OUTLINE

1. The homogeneous Universe
   (a) Distance
   (b) Dynamics
2. Dark Matter Structures
   (a) Linear evolution of density perturbations
   (b) Spherical collapse model for non-linear evolution
   (c) Excursion-set formalism and halo mass functions
   (d) Lagrangian perturbation theory: Zel'dovich approximation
   (e) N-body simulations
   (f) The halo model
   (g) Mid-term exam: writing the mass function calculation
3. Baryonic Structures
   (a) The formation of galaxies
      i. Linear evolution with pressure
      ii. Cosmological Jeans mass
      iii. Thermal evolution of collapsing gas
      iv. The first stars and black holes
      v. Analytic models of galaxy evolution and star formation
      vi. Empirical trends of galaxy formation
      vii. Radiative transfer
   (b) The intergalactic medium
      i. Ionization evolution: the Epoch of Reionization
      ii. Density evolution and HI substructure
      iii. Thermal evolution
      iv. The cosmic 21-cm signal
   (c) Final exam

1. Intro

-focus on practical knowledge, useful for dark ages, first stars and galaxies, and reionization. this means. skip GR, focus on Newtonian physics + SR most practical, and easier intuition - computer resources? - discuss mid-term and final

2. The homogeneous Universe

2.1. Distances

One of the fundamental concepts in cosmology is "distance". Virtually no measurements would be possible without distances and related quantities. One can compute the distance between two points, A and B, by defining a coordinate system and an associated metric. In classical Euclidean space, a common coordinate system is the Cartesian (see Fig. 1).

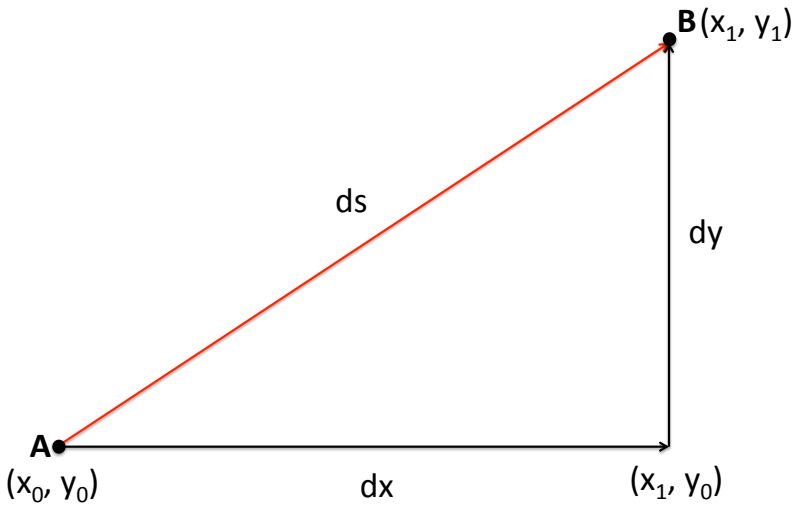

Fig. 1.— Distance in a Cartesian coordinate system in Euclidean space.

Adopting the Cartesian basis vectors $(\hat{\mathbf{i}}_\mathbf{x}, \hat{\mathbf{i}}_\mathbf{y})$, we can label the points A and B as $(x_0\hat{\mathbf{i}}_\mathbf{x}, y_0\hat{\mathbf{i}}_\mathbf{y})$ and $(x_1\hat{\mathbf{i}}_\mathbf{x}, y_1\hat{\mathbf{i}}_\mathbf{y})$ respectively. This allows us to compute the distance between A and B, denoted as ds, using the classical Pythagorean theorem:

$$ds^2 = dx^2 + dy^2 \tag{1}$$

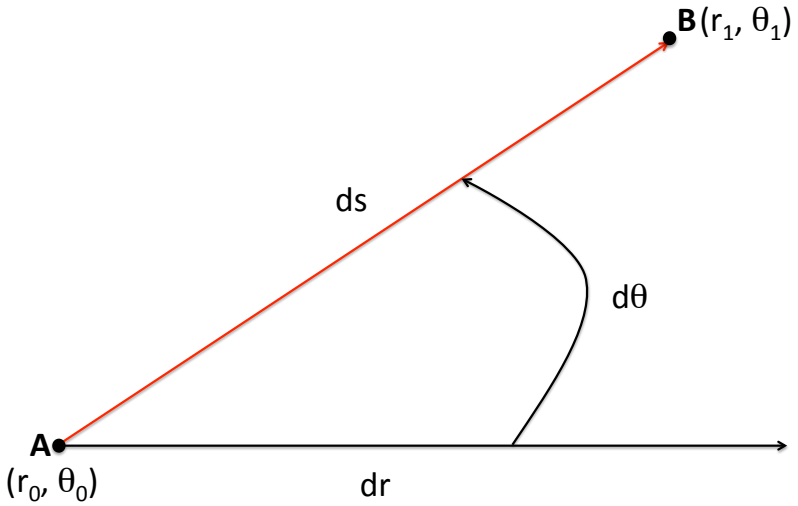

Fig. 2.— Distance in a Polar coordinate system in Euclidean space.

