HOW THE DENSITY ENVIRONMENT CHANGES THE INFLUENCE OF THE DARK MATTER–BARYON STREAMING VELOCITY ON COSMOLOGICAL STRUCTURE FORMATION

KYUNGJIN AHN
Department of Earth Sciences, Chosun University, Gwangju 61452, Korea; kjahn@chosun.ac.kr
Received 2016 March 30; revised 2016 July 20; accepted 2016 July 31; published 2016 October 13

ABSTRACT

We study the dynamical effect of the relative velocity between dark matter and baryonic fluids, which remained supersonic after the epoch of recombination. The impact of this supersonic motion on the formation of cosmological structures was first formulated by Tseliakhovich & Hirata, in terms of the linear theory of small-scale fluctuations coupled to large-scale, relative velocities in mean-density regions. In their formalism, they limited the large-scale density environment to be that of the global mean density. We improve on their formulation by allowing variation in the density environment as well as the relative velocities. This leads to a new type of coupling between large-scale and small-scale modes. We find that the small-scale fluctuation grows in a biased way: faster in the overdense environment and slower in the underdense environment. We also find that the net effect on the global power spectrum of the density fluctuation is to boost its overall amplitude from the prediction by Tseliakhovich & Hirata. Correspondingly, the conditional mass function of cosmological halos and the halo bias parameter are both affected in a similar way. The discrepancy between our prediction and that of Tseliakhovich & Hirata is significant, and therefore, the related cosmology and high-redshift astrophysics should be revisited. The mathematical formalism of this study can be used for generating cosmological initial conditions of small-scale perturbations in generic, overdense (underdense) background patches.

Key words: cosmology: theory – dark ages, reionization, first stars – surveys

1. INTRODUCTION

The Λ-cold dark matter (ΛCDM) scenario, combined with the theory of cosmic inflation, is the successful, contemporary model describing the past and present of our universe, consistent with a wide range of observations. In this scenario, cosmological structures grow out of an extremely uniform density field but with tiny fluctuations that are seeded by cosmic inflation. The growth of the CDM density fluctuations and that of baryon density fluctuations are not in perfect synchronization, because baryons were tightly coupled to photons before the epoch of recombination and thus their motion was different from the motion of CDM which only reacts to gravity. Only after recombination did baryons gradually decouple from photons, and follow the motion of the CDM under gravity.

Cosmological observations have verified the ΛCDM scenario at scales large enough to make the baryonic physics almost irrelevant (e.g., Komatsu et al. 2011; Reichardt et al. 2012; Planck Collaboration et al. 2015). However, once in the regime where the baryonic physics becomes important, the growth of baryon fluctuations is affected by hydrodynamics and the growth of the CDM is affected by the gravitational feedback from the baryon fluctuations. Some of the usual assumptions that are made for treating very large scales, therefore, should be treated with caution or modified when treating relatively small scales. For example, Naoz & Barkana (2005) improved on the previous estimation of the linear density power spectrum at small scales by replacing the usual assumption made in cosmology, that the sound speed of baryons is uniform in space, with the fact that the sound speed fluctuates in space at small scales. They showed that more than ~10% change occurs in the baryon density power spectrum and even more in the baryon temperature power spectrum. The mere inclusion of the sub-dominant, yet non-negligible, baryonic component in the analysis changes the prediction on the matter density power spectrum by a few percent even at large scales, as was shown in the framework of high-order perturbation theory (Shoji & Komatsu 2009; Somogyi & Smith 2010).

Similarly, the relative velocity (“streaming velocity”) between baryons and the CDM after recombination should also be considered carefully in cosmology. Tseliakhovich & Hirata (2010, TH hereafter), for the first time, properly calculated the growth of small-scale density fluctuations under the influence of the streaming velocity. Small-scale fluctuations are coupled to large-scale streaming velocity fields which are coherent over a scale of a few comoving Mpc. This has its own spatial fluctuation with ~30 km s⁻¹ standard deviation at the epoch of recombination, and then decays in proportion to the inverse of the scale factor a. Its impact on the small-scale structure formation is non-negligible because the streaming velocity remains supersonic (until the intergalactic medium (IGM) is strongly heated). This then leads to the suppression of small-scale matter density fluctuations. The wave-modes that are the most strongly affected are around k ∼ 200 Mpc⁻¹, and the impact is on the mass power spectrum, the conditional mass function, and the halo bias parameter, to name a few (TH).

Subsequent studies have considered the impact of the streaming velocity from the perspective of both cosmology and astrophysics. The boost in amplitude and the shift of the peak of the baryonic acoustic oscillation (BAO) feature due to the streaming velocity, in gas intensity mappings or galaxy surveys, have been intensively investigated (Dalal et al. 2010; Yoo et al. 2011; McQuinn & O’Leary 2012; Blazek et al. 2015; Lewandowski et al. 2015; Slepian & Eisenstein 2015; Schmidt 2016). They find that both high-redshift (e.g., Dalal et al. 2010; McQuinn & O’Leary 2012) and low-redshift (e.g., Yoo et al. 2011) surveys will be affected. The impact of the
streaming velocity on the BAO may be separated out from the impact of the matter density itself (Slepian & Eisenstein 2015), which is important because otherwise it will become another nuisance parameter in cosmology. The astrophysical impact of the streaming velocity has been investigated with focus on the formation of nonlinear structure such as cosmological halos and stellar objects (Greif et al. 2011; Maio et al. 2011; Stacy et al. 2011; Tseliakhovich et al. 2011; Fialkov et al. 2012; Naoz et al. 2012; O’Leary & McQuinn 2012; Bovy & Dvorkin 2013; Richardson et al. 2013; Tanaka et al. 2013; Naoz & Narayan 2014; Popa et al. 2015; Asaba et al. 2016). Most studies indicate that the formation of minihalos (roughly in the mass range \( M = [10^{7}--10^{8}] M_\odot \)) and the formation of stellar objects within them are suppressed. It may induce baryon-dominated objects such as globular clusters (Naoz & Narayan 2014; Popa et al. 2015), but the actual star formation process leading to globular clusters has yet to be simulated. It may be responsible even for the generation of the primordial magnetic field (Naoz & Narayan 2013). Because minihalos are the most strongly affected among cosmological halos and are responsible for the early phase of the cosmic reionization process, how they change the high-redshift 21 cm background is also of prime interest (McQuinn & O’Leary 2012; Visbal et al. 2012; Fialkov et al. 2013). It is noteworthy that some of these numerical simulation results (Greif et al. 2011; Maio et al. 2011; Stacy et al. 2011), which are based on the initial condition generated by the usual Boltzmann solver such as the Code for Anisotropies in the Microwave Background (CAMB: Lewis et al. 2000), need to be re-examined. This is because the impact of the streaming velocity is cumulative and inherent even at \( z \sim 200 \) (McQuinn & O’Leary 2012), at which or later these simulations start with a streaming velocity implemented by hand.

The original formalism by TH has been re-investigated in terms of high-order perturbation theory in wave-number space (“k-space” henceforth), and some “missing terms” previously neglected were found important (Blazek et al. 2015; Schmidt 2016). Basically, for the large-scale modes responsible for the streaming velocity (\( k \sim [0.01--1]/Mpc \)), TH used a trivial solution for the evolution of the streaming velocity (\( V_{bc} \)) and the density (\( \Delta_c \) and \( \Delta_b \)) being overdensities of the CDM and baryons at large scale, respectively) environment: \( V_{bc} \propto a^{-1}, \Delta_c = \Delta_b = 0 \). Treating this as a new zeroth-order solution to the perturbation equations, they then examined how the perturbation at small scales grows. In doing so, they treated \( V_{bc} \) as a spatial quantity and perturbation variables at small scales as \( k \)-space quantities. This is basically a high-order perturbation theory, coupling the large-scale mode (\( V_{bc} \)) and the small-scale modes (\( \delta_c \) and \( \delta_b \), small scale CDM and baryonic overdensities, respectively). However, \( V_{bc} \) is closely linked to the fluctuation of \( \Delta_c \) and \( \Delta_b \) through the density continuity equation, and thus the trivial solution adopted by TH cannot be used for generically overdense and underdense regions. The continuity equation connects the divergence of \( V_{bc} \) to \( \Delta_c \) and \( \Delta_b \), and Schmidt (2016) finds that the divergence of \( V_{bc} \) is indeed an important term that one should not ignore. Blazek et al. (2015) also works on the generic basis of non-zero \( \Delta_c \) and \( \Delta_b \).

We improve on the formalism of TH by also considering the non-zero overdensities. To this end, unlike Blazek et al. (2015) and Schmidt (2016), we inherit the original method by TH and focus on the impact on small-scale modes: large-scale \( V_{bc} \) is treated as the spatial quantity and small-scale overdensities \( \delta_c \) and \( \delta_b \) are treated as the \( k \)-space quantity in the perturbation analysis. Most importantly, we explicitly include generically “non-zero” \( \Delta_c \) and \( \Delta_b \) as a new set of spatial quantities, and consequently the divergence of CDM and baryon velocities as well. We find that this leads to a set of mode–mode coupling terms, including the velocity divergence–density coupling. We carefully include all the coupling terms to the leading order in our perturbation analysis. We also include the baryonic physics, namely fluctuations in the sound speed (Naoz & Barkana 2005), the gas temperature and the photon temperature. Our formalism is also suitable for generating initial conditions for \( N \)-body+hydro numerical simulations. Because the new set of mode–mode couplings is imprinted in the initial condition, the initial condition generator considering the streaming-velocity effect by O’Leary & McQuinn (2012), CICSASS, should also be improved on if one were to numerically simulate the structure formation inside overdense or underdense regions.

