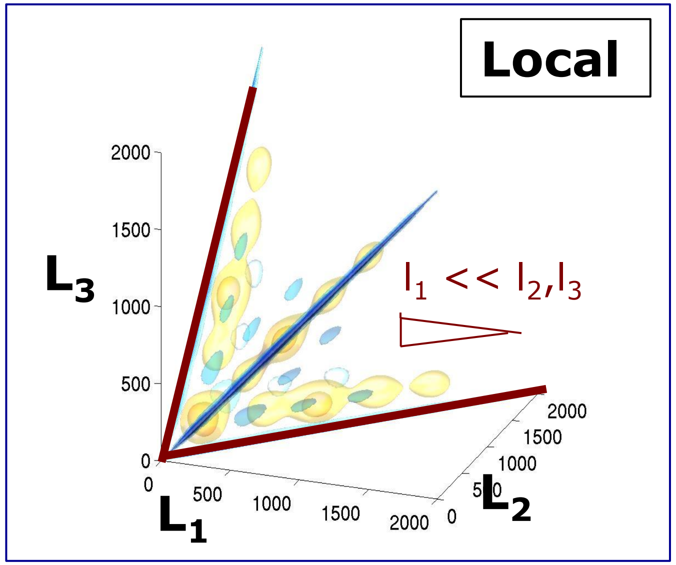
- The NG amplitude is then obtained as a scalar product between the theory coefficients α_i and the estimates β_i

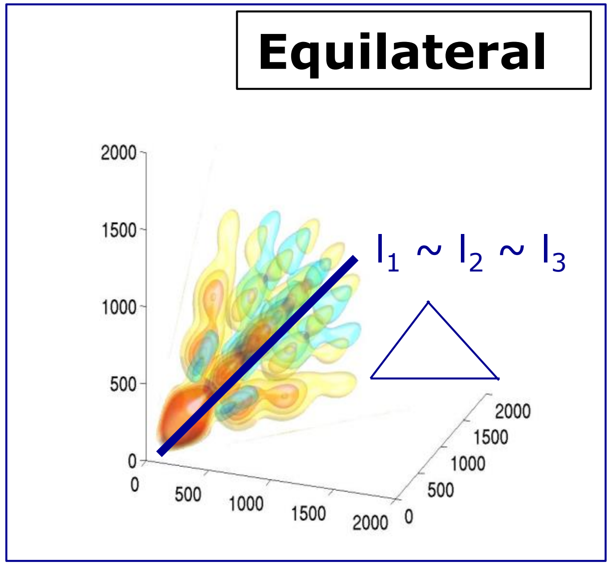
$$f_{NL} = \frac{1}{N} \sum_n a_n b_n$$

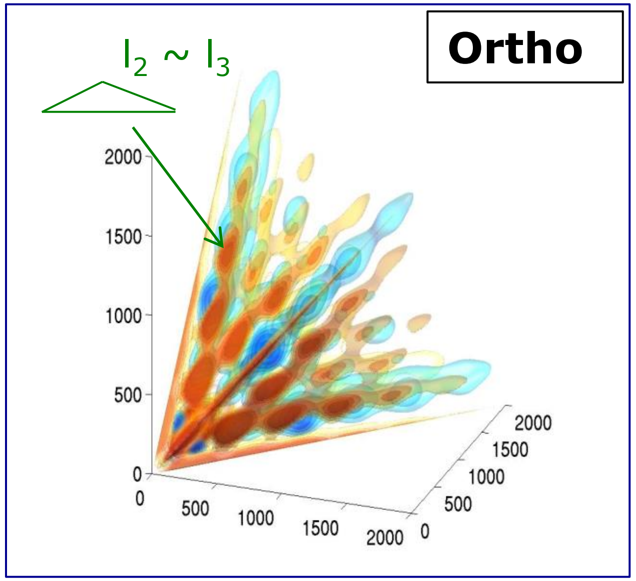
$$N = \frac{1}{6} \sum_n a_n^2$$

J. Fergusson, ML, P. Shellard 2009, 2010, arXiv: 0912.5516, 1006.1642
J. Fergusson, P. Shellard, 2011, arXiv: 1105.2791,
M. Shiraishi, ML, J. Fergusson 2014, arXiv: 1403.4222, 1409.0265
J. Fergusson 2014, arXiv:1403.7949

Inflationary bispectrum templates







- Multi-field
- Curvaton
- Ekpyrotic/cyclic

- Non-canonical kinetic terms (K-inflation, DBI)
- Higher derivative terms (Ghost Inflation)
- EFT

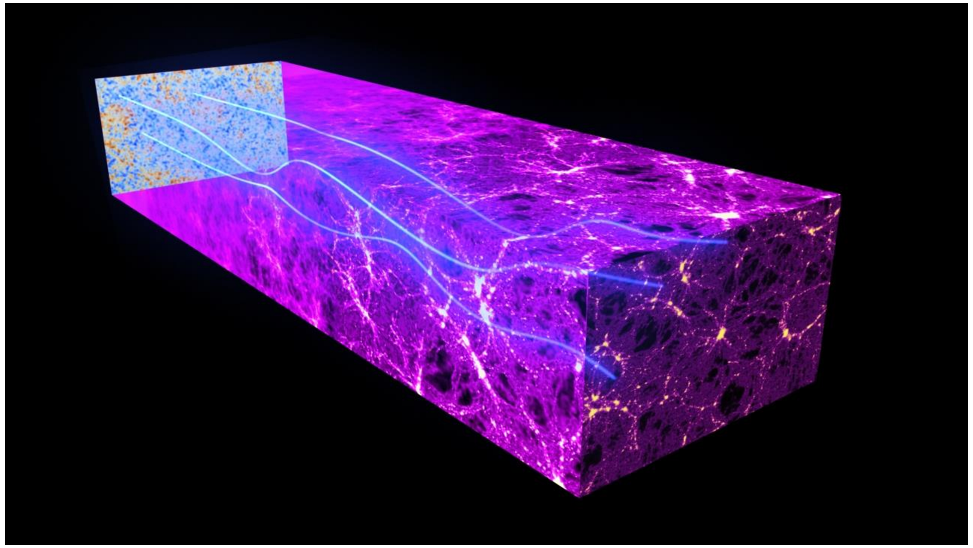
- Variants of non canonical kinetic terms and higher derivatives
- EFT

Planck constraints

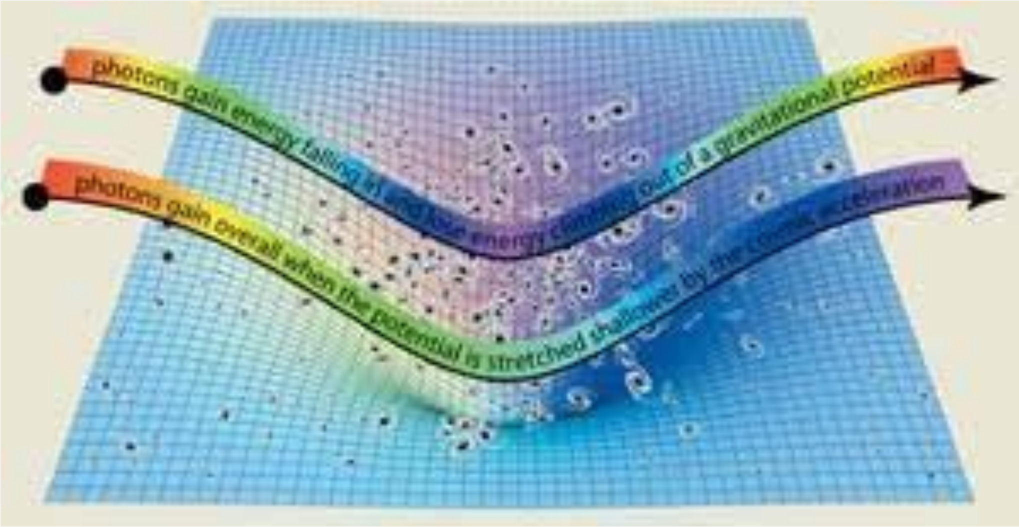
Shape and method	f_{NL}(KSW) Independent	ISW-lensing subtracted
`SMICA` (T)		
Local	6.7 ± 5.6	**−0.5 ± 5.6**
Equilateral	4.0 ± 67	**4.7 ± 67**
Orthogonal	−38 ± 37	**−15 ± 37**
`SMICA` (T+E)		
Local	4.1 ± 5.1	**−0.9 ± 5.1**
Equilateral	−25 ± 47	**−26 ± 47**
Orthogonal	−47 ± 24	**−38 ± 24**