However, there is nothing intrinsic about the choice of coordinate system. We could just as easy change the basis vectors to $(\hat{\mathbf{i}}_\mathbf{r}, \hat{\mathbf{i}}_\theta)$, and adopt the Polar coordinate system (see Fig. 2). In this coordinate system, the distance now is computed according to:

$$ds^2 = dr^2 + r^2 d\theta^2 \tag{2}$$

All we have done is changed the language we use to describe the location of A and B, e.g. point A is now identified as $(r_0\hat{\mathbf{i}}_\mathbf{r}, \theta_0\hat{\mathbf{i}}_\theta)$. Neither points have actually moved. Distance is therefore a so-called scalar invariant. Only the way of computing it depends on the coordinate system. This 'way of computing' is called a metric, $\mathbf{g_{ij}}$. Specifically, we can compute the distance element ds between points at dX^i and dX^j with

$$ds^2 = \sum_{i,j} g_{ij} dX^i dX^j \tag{3}$$

In a Cartesian basis set with $dX = (dx, dy)$, we have:

$$\mathbf{g_{ij}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \tag{4}$$

And in a polar basis set with $dX = (dr, d\theta)$, we have:

$$\mathbf{g_{ij}} = \begin{pmatrix} 1 & 0 \\ 0 & r^2 \end{pmatrix} \tag{5}$$

For spacetime: $ds^2 = \sum_{\mu,\nu} g_{\mu\nu} dX^\mu dX^\nu$, with the Minkowski metric:

$$\mathbf{g}_{\mu\nu} = \begin{pmatrix} -c^2 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{6}$$

Note that if the basis vectors are chosen to be orthogonal, the off-diagonal elements are zero.

What happens if space is expanding? If we assume that the expansion is uniform and isotropic, we obtain the so-called Friedmann Lemaitre Robertson Walker (FLRW) metric for a flat spacetime:

$$\mathbf{g}_{\mu\nu} = \begin{pmatrix} -c^2 & 0 & 0 & 0 \\ 0 & a^2(t) & 0 & 0 \\ 0 & 0 & a^2(t) & 0 \\ 0 & 0 & 0 & a^2(t) \end{pmatrix} \tag{7}$$

Here $a(t)$ is the so-called expansion factor, corresponding to the ratio of the physical separation between two events which are at fixed coordinates (so-called "comoving coordinates"; more on this later). The common convention is to define $a(t)$ such that at present day $a(t = t_0) = 1$ and $a(t = 0) = 0$.

The FLRW metric is more commonly written in spherical comoving coordinates[1]:

$$ds^2 = -c^2 dt^2 + a^2(t) \left[\frac{dR^2}{1 - kR^2} + R^2 d\Omega^2 \right] , \tag{8}$$

where k is the so-called curvature term, and can take the values of $k = -1, 0, 1$ for a closed, flat, open spacetime respectively (more on this later). Generally, we are interested in the separation of two events.

[1]Unless stated otherwise, comoving coordinates will generally be referenced with capital letters or with a subscript "c". Analogously, proper or physical distances will be denoted with small letters or a subscript "p". We adopt the usual convention of $a(t = t_0) = 1$, such that $dr = a(t)dR$.

In this case, we have the freedom to orient our coordinate system radially such that the angular term is zero, i.e. $d\Omega = 0$:

$$ds^2 = -c^2dt^2 + a^2(t)\frac{dR^2}{1-kR^2}. \tag{9}$$

Now let us consider a photon emitted by a stationary (in the comoving frame) source at coordinates (R_1, t_1) and received at coordinates (R_0, t_0). Recalling that light follows a null geodesic, $ds = 0$, we can equate the time and radial components:

$$c^2dt^2 = a^2(t)\frac{dR^2}{1-kR^2} \ . \tag{10}$$

Dividing through by $a^2(t)$ and taking the square root, we obtain:

$$\frac{cdt}{a(t)} = \frac{dR}{\sqrt{1-kR^2}} \ . \tag{11}$$

All of the temporal dependence is on the LHS, while the spatial dependence is on the RHS. We can then relate the spatial and temporal separations of the events (R_1, t_1) and (R_0, t_0) by integrating the above differentials:

$$\int_{t_1}^{t_0} \frac{cdt}{a(t)} = \int_{R_1}^{R_0} \frac{dR}{\sqrt{1-kR^2}} \ . \tag{12}$$

Now suppose that this stationary source (again at R_1) emits two photons, one at t_1 and the other at $t_1 + dt_1$, received by us at t_0 and $t_0 + dt_0$. Since the source is not moving spatially, the RHS in eq. (12) is the same for both photons. Since the spatial separation is the same, we can also equate the temporal separations of the two photons (also dividing out the speed of light):

$$\int_{t_1}^{t_0} \frac{dt}{a(t)} = \int_{t_1+dt_1}^{t_0+dt_0} \frac{dt}{a(t)} \ . \tag{13}$$

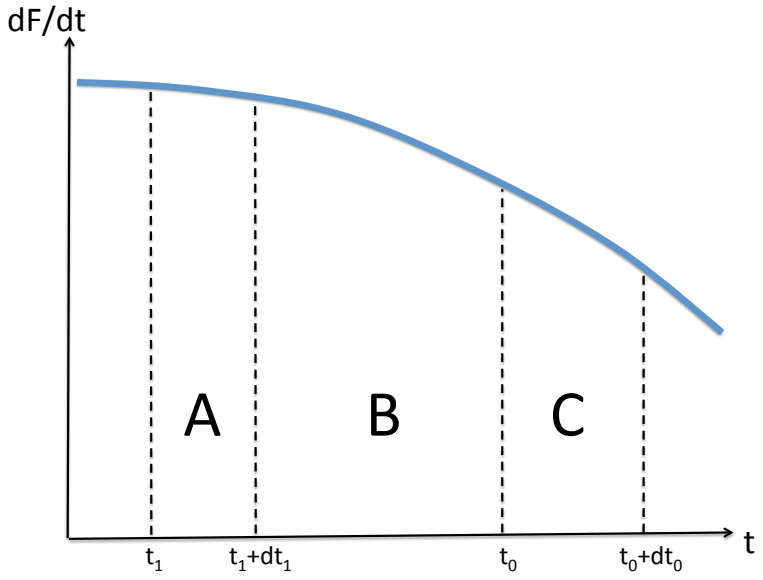

Fig. 3.— An arbitrary differentiable function, with the following definite integrals: $A \equiv \int_{t_1}^{t_1+dt_1} dF$, $B \equiv \int_{t_1+dt_1}^{t_0} dF$, $C \equiv \int_{t_0}^{t_0+dt_0} dF$.