This paper is organized as follows. We lay out the basic formalism and describe the statistics of large-scale fluctuations in Section 2. In Section 3, we show results on the matter density power spectrum, the conditional halo abundance, and the halo bias as applications of the formalism. We conclude this work in Section 4 with a summary, discussion and future prospects. Some details omitted from the main text are described in Appendices A and B.

2. FORMALISM AND NUMERICAL METHOD

2.1. Fluctuation under Non-zero Overdensity and Relative Velocity: Perturbation Formalism

We start from a set of equations for perturbations of relevant physical variables. All the \( k \) modes of interest are in sub-horizon scale such that Newtonian perturbation theory holds. Let us define the overdensity \( \delta_j \equiv (\rho_j - \bar{\rho}_j)/\bar{\rho}_j \), where the subscript \( j = \{c, b\} \) denotes either the CDM (\( c \)) or the baryonic (\( b \)) component, and \( \bar{\rho}_j = (1/a)\nabla \cdot v_j \) where \( a \) is the scale factor and \( v_j \) is the proper peculiar velocity of component \( j \). When the universe is in the regime where we can ignore fluctuations of photons and neutrinos due to their rapid diffusion after recombination, we have (e.g., Bernardeau et al. 2002; TH)

\[
\begin{align*}
\frac{\partial \delta_c}{\partial t} &= -a^{-1} \nabla \cdot v_c - a^{-1} (1 + \delta_c) \nabla \cdot \nu_c,
\frac{\partial \delta_b}{\partial t} &= -a^{-1} (\nabla \cdot \nu_c) \nu_c - a^{-1} \Delta \phi - H \nu_c,
\frac{\partial \delta_c}{\partial t} &= -a^{-1} \phi_c \frac{\nabla \delta_c}{\nabla \phi} - a^{-1} (1 + \delta_c) \nabla \cdot \nu_b,
\frac{\partial \delta_b}{\partial t} &= -a^{-1} \psi_b \frac{\nabla \delta_b}{\nabla \psi} - a^{-1} (1 + \delta_b) \nabla \cdot \nu_b,
\nabla^2 \phi &= 4\pi Ga^2 \rho_m \delta_m, \tag{1}
\end{align*}
\]

where \( a \) is the scale factor, \( \nabla \) is the gradient in the comoving frame, \( \rho_j = \rho_c \delta_c + \rho_b \delta_b \), \( \delta_m = \delta_c + \delta_b \), \( f_c = \Omega_{c,0}/(\Omega_{c,0} + \Omega_{b,0}) \), \( f_b = \Omega_{b,0}/(\Omega_{c,0} + \Omega_{b,0}) \), \( \Omega_c = \rho_{\gamma,0}/\rho_m \) is the present-day mean density of component \( j \) in the unit of the critical density, \( c_s \) is the sound speed, and \( H \) is the Hubble constant at a given redshift. Even though a usual approximation
for $c_s$ is a spatially uniform one given by

$$c_s^2 = \frac{k_B}{\mu m_H} \left( 1 - \frac{1}{3} \frac{\partial \log T}{\partial \log a} \right)$$

(2)

which assumes a mean-density environment undergoing Hubble expansion with the mean baryon temperature $T_b$, for high-$k$ modes a more accurate treatment is required (Naoz & Barkana 2005). This requires replacing the pressure term$^1$,

$$c_s^2 \nabla \delta_b \rightarrow \frac{k_B T}{\mu m_H} \nabla (\delta_b + \delta_T + \delta_b \delta_T),$$

(3)

in Equation (1) and considering another rate equation

$$\frac{\partial \delta_T}{\partial t} = \frac{2}{3} \frac{\partial \delta_b}{\partial t} + \delta_b \langle \alpha / \langle \rangle \rangle - 4 \left\{ \delta_T \left( 5 \frac{T}{T} - 4 \right) - \frac{\delta_T}{T} \right\}$$

(4)

where $\delta_T$ and $\delta_T$ are temperature fluctuations of the baryon and the photon, respectively, $T_b = 2.725 \, K (1 + z)$ is the mean photon temperature, $x_b(t)$ is the mean electron fraction at time $t$ and $t_c = 1.17 \times 10^{12}$ years. For $T_b$, we use the fitting formula of TH.

Equation (1) allows a trivial solution: $\delta_c = \delta_b = \phi = 0$, $x_b = x_{bc} = a(a/a_i)^{-1}$ and $b_b = v_b = v_{bc}(a/a_i)^{-1}$, where $a_i$ is the initial scale factor. TH took this as the zeroth-order solution of a spatial patch and developed a linear perturbation theory of small-scale modes inside the patch. The physical process is in principle a coupling of small-$k$ and large-$k$ modes (mode-mode coupling), which is beyond the linear theory where all modes are assumed to be mutually independent. TH used the fact that the relative velocity $V_{bc} = v_b - v_c$ is coherent over the length scale of a few comoving Mpc (contributed by modes with wave numbers in the range $0.01 \leq k \leq (k/{Mpc})^{-1}$), and averaged the “local” power spectra of density fluctuations over many such patches with varying $V_{bc}$.

Even though a trivial solution exists, there also exists a nontrivial solution to Equation (1), exact to the first order. This nontrivial solution is suited to describing the physics inside patches with non-zero overdensity (see also Blazek et al. 2015). In order to obtain the nontrivial solution, we first linearize Equation (1):

$$\frac{\partial \delta_c}{\partial t} = -\theta_c,$$

$$\frac{\partial \theta_c}{\partial t} = \frac{3}{2} H^2 \Omega_m (f_c \delta_c + f_b \delta_b) - 2 \theta_c,$$

$$\frac{\partial \delta_b}{\partial t} = -\theta_b,$$

$$\frac{\partial \theta_b}{\partial t} = -\frac{3}{2} H^2 \Omega_m (f_c \delta_c + f_b \delta_b) - 2 \theta_b,$$

(5)

where $\Omega_m \equiv \rho_m(a)/\rho_{crit}(a)$ is the matter content with respect to the critical density $\rho_{crit}$ at $a$, and we have ignored second-order terms and also the pressure term $a^{-1} c_s^2 \nabla \delta_b$. This is indeed a valid approximation in the wave number range $(0.01 \leq k \leq (k/{Mpc})^{-1}) \leq 1$ relevant to the coherent $V_{bc}$ (TH), where the second-order terms remain much smaller than the first-order terms and the baryonic sound speed keeps decreasing from $c_s \sim 6 \, km/s$ after recombination to make the pressure term negligible. Then, Equation (5) can be rewritten as

$$\frac{\partial \delta_c}{\partial t} = -\theta_c,$$

$$\frac{\partial \theta_c}{\partial t} = -\frac{3}{2} H^2 \Omega_m \delta_c - 2 \theta_c,$$

$$\frac{\partial \delta_b}{\partial t} = -\theta_b,$$

$$\frac{\partial \theta_b}{\partial t} = -2 \theta_b,$$

(6)

where $\delta_c = f_c \delta_c + f_b \delta_b$, $\theta_c = f_c \theta_c + f_b \theta_b$, $\delta_b = \delta_c - \delta_b$, and $\theta = \theta_c - \theta_b$. In the matter-dominated ($\Omega_m = 1$) flat universe, $\delta_c$ allows both the growing mode ($\delta_c \propto a$, $\theta_c \propto a^{-1/2}$), $v_c = f_c v_c + f_b v_b \propto a^{1/2}$ and the decaying mode ($\delta_c \propto a^{-3/2}$, $\theta_c \propto a^{-3}$, $v_c \propto a^{-2}$), $\delta_b$ allows a slowly decaying (“streaming”) mode ($\delta_b \propto a^{-1}$, $\theta_b = \theta_c - \theta_d$) and a compensated mode ($\delta_b = \text{constant}$, $\theta_b = 0$, $v_b = 0$). During $1000 \leq z \leq 50$, the non-negligible amount of the radiation component (CMB and neutrinos) makes $\Omega_m \neq 1$, and most of the simple analytical forms above become no longer intact except for $\{\theta_c, \theta_b\}$.

