Chapter 7

Non-linear filtering of large-scale structure samples

Contents

> "While o'er him fast, through sail and shroud,
> The wreathing fires made way.
> They wrapt the ship in splendour wild,
> They caught the flag on high,
> And streamed above the gallant child,
> Like banners in the sky."
> — Felicia Hemans (1826), *Casabianca*

Abstract

Due to the approximate 2LPT model implemented in the BORG algorithm, inferred large-scale structure samples are only correct in the linear and mildly non-linear regime of structure formation. This chapter describes subsequent improvement of such samples at non-linear scales, via an operation that we refer to as "non-linear filtering". This process does not replace fully non-linear large-scale structure inference, but rather fills small scales with physically reasonable information. Several approaches to non-linear filtering are considered and discussed.

This chapter discusses the generation of non-linear, constrained realizations of the late-time large-scale structure via an operation that we call "filtering" of BORG samples. It is structured as follows. We give motivation for non-linear filtering and describe two different approaches (direct improvement of final conditions, and constrained simulations) in section 7.1. For later use in chapter 8, we describe a set of samples optimally filtered with Gadget in section 7.2. In section 7.3, we describe the efficient COLA scheme for fast production of non-linear large-scale structure realizations, and apply it to generate a large ensemble of samples, used in chapter 9.

7.1 Introduction

7.1.1 Motivation for non-linear filtering of large-scale structure samples

As noted in section 4.2.1.2, the likelihood for Bayesian large-scale structure inference involves a structure formation model to translate from the initial to the final density field:

$$\delta^{\mathrm{i}} \mapsto \delta^{\mathrm{f}} = \mathcal{G}(\delta^{\mathrm{i}}, a). \tag{7.1}$$

Ideally, this step should involve a numerical model that fully accounts for the non-linearities of the Vlasov-Poisson system, which describes structure formation (see chapter 1). Unfortunately, this is not currently computationally tractable. For this reason, BORG uses 2LPT as a proxy for gravitational dynamics.[1]

Nevertheless, the description of particular patterns of the cosmic web (as presented in part IV of this thesis) requires description of the LSS not only correct at the scales correctly described by 2LPT ($k \lesssim 0.1$ Mpc$/h$, see chapter 2), but also physically reasonable at smaller scales, up to $k \sim 1$ Mpc$/h$. At this point, it is also useful to recall that the number of Fourier modes usable for cosmology scales as the cube of the smallest accessible mode, k^3.

For these reasons, data-constrained, non-linear realizations of the LSS have a large variety of applications. As noted before, constraining small, non-linear scales within the inference framework is not yet possible; therefore, such realizations will rely on fusing data-constrained large scales and unconstrained small scales that only reflect our theoretical understanding of structure formation. Throughout this thesis, we refer to the production of data-constrained, non-linear realizations, on the basis of BORG large-scale structure samples, as *non-linear filtering*.

7.1.2 Filtering in the final conditions

One possible way to perform non-linear filtering is to directly improve the final conditions produced as BORG outputs. The technique of remapping Lagrangian perturbation theory can be useful in this context: as demonstrated in chapter 6, it cheaply yields improvements of density fields in the mildly non-linear regime. A particular advantage of remapping is its very low computational cost, which allows to process a large number of samples.[2] As seen in chapters 4 and 5, this is crucial for adequate uncertainty quantification.

7.1.3 Filtering via constrained simulations

Another idea is to capitalize on the inference of the initial conditions by BORG. Starting from inferred density fields, which contain the data constraints (see in particular section 5.3.3 for a discussion of information transport), it is possible to go forward in time using an alternative structure formation model, noted $\mathcal{G}_{\rm NL}$, that improves upon $\mathcal{G}$ for the description of small scales structures:

$$\delta^{\rm i} \mapsto \delta^{\rm f}_{\rm NL} = \mathcal{G}_{\rm NL}(\delta^{\rm i}, a). \tag{7.2}$$

This process is known in the literature as running *constrained simulations*. Final density fields $\delta^{\rm f}_{\rm NL}$ constructed in this way agree with corresponding BORG final conditions $\delta^{\rm f}$ at large scales, but are also physically reasonable at smaller scales, up to the validity limit of $\mathcal{G}_{\rm NL}$.