We can rearrange the limits of integration by noting that for any differentiable function, the definite integrals defined in Fig. 3 must obey the following: if A+B (i.e. the LHS of eq. 13) is equal to B+C (i.e. the RHS of eq. 13), then A = C. Therefore we have:

$$\int_{t_1}^{t_1+dt_1} \frac{dt}{a(t)} = \int_{t_0}^{t_0+dt_0} \frac{dt}{a(t)} \ . \tag{14}$$

If we choose a small $dt \ll a/\dot{a}$, we can treat the expansion factor as a constant over the above definite integrals, obtaining:

$$\frac{dt_1}{a(t_1)} = \frac{dt_0}{a(t_0)}$$

$$\boxed{dt_0 = \frac{a(t_0)}{a(t_1)} dt_1} \tag{15}$$

When the photons are observed at the present day, $a(t_0) = 1$ and we have $dt_0 = dt_1/a(t_1)$. Equation (15) shows the principle of cosmological time dilation; for an expanding spacetime, a time interval in the past will be shorter than it appears in the present by a factor equal to the ratio of the expansion factors.

If we think of the t_1 and $t_1 + dt_1$ events as the separation of peaks of an electromagnetic wavefront, we arrive at the (non-rigorous[2]) definition of cosmological redshift. Taking the emitted photon frequency, $\nu_1 \propto (dt_1)^{-1}$, and the received photon frequency, $\nu_0 \propto (dt_0)^{-1}$, we obtain:

$$\boxed{\nu_0 = \frac{a(t_1)}{a(t_0)} \nu_1} \, . \tag{16}$$

In an expanding Universe, the energy of a photon scales inversely with the expansion factor. Cosmological redshift is usually defined using the difference of the observed and emitted frequencies:

$$z(t_1) \equiv \frac{\nu_1 - \nu_0}{\nu_0} \, . \tag{17}$$

Substituting in eq. (16), and evaluating at the present day $a(t_0) = 1$:

$$\boxed{1 + z(t_1) \equiv a(t_1)^{-1}}. \tag{18}$$

The redshift, z, is used more often than the expansion factor in astronomy literature.

Finally, there are two distance measures which are used often enough that they have their own definitions: *the luminosity distance* and *the angular diameter distance*. The luminosity distance is motivated by the question, "What is the flux (in erg s^{-1} cm^{-2}) arriving at redshift z_0 from a source at redshift z_1 that has an intrinsic luminosity of $L_{\rm int}$ (in erg s^{-1})?" The flux at z_0 can be expressed as:

$$f(z_0) = L_{\rm int} \left[\frac{1+z_0}{1+z_1}\right] \left[\frac{1+z_0}{1+z_1}\right] \left[\frac{(1+z_0)^2}{4\pi R^2}\right] \, . \tag{19}$$

Here the second term on the RHS accounts for the energy loss from cosmological redshifting, $E = h\nu \propto (1+z)$, the third term accounts for time dilation [since flux is a rate and $dt^{-1} \propto (1+z)$], and the fourth term is the proper surface area of a sphere whose comoving radius, R, extends from z_1 to z_0 [the $(1+z_0)^{-2}$ term converts the comoving area to proper units]. The luminosity distance, d_L, is obtained if we set $z_0 = 0$ (i.e. present day observations):

$$f(z_0 = 0) = \frac{L_{\rm int}}{4\pi R^2 (1+z_1)^2} \equiv \frac{L_{\rm int}}{4\pi d_L^2} \, , \tag{20}$$

with the final form taken to match the classical flux equation in Euclidean space. Thus the luminosity distance to an object at redshift z is just the comoving distance to that object multiplied by $1 + z$:

$$\boxed{d_L = R(1+z)} \, . \tag{21}$$

The angular diameter distance is also defined in analogy to Euclidean space, being the ratio of the physical size of an object at redshift z to the subtended angle: $d_A \equiv r_A/\theta$. Looking at Fig. 4, we note that

$$\frac{R_A}{R} = \tan\theta \approx \theta \, , \tag{22}$$

[2]More rigorously, it can be shown that the momentum $|p| \propto a^{-1}$ so the energy $E = |p|c = h\nu \propto a^{-1}$ (e.g. Padmanabhan 1993).