Using this mode decomposition, the large-scale perturbations evolve in the following form:

$$\Delta_c(a) = \{\Delta_{gro} D^c(a) + \Delta_{dec} D^d(a)\}$$

$$\Delta_b(a) = \{\Delta_{gro} D^b(a) + \Delta_{dec} D^d(a)\}$$

$$\Theta_c(a) = -a H \left\{ \Delta_{gro} \frac{d D^c(a)}{da} + \Delta_{dec} \frac{d D^d(a)}{da} \right\} + f_b \left( \Theta_{c,1} - \Theta_b \right) \left( \frac{a}{a_{i,1}} \right)^{-2},$$

$$\Theta_b(a) = -a H \left\{ \Delta_{gro} \frac{d D^b(a)}{da} + \Delta_{dec} \frac{d D^d(a)}{da} \right\} - f_c \left( \Theta_{c,1} - \Theta_b \right) \left( \frac{a}{a_{i,1}} \right)^{-2},$$

$$V_{bc}(a) = (V_{c,1} - V_{bc}) \left( \frac{a}{a_{i,1}} \right)^{-1},$$

(7)

where we used upper-case letters to denote the “background” fluctuations for each patch of a few comoving Mpc, over which we will develop the small-scale perturbation. $\Delta_{gro}$, $\Delta_{dec}$, $\Delta_{com}$, and $\Delta_{str}$ are the initial (at $z = 1000$) values of the growing, decaying, compensated, and streaming modes, respectively. $D^c(a)$, $D^d(a)$, and $D^d(a)$ are growth factors of the growing, decaying, and streaming modes, respectively, and they are all approximated as $D^c = D^d = D^d = 1$ at $z = 1000$. We describe the details of these modes in Appendix A. It is noteworthy, as is well known already, that both CDM and baryonic components tend to approach the same asymptotes $\Delta = \Delta_{gro} D^c(a)$ and $\Theta = -a H \Delta_{gro} D^c(a)/da$, indicating that baryons tend to

---

$^1$ We keep the cross-term $\delta_b \delta_T$ in Equation (3) in order to find the correct coupling of high-$k$ and low-$k$ modes in Equations (10) and (11).
move together with CDMS in time. More interestingly, $V_{bc}$ decays as $a^{-1}$ throughout the evolution at any overdensity environment even though $V_c$ and $V_b$ grow roughly as $a^{1/2}$ individually (except in regions with $\Delta_b = \Delta_b = 0$ where $V_c$ and $V_b$ decay as $a^{-1}$). This fact may seem to make the analysis by TH valid in generic overdensity environments to some extent: TH relied on the trivial solution $\Delta_a = \Delta_b = 0$, in which all velocity components decay in time such that $V_c \propto a^{-1}$, $V_b \propto a^{-1}$, and most importantly $V_{bc} \propto a^{-1}$. Because the suppression of the matter–density fluctuations ($f_c \delta_c + f_b \delta_b$) depends not on individual velocity components but on $V_{bc}$ only, different temporal behavior of individual velocity components among the trivial and generic solutions does not matter. Nevertheless, quantitative prediction by TH will be questioned in Section 3.1, because we use nontrivial solutions (Equation 7) which result in the new type of coupling of high-$k$ and low-$k$ modes in general. We also require the evolution of $\Delta_T$, which is given by Equation (4):

$$\frac{\partial \Delta_T}{\partial t} = \frac{2}{3} \frac{\partial \Delta_b}{\partial t} - \frac{x_r(t)}{t_r} a^{-4} \frac{T_i}{T} \Delta_T,$$

(8)

which evolves in conjunction with Equation (1). Here we neglect $\Delta_T$ term due to its smallness (see also the following discussion), even though we do not neglect $\Delta_T(t)$ when initializing $\Delta_T(a)$ (Section 2.2). We have found a useful fitting formula for $\Delta_T(a)$ for patches with volume (4 Mpc)$^3$:

$$\Delta_T(a) = \text{sign}(T_{r, A}) \exp[\alpha (\log a + 2.8)^{0.33}]$$
$$\exp[(\Delta_T, A)^{0.2591} |\Delta_{T, A}|^{0.32591}],$$

(9)

which provides a good fit to $\Delta_T$ at $300 \lesssim z \lesssim 5$, and for high-redshift ranges we simply ignore $\Delta_T$ altogether because of the smallness of $\Delta_T$ in general. Here $\Delta_{T, A} \equiv \Delta_T(a = 0.01) = 0.279 \Delta_0(a = 0.01)$, $\alpha \equiv \log_{10}(\Delta_{T, B}/\Delta_{T, A})/0.2805$ and $\Delta_{T, B} \equiv \Delta_T(a = 0.1) = 0.599 \Delta_0(a = 0.1)$, and an almost complete coupling of $\Delta_T(a)$ to $\Delta_T(a)$ at $300 \lesssim z \lesssim 300$ yields this simple, empirical relation to $\Delta_b(a)$. We describe this fitting formula in more detail in Appendix B. The decoupling of $\Delta_T(a)$ from the CMB is much earlier than the mean value experiences, which occurs at $z \approx 150$, because the larger the value of $k$, the earlier the decoupling occurs (see, e.g., Figure 1 of Naoz & Barkana 2005). Including $\Delta_T$ explicitly may delay this decoupling to some extent, but we leave such an accurate calculation to future work. Using the evolution equation for $\Delta_b(a)$ in Equations (7), (9) is determined solely by local values of the four modes.

Now we expand Equations (1), (3) and (4) to linear order, taking Equation (7) as the zeroth-order solution. We first define the net density, the net velocity (of the fluid component $i$), the net gravitational potential and the net baryon temperature as $\rho_i(a, x) = \bar{\rho}_i(a) (1 + \Delta_i(a, X) + \delta_i(a, x)), \nabla_{net}(a, x) = H \Delta_X + V_i(a, X) + \nabla_i(a, x), \dot{\phi}_{net}(a, x) = -(a/2) \dot{\phi}(Ha/a) + \dot{\phi}(a, x)$, and $\bar{\rho}_i = \bar{\rho}_i + \bar{T}_i (1 + \Delta_T + \delta_T)$, respectively. Here $X$ and $x$ denote the comoving-coordinate position of the center of a background patch and that of a small-scale fluid component, respectively, and we use lower-case letters for small-scale fluctuations. $\bar{T}$

is the gravitational potential sourced only by the background fluctuations such that $\nabla^2 \Phi = 4 \pi G a^2 \Omega_m (f_c \delta_c + f_b \delta_b)$ (e.g., Bernardeau et al. 2002). Similarly, $\nabla^2 \phi = 4 \pi G a^2 \Omega_m (f_c \delta_c + f_b \delta_b)$. We then have

$$\frac{\partial \delta_c}{\partial t} = -a^{-1} V_c \cdot \nabla \delta_c - (1 + \Delta_c) \dot{\theta}_c - \theta_c \delta_c,$$

$$\frac{\partial \delta_b}{\partial t} = -a^{-1} (V_b \cdot \nabla) V_b - a^{-1} \nabla \phi - \dot{H} V_c$$
$$- [a^{-1} (V_b \cdot \nabla) V_b],$$

$$\frac{\partial \delta_b}{\partial t} = -a^{-1} (V_b \cdot \nabla) V_b - a^{-1} \nabla \phi - \dot{H} V_b$$
$$- a^{-1} \frac{k_b T}{H_{\mu H}} \nabla \Delta_T - a^{-1} \Delta_T \delta_T + (1 + \Delta_T) \delta_b$$
$$- [a^{-1} (V_b \cdot \nabla) V_b],$$

where we mark the terms that can be further ignored in square brackets. First, we can safely ignore any terms containing $\delta_T$ and $\Delta_T$, because the former is negligible at $z \lesssim 1000$ compared to $\delta_T$ (see Figure 1) and the latter is just too small ($\Delta_T \lesssim 10^{-3}$ at $z = 1000$ and decaying in time) to produce any appreciable impact on the baryon temperature of high-$k$ modes. Second, the justification for ignoring $a^{-1} (V_b \cdot \nabla) V_b$ is easily seen from the viewpoint of the $k$-space. With the Fourier expansion $A(x) = \sum_k A(k) \exp(i k \cdot x)$ of a quantity $A(x)$

![Figure 1. Fluctuations of the CDM density (short-dashed, black), the baryon density (solid, blue), the baryon temperature (long-dashed, cyan), and the photon temperature (dotted, red) at $z = 1000$.](image-url)
in an actual, real space (“r space” henceforth), the background velocities have \( \mathbf{V}_i(x) = \sum_k V_i(K) \exp(ik \cdot x) \) but with the condition \( K \ll 1 \text{ Mpc}^{-1} \). In contrast, the small-scale modes fluctuating against the background patches have intrinsically larger wave number \( k \), or \( K \ll k \). Then, at each \( K \), \( \langle |\nabla | \mathbf{V}_i(K) \rangle \sim k_1 \mathbf{V}_i \ll \langle |\nabla | \mathbf{V}_i \rangle \sim k_2 \mathbf{V}_i \). Similarly, \( \langle |\nabla | \mathbf{U} \rangle \sim |\nabla | \mathbf{U} \rangle \sim k_\delta \mathbf{U} \). We finally note that we do not include the coupling terms between similar wave numbers in Equation (10), which will involve quadratic and higher-order polynomials of \( \Delta \) and \( \delta \). Therefore, the validity of Equation (10) will break down in the nonlinear regime. Nevertheless, our approach is more suitable for a crude estimation of the conditional halo mass function (Section 3.2) in terms of the extended Press–Schechter formalism, which is based on the mapping of the linear density growth to the nonlinear growth of the e.g., top-hat density perturbation.