In this picture, interesting questions are the determination of the smallest scale influenced by the data and the characterization of the reliability of structures extrapolated in unobserved regions, at high redshift or near survey boundaries. An upcoming publication will investigate the validity of constrained simulations, in particular the strength of data constraints in domains or at scales that have not been considered in the inference scheme.

In the following, we examine two particular cases for $\mathcal{G}_{\rm NL}$, corresponding to the GADGET-2 cosmological code (section 7.2) and to the fast COLA scheme (section 7.3).

7.2 Fully non-linear filtering with Gadget

This section draws from section II.B. in Leclercq *et al.* (2015).

Optimal non-linear filtering of BORG results is achieved when $\mathcal{G}_{\rm NL}$ fully accounts for non-linear gravitational dynamics. This is the case when a cosmological simulation code is used. For the purpose of this thesis, we consider that non-linear filtering of BORG results with the GADGET-2 cosmological code (Springel, Yoshida & White, 2001; Springel, 2005) is optimal.

For a variety of later uses, in particular for inference of dark matter voids in the Sloan volume (chapter 8), we generate a set of such optimally filtered, data-constrained realizations of the present large-scale structure.

[1] For the record, a BORG run, using 2LPT, takes of the order of a year (wall-clock time).

[2] The computational cost for remapping all the outputs of a BORG run, about 10,000 samples, would be comparable to a few full-gravity dark matter simulations using GADGET-2.

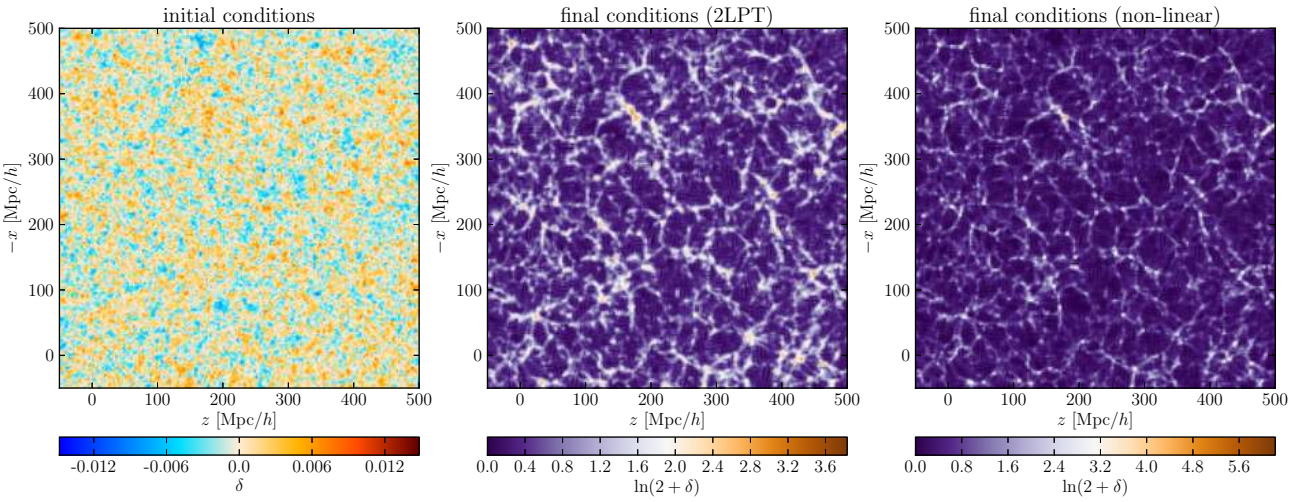

Figure 7.1: Non-linear filtering of BORG results. Slices through one sample of initial (left panel) and final density fields (middle panel) inferred by BORG. The final density field (middle panel) is a prediction of the 2LPT model used by BORG. On the right panel, a slice through the data-constrained realization obtained with the same sample via non-linear filtering (fully non-linear gravitational structure formation starting from the same initial conditions) is shown.

To do so, we rely on a subset of statistically independent initial conditions realizations, provided by Jasche, Leclercq & Wandelt (2015) (see chapter 5). The initial density field, defined on a cubic equidistant grid with side length of 750 Mpc/h and 256^3 voxels, is populated by 512^3 dark matter particles placed on a regular Lagrangian grid. The particles are evolved with 2LPT to the redshift of $z = 69$, followed by a propagation with GADGET-2 from $z = 69$ to $z = 0$. In this fashion, we generate fully non-linear, data-constrained reconstructions of the present-day large-scale dark matter distribution.