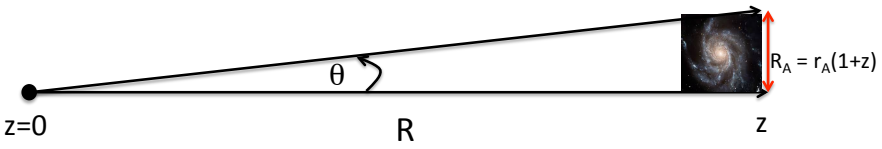

Fig. 4.— Schematic illustrating the definition of angular diameter distance.

with the final step assuming the small angle approximation. With the relations $\theta \approx R_A/R$ and $r_A = R_A/(1+z)$ in hand, we can express the angular diameter distance in terms of the comoving distance to z:

$$d_A \equiv \frac{r_A}{\theta} = \frac{R_A}{(1+z)}\frac{R}{R_A}$$

$$\boxed{d_A = \frac{R}{(1+z)}} \tag{23}$$

2.2. Dynamics

The evolution of the homogeneous (i.e. average) Universe is described through the Friedman equations (FEs), derived from Einstein's equations assuming an isotropic and homogeneous stress-energy tensor. Letting $c = 1$, the two independent FEs are:

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}(\rho + 3p) \tag{24}$$

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - ka^{-2} \tag{25}$$

Here ρ is the energy density and p the pressure. The rigorous derivation of these equations from general relativity will not be covered here; however, to build some intuition it is common to express the analogous dynamics with a Newtonian analogy. We will use this same set-up in the following sections to follow the evolution of (sub-horizon) density perturbations.

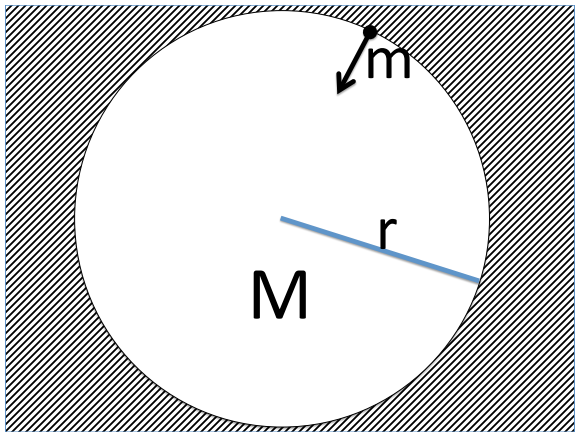

Fig. 5.— An arbitrarily chosen spherical region of mass M and proper radius r. A Newtonian analogy to the FEs is obtained by studying the equations of motion and energy conservation of a test particle of mass m.

To do so, we take a spherical region in a homogeneous background, shown in Fig. 5. A test particle at some proper distance r from the center will only be affected by the enclosed mass, M. In the Newtonian regime, its motion will follow:

$$\cancel{m}\ddot{r} = -\frac{GM\cancel{m}}{r^2} \ . \tag{26}$$

Because the choice of the region size is arbitrary, we can make the substitution $r \to a$, which encodes the temporal evolution of any physical scale, with an arbitrary normalization. We then have:

$$\ddot{a} = -\frac{G}{a^2}\left[\frac{4}{3}\pi a^3 \rho\right]$$

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\rho \tag{27}$$

Equation (27) looks like the first FE, without the pressure term.[3]

Similarly, we recover the second FE by writing the energy conservation equation, with the sum of the kinetic and potential energy of the particle equaling some constant:

$$\frac{1}{2}\cancel{m}\dot{r}^2 - \frac{GM\cancel{m}}{r} = \text{constant} \ . \tag{28}$$

[3]Note that for relativistic fluids the pressure can be related to the energy density as $p_\gamma = \frac{1}{3}\rho_\gamma \ \to \rho_\gamma = 3p_\gamma$, thus recovering the pre-factor for the pressure in the FE.

Again making the substitution $r \to a$,

$$\frac{1}{2}\dot{a}^2 - \frac{G}{a}\left(\frac{4}{3}\pi a^3 \rho\right) = \mathcal{K}$$
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho + Ka^{-2} \tag{29}$$

Within this analogy, we can note that the curvature is related to negative energy. A flat topology, with $K = 0 \implies k = 0$, implies that the kinetic and potential energy are equal, so an initial expansion will stop at a time $\to \infty$. A closed topology, $K < 0 \implies k > 0$, implies that the potential energy is larger, so an initial expansion will stop and reverse itself. An open topology, $K > 0 \implies k < 0$, implies that the kinetic energy is larger, so an initial expansion will never stop.

Now, let us evaluate this second FE at the present epoch (values at the present epoch will be denoted with subscripts '0').

$$k\cancel{a_0^{-2}}^{1} = \frac{8\pi G}{3}\rho_0 - \left(\frac{\dot{a}_0}{a_0}\right)^2 \tag{30}$$

Defining the so-called Hubble constant as $H \equiv \dot{a}/a$, we have

$$k = \frac{8\pi G}{3}\rho_0 - H_0^2$$
$$\equiv H_0^2(\Omega_0 - 1) \ , \tag{31}$$

where the present-day energy density, Ω_0 is expressed in terms of a "critical" density required to make the topology flat:

$$\Omega_0 \equiv \frac{\rho_0}{\rho_{\rm crit,0}} \tag{32}$$
$$\rho_{\rm crit,0} = \frac{3H_0^2}{8\pi G} \tag{33}$$

The critical density is fairly modest: $\rho_{\rm crit,0} \approx 10^{-29}$ g cm^{-3}, by current estimates. Using these definitions, a flat Universe implies $k = 0 \implies \Omega = 1$, a closed Universe implies $k > 0 \implies \Omega > 1$, and an open Universe implies $k < 0 \implies \Omega < 1$.