Planck 2018 results. IX, arXiv:1905.05697

ISW-lensing bispectrum



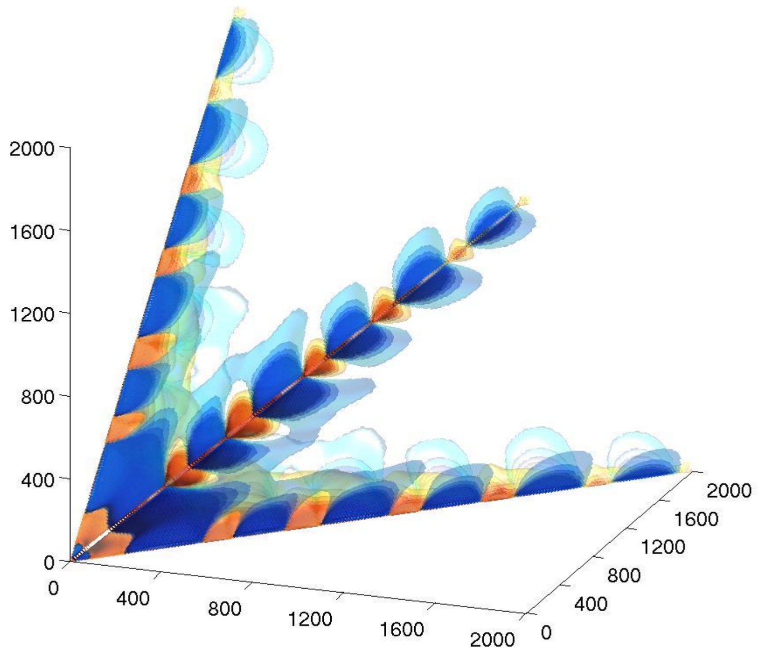
CMB lensing: photons geodesics are deflected due to LSS.



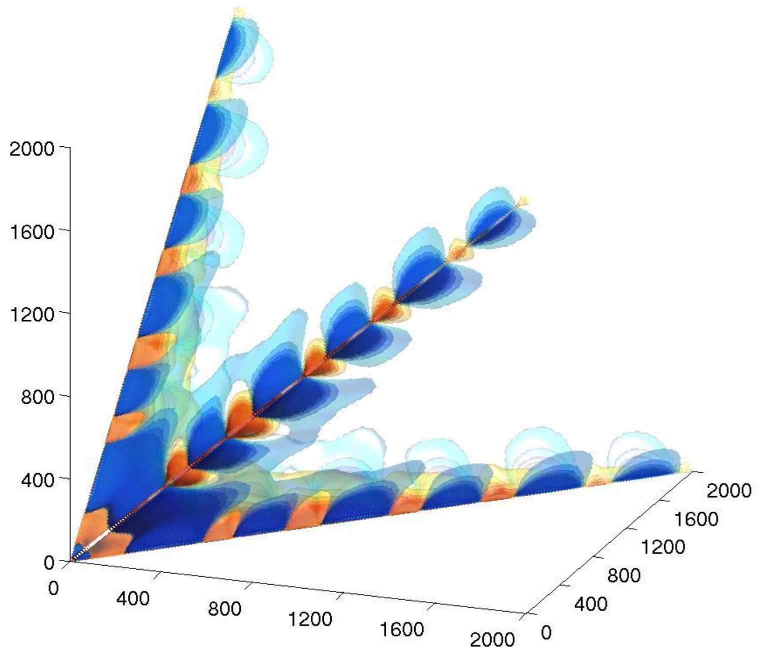
ISW effect: late-time acceleration (cosmological costant/dark energy) slows-down structure growth, time evolving potential generates differential redshift/blueshift

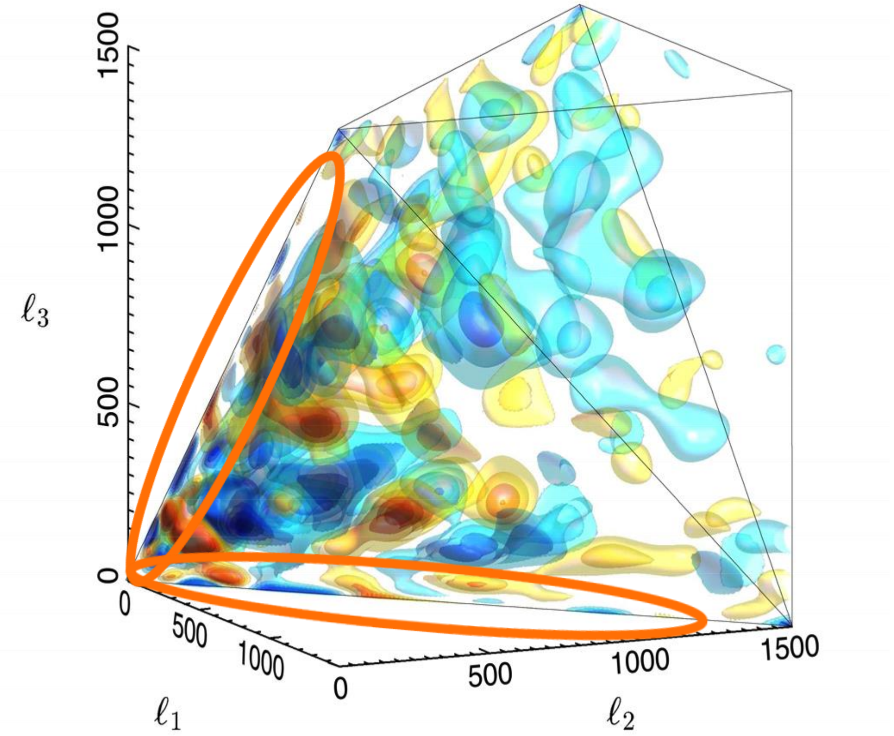
ISW and lensing are correlated (both produced by structures at low redshift) and generate a non-vanishing bispectrum in the CMB

ISW-lensing bispectrum



ISW-lensing bispectrum





ISW-lensing detected at $\sim 4\sigma$ C.L. in Planck data. Signal amplitude fully consistent with ΛCDM. Independent probe of cosmic-acceleration.

Non-Gaussianity in Large Scale Structure

- The study of NG features in LSS allows us in principle too:
  - Test models of gravitational collapse on non-linear cosmological scales
  - Improve power spectrum contraints on cosmological parameters
  - Constrain primordial NG, improving over CMB bounds. Favourable signal scaling due to more bispectrum triangles in the 3D galaxy density field, including small scales, than in the 2D CMB anisotropy field.
- Many complications, compared to previous CMB analysis:
  - In the strongly non-linear regime, there is potential information in all higher-order cumulants
  - Coupling between non-linear scales makes creates a complex, hard to model, covariance structure between bispectrum triangles and/or other statistics. In principle, NG likelihoods
  - If one is specifically interested in primordial NG, this is now a tiny bispectrum signal, about 1000 time smaller than the NG signature from gravitational instability

“Simulation-based” inference

- One way to address the difficulty in analytically modeling the strongly non-linear regime is to rely on large sets mock realizations of the matter/halo/galaxy density field.

1. Generate tens of thousands of realizations of the density field, for a fiducial cosmological model, which should be close to the actual maximum likelihood.