The evolution of small-scale perturbations, therefore, is coupled to large-scale perturbations on which they are sitting. Ignoring the terms in square brackets, and shifting the viewpoint to the CDM rest frame in which \( V_c = 0 \) (as in O’Leary & McQuinn 2012; TH chose the baryon rest frame), Equation (10), in the \( k \)-space, finally becomes

\[
\frac{\partial \delta_c}{\partial t} = -(1 + \Delta_c) \theta_c - \Theta_c \delta_c,
\]

\[
\frac{\partial \delta_b}{\partial t} = -\frac{3}{2} H^2 \Omega_m (f_c \delta_c + f_b \delta_b) - 2H \theta_c,
\]

\[
\frac{\partial \delta_b}{\partial t} = -ia^{-1} \mathbf{V}_b \cdot \mathbf{k} \delta_b - (1 + \Delta_b) \theta_b - \Theta_b \delta_b,
\]

\[
\frac{\partial \delta_b}{\partial t} = -ia^{-1} \mathbf{V}_b \cdot \mathbf{k} \theta_b - \frac{3}{2} H^2 \Omega_m (f_c \delta_c + f_b \delta_b) - 2H \theta_b,
\]

\[
+ a^{-2} k_B T_{\mu \mu 1}^2 k^2 \{1 + \Delta_b \} \delta_b + (1 + \Delta_b) \delta_b,
\]

\[
\frac{\partial \delta_T}{\partial t} = 2 \left\{ \frac{\partial \delta_b}{\partial t} + \frac{\partial \Delta_b}{\partial t} (\delta_T - \delta_b) + \frac{\partial \delta_b}{\partial t} (\Delta_T - \Delta_b) \right\}
\]

\[
- \frac{\chi}{t_\gamma} a^{-4} \frac{T_{\mu \mu 1}^2}{T} \frac{\partial \delta_T}{\partial t},
\]

(11)

where \( \delta_j \), \( \theta_j \) and \( \delta_T \) now denote fluctuations in \( k \)-space while \( \Delta_j, \Theta_j \) and \( V_j \) are fluctuations of a given patch at \( (a, X) \) in \( r \) space, given by Equation (7).

2.2. Evolution of Perturbation Inside Patches: Numerical Scheme

Evolution of small-scale perturbations can be calculated by integrating the rate equation (Equation (11)) from some initial redshift, preferentially not too long after the recombination epoch when the relative motion has not yet influenced the evolution. We take \( z_i = 1000 \) as the initial redshift. The initial condition should be generated for both the background quantities and the small-scale modes. For the background, as perturbations in Equation (11) are \( r \)-space quantities whose distributions are all Gaussian, one needs to sample these values in the \( r \)-space accordingly. For the small-scale modes, one just needs to track the evolution of the average value in the \( k \)-space.

Let us first describe the statistics of background patches that we expect. TH calculated the evolution of small-scale \( k \geq 10 \) fluctuations under different background patches but only of \( \Delta_c = \Delta_b = 0 \), and defined a “local power spectrum” \( P_{\text{loc,m}} (k; V_{bc}) \) averaged out over all possible opening angles between \( V_{bc} \) and \( k \). In our case, there are extra dimensions to consider which are \( \Delta_c, \Delta_b, \Theta_c, \Theta_b \). In addition, their initial values at \( a = a_i \) completely compose the ensemble at any time through Equation (7). At the minimal level², it would suffice to just consider variation of \( \Delta_c \) in addition to \( V_{bc} \) such that the local power spectrum is an explicit function of the two background quantities, or \( P_{\text{loc,m}} = P_{\text{loc,m}} (k; V_{bc} (a_i); \Delta_c (a_i)) \).

For the initial condition for background patches, we generate 3D maps of \( \Delta_c, \Delta_b, \Theta_c, \Theta_b, V_c, V_b \), and \( \Delta_T \) at \( z = 1000 \) on 151³ uniform grid cells inside a cubical volume of \( V_{box} = (604 \text{ Mpc})^3 \). We generate fluctuations of discrete modes that are randomized as

\[
\text{Re} (\Delta_k) = G_1 N^3 \left( \frac{P(k)}{2V_{box}} \right)^{1/2} \text{sign}[\mathbf{T} \mathbf{F}(\Delta_k)],
\]

\[
\text{Im} (\Delta_k) = G_2 N^3 \left( \frac{P(k)}{2V_{box}} \right)^{1/2} \text{sign}[\mathbf{T} \mathbf{F}(\Delta_k)],
\]

(12)

for given \( k \), where \( G_1 \) and \( G_2 \) are random numbers drawn from mutually independent Gaussian distributions with mean 0 and standard deviation 1, \( \Delta_k \) stands for any kind of \( k \)-space fluctuation, and \( \mathbf{T} \mathbf{F} \) is the transfer function of \( \Delta_k \), whose sign should be multiplied because some \( \Delta_k \) oscillate around zero in \( k \). The configuration is roughly equivalent to applying a smoothing filter of length \( (604/151) = 4 \text{ Mpc} \). In practice, we use CAMB (Lewis et al. 2000) for \( \Delta(k, a_i) \), and use the continuity equations \( \partial \Delta / \partial t = -\Theta \) for \( \Theta(k, a_i) \), with the help of two CAMB transfer-function outputs at mutually nearby redshifts for time differentiation. \( V_j(k, a_i) \) is obtained from the relation \( V_j = -i(a k / k^2)^2 \frac{\partial \delta}{\partial t} \), \( \Delta_T (k, a_i) \) is fixed by following the scheme of Naoz & Barkana (2005): we require \( \partial \Delta_T / \partial t = \partial \Delta_T / \partial t \) at the initial redshift in Equation (4), which results in

\[
\Delta_T = \Delta_T \left( 5 - \frac{4T}{T_c} \right) + \frac{t_\gamma}{x_{\gamma}(t)} a_i \left( \frac{2 \partial \Delta_b}{\partial t} - \frac{\partial \Delta_T}{\partial t} \right)
\]

(13)

where all quantities are evaluated at \( a_i \), especially with the help of Equation (7) for \( \partial \Delta_b / \partial t \) and two adjacent CAMB transfer-function outputs for \( \partial \Delta_T (k, a_i) / \partial t = (1/4) \partial \Delta_T (k, a_i) / \partial t \). Finally, all these \( k \)-space fluctuations are Fourier-transformed to obtain \( r \)-space fluctuations.

Three-dimensional maps and two-dimensional histograms of several initial quantities are presented in Figure 2. Fields of \( \Delta_c \) and \( V_{bc} = -V_{bc} \) on a part of a slice of the box at \( z = z_i = 1000 \) are shown in Figure 2(a). As expected, the velocity field converges on overdense regions and diverges on underdense regions. We find that in most patches \( V_c \) dominates over \( V_b \), and

² For a more accurate treatment or for a specific patch of interest, one should also consider variations in other variables, such as \( P_{\text{loc,m}} = P_{\text{loc,m}} (k; V_{bc}; \Delta_c; \Delta_b; \Theta_c; \Theta_b; \Delta_T) \). As the correlation between \( V_c (\Delta_c) \) and \( V_c (\Delta_c) \) becomes tighter in time, the initial variation gets gradually diluted, which roughly justifies our restricting the parameter space only to \( V_{bc} \) and \( \Delta_c \).
thus the map of $V_c$ looks very similar to Figure 2(a). This occurs because baryons lag behind CDMs due to their coupling to the CMB. $\Delta_b$ (Figure 2(b)) is coupled to $\Delta_T$ (Figure 2(c)) more strongly than to $\Delta_c$. $\Delta_c$ and $\Delta_T$ (similarly $\Delta_b$ and $\Delta_T$) are almost perfectly correlated (Figure 2(d)). $\Delta_c$ and $\Delta_b$ are very loosely correlated due to the tight coupling of baryons to photons at the redshift (Figure 2(e)), but the correlation becomes tighter in time. $\Delta_c$ and $V_{bc}$ are not correlated (Figure 2(f)). Because of this, the probability distribution function (PDF) $P$ is simply a multiplication of PDFs $P(k; V_{bc})$ and $P(k; \Delta_c)$, at the minimal level. Due to Gaussianity at $z_i$, we have

$$P(k; \Delta_c) = \frac{1}{\sqrt{2\pi} \sigma_{\Delta_c}} \exp \left[ -\frac{\Delta_c^2}{2\sigma_{\Delta_c}^2} \right],$$

$$P(k; V_{bc}) = \frac{2}{\sqrt{\pi}} \frac{V_{bc}}{\sigma_{V_{bc}}} \exp \left[ -\frac{V_{bc}^2}{2\sigma_{V_{bc}}^2} \right],$$

where $\sigma_{\Delta_c}$ and $\sigma_{V_{bc}}$ are the standard deviations of $\Delta_c$ and $V_{bc}$ projected onto one Cartesian-coordinate axis, respectively. With our setup, we find that $\sigma_{\Delta_c} = 0.0042$ and $\sigma_{V_{bc}} = 17.8$ km s$^{-1}$ at $z_i$ (or the root-mean-square of $V_{bc}$ is $\sqrt{3} \sigma_{V_{bc}} = 30.9$ km s$^{-1}$). At the minimal level of only allowing the variance in $\Delta_c$ and $V_{bc}$, the average power spectrum $P_m(k)$ will then be given by the ensemble average

$$P_m(k) = \int_0^\infty dV_{bc} \int_0^\infty d\Delta_c P(k; V_{bc})P(k; \Delta_c) \times P_{loc,m}(k; V_{bc}; \Delta_c; \Delta_b; \Theta_c; \Theta_b; \Delta_T; \Delta_T),$$

where the PDFs and integral arguments are those at $z_i$. Of course, a more accurate and straightforward way is to just ensemble-average $P_{loc,m}(k; V_{bc}; \Delta_c; \Delta_b; \Theta_c; \Theta_b; \Delta_T; \Delta_T)$ over the patches from a large-box realization, because for example $\Delta_c$ and $\Delta_b$ are too poorly correlated at $z_i$.