What do we think contributes to Ω? The currently-favored cosmology (e.g. Planck Collaboration 2018) is flat ($k = 0$) with three main energy components: (i) radiation, with $\Omega_{\gamma,0} \approx 10^{-4}$; (ii) matter, with $\Omega_{m,0} \approx 0.3$; and (iii) a cosmological constant with $\Omega_{\Lambda,0} \approx 0.7$. We see that although currently the cosmological constant has the largest energy density, the matter energy density is of the same order. This is known as the "coincidence problem", and is a matter of some debate in physics/philosophy. The corresponding energy densities evolve as $\rho_\gamma \propto a^{-4}$ (accounting for photon number density and energy redshifting); $\rho_m \propto a^{-3}$, and $\rho_\Lambda \propto a^0$. We can then ask what form of energy dominated in the past. If we set:

$$\Omega_m = \Omega_\Lambda$$
$$\frac{\rho_{m,0}(1+z)^3}{\rho_{\rm crit}} = \frac{\rho_{\Lambda,0}}{\rho_{\rm crit}}$$

multiplying through by $\rho_{\rm crit}/\rho_{\rm crit,0}$:

$$\Omega_{m,0}(1+z)^3 = \Omega_{\Lambda,0}$$
$$0.3(1+z)^3 = 0.7 \tag{34}$$

Solving for the redshift, we get that before $z \approx 0.3$, the matter density was greater than the cosmological constant.

Performing a similar exercise for matter and radiation:

$$\begin{aligned} \Omega_m &= \Omega_\gamma \\ \frac{\rho_{m,0}(1+z)^3}{\rho_{\rm crit}} &= \frac{\rho_{\gamma,0}(1+z)^4}{\rho_{\rm crit}} \\ \Omega_{m,0}(1+z)^3 &= \Omega_{\gamma,0}(1+z)^4 \\ 0.3(1+z)^3 &= 10^{-4}(1+z)^4 \end{aligned} \tag{35}$$

we get that before $z \approx$ few $\times$ 1000, the Universe was dominated by radiation. Thus we can classify the history of the Universe by its dominant energy component:

$$\begin{aligned} z &\lesssim 0.3 && \Longrightarrow \Lambda \text{ Dominated} \\ 0.3 \lesssim z &\lesssim \text{few} \times 1000 && \Longrightarrow \text{Matter Dominated (MD)} \\ z &\gtrsim \text{few} \times 1000 && \Longrightarrow \text{Radiation Dominated (RD)} \end{aligned}$$

We see that for the formation of early structures (e.g. $5 \lesssim z \lesssim 40$), *the MD regime is the most relevant.* For the remainder of the course, we will be working in the MD regime: $\Omega = \Omega_m = 1$.

3. Dark Matter structures: pressureless collapse

3.1. Linear evolution of gravitational instabilities

Thankfully (for the sake of our existence) the Universe is only homogeneous on large scales. The perturbations of the CMB are of order $\sim 10^{-5}$ – 10^{-6} (e.g. Komatsu et al. 2011)[4]. How do the initial seed perturbations start growing?

We shall first concern ourselves with pressure-less collapse. Physically, this means that we are studying either scales larger than the Jeans length, and/or dark matter (DM) evolution. In addition to ignoring pressure, we will work in the Newtonian regime. This corresponds to assuming: (i) the scales of perturbations are $\ll$ the horizon (cH^{-1}); and (ii) particles are non-relativistic.

Locally, we can write the usual fluid equations (mass conservation, Euler and Poisson) in proper coordinates (explicitly denoted with a subscript 'p'):

$$\frac{\partial \rho}{\partial t} + \nabla_{\mathbf{p}} \cdot (\rho \mathbf{v_p}) = 0 \tag{36}$$

$$\frac{\partial \mathbf{v_p}}{\partial t} + (\mathbf{v_p} \cdot \nabla_{\mathbf{p}})\mathbf{v_p} = -\nabla_{\mathbf{p}}\Phi \tag{37}$$

$$\nabla_p^2 \Phi = 4\pi \mathrm{G}\rho \tag{38}$$

We begin by transforming these to our comoving coordinate system. We have the following scalings:

Distance:

$$\mathbf{r_p} = a\mathbf{R} \tag{39}$$

Velocity:

$$\begin{aligned} \mathbf{V} &= \dot{\mathbf{R}} = a^{-1}\dot{\mathbf{r}}_{\mathbf{p}} - a^{-2}\dot{a}\mathbf{r_p} \\ \mathbf{v_p} &= \dot{\mathbf{r}}_{\mathbf{p}} = \dot{a}\mathbf{R} + a\mathbf{V} \equiv \dot{a}\mathbf{R} + \mathbf{v}_{\rm pec} \ , \end{aligned} \tag{40}$$

where the two final terms correspond to the Hubble flow and the peculiar velocity, which is kept in proper units because of (somewhat arbitrary) convention.

Gradient:

$$\nabla_{\mathbf{p}} = a^{-1}\nabla \tag{41}$$

And converting the fluid derivative from Eulerian, $\partial/\partial t$, to Lagrangian, d/dt:

$$\begin{aligned} \frac{d}{dt} &= \frac{\partial}{\partial t} + \mathbf{v_p} \cdot \nabla_{\mathbf{p}} \\ &= \frac{\partial}{\partial t} + (\dot{a}\mathbf{R} + \mathbf{v}_{\rm pec}) \cdot a^{-1}\nabla \\ &= \frac{\partial}{\partial t} + a^{-1}\dot{a}\mathbf{R} \cdot \nabla + \cancelto{0}{a^{-1}\mathbf{v}_{\rm pec} \cdot \nabla} \\ \frac{\partial}{\partial t} &= \frac{d}{dt} - a^{-1}\dot{a}\mathbf{R} \cdot \nabla \end{aligned} \tag{42}$$

where we use the fact that by construction the dot product of the peculiar velocity and *comoving* gradient goes to zero.