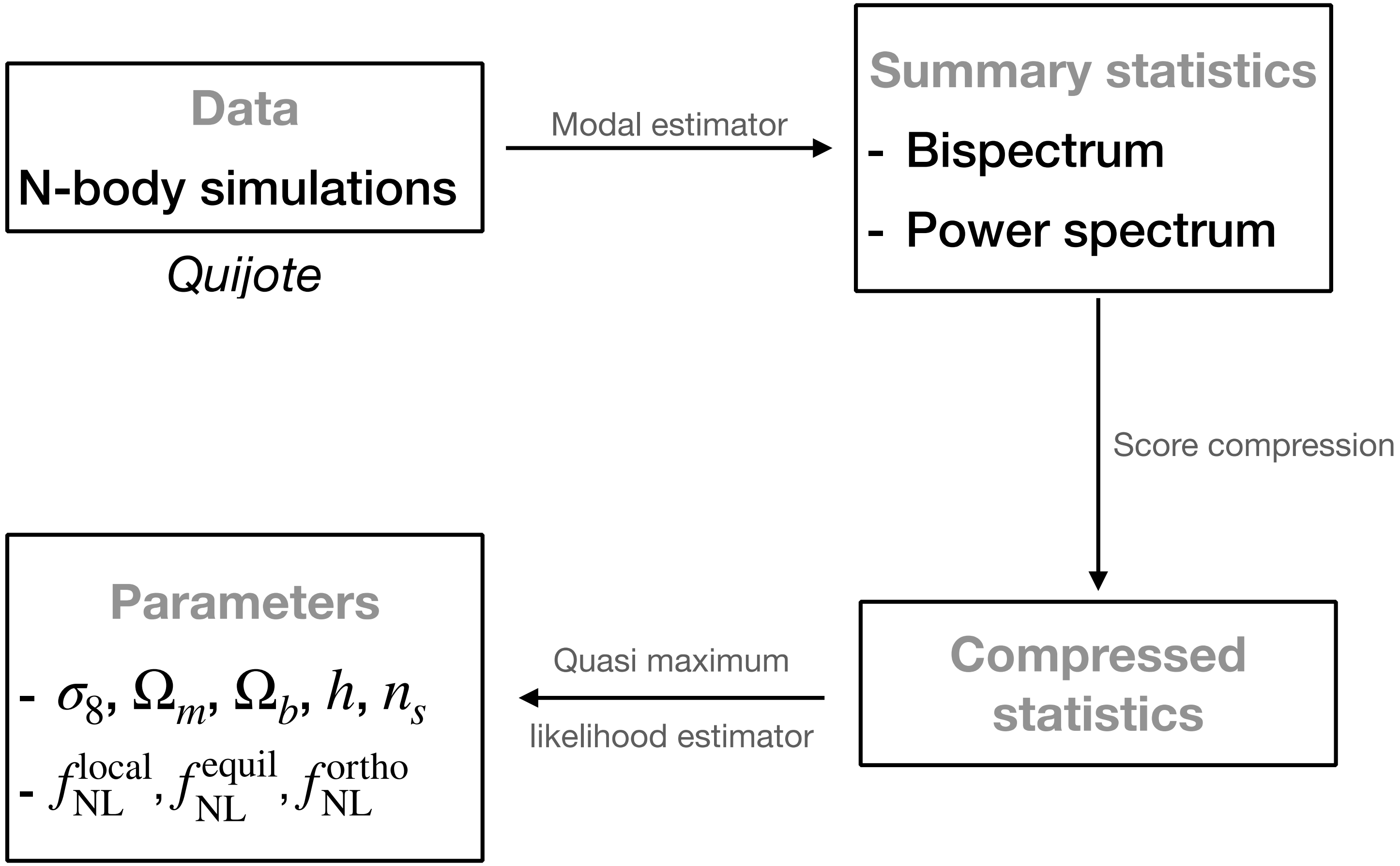
2. Choose a set of summary statistics that retain as much information as possible about your parameters, while compressing the data. Extract these statistics for each realization in your simulated dataset.

3. Compute the covariance of your summaries and the response to changes in parameters, via Monte Carlo average over the mocks.

4. Look for a further data compression scheme for your summary statistics, as lossless as possible for the parameters of interest. Typically, you compress all your starting modes into a set of N numbers, where N is the number of parameters

5. Build the covariance matrices and the find the response of your compressed statistics to changes of parameters. Use these quantities to estimate parameters.

Joint power spectrum – bispectrum estimation of cosmological parameters



Data: N-body simulations. Quijote suite

- Quijote simulations, Gaussian initial conditions (f_{NL} =0)

  https://quijote-simulations.readthedocs.io/ (***F.Villaescusa Navarro***)

  - Large suite of 44000 N-body realizations with 512^3 particles in a 1 Gpc/h side box, *Planck* fiducial cosmology
  - 8000 simulations were used to compute covariances
  - different sets of 500 simulations were used to compute numerical derivatives w.r.t. cosmological parameters $(\sigma_8, \Omega_m, \Omega_b, n_s, h)$

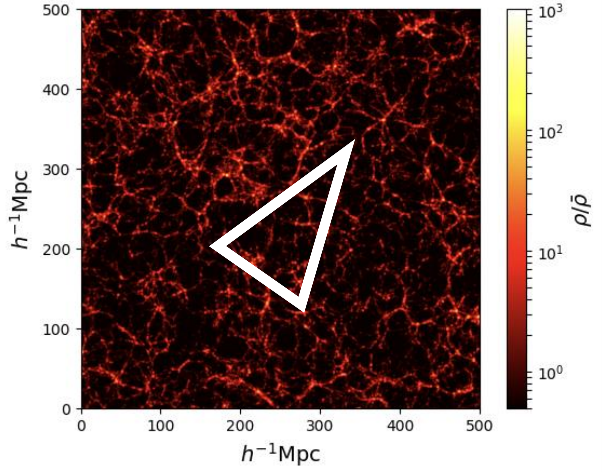
- Quijote simulations, non-Gaussian

  - Sets of 500 simulations with primordial NG conditions: local, equilateral, orthogonal
  - Numerical derivatives $(f_{NL}^{loc}, f_{NL}^{eq}, f_{NL}^{ortho})$

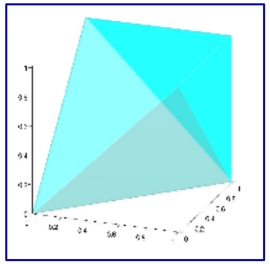
Summary statistics

- For PNG parameters (f_{NL}) we know that power spectrum and bispectrum retain most information
- The optimal choice of summaries for late-time NG is instead an open problem. In our first analysis we start from power spectrum + bispectrum.
- Need *fast* algorithm to compute bispectrum and an *efficient pre-compression step*. Extended modal algorithm from CMB
- We can compress the Quijote bispectrum information, up to k_{max} = 0.5 h/Mpc, using ~ 100 modes