As discussed in section 7.1, final conditions inferred by BORG are accurate only at linear and mildly non-linear scales. Application of fully non-linear dynamics to the corresponding initial conditions acts as an additional filtering step, extrapolating predictions to unconstrained non-linear regimes. In a Bayesian approach, this new information can then be tested with complementary observations in the actual sky for updating our knowledge on the Universe.

An illustration of the non-linear filtering procedure is presented in figure 7.1.[3] By comparing initial and final density fields, one can see correspondences between structures in the present Universe and their origins. Comparing final density fields before and after filtering (middle and left panels), one can check the conformity of the linear and mildly non-linear structures at large and intermediate scales, correctly predicted by 2LPT. Small-scale structures, corresponding to the deeply non-linear regime, are much better represented after non-linear filtering (resulting particularly in sharper filaments and clusters). N-body dynamics also resolves much more finely the substructure of voids – known to suffer from spurious artifacts in 2LPT, namely the presence of peaky, overdense spots where there should be deep voids (Sahni & Shandarin, 1996; Neyrinck, 2013; Leclercq *et al.*, 2013; see also chapter 2) – which is of particular relevance for the purpose of inferring dark matter voids (see chapter 8).

The improvement introduced by non-linear filtering at the level of two-point statistics is presented in figure 7.2, where we plot the power spectra of dark matter density fields at $z = 0$. The agreement between unconstrained and constrained realizations at all scales can be checked. The plot also shows that our set of constrained reconstructions contains the additional power expected in the non-linear regime[4], up to $k \approx 0.4$ (Mpc/h)$^{-1}$.

[3] In figure 7.1 and in all slice plots of the rest of this thesis, we keep the coordinate system of Jasche, Leclercq & Wandelt (2015), also used in chapter 5.

[4] Note that the lack of small scale power in GADGET and COLA with respect to theoretical predictions, for $k \gtrsim 0.5$ (Mpc/h)$^{-1}$, is a gridding artifact due to the finite mesh size used for the analysis. This value corresponds to around one quarter of the Nyquist wavenumber.

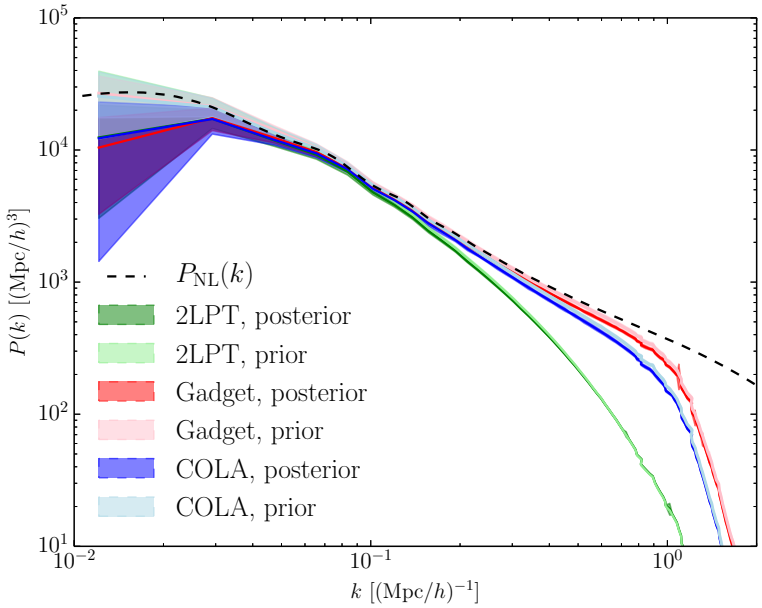

Figure 7.2: Power spectra of dark matter density fields at redshift zero, computed with a mesh size of 3 Mpc/h. The particle distributions are determined using: 1,000 unconstrained 2LPT realizations ("2LPT, prior"), 4,473 constrained 2LPT samples inferred by BORG ("2LPT, posterior"), 11 unconstrained GADGET-2 realizations ("Gadget, prior"), 11 constrained samples inferred by BORG and filtered with GADGET-2 ("Gadget, posterior"), 1,000 unconstrained COLA realizations ("COLA, prior"), 1,097 constrained samples inferred by BORG and filtered with COLA ("COLA, posterior"). The solid lines correspond to the mean among all realizations used in this work, and the shaded regions correspond to the 2-σ credible interval estimated from the standard error of the mean. The dashed black curve represents $P_{\rm NL}(k)$, the theoretical power spectrum expected at $z = 0$ from high-resolution N-body simulations.