Finally, we write the density in terms of the comoving (present day) density:

$$\rho = \rho_0 a^{-3} \tag{43}$$

[4]Note that these correspond to acoustic waves in the photon-baryon plasma. Dark matter decouples earlier and has a "head-start" in the formation of structure. We shall return to this point later.

Now let's "seed" our structure by introducing a small perturbation in the matter density[5], $\delta \ll 1$, around the background mean density, $\bar{\rho}$:

$$\rho = \bar{\rho}(1+\delta) = \bar{\rho}_0 a^{-3}(1+\delta) \ , \tag{44}$$

where $\bar{\rho}_0$ is the comoving (present-day) density.

Substituting the above into eq. (36):

$$\left(\frac{d}{dt} - a^{-1}\dot{a}\mathbf{R}\cdot\nabla\right)[\cancel{\bar{\rho}}_0 a^{-3}(1+\delta)] + a^{-1}\nabla\cdot[\cancel{\bar{\rho}}_0 a^{-3}(1+\delta)(\dot{a}\mathbf{R}+\mathbf{v}_{\rm pec})] = 0$$
$$-3a^{-4}\dot{a}(1+\delta) + a^{-3}\dot{\delta} - a^{-4}\dot{a}\mathbf{R}\cdot\nabla(1+\delta) + a^{-4}\dot{a}\nabla\cdot[(1+\delta)\mathbf{R}] + a^{-4}\nabla\cdot[(1+\delta)\mathbf{v}_{\rm pec}] = 0$$

using $\nabla\cdot[(1+\delta)\mathbf{R}] = (1+\delta)\nabla\cdot\mathbf{R} + \mathbf{R}\cdot\nabla(1+\delta)$ and $\nabla\cdot\mathbf{R} = 3$, and multiplying through by a^3, we can expand and cancel:

$$\cancel{-3a^{-1}\dot{a}(1+\delta)} + \dot{\delta} - \cancel{a^{-1}\dot{a}\mathbf{R}\cdot\nabla(1+\delta)} +$$
$$\cancel{a^{-1}\dot{a}(1+\delta)\nabla\cdot\mathbf{R}}^{3} + \cancel{a^{-1}\dot{a}\mathbf{R}\cdot\nabla(1+\delta)} + a^{-1}\nabla\cdot[(1+\delta)\mathbf{v}_{\rm pec}] = 0 \tag{45}$$

thus we finally obtain the Lagrangian comoving form of mass conservation (eq. 36):

$$\dot{\delta} + a^{-1}\nabla\cdot[(1+\delta)\mathbf{v}_{\rm pec}] = 0 \ . \tag{46}$$

Keeping only the lowest order terms (i.e. dropping $\delta\mathbf{v}_{\rm pec}$):

$$\dot{\delta} + a^{-1}\nabla\cdot\mathbf{v}_{\rm pec} = 0 \tag{47}$$

We can repeat this substitution exercise into eq. (37-38), and keeping only the perturbed part of the potential, we get:

$$\dot{\mathbf{v}}_{\rm pec} + \frac{\dot{a}}{a}\mathbf{v}_{\rm pec} + a^{-1}\nabla\phi = 0 \tag{48}$$
$$\nabla^2\phi = 4\pi \mathrm{G}\bar{\rho}_0 a^{-1}\delta \tag{49}$$

Finally, we can combine these three equations into a single equation for the evolution of the overdensity, δ, in the linear regime. We take the time derivative of eq. (47):

$$\ddot{\delta} - \frac{\dot{a}}{a^2}\nabla\cdot\mathbf{v}_{\rm pec} + a^{-1}\nabla\cdot\dot{\mathbf{v}}_{\rm pec} = 0 \tag{50}$$

and $a^{-1}\nabla\cdot$ of eq. (48):

$$a^{-1}\nabla\cdot\dot{\mathbf{v}}_{\rm pec} + \frac{\dot{a}}{a^2}\nabla\cdot\mathbf{v}_{\rm pec} + a^{-2}\nabla^2\phi = 0 \tag{51}$$

Now subtracting (51) from (50), and substituting in the Poisson eq. (49):

$$\ddot{\delta} - 2\frac{\dot{a}}{a^2}\nabla\cdot\mathbf{v}_{\rm pec} - 4\pi\mathrm{G}\bar{\rho}_0 a^{-3}\delta = 0 \tag{52}$$

And finally replacing the peculiar velocity with the relation in (47), we obtain the evolution of the overdensity:

$$\boxed{\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} - 4\pi\mathrm{G}\bar{\rho}_0 a^{-3}\delta = 0} \ . \tag{53}$$

[5]In principle, one can perturb other quantities as well (see e.g. Tseliakhovich & Hirata 2010 and references therein). However the matter perturbations are dominant.

From the above, we can identify the second term as the Hubble expansion, which initially damps the growth of perturbations. During matter domination, the second Friedman equation gives us: $a \propto t^{2/3}$ and $\dot{a} \propto t^{-1/3}$.[6] Replacing the scale factor with these relations yields:

$$\ddot{\delta} + C_1 t^{-1}\dot{\delta} - C_2 t^{-2}\delta = 0 \ , \tag{55}$$

where C_1 and C_2 are constants. This differential equation has a solution of the form:

$$\delta = At^{2/3} + Bt^{-1} \ , \tag{56}$$

Evaluating the constants, one obtains $A = 3/5$ and $B = 2/5$. The terms on the RHS are called the "growing" and "decaying" modes, respectively. The growing mode (increasing with time), dominates structure formation. Thus we finally come to the important relation governing the evolution of density perturbations in the linear regime during matter domination:

$$\boxed{\delta \propto t^{2/3} \propto a} \ . \tag{57}$$

[6]This can be obtained from the second FE:

$$\begin{aligned} (\dot{a}/a)^2 &\propto \rho \\ \dot{a}^2 &\propto a^2 a^{-3} \\ da/dt &\propto a^{-1/2} \\ \int da\sqrt{a} &\propto \int dt \\ a^{3/2} &\propto t \\ a \propto t^{2/3} \quad &\text{and} \quad \dot{a} \propto t^{-1/3} \end{aligned} \tag{54}$$

where the energy density scallings are valid during MD.