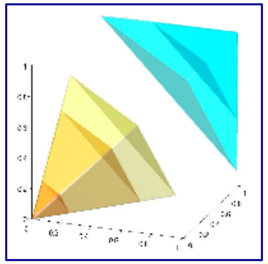
Summary statistics: bispectrum modes



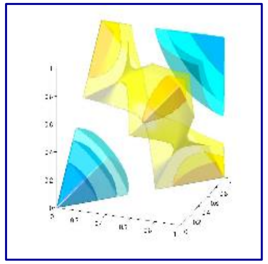
$= \beta_0$



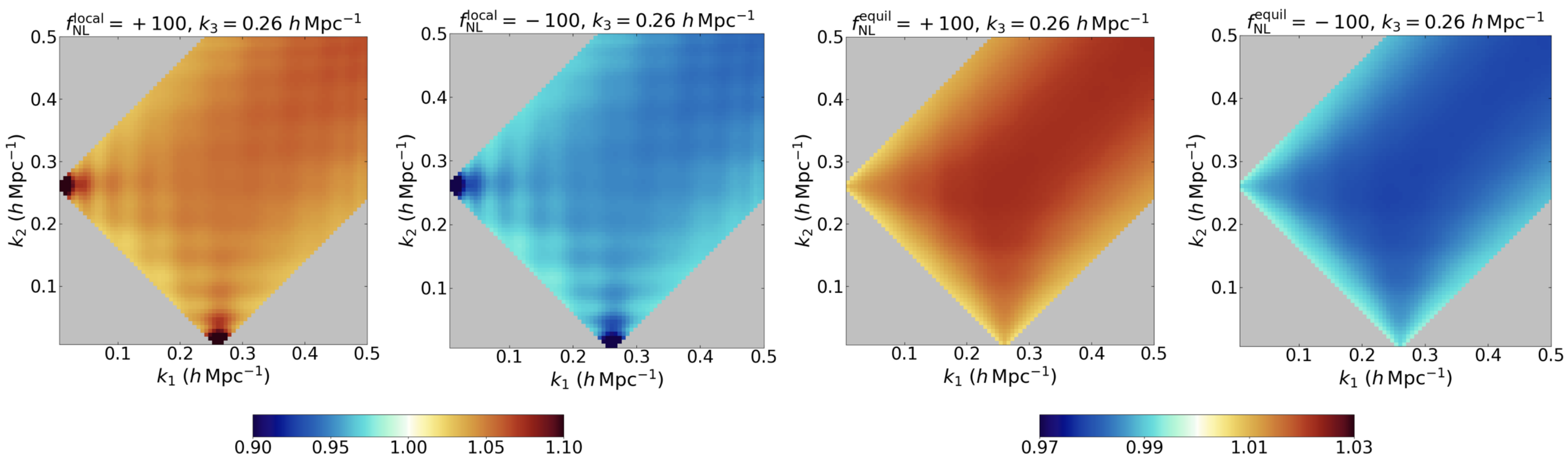
$+ \beta_1$



$+ \beta_2$



$+ \ldots$



Score function compression

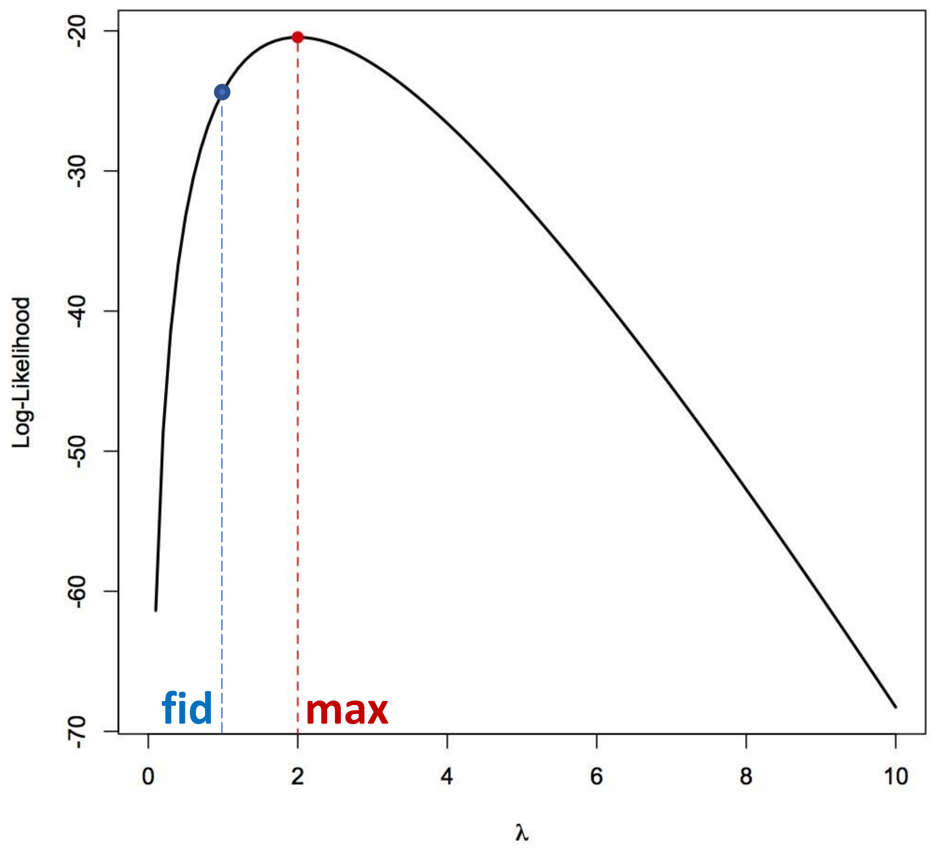
If the chosen fiducial point is close to the maximum likelihood, expand:

$$\mathcal{L} = \mathcal{L}_* + \nabla_\Theta \mathcal{L}_* - \delta\Theta \langle \nabla\nabla \mathrm{L} \rangle_* \delta\Theta + \cdots$$

Score function

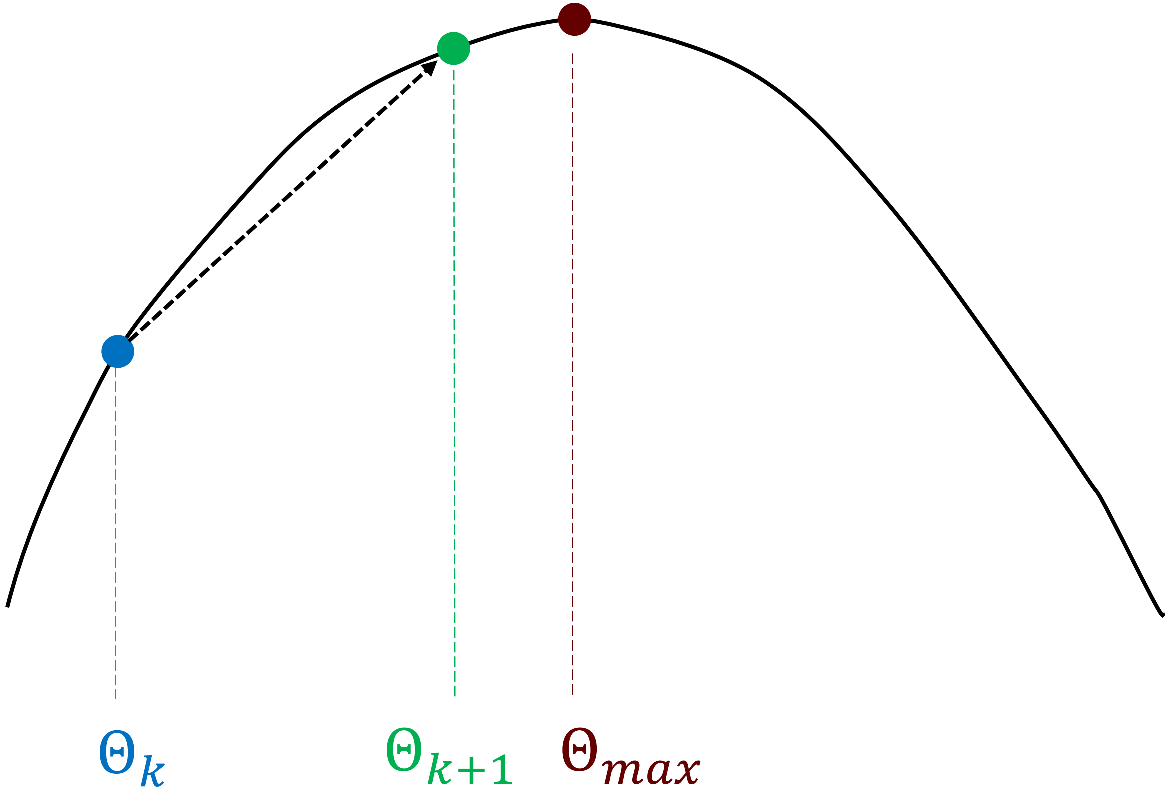
average curvature

The only parameter dependent part is the score function => compression in N statistics (N = number of params)