We numerically integrate Equation (11) to examine the evolution of small-scale (high-$k$) fluctuations at any overdense (underdense) patch to the linear order, with the help of Equations (7) and (9) for the evolution of background quantities. In practice, we used the ODE45 modules of MATLAB® (2015b, The MathWorks, Inc., Natick, MA, USA) and of GNU Octave, which use the fourth-order Runge–Kutta method, with relative tolerance $10^{-4}$ and absolute tolerance $10^{-12}$ at $a_i$. During the evolution, the number of integration steps is highest for $\delta_T$, because its amplitude changes from the initial, very small values around $\delta_T$ to final, much larger values close to $\delta_b$. Therefore, taking sub-steps for $\delta_T$ while coarser steps for other $\delta_k$ is expected to boost the computational efficiency, even though we have not yet implemented the method in our computation. The end result
is then ensemble-averaged over varying $\Delta_z$ and $V_{bc}$ to obtain $P_m(k)$ (Equation (15)).

3. RESULTS

3.1. Power Spectrum of the Matter Density

We first examine how the evolution of $P_{\text{loc,}m}(k; V_{bc}; \Delta_z)$ depends on the density environment, and compare the result to the prediction by TH. Figure 3 shows the evolution of $\Delta_{\text{loc,}m}^2 = k^3P_{\text{loc,}m}/2\pi^2$ of three arbitrarily chosen wave numbers ($k = \{33, 150, 2000\}$ Mpc$^{-1}$) when $V_{bc} = 22$ km s$^{-1}(a/a_0)^{-1}$ in different density environments ($\Delta_z(a_t) = \{-0.01, -0.005, 0, 0.005, 0.01\}$). Note again that $\sigma_{\Delta_z} = 0.0042$ at $z = 1000$, and thus these samples correspond to $\pm 2.4\sigma_{\Delta_z}$ and $\pm 1.2\sigma_{\Delta_z}$.

First, as expected, the growth of small-scale fluctuations is biased when $\Delta_z > 0$ and anti-biased when $\Delta_z < 0$, with respect to the mean-density case, or the prediction by TH. Second, when $\Delta_z(a_t)$ are equal in amplitude but opposite in sign, the deviations of $P_{\text{loc,}m}(k; V_{bc}; \Delta_z)$ from $P_{\text{loc,}m}(k; V_{bc}; \Delta_z = 0)$ reveal the same trend but only until $z \approx 850$ when $|\Delta_z(a_t)| = 0.01$ and $z \approx 730$ when $|\Delta_z(a_t)| = 0.005$. Afterwards, the bias and the anti-bias are not balanced by more than 1% and this off-balance keeps growing in time. The higher the value of $|\Delta_z(a_t)|$, the earlier this imbalance starts. Third, the fractional deviation from the mean-density case is almost universal regardless of the value of $k$, and thus the timing of the imbalance is approximately a function only of $|\Delta_z(a_t)|$. Finally, in some high-$\Delta_z$ patches, our formalism would provide sufficient accuracy.

How large-scale overdensity impacts the evolution of small-scale inhomogeneities is reflected in the density continuity equation. In overdense background patches, $\delta_i$ grows in time and $\Theta_i < 0$. Then, in Equation (11), $-\delta_i \Theta_i$ and $-\Theta_i \delta_i$ work as sources terms in addition to $-\Theta_i$ to the growth of $\delta_i$. This will boost the growth rate of both overdense ($\delta_i > 0, \Theta_i < 0$) and underdense ($\delta_i < 0, \Theta_i > 0$) modes. In contrast, in underdense background patches, these terms suppress the growth rate of both overdense and underdense modes. The distribution of $\Delta_z(a_t)$ is Gaussian, and therefore for each “bias” case with $\Delta_z > 0$ there exists an “anti-bias” case with $\Delta_z < 0$. Nevertheless, the imbalance described above is expected to boost the average power spectrum from that by TH.

The overall effect of including the Gaussian distribution of $\Delta_z$ is thus to mitigate the negative impact by the relative velocity, predicted by TH, to some extent. In addition, the universality of the imbalance in $k$ boosts $P_m(k$) even in the $k$ range ($10 \lesssim k \lesssim 100$ Mpc and $k \gtrsim 1000$ Mpc) where the power spectrum is almost unaffected by non-zero $V_{bc}$ (Figure 4), shown in terms of the $k$-space matter-density variance $\Delta_{\text{loc,}m}^2 = k^3P_m(k)/2\pi^2$. Note that the result for $k \lesssim 10$ Mpc cannot be trusted, because our perturbation theory is based on the condition that large-scale modes ($0.01 \lesssim K$ Mpc $\lesssim 1$) are well separated from small-scale modes in scale. The discrepancy of $P_m(k)$ including non-zero $\Delta$ from the prediction by TH is negligible at $z \gtrsim 45$, but later the discrepancy grows in time. Of course, individual patches may experience a discrepancy in $P_{\text{loc,}m}(k; V_{bc}; \Delta_z)$ much earlier than this epoch (Figure 3).

3.2. Halo Abundance

Understanding the abundance and the spatial distribution of cosmological halos is crucial in modern astrophysics and cosmology. In this section, we examine the halo abundance

$\text{Figure 3. Growth of } \Delta_{\text{loc,}m}^2 = k^3P_{\text{loc,}m}(k; V_{bc}; \Delta_z)/(2\pi^2) \text{ with wave numbers } k = \{33, 150, 2000\} \text{ Mpc}^{-1} \text{ and } V_{bc}(z = 1000) = 22 \text{ km s}^{-1} \text{ in initially overdense (red, long-dashed and orange, dotted-dashed), mean-density (black, solid), and underdense (cyan, short-dashed and blue, dotted) regions. Initial CDM overdensities are chosen to be } \Delta_z(a_t) = \{-0.01, -0.005, 0, 0.005, 0.01\}. \text{ The mean-density case, or } P_{\text{loc,}m}(k; V_{bc}; \Delta_z = 0), \text{ corresponds to } P_{\text{bc,}m}(k; V_{bc}) \text{ by TH. Fractional differences } (P_{\text{loc,}m}(k; V_{bc}; \Delta_z) - P_{\text{bc,}m}(k; V_{bc}; \Delta_z = 0))/P_{\text{bc,}m}(k; V_{bc}; \Delta_z = 0) \text{ in } \% \text{ are plotted in the bottom sub-panels, with the line-type convention same as in the top sub-panels.}$
both in the local and the global sense, just as we did for the matter power spectrum. Let us first revisit the calculation by TH. They adopted the extended Press–Schechter formalism and calculated the local halo abundance, in terms of the conditional mass function, using the peak-background split scheme. A patch with \( \Delta \) (let us use this notation for the matter overdensity, to avoid confusion with the \( k \)-space matter–density variance \( D_m^2 \)) and \( V_{bc} \) will have the number of halos per unit Eulerian comoving volume per \( M \) given by

\[
\frac{dn}{dM}(M|\Delta, V_{bc}) = \sqrt{\frac{2}{\pi}} \frac{\bar{\rho}_m \delta_{\text{crit}} - \Delta}{\sigma^2} \left( 1 + \Delta \right) d\sigma \times \exp\left(-\frac{(\delta_{\text{crit}} - \Delta)^2}{2\sigma^2}\right),
\]

(16)

where \( \delta_{\text{crit}} \) is the critical overdensity of spherical collapse, and \( \sigma^2 \) is the variance of density field smoothed with the window function corresponding to mass \( M \),

\[
\sigma^2(M, V_{bc}) = \int \Delta_m^2(k, V_{bc}) W_M^2 d\ln k.
\]

(17)