3.2. Spherical collapse model

We consider spherical perturbations in a uniform background (e.g. Fig. 6). As above, we only consider perturbations in a non-relativistic fluid, on scales much smaller than the Hubble length. This allows us to use Newtonian physics. Our set-up is the same as we used to interpret the Friedman equations in §2.2, except that the enclosed density is now initially $\rho_i = \bar{\rho}_i(1+\delta_i)$. *We will work in physical units in this section*, since the dynamics of the shell are more intuitive in physical units.

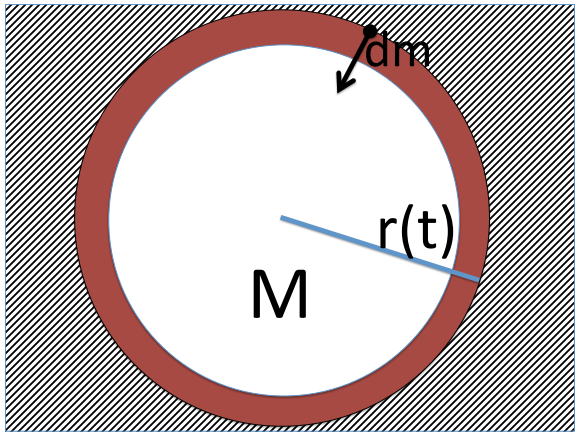

Fig. 6.— A linear spherical perturbation in the matter density, with mass M and mean overdensity $(1+\delta) = \rho/\bar{\rho}$. We derive the evolution of the enclosing shell.

The mean density evolves as $\bar{\rho} \propto (1+z)^3 \propto t^{-2}$. Like we did above, we again write the equation of energy conservation per unit mass for the mass shell:

$$\mathrm{KE} + \mathrm{PE} = E_{\rm tot} \tag{58}$$

$$\frac{1}{2}\dot{r}^2 - \frac{\mathrm{G}M}{r} = E_{\rm tot} \tag{59}$$

Assuming the initial (as some sufficiently early time) bulk velocity is correlated on scales larger than the perturbation, the initial (denoted by subscript 'i') proper velocity of the shell is just the Hubble flow: $\dot{r}_i = (\dot{a}_i/a_i)r_i = H_i r_i$. We can write the corresponding kinetic energy as:

$$\mathrm{KE}_i = \frac{H_i^2 r_i^2}{2} \ , \tag{60}$$

and the potential energy as

$$\mathrm{PE}_i = -\frac{\mathrm{G}M}{r_i} = -\frac{\mathrm{G}}{r_i}\frac{4}{3}\pi r_i^3 \bar{\rho}_i(1+\delta_i) \ . \tag{61}$$

remembering the definition $\Omega_i = \bar{\rho}_i/\rho_{\rm crit,i} = \bar{\rho}_i(8\pi\mathrm{G})/(3H_i^2)$, we can write:

$$\begin{aligned}\mathrm{PE}_i &= -\Omega_i \left(\frac{3H_i^2}{8\pi\mathrm{G}}\right)\frac{4}{3}\pi\mathrm{G}r_i^2(1+\delta_i) \\ &= -\frac{1}{2}\Omega_i H_i^2 r_i^2(1+\delta_i) \ ,\end{aligned} \tag{62}$$

and we identify the enclosed mass as:

$$M = \frac{1}{2\mathrm{G}}\Omega_i H_i^2 r_i^3(1+\delta_i) \ . \tag{63}$$

Note that in this approximation, the enclosed mass remains constant, as the mass shells only cross after collapse.

With the above terms, we can write the total energy (which does not change with time) as:

$$E_{\rm tot} = \frac{H_i^2 r_i^2}{2}[1 - \Omega_i(1+\delta_i)] \ . \tag{64}$$

So the perturbation collapses if $E_{\rm tot} < 0$, which implies:

$$\boxed{(1+\delta_i) > \Omega_i^{-1}} \tag{65}$$

Here we see explicitly that in a closed or a flat Universe all overdensities will eventually collapse. In an open Universe, only some overdensities satisfying the criterion in eq. (65) will collapse.