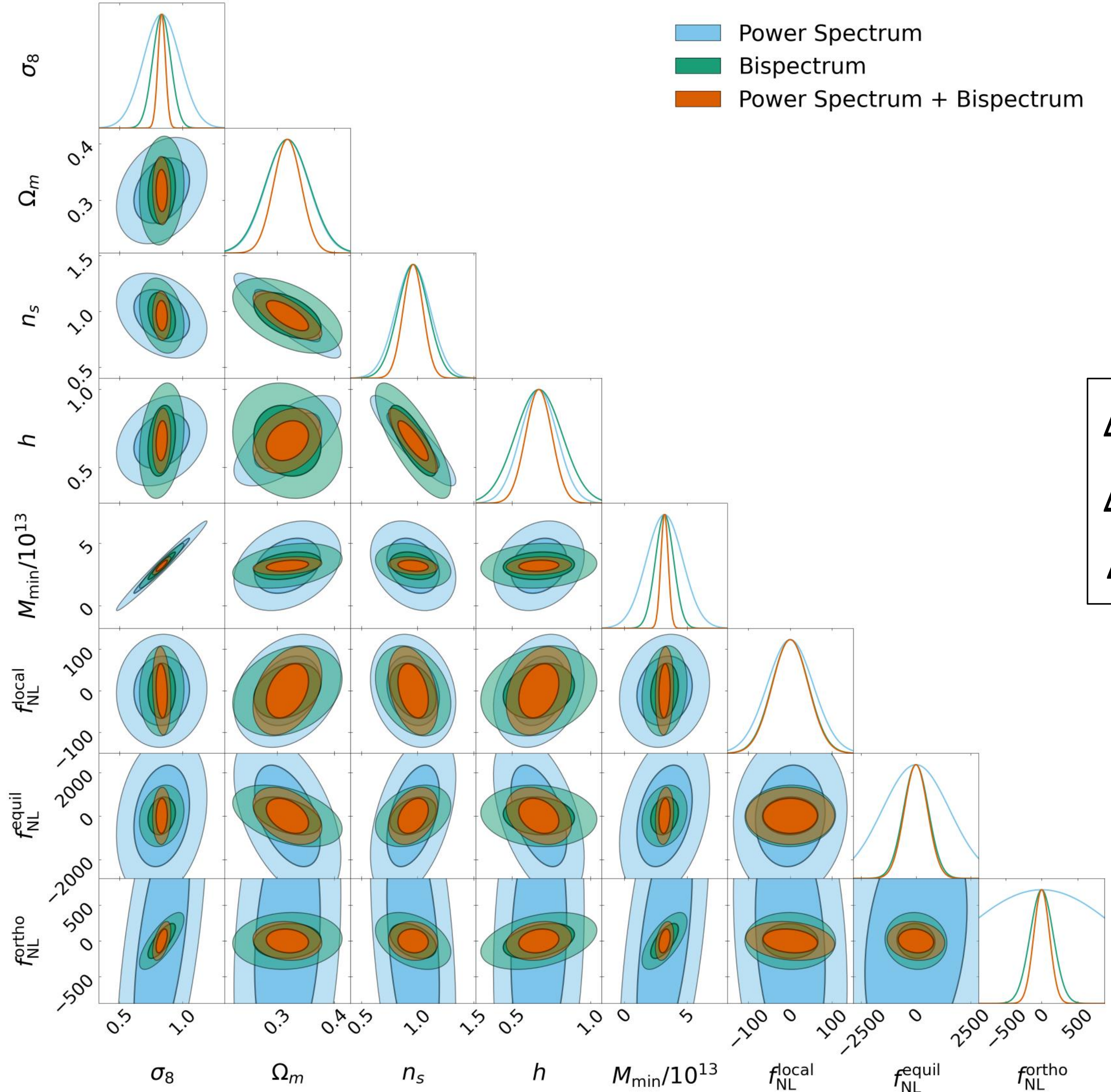
Parameter estimation

1. Compute summary statistic (e.g. bispectrum modes β_n)
2. Compute covariance via MC average (>40000 sims)
3. Compute numerical derivatives of summaries w.r.t parameters (500 simulations per parameter value)
4. Use the above to build the score function and compressed statistics
5. Build estimator
   $\widehat{\Theta}_{k+1} = \widehat{\Theta}_k + F_k^{-1} \nabla \mathcal{L}_k$



✓ **Jung, Karagiannis, Liguori, Baldi, Coulton, Jamieson, Verde, Villaescusa-Navarro, Wandelt, (2022a, 2022b), https://arxiv.org/abs/2211.07565, https://arxiv.org/abs/2206.01624**

✓ **Coulton, Villaescusa-Navarro, Jamieson, Baldi, Jung, Karagiannis, Liguori, Verde, Wandelt, (2022a, 2022b), https://arxiv.org/abs/2206.01619, https://arxiv.org/abs/2206.01619**



$$\Delta(b_\phi f_{NL}^{loc}) = 45$$

$$\Delta\left(f_{NL}^{equil}\right) = 570$$

$$\Delta\left(f_{NL}^{ortho}\right) = 110$$

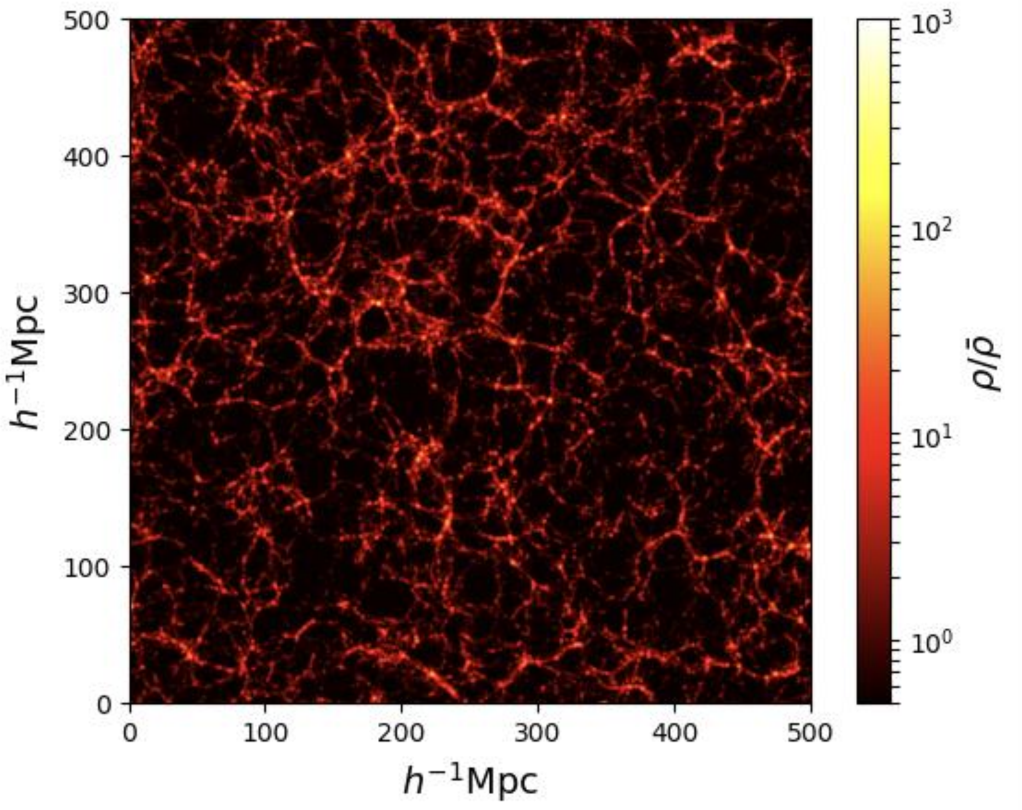
Field level analysis

- In the strongly non-linear regime there is NG information beyond the bispectrum. Higher order correlation functions might also not be the best suited statistics to extract all of it.

- One open line of research is therefore the search for additional summary statistics for optimal data compression

- Or, skip summary statistics and go for field level analysis

$\boldsymbol{\rho(\vec{x})} \rightarrow$ **parameters**



Neural network

f_{NL}

Graph Neural Network

- Dark matter halos are nodes in a graph. Each halo is labeled by a vector defining its physical properties (mass, position, velocity, concentration…) [P. Villanueva-Domingo and F. Villaescusa Navarro 2022, P. Villanueva Domingo et al. 2021, H. Shao et al. 2023]

- Nearby halos are connected by edges

- Update properties of a node using those of nearby nodes via MLP

- Classify the graph: associate the properties of the various nodes to some overall label (parameters to measure)