7.3 Fast non-linear filtering with COLA

For means of uncertainty quantification within large-scale structure inference, it is necessary to process a large number of samples. Unfortunately, optimal non-linear filtering with GADGET-2 is too expensive for the $\sim 10,000$ samples of a single BORG run. However, an approximate model for non-linear structure formation, correct up to scales of a few Mpc/h, is enough for our purposes, as long as the approximation error is controlled and quantified.

7.3.1 The COLA method

The COLA (COmoving Lagrangian Acceleration, Tassev, Zaldarriaga & Eisenstein, 2013; Tassev *et al.*, 2015) technique offers a cheap way to perform non-linear filtering of a large number of BORG samples. A particular advantage (in opposition to standard particle-mesh codes) is its flexibility in trading accuracy at small scales for computational speed, without sacrificing accuracy at the largest scales.

The general idea of COLA is to use our analytic understanding of structure formation at large scales via LPT, and to solve numerically only for a subdominant contribution describing small scales. Specifically, Tassev & Zaldarriaga (2012c) propose to expand the Lagrangian displacement of particles as

$$\boldsymbol{\Psi}(\mathbf{x}, \tau) = \boldsymbol{\Psi}_{\rm LPT}(\mathbf{x}, \tau) + \boldsymbol{\Psi}_{\rm MC}(\mathbf{x}, \tau) \tag{7.3}$$

where $\boldsymbol{\Psi}_{\rm LPT}(\mathbf{x}, \tau)$ is the analytic displacement prescribed by LPT[5] (the ZA or 2LPT, see chapter 2) and $\boldsymbol{\Psi}_{\rm MC}(\mathbf{x}, \tau) \equiv \boldsymbol{\Psi}(\mathbf{x}, \tau) - \boldsymbol{\Psi}_{\rm LPT}(\mathbf{x}, \tau)$ is the "mode-coupling residual". Using this Ansatz, the Eulerian position is $\mathbf{x} = \mathbf{q} + \boldsymbol{\Psi}_{\rm LPT} + \boldsymbol{\Psi}_{\rm MC}$, and the equation of motion, which reads schematically (omitting constants and Hubble expansion; see equation (1.74))

$$\frac{\mathrm{d}^2 \mathbf{x}}{\mathrm{d}\tau^2} = -\nabla_{\mathbf{x}} \Phi, \tag{7.4}$$

[5] Following Tassev & Zaldarriaga (2012c), this first term can be written more generally in Fourier space as $\boldsymbol{\Psi}_\star(\mathbf{k}, \tau) = R_{\rm LPT}(k, \tau)\, \boldsymbol{\Psi}_{\rm LPT}(\mathbf{k}, \tau)$, where $R_{\rm LPT}(k, \tau)$ is a transfer function that we ignore here for simplicity.