One should note that \( \sigma^2(M, V_{bc}) \) should be that of high-\( k \) modes only, or more accurately a reduced value \( \sigma^2(M, V_{bc}) - \sigma_{\text{patch}}^2 \) where \( \sigma_{\text{patch}}^2 \) is the variance of the density field smoothed with the window function corresponding to the mass of the patch\(^3\) (e.g., Bond et al. 1991; Mo & White 1996; Ahn et al. 2015). One can instead put a lower bound \( k_{\text{min}} \) in the integral of Equation (17), which would be identical to the reduced variance if a sharp \( k \)-space window function is used. The global mass function \( \frac{dn}{dM}_g \) is simply an average of the local mass function, \( \frac{dn}{dM}_l \), over the ensemble of patches. When \( \frac{dn}{dM}(M|\Delta, V_{bc})/dM \) is averaged only over \( V_{bc} \) for a given \( \Delta \), which is equivalent to visiting only those patches with the same \( \Delta \) and taking the average, it leads to the conditional mass function

\[
\sigma^2(M, V_{bc}, \Delta) = \int_{k_{\text{min}}}^{k_{\text{max}}} \Delta_m^2(k, V_{bc}, \Delta) d\ln k,
\]

(18)

which will enter Equation (16). Here we used the sharp \( k \)-space filter, and thus \( k_{\text{max}} = [6\pi^2 \bar{\rho}_m/M_{\text{patch}}]^{1/3} \). An overdense patch will then have a boost in \( \frac{dn}{dM}(M|\Delta, V_{bc})/dM \) from the value by TH because \( P_n(k, V_{bc}, \Delta > 0) > P_n(M, V_{bc}, \Delta = 0) \) and thus

\(^3\) It is not clear whether TH used this reduced variance. In addition, a factor of 2 should be multiplied to Equation (18) of TH.
It is important to compare our findings to the usual peak-background split scheme and that of TH. In the "standard" scheme, if the density field is purely Gaussian, all the wave modes are assumed mutually independent in the linear regime. Therefore, the local, high-\(k\) modes have a universal\(^4\) variance \(\sigma_k^2\), whether or not they are placed inside a patch with non-zero \(\Delta\). How the halo formation is biased in an overdense region is simply through the shift in the density \((+\Delta)\). TH then realized that the variance is not universal but should depend on \(V_{bc}\). Because non-zero \(V_{bc}\) tends to suppress \(\sigma_k^2\), the odds to cross \(\delta_{crit}\) decrease relative to the standard picture. We find that there is another dependency of the variance, which is \(\Delta\). In other words, we find that there are two biasing effects in an overdense region compared to a mean-density region: getting closer to \(\delta_{crit}\) because of the shift in the density \((+\Delta)\) also in the standard scheme), and having a larger degree of fluctuation in \(\delta_k\) due to the mode-mode coupling (e.g., source terms \(-\Delta_{bc}\theta\) and \(-\Theta_{bc}\delta\) in \(\partial\delta_c/\partial t\) in Equation (11), which is a new finding). Therefore, by not fully implementing the effect of non-zero \(\Delta\), TH in effect underestimate and overestimate the halo mass functions in overdense and underdense regions, respectively.

We note that this additional bias effect should be present even in the standard picture with \(V_{bc} = 0\), because this is due to the natural coupling between the large-scale and small-scale density perturbations. In this case, however, we are not sure about the quantitative validity of the extended Press–Schechter formalism on the conditional mass function (Equation (16)), which is based on the linear theory guaranteeing Gaussianity at any filtering scales without the mode-mode couplings. Qualitatively, we believe that the boost of local \(\delta_k\) and \(\sigma_k^2\) under \(\Delta\) should boost the conditional mass function to the level estimated by Equations (16) and (18) as described above in any

---

\(^4\)Rigorously speaking, it is not perfectly universal because the lower bound changes slightly in \(\Delta\) as \(M_{patch} = \rho_m(1 + \Delta)\).
case. We postpone a further investigation of this issue, which can be clarified with numerical simulations of the halo formation under different $\Delta$.

Discrepancy between the conditional mass functions by this work and by TH is significant if we focus on individual patches. Figure 5 illustrates how our prediction differs from that of TH. For example, at $z \sim 44$–19, under $\Delta_c(a_i) = 0.005$, we predict a [100–200]% boost in $(dn/dM)_\Delta$ compared to the values of TH (let us denote them by $(dn/dM)_{\Delta, TH}$). For $\Delta_c(a_i) = -0.005$, we predict a 90% or more decrease in $(dn/dM)_\Delta$ compared to the values of TH. This is the obvious result of the mode–mode coupling of $\Delta$ and $\delta$ described above. It is also noteworthy that the discrepancy is the largest for the rarest halos: first, at any redshift, the discrepancy increases as the halo mass increases and, second, for any given halo mass, the discrepancy decreases in time.

The discrepancy among the conditional mass functions of this work, of TH, and of the standard picture also influences $(dn/dM)_k$. Let us just take the example of $z = 19$ (Figure 6). Compared to the standard prediction ($V_{bc} = 0$), conditional mass functions of TH stay lower regardless of $\Delta$. This is due to the suppression of structure formation by the relative velocity. In contrast, $(dn/dM)_\Delta$ in this work is either higher or lower than the standard prediction $(\equiv (dn/dM)_{V_{bc}=0})$ depending on the halo mass and $\Delta$. At $z = 19$, for $\Delta_c(a_i) = -0.005$, $(dn/dM)_\Delta < (dn/dM)_{V_{bc}=0}$ and for $\Delta_c(a_i) = 0.005$, $(dn/dM)_\Delta > (dn/dM)_{V_{bc}=0}$ for any halo mass. The tendency for $(dn/dM)_\Delta$ to overshoot $(dn/dM)_{V_{bc}=0}$ when $\Delta > 0$ is not generic, because the influence of the positive $\Delta$ on the structure formation appears only late in its evolution (see Figures 3 and 4). At any rate, the contrast in $(dn/dM)_\Delta$ among overdense and underdense regions is increased compared to TH and the standard picture, and the net effect, e.g., at $z = 19$ is to boost $(dn/dM)_k$ from $(dn/dM)_{V_{bc}=0}$. At much higher redshifts, our $(dn/dM)_k$ is almost indistinguishable from that of TH, which undershoots $(dn/dM)_{V_{bc}=0}$.

We note that $(dn/dM)_\Delta$ and $(dn/dM)_k$ have the usual problem of not correctly predicting the actual mass function, if one sticks to the original extended Press–Schechter formalism. Minihalos at high redshifts are usually underestimated by the extended Press–Schechter formalism. The usual peak-background split method suffers from large discrepancies between its prediction and the $N$-body simulation results for rare halos in general. In this case, a hybrid method to connect the peak-background-split halo bias parameter to the better-fitting mean mass function types (e.g., Barkana & Loeb 2004; Ahn et al. 2015) is much more appropriate. We will apply this method in the future for a better estimation of the conditional mass function and the $k$-space halo bias parameter (Section 3.3).

3.3. Halo Bias and Stochasticity

The conditional mass function we examined in Section 3.2 is an indicator of how halo formation is biased toward overdense regions. The halo bias can be viewed also in the $k$-space. This is a crucial parameter in cosmology when trying to probe the fluctuation of the matter density from surveys of galaxies through, for example, the power spectrum analysis.

The halo bias parameter in $k$-space is defined as

$$b(k) = \left( \frac{P_h(k)}{P_b(k)} \right)^{1/2},$$

where $P_h(k)$ is the power spectrum of halo overabundance

$$\delta_n(M, x) = \frac{(dn/dM)_\Delta(M, x) - (dn/dM)_g}{(dn/dM)_g}(M).$$
As was pointed out by TH, $b(k)$ oscillates in $k$ due to the BAO, which is tied to the modulation of the streaming velocity, the density fluctuation, and the baryon fraction in the existence of the compensated mode (Barkana & Loeb 2011). This makes it difficult to deduce $P_{m}(k)$ from $P_{b}(k)$, not to mention from the galaxy surveys where galaxy formation mechanism, strongly influenced by baryonic physics, is another nuisance parameter (but see Slepian & Eisenstein 2015 for how to separate out the streaming velocity effect). For the halo mass range treated in this paper, $b(k)$ should modulate the distribution of the first stars which grow predominantly inside minihalos. As was also noted by TH, the difference in $b(k)$ should also influence the formation of much larger-halo halos, which are used for galaxy surveys. It is also possible that the nonlinear effect described by O’Leary & McQuinn (2012), or the heating of the IGM due to the velocity difference between CDM and baryons, is modulated in space depending on the overdensity. Then the power spectrum in the 21 cm background, which may be dominated by the velocity fields if the heating is efficient (McQuinn & O’Leary 2012), is likely to be boosted. Because such a power spectrum shows a very clear BAO feature and the 21 cm observation usually suffers from low sensitivity, the signal boosted even more from the prediction by McQuinn & O’Leary (2012) will be a very promising target for the high-redshift 21 cm cosmology.

Let us briefly discuss the halo stochasticity. This is defined as

$$\chi = \frac{P_{Hm}(k)}{P_{b}(k)P_{m}(k)},$$

where $P_{Hm}(k)$ is the cross-power spectrum between the halo density and the matter density. Both in TH and in this work, stochasticity is caused by the fluctuation in $(dn/dM)_\Delta$ because patches with the same $\Delta$ can have different $V_{bc}$ which affect $(dn/dM)_\Delta$. One should note that the fluctuation should be caused also by the sampling variance. The conditional mass functions usually show super-Poissonian distributions in $(dn/dM)_\Delta$ even in the standard picture with $V_{bc} = 0$ (e.g., Saslaw & Hamilton 1984; Sheth 1995; Neyrinck et al. 2014; Ahn et al. 2015), and obviously this should give rise to stochasticity in addition to that caused by the varying $V_{bc}$. Because we use the “mean” conditional mass function just as TH did, in both works $\chi$ does not reflect the sampling variance. Including this effect requires the calculation of the sub-cell correlation function (Ahn et al. 2015), which we defer to future work.