- Moment network: predicts posterior mean and variance

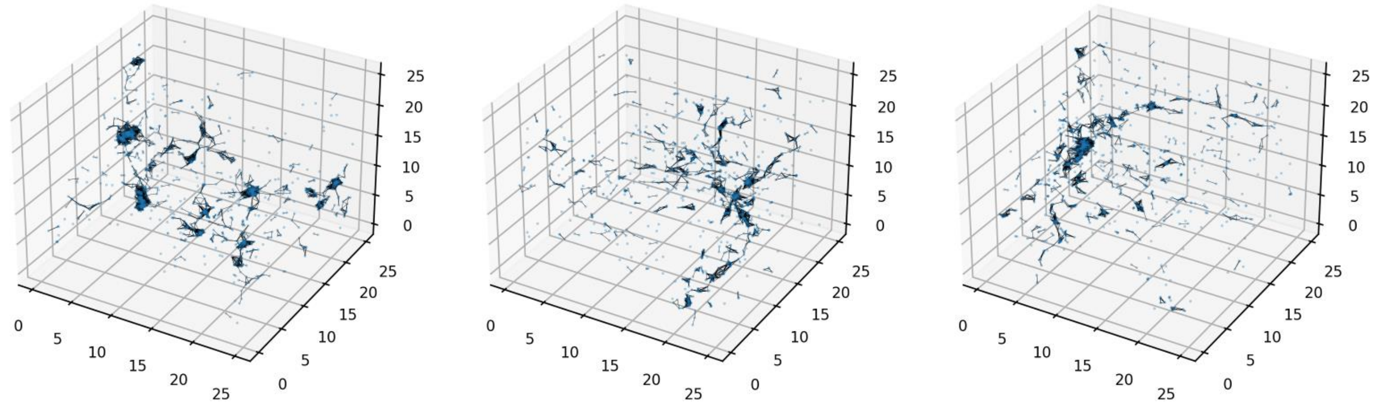


Figure from arXiv: 2204.13713

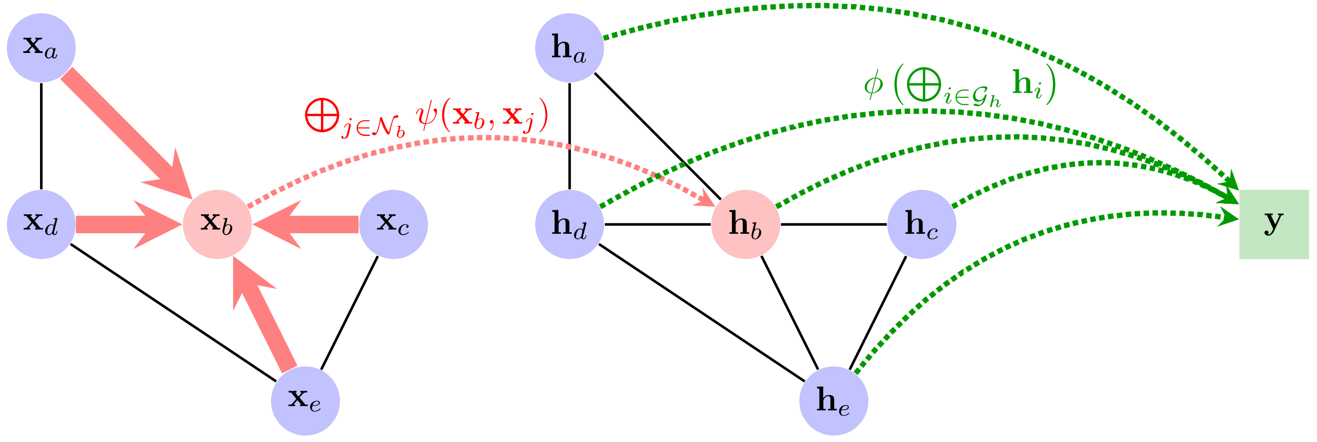


Figure from arxXiv: 2111.08683

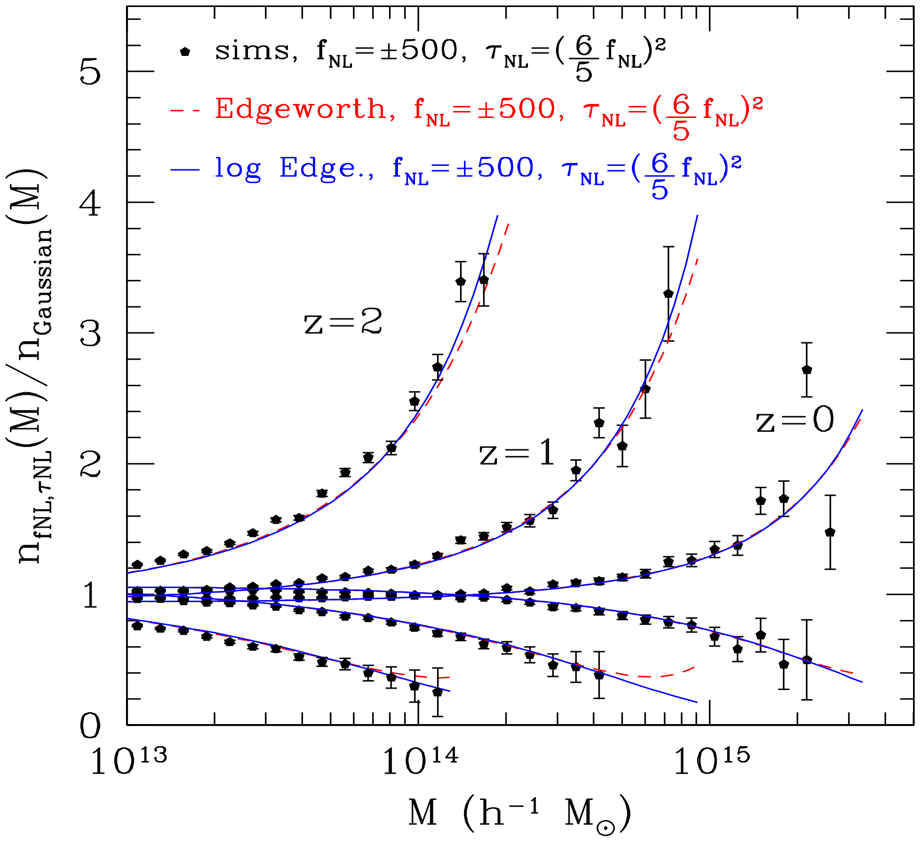
Preliminary field level analysis, HMF and NG

- A GNN was trained on a set of 1000 simulations with $-300 < f_{NL} < 300$. *All other parameters fixed*. Final error $\sigma_{fNL} \sim 35$

- A *nearly identical* performance was obtained by removing all information on the position (hence, clustering) of the halos from the NN. *All the important information, in this exercise, comes from HMF*

- The sensitivity of the HMF on primordial NG was known in previous literature. Non-Gaussian initial conditions skew the distribution of the initial perturbation field and increase the probability of forming high mass halos.

- Parameter degeneracies are crucial. We investigated the impact of the HMF by including it as an additional summary statistic in the previous analysis



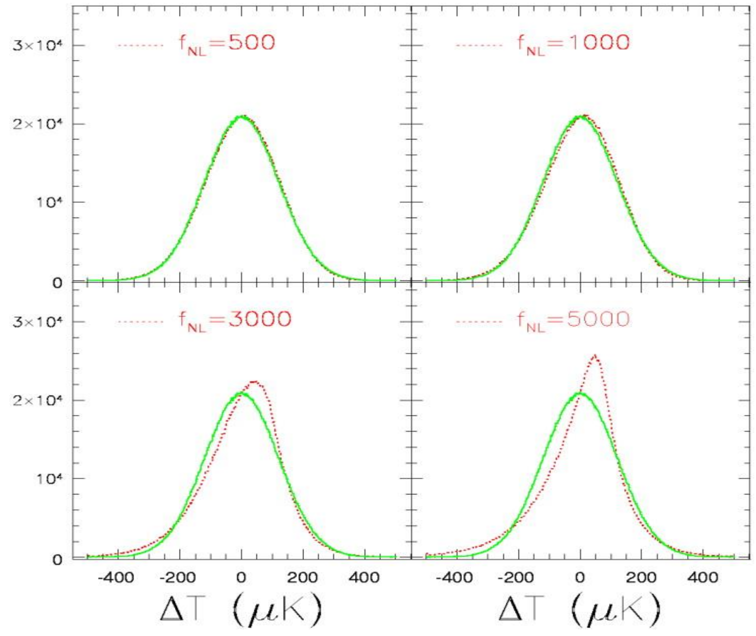
Loverde and Smith 2011

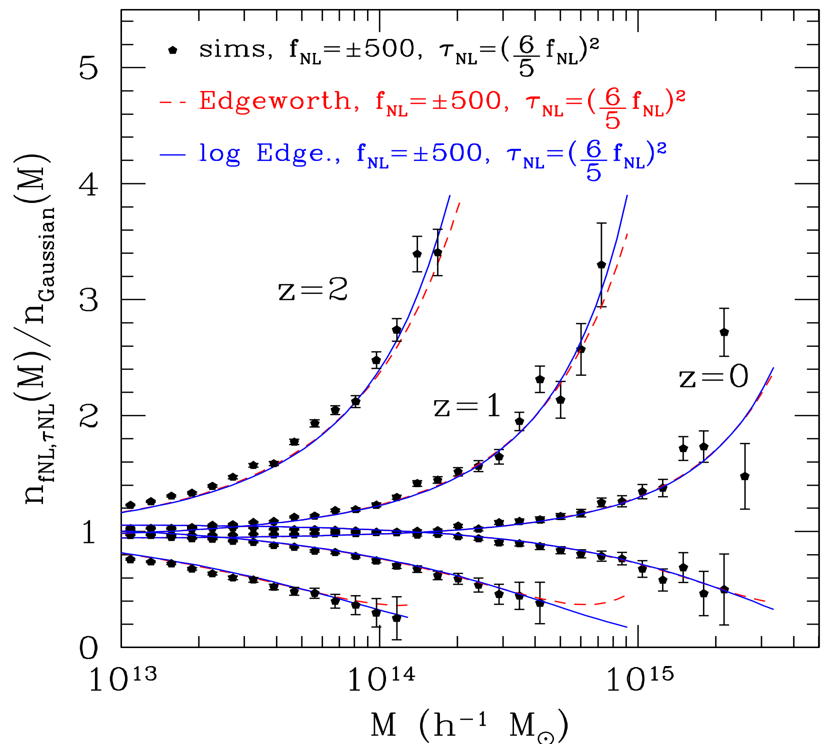
Halo mass function and NG

- Preliminary result from NN: when we fix all parameters except local f_{NL} , an improvement by a factor ~ 2 is achieved by using *only* the masses of different halos (no clustering information!)