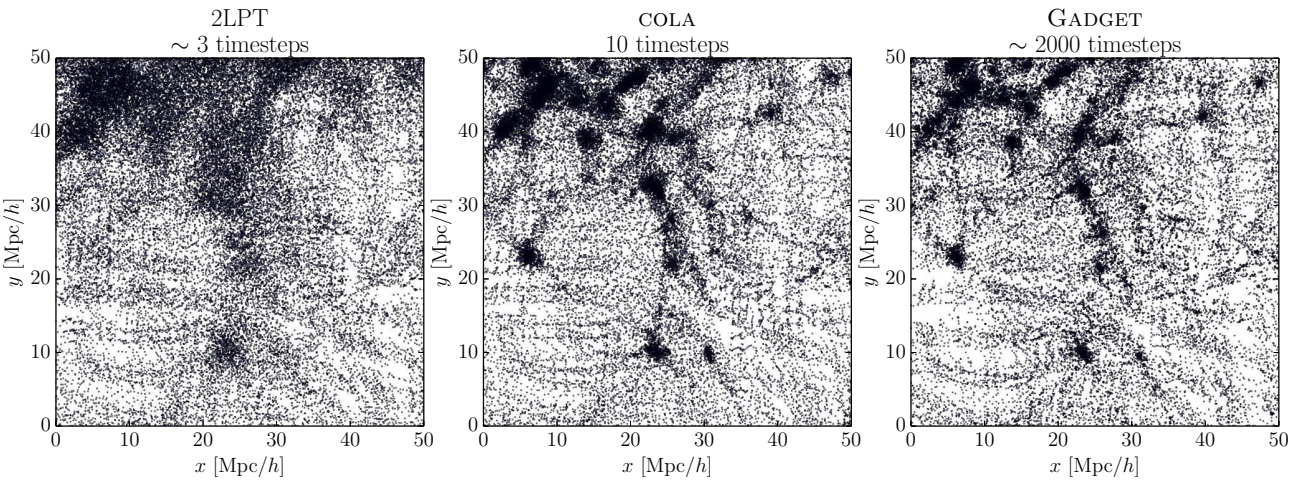

Figure 7.3: Slices through three particle realizations evolved from the same initial conditions up to $z = 0$. The particles are shown as black points. Each slice is 20 Mpc/h thick and 50 Mpc/h on the side. The left panel shows the 2LPT approximation, of computational cost roughly equivalent to 3 timesteps of a N-body code. The right panel shows the reference result obtained from GADGET-2 after $\sim$ 2000 timesteps, starting from 2LPT initial conditions at $z = 69$. The middle panel shows the result obtained with COLA with 10 timesteps, starting from 2LPT initial conditions at $z = 9$.

can be rewritten in a frame comoving with "LPT observers", whose trajectories are given by $\mathbf{\Psi}_{\rm LPT}$, as

$$\frac{\mathrm{d}^2\mathbf{\Psi}_{\rm MC}}{\mathrm{d}\tau^2} = -\nabla_{\mathbf{x}}\Phi - \frac{\mathrm{d}^2\mathbf{\Psi}_{\rm LPT}}{\mathrm{d}\tau^2}. \tag{7.5}$$

In analogy with classical mechanics, $\mathrm{d}^2\mathbf{\Psi}_{\rm LPT}/\mathrm{d}\tau^2$ can be thought of as a fictitious force acting on particles, coming from the fact that we are working in a non-inertial frame of reference.

The standard approach in PM codes (see appendix B) is to discretize the second-derivative time operator in equation (7.4). At large scales, this is nothing more than solving for the linear growth factor. Thereforce, if few timesteps are used in PM codes, the large-scale structure will be miscalculated only because of a faulty estimation of the growth factor, the exact value of which being well-known.

In contrast, the COLA method uses a numerical discretization of the operator $\mathrm{d}^2/\mathrm{d}\tau^2$ only on the left-hand side of equation (7.5) and exploits the exact analytic expression for the fictitious force, $\mathrm{d}^2\mathbf{\Psi}_{\rm LPT}/\mathrm{d}\tau^2$. The equation solved by COLA, equation (7.5), is obviously equivalent to (7.4). However, as demonstrated by Tassev, Zaldarriaga & Eisenstein (2013), using this framework requires a smaller number of timesteps to recover accurate particle realizations. In particular, they show that as few as 10 timesteps from $z = 9$ to $z = 0$ are sufficient to obtain an accurate description of halo statistics up to halos of mass 10^{11} $\mathrm{M}_\odot/h$, without resolving the internal dynamics of halos. Concerning the description of the large-scale matter density field, 10 COLA timesteps achieve better than 95% cross-correlation with the true result up $k \sim 2$ Mpc/h.

As an illustration of the performance of COLA, we show slices through corresponding 2LPT, COLA and GADGET particle realizations in figure 7.3. The simulations contain 512^3 particles in a 750 Mpc/h cubic box with periodic boundary conditions. Forces are calculated on a PM grid of 512^3 cells. The initial conditions are generated with 2LPT at a redshift of $z = 69$ for GADGET and $z = 9$ for COLA.