4. DISCUSSION

We investigated the impact of the relative velocity (streaming velocity) between CDM and baryons on small-scale structure formation. TH first studied this effect by adopting a trivial solution to the large-scale velocity and density fields. Because velocity fields are correlated with density fields, however, such a trivial solution cannot accurately describe the physics in regions with non-zero overdensity. We thus improved on the work of TH by implementing a non-trivial solution to the large-scale velocity and density fields, and we find that this causes a new type of coupling between large-scale and small-scale modes. This results in boosting the small-scale structure formation in overdense regions and suppressing that in underdense regions, aside from the suppression originating from the streaming velocity. The net effect on the structure formation is to boost the overall fluctuation, in terms of $P_{m}(k)$, and thus the “negative” effect noted by TH is mitigated to some extent. Depending on the wave mode $(k)$ and the observing redshift, $P_{m}(k)$ can even be larger than that in the standard picture with $V_{bc} = 0$. The conditional halo mass function and the halo bias are also affected in similar ways.

The results of this work show that the formation and evolution of small-scale structures depend strongly on not only the streaming velocity but also the density environment. The most important aspect of our work is that in contrast to TH, who predict that regardless of the underlying density the local matter power spectrum of small-scale structures will be identical as long as $V_{bc}$ is the same, the underlying large-scale ($\sim$ a few Mpc) overdensity is another key parameter in addition to $V_{bc}$. This then requires re-examining previous work based on the formalism by TH. We already showed that $P_{loc,m}(k)$, $P_{m}(k)$, $(dn/dM)_\Delta$, $(dn/dM)_b$, and $b(k)$ are affected. If we were to simulate the nonlinear evolution of density perturbations and the structure formation in $\sim$4 Mpc patches, we should generate initial conditions based on this work. As is seen in Figure 3, we cannot neglect the impact of overdensity even when the simulation starts at e.g., $z = 200$, because at $\sim1\sigma_{\Delta_c}$ the

---

5 Our box size is barely larger than the BAO scale, and is thus too small to accurately estimate $b(k)$ at low $k$. We will increase the size of the box in future work for a better estimation.
discrepancy between our prediction and that of TH is already a few percent at that redshift.

Both the previous semi-analytical work (Tseliakhovich et al. 2011; Fialkov et al. 2012; McQuinn & O’Leary 2012; Bovy & Dvorkin 2013; Naoz & Narayan 2014; Asaba et al. 2016) and the semi-numerical work (e.g., Visbal et al. 2012; Fialkov et al. 2013) should be re-examined. Attempts to numerically simulate the nonlinear evolution of small-scale structures have been mostly limited to the physics inside mean-density regions (Greif et al. 2011; Maio et al. 2011; Stacy et al. 2011; McQuinn & O’Leary 2012) or special, isolated regions (Tanaka & Li 2014). These numerical simulations thus need to be extended to incorporate Δ which varies in space. In doing so, a reasonable method would be to use adaptive mesh refinement codes with nested grids, so that one or a few interesting regions (~4 Mpc patches with Δ = ±10δ, for example) are treated with fine meshes and other regions with coarse meshes for computational efficiency.

We also showed that cosmology through galaxy surveys should carefully consider the impact of the mode–mode coupling, because the halo bias (and galaxy bias as well) would be boosted from not only the standard prediction with V_{s} = 0 but also the prediction of TH. Cosmology with intensity mapping may also be affected. The post-reionization intensity mapping targets the large-angle, diffuse 21 cm background from neutral hydrogen atoms inside galaxies (Chang et al. 2008; Abdalla et al. 2010; Bandura et al. 2014; Xu et al. 2015). Because any galaxies, small or large, contribute to this cumulative 21 cm background, such observations will be affected by the streaming velocity through b(k). The pre-reionization intensity mapping targets the large-angle, diffuse 21 cm background from the intergalactic neutral hydrogen atoms (Scott & Rees 1990; Bharadwaj & Bai 2004; Loeb & Zaldarriaga 2004; Barkana & Loeb 2005; McQuinn et al. 2006; Mao et al. 2012; Shapiro & Abel 2013). Even though this is free from the galaxy bias, the streaming velocity may act as a heating mechanism and boost the power spectrum of the velocity field (McQuinn & O’Leary 2012), and therefore the new findings of our work should be incorporated.

Application to the study of the cosmic reionization process is of prime interest in terms of high-redshift astrophysics. The complex nature of the process usually requires numerical simulations, and they are performed through either efficient semi-numerical methods (Furlanetto et al. 2004; McQuinn et al. 2007; Mesinger et al. 2011; Alvarez & Abel 2012) or fully numerical methods (e.g., Gnedin & Abel 2001; Razoumov et al. 2002; Maselli et al. 2003; Mellema et al. 2006; Baek et al. 2009; Wise & Abel 2011). The early phase of cosmic reionization must have been driven by the first stars, possibly forming first in minihalos, as these are the first luminous objects in the universe. A very important factor that modulates the formation of the first stars inside minihalos is the Lyman–Werner intensity, which has been properly treated in simulations in a box large enough for statistical reliability but implementing subgrid physics for the first star formation inside minihalos (Ahn et al. 2012; Fialkov et al. 2013). To predict reionization scenarios to our best knowledge, especially in its early phase, this work should be properly incorporated because the star formation inside minihalos is strongly modulated by the streaming velocity as well.

The results of this paper have room for further improvement. This paper is based on the presumption that the fluctuations at a few Mpc scale remain linear even at later epochs. However, high-density regions will reach the nonlinear regime earlier than the others, and then our formalism will break down in such regions. The halo bias is more strongly pronounced in the nonlinear patches (e.g., Ahn et al. 2015) than in the linear theory, and thus one should use the actual values of the overdensity in such circumstances. One could achieve this goal by adopting the quasi-nonlinear calculation (e.g., 2LPT by Crocce et al. 2006), adopting the top-hat collapse model as in Mo & White (1996) and Ahn et al. (2015), or for the best accuracy running N-body+hydro simulations which resolve the density fluctuation at ~Mpc scale. Then, in each patch of a few comoving Mpc, Equation (11) can be integrated with the newly computed values of Δ. Wave modes in the range k ≈ [1–10]/Mpc are not accurately treated, because we based our formalism on the separability of the large-scale modes (k ≤ 1/Mpc) and the small-scale modes (k ≥ 10/Mpc). The code we used will be released for public use, but it requires technical improvements such as allowing parallel computation and porting to more generic computation languages. We will maintain and control its development through the website http://www.chosun.ac.kr/kjahn.

We thank P.R. Shapiro, F. Schmidt and R. Barkana for helpful discussions. We also thank the anonymous referee for the clear report which led to a significant, quantitative improvement of the paper. This work was supported by a research grant from Chosun University (2016).

APPENDIX A
NORMAL MODES FOR THE LARGE-SCALE FLUCTUATIONS

When the fluctuation of radiation components is neglected, the growth of large-scale density and velocity fluctuations are well approximated by Equations (5) and (6). Their evolution can then be described by the four normal modes described in Section 2.1. The growing and decaying modes are the two solutions to the second-order equation

$$\frac{d^2\Delta_+}{dt^2} + 2H\frac{d\Delta_+}{dt} - \frac{3}{2}H^2\Omega_m\Delta_+ = 0, \quad (22)$$

where Δ_+ = f_c Δ_c + f_b Δ_b, and can be written also as

$$\frac{d^2\Delta_+}{da^2} + \left(\frac{3}{a} + \frac{d\ln H}{da}\right)\frac{d\Delta_+}{da} - \frac{3}{2a^2}\Omega_m\Delta_+ = 0. \quad (23)$$

Similarly, the compensated and streaming modes are the two solutions to

$$\frac{d^2\Delta_-}{dt^2} + 2H\frac{d\Delta_-}{dt} = 0, \quad (24)$$

where Δ_- = Δ_c − Δ_b, and can be written also as

$$\frac{d^2\Delta_-}{da^2} + \left(\frac{3}{a} + \frac{d\ln H}{da}\right)\frac{d\Delta_-}{da} = 0. \quad (25)$$