- That would mean that the most relevant summary statistic is the histogram N_{halos} vs. Mass, i.e. the Halo Mass Function (HMF)

- The sensitivity of the HMF on primordial NG was known in previous literature. Non-Gaussian initial conditions skew the distribution of the initial perturbation field and increase the probability of forming high mass halos. However, parameter degeneracies are crucial.

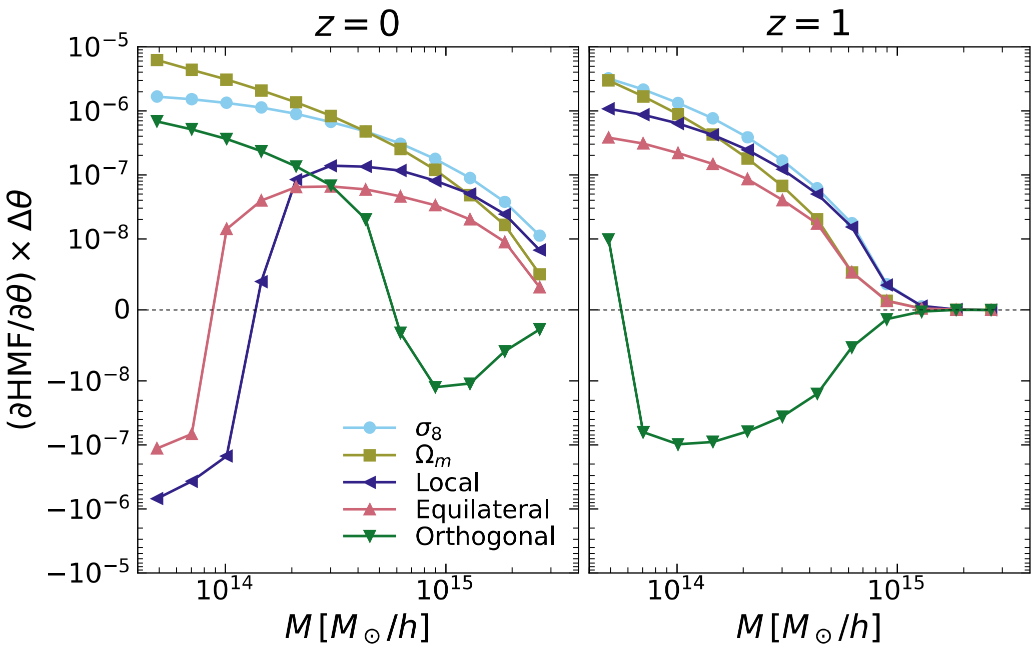


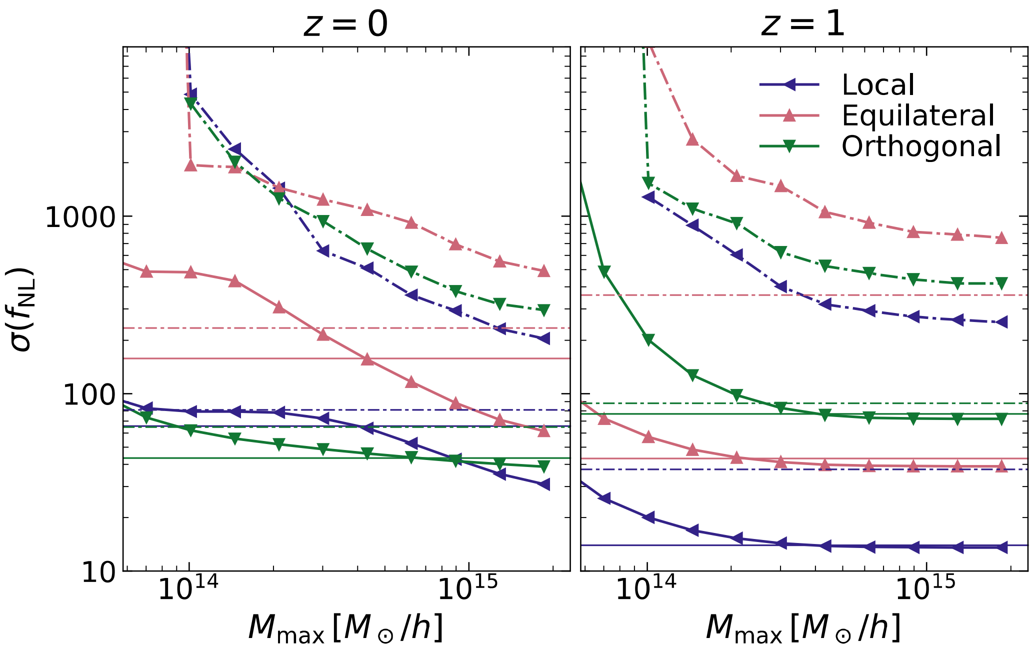


HMF as summary statistic

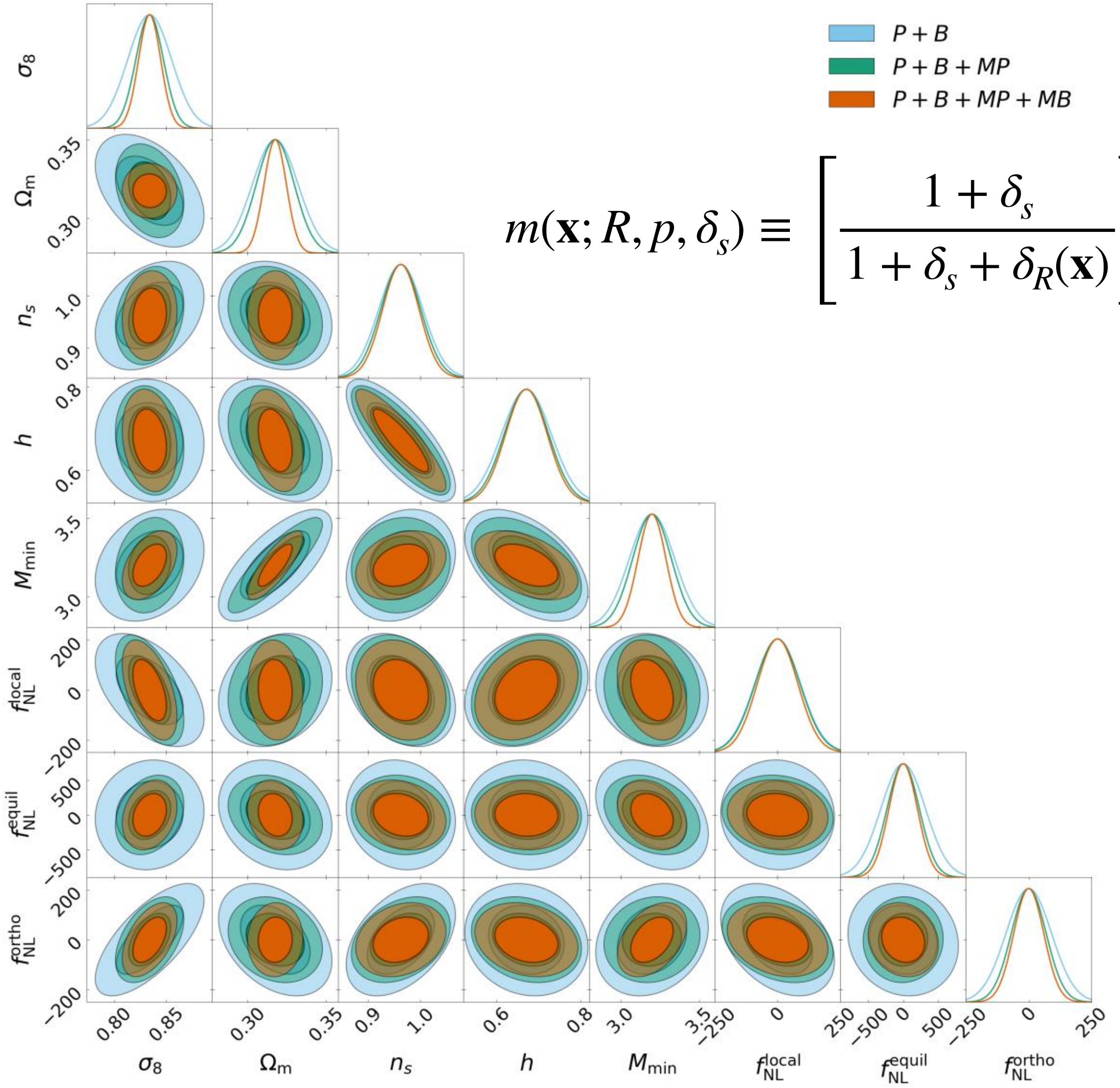
- We measured the HMF in 15 logarithmic bins with hallo masses in the range $2.0\ \times 10^{13}\,\frac{M_\odot}{h} < M < 4.6\ \times 10^{15}\ \frac{M_\odot}{h}$