7.3.2 Non-linear BORG-COLA realizations

This section draws from section II.B. in Leclercq, Jasche & Wandelt (2015c).

In chapter 9, we use an ensemble of 1,097 large-scale structure realizations produced via non-linear filtering of BORG samples with COLA. The initial density field, defined on a cubic equidistant grid with side length of 750 Mpc/h and 256^3 voxels, is populated by 512^3 particles placed on a regular Lagrangian lattice. The particles are evolved with 2LPT to the redshift of $z = 69$ and with COLA from $z = 69$ to $z = 0$. The final density field is constructed by binning the particles with a CiC method on a 256^3-voxel grid. This choice corresponds to a resolution of around 3 Mpc/h for all the maps described in chapter 9. In this fashion, we generate a

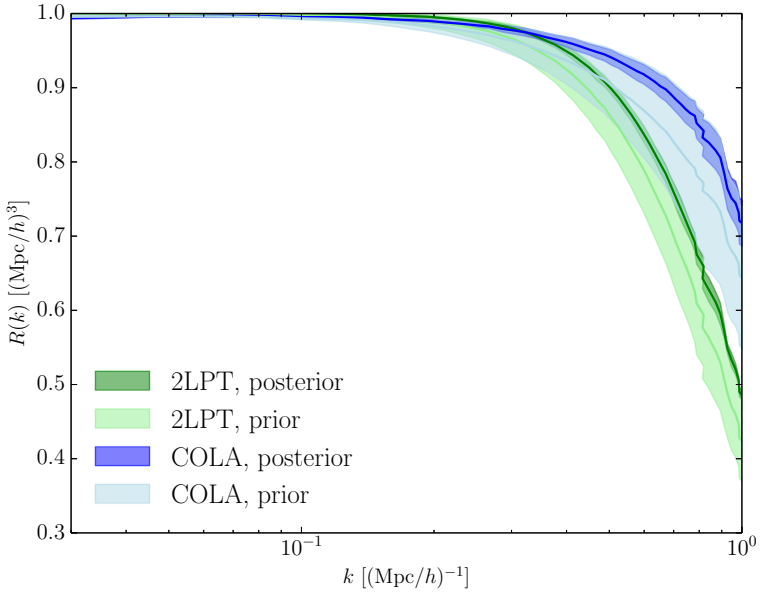

Figure 7.4: Cross-correlations between density fields at redshift zero, computed with a mesh size of 3 Mpc/h. The reference fields are the result of GADGET-2. The lines correspond to the cross-correlation between unconstrained 2LPT realizations and corresponding simulations ("2LPT, prior"), constrained 2LPT samples inferred by BORG and corresponding optimal filtering ("2LPT, posterior"), unconstrained COLA realizations and corresponding simulations ("COLA, prior"), constrained BORG-COLA samples and corresponding optimal filtering ("COLA, posterior"). In each case, we use 11 constrained or unconstrained realizations. The solid lines correspond to the mean among all realizations used in this work, and the shaded regions correspond to the 2-σ credible interval estimated from the standard error of the mean.

large set of data-constrained reconstructions of the present-day dark matter distribution (see also Lavaux, 2010; Kitaura, 2013; Heß, Kitaura & Gottlöber, 2013; Nuza *et al.*, 2014). To ensure sufficient accuracy, 30 timesteps logarithmically-spaced in the scale factor are used for the evolution with COLA.

COLA enables us to cheaply generate non-linear density fields at the required accuracy, as we now show. The power spectrum of non-linear BORG-COLA realizations is shown in figure 7.2 in comparison to that of unconstrained realizations and to samples optimally filtered with GADGET-2. In figure 7.4, we plot the cross-correlation between approximate density fields (predicted by 2LPT or by COLA) and the result of GADGET-2, for both unconstrained and constrained realizations. On these plots, it can be checked that our constrained samples, inferred by BORG and filtered with COLA, contain the additional power expected in the non-linear regime and cross-correlate at better that 95% accuracy with the corresponding fully non-linear realizations, up to $k \approx 0.4$ Mpc/h. Therefore, as for unconstrained simulations, our setup yields vanishing difference between the representation of constrained density fields with COLA and with GADGET-2, at the scales of interest of this work.