We now take the convention of writing each mode as the product of its initial value at z = 1000 and its growth factor: Δ^x (z) = Δ(z)D^x (z), Δ^x (z) = Δ(z)D^x (z), Δ^± (z) = Δ(z)D^± (z), and Δ_(z) = Δ(z)D_+ (z), denoting the growing, decaying, compensated, and streaming modes, respectively. These modes comprise Δ_+ (z) and Δ_-(z), as Δ_+ (z) = Δ(z)D^+ (z) + Δ(z)D^- (z) and Δ_-(z) = Δ(z)D^- (z) + Δ(z)D^+ (z), with the normalization D^± = D^± = D^± = 1 at z = 1000.
The first task in finding these modes is to calculate the growth factors. Because the Hubble constant \( H(a) \) (=\( H_0 \sqrt{\Omega_{r,0}a^{-4} + \Omega_{m,0}a^{-3} + \Omega_{\Lambda,0}} \)) and \( \Omega_m(a) \) have non-negligible radiation components during the period of interest, \( 1000 \gtrsim z \gtrsim 50 \), growth factors should be calculated numerically. We factor out deviations from the analytical form valid during the matter-dominated (\( \Omega_m = 1 \)) era, as \( D_g(a) = \frac{a}{a_i} F_g(a) \), \( D^d(a) = (a/a_i)^{-3/2} F^d(a) \), and \( D^s(a) = (a/a_i)^{-1/2} F^s(a) \), and then solve for the order-of-unity values of \( F^g \), \( F^d \), and \( F^s \). They are determined by

\[
\frac{d^2 F^g}{da^2} + \left( \frac{5}{a} + \frac{d \ln H}{da} \right) \frac{d F^g}{da} - \frac{1}{a} \left\{ 3 \left( \frac{\Omega_m}{2} - 1 \right) - \frac{d \ln H}{da} \right\} \times F^g = 0, 
\tag{26}
\]

\[
\frac{d^2 F^d}{da^2} + \frac{d \ln H}{da} \frac{d F^d}{da} \frac{d \ln H}{da} \left\{ \frac{3}{a} \left( \frac{\Omega_m}{2} + \frac{1}{4} \right) + \frac{3}{2} \frac{d \ln H}{da} \right\} \times F^d = 0, 
\tag{27}
\]

and

\[
\frac{d^2 F^s}{da^2} + \left( \frac{2}{a} + \frac{d \ln H}{da} \right) \frac{d F^s}{da} - \frac{1}{a} \left\{ \frac{3}{4} a + \frac{1}{2} \frac{d \ln H}{da} \right\} \times F^s = 0. 
\tag{28}
\]

In practice, the numerical integration of Equations (26)–(28) is started at \( z = 10 \) (backward for \( z \geq 10 \) and forward for \( z \leq 10 \)), with the condition that \( dF^g,d,^s/dz = 0 \) at \( z = 10 \) because it is a matter-dominated epoch, and \( F^g,d,^s = 1 \) at \( z = 1000 \) because of our normalization convention. One can instead start the integration from the radiation-dominated epoch (just numerically assuming that Equations (26)–(28) are all valid at \( z \approx 10,000 \) and using the asymptotes for \( dF^g,d,^s/dz \) and \( F^g,d,^s \) at that time), but we find that Equations (27) and (28) become stiff if integrated forward in increasing \( a \) at high \( z \). We show the growth factors found this way in Figure 8.

Now we can find the initial values of these modes by using the growth factors found above on the transfer function outputs from CAMB. We algebraically relate two redshift outputs, at \( z_1 = z_i = 1000 \) and \( z_2 = 800 \) in practice, to find these modes:

\[
\Delta_{\text{gro}} = \frac{\Delta_+(a_2) - D^d(a_2) \Delta_+(a_1)}{D^g(a_2) - D^d(a_2)}, \\
\Delta_{\text{dec}} = \Delta_+(a_1) - \Delta_{\text{gro}}, \\
\Delta_{\text{str}} = \frac{\Delta_-(a_2) - \Delta_-(a_1)}{D^s(a_2) - 1}, \\
\Delta_{\text{com}} = \Delta_+(a_1) - \Delta_{\text{str}}, 
\tag{29}
\]

where \( \Delta_\pm(a_1,z) \) are those from CAMB. The modes found this way are shown in Figure 9.

A few things are notable. An unperturbed Hubble flow requires \( \Delta_{\text{gro}}/\Delta_{\text{dec}} = 3/2 \), while we find that at \( z = 1000 \) \( \Delta_{\text{gro}}/\Delta_{\text{dec}} \approx 27 \) and is not constant over \( k \). Even though one can choose different redshifts to extract these modes and they should not change in principle, the resulting modes at \( z_i \) vary depending on the choice of the redshifts. We believe that this is partly due to our neglect of the fluctuations of radiation components, because they can affect the evolution of density fluctuations. We find that our current choice of \( z_1 \) and \( z_2 \) is optimal for \( k \lesssim 0.01 \) Mpc: using these modes and evolving them with the growth factors, we find a good match between \( \{ \Delta_+, \Delta_- \} \) “constructed” by using these modes and those

Figure 8. Evolution of growth factors, for the growing (black, solid), decaying (blue, dotted), and streaming (red, dashed) modes. When \( F \) (left panel) is multiplied to each analytic power-law evolution valid for the matter-dominated era, it results in the actual growth factor \( D \) (right panel).

Figure 9. Growing, decaying, compensated, and streaming modes at \( z = 1000 \), plotted in varying line widths. The negative values are plotted by a dashed line for the decaying mode and a dotted line for the streaming mode, after flipping the sign.
calculated by CAMB, at any redshifts with at most a several percent error. We show their comparison in Figure 10.

Schmidt (2016) follows a similar approach for the mode extraction but uses CAMB transfer function outputs at \( z \approx 0 \). Their focus is on the low-redshift galaxy surveys, and thus the accuracy is required mostly at and near the present. In our case, accuracy in \( \Delta \), and \( \Theta \), is required mostly at a high redshift range, 1000 \( \lesssim z \lesssim 50 \), because the small-scale modes are only significantly affected by large-scale modes from the epoch of recombination and the linear perturbation analysis on small-scale modes breaks down later when they become nonlinear.

Figure 11. Actual evolution of \( \Delta_T \) of three arbitrarily chosen spatial patches (4 comoving Mpc) with varying \( \Delta_c(a_0) \) (thick; solid, dashed, dotted-dashed). Overlaid are the corresponding fitting functions given by Equation (9), in thin dotted lines.

Therefore, one expects a strong correlation between \( \Delta_T \) and \( \Delta_0 \) long after recombination. For example, at \( a = 0.01 \) and 0.1, they follow the linear relation: \( \Delta_T/\Delta_0 = 0.239 \) at \( a = 0.01 \), and \( \Delta_T/\Delta_0 = 0.586 \) at \( a = 0.1 \). Regardless of the variance in \( \Delta_T \), therefore, one can find a fitting formula for \( \Delta_T(a) \) after its decoupling from the CMB temperature. The fitting formula is given by Equation (9). We note that Equation (9) is valid only when the patch size is 4 comoving Mpc. For patches in different size, we believe that a generic form of a two-parameter fit, \[ \Delta_T(a) = \text{sign}(\Delta_T) \text{dex}[\alpha (\log_{10} a + \beta)^{\gamma} - Y], \] \[ \alpha = \frac{\log_{10}(\Delta_T, b/\Delta_T, a)}{(\beta - 1)^{\gamma} - (\beta - 2)^{\gamma}}, \] \[ Y = \frac{(\beta - 2)^{\gamma} \log_{10} |\Delta_T, b| - (\beta - 1)^{\gamma} \log_{10} |\Delta_T, a|}{(\beta - 1)^{\gamma} - (\beta - 2)^{\gamma}}, \] (30)

will serve as a good fit regardless of the patch size. The linear relation between \( \Delta_T \) and \( \Delta_0 \) at \( a = 0.01 \) and 0.1 can be found by comparing the two quantities. In the case of the 4 comoving Mpc patch, we find that \( \beta = 2.8 \) and \( \gamma = 0.33 \) (Equation (9)) provides an excellent fit to \( \Delta_T \) when \( \Delta_T \gtrsim 10^{-4} \), with the linear relations \( \Delta_T/\Delta_0 = 0.279 \) at \( a = 0.01 \) and \( \Delta_T/\Delta_0 = 0.599 \) at \( a = 0.1 \). This is demonstrated in Figure 11, where the evolution of different \( \Delta_T \) is shown depending on the initial \( \Delta_c \), together with the corresponding fits.

REFERENCES

Abdalla, F. B., Blake, C., & Rawlings, S. 2010, MNRAS, 401, 743
Ahn, K., Iliev, I. T., Shapiro, P. R., et al. 2012, ApJL, 756, L16
Ahn, K., Iliev, I. T., Shapiro, P. R., & Srisawat, C. 2015, MNRAS, 450, 1486
Alvarez, M. A., & Abel, T. 2012, ApJ, 747, 126
Asaba, S., Ichiki, K., & Tashiro, H. 2016, PhRvD, 93, 023518
Baek, S., Di Matteo, P., Semelin, B., Combes, F., & Revaz, Y. 2009, A&A, 495, 389
Bandura, K., Addison, G. E., Amiri, M., et al. 2014, Proc. SPIE, 9145, 914522
Barkana, R., & Loeb, A. 2004, ApJ, 609, 474

APPENDIX B
FITTING FORMULA FOR THE LARGE-SCALE TEMPERATURE FLUCTUATION

For the evolution of \( \Delta_T \), we integrate Equation (8) on each spatial patch. Early on, its value is strongly affected by the initial fluctuation of the CMB temperature. Later, it decouples from the CMB fluctuation and is determined mostly by \( \Delta_0 \).