- A preliminary analysis confirms the GNN findings but also shows as expected that degeneracies with σ_8, Ω_m are large

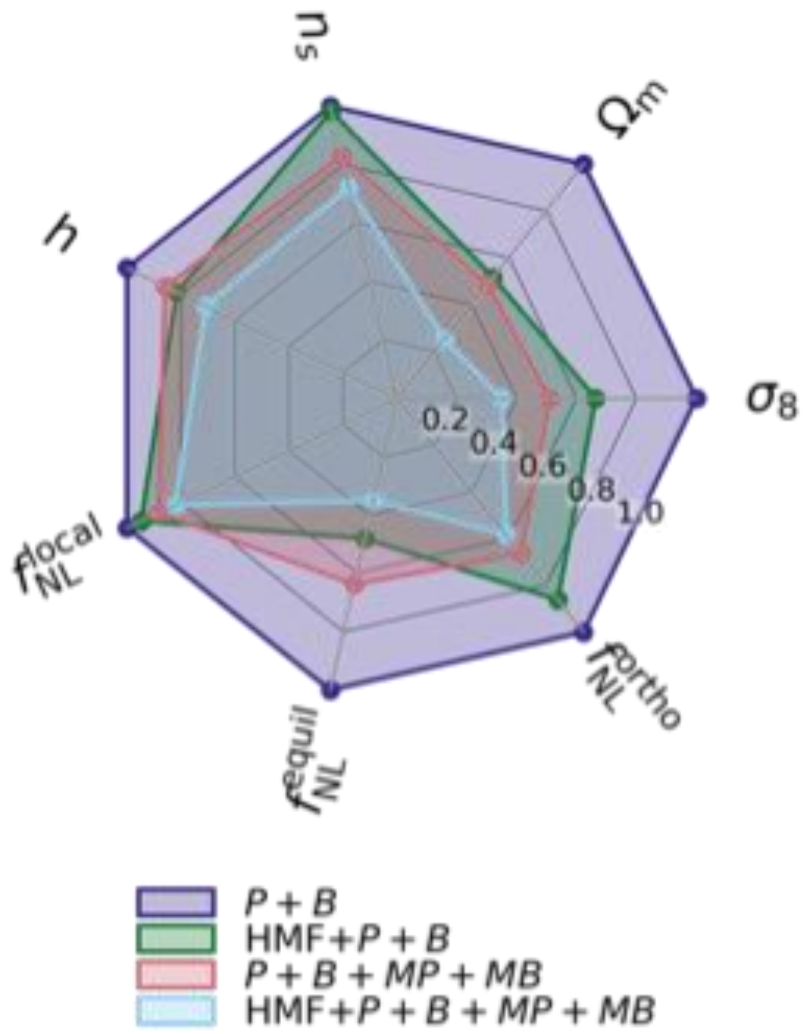




Preliminary: marked correlators



$$m(\mathbf{x}; R, p, \delta_s) \equiv \left[\frac{1+\delta_s}{1+\delta_s+\delta_R(\mathbf{x})}\right]$$



Preliminary: Molino galaxies

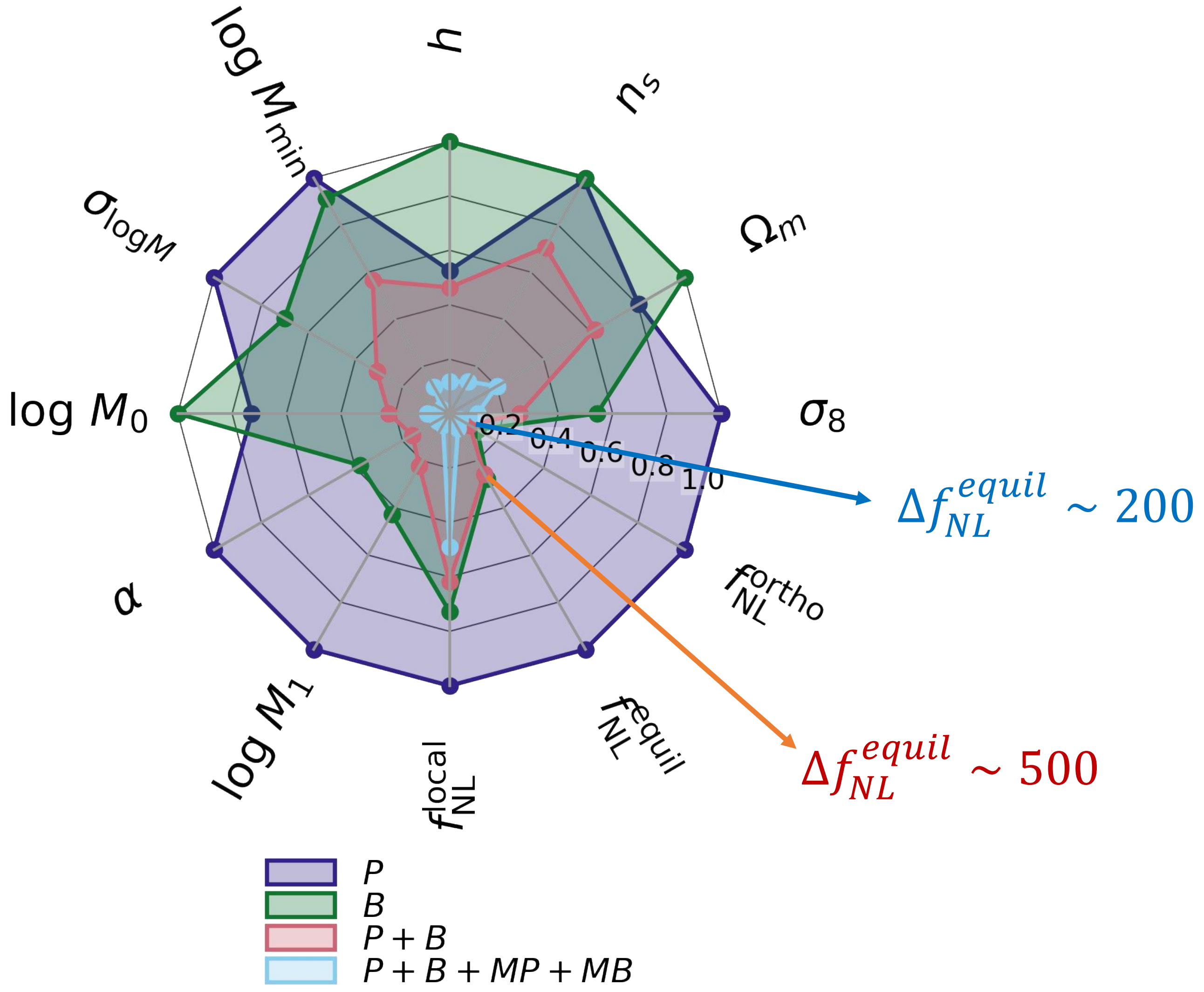
Molino galaxy catalogues

Hahn & Villaescusa-Navarro (2012.02200)

Constructed from the Quijote N-body simulations using the HOD model

Zheng, Coil & Zehavi (astro-ph/0703457)

5 parameters to describe galaxy bias



Conclusions

- Cosmological non-Gaussianity opens an important observational window, allowing us to tighten our measurements of cosmological parameters, test gravity on non-linear scales and strongly constrain Inflationary models

- The study of non-Gaussianity with the current and forthcoming big cosmological datasets is a tough statistical challenge

- In the CMB, we have constrained PNG at 0.1% level, using hundreds of millions of bispectrum configurations, via optimized compression procedures. This allowed us to constrain in turn many inflationary scenarios, but no PNG detection

- LSS observations open new big opportunities (3D vs 2D fields, many more modes) and challenges (strong NG regime). New tools and developments in Likelihood Free Inference and machine learning are currently being investigated with promising results: large gains in constraining power using small scales, hard to model analitically