What is the maximum proper radius when the shell starts to fall back or turn around, $r_{\rm turn}$? Again, we can write energy conservation at the turn around point:

$$\begin{aligned} \frac{1}{2}\cancelto{0}{\dot{r}^2_{\rm turn}} - \frac{\rm G M}{r_{\rm turn}} &= E_{\rm tot} \\ -\frac{\rm G M}{r_{\rm turn}} &= \frac{H_i^2 r_i^2}{2}[1-\Omega_i(1+\delta_i)] \ . \end{aligned}$$

Putting-in the enclosed mass from eq. (63),

$$\begin{aligned} -\frac{\rm G}{r_{\rm turn}}\left[\frac{1}{2\rm G}\Omega_i H_i^2 r_i^3 (1+\delta_i)\right] &= \frac{H_i^2 r_i^2}{2}[1-\Omega_i(1+\delta_i)] \\ r_{\rm turn} = -\frac{\Omega_i r_i (1+\delta_i)}{1-\Omega_i(1+\delta_i)} &= \frac{r_i(1+\delta_i)}{-\Omega_i^{-1}+(1+\delta_i)} \\ r_{\rm turn} = \frac{r_i(1+\delta_i)}{\delta_i - (\Omega^{-1}-1)} &\approx \frac{r_i}{\delta_i} \ , \end{aligned} \tag{66}$$

where the last approximation assumes matter domination, and only keeps the highest order in $\delta_i \ll 1$. Again, we see that as $\delta_i \to \Omega_i^{-1} - 1$, $r_{\rm turn} \to \infty$.

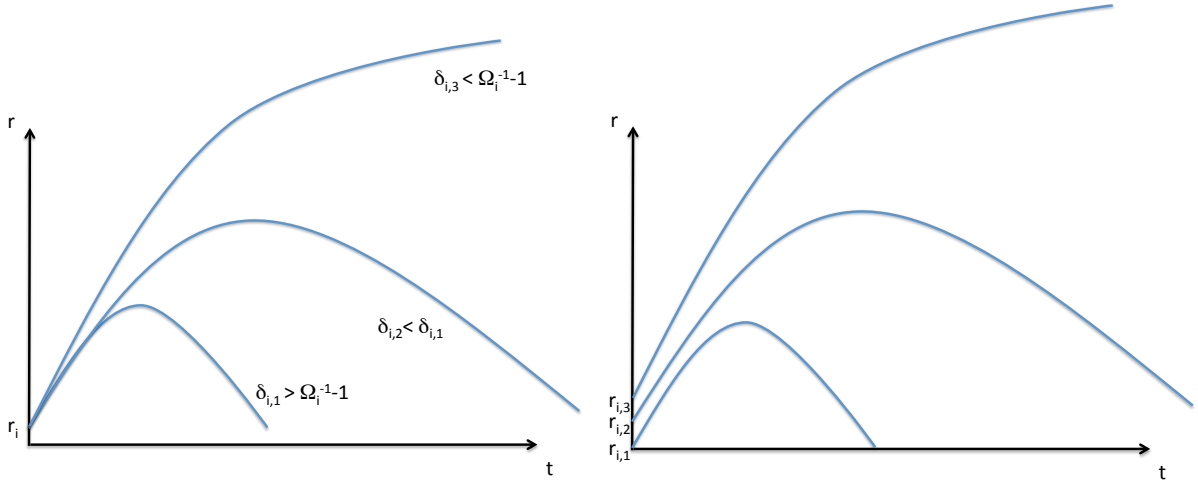

Fig. 7.— *Left:* Evolution of mass shells enclosing disparate initial perturbation amplitudes. *Right:* Evolution of mass shells enclosing disparate initial perturbation amplitudes, around the same point, with the perturbation decreasing with distance.

More over-dense shells collapse at smaller radii, with smaller maximum extents before turnaround. This is shown in the left panel of Fig. 7. Under-dense regions never collapse. Or, alternatively as shown in the right panel, one can think of shells surrounding a single point, where the mean enclosed perturbation decreases with initial size (we will get to what this means later). In the limit of $r_i \to \infty$, the perturbation is the Universe, and the max extent is reached at $t = \infty$, assuming MD.

The full evolution of the shell radius for collapsing shell from the above energy conservation equations

has the functional form:

$$r = A(1 - \cos\theta) \tag{67}$$
$$t = B(\theta - \sin\theta) \tag{68}$$
$$A^3 = \mathrm{G}MB^2 \ , \tag{69}$$

where θ is a parametrization of cosmic time and it increases with increasing time. Armed with this functional form, we can compute the evolution of the density perturbation,

$$1 + \delta(r,t) = \frac{\rho(r,t)}{\bar{\rho}(t)} = M\left(\frac{4}{3}\pi r^3\right)^{-1}\bar{\rho}(t)^{-1} \ . \tag{70}$$

The mean density in a $\Omega_i = 1$ (MD) Universe evolves as (remembering that during MD $2/(3H_i) = t_i$ and $(t/t_i)^{2/3} = a/a_i$):

$$\begin{aligned}\bar{\rho}(t) = \bar{\rho}_i\left(\frac{a}{a_i}\right)^{-3} &= \cancel{\Omega_i}^{1}\rho_{\mathrm{crit,i}}\left(\frac{a}{a_i}\right)^{-3} = \frac{3H_i^2}{8\pi\mathrm{G}}\left(\frac{t}{t_i}\right)^{-2} \\ &= \frac{3H_i^2}{8\pi\mathrm{G}}\left(\frac{2}{3}H_i^{-1}\right)^2 t^{-2} \ = \ \frac{4}{24\pi\mathrm{G}t^2} \ = \ \frac{1}{6\pi\mathrm{G}t^2} \ .\end{aligned} \tag{71}$$

Substituting eq. (71) and eq. (67)–(69), into eq. (70):

$$\begin{aligned}1 + \delta(t) &= \frac{A^3}{\mathrm{G}B^2}\left[\frac{4}{3}\pi A^3(1-\cos\theta)^3\right]^{-1} 6\pi\mathrm{G}B^2(\theta - \sin\theta)^2 \\ &= \frac{9}{2}\frac{(\theta - \sin\theta)^2}{(1-\cos\theta)^3} \ .\end{aligned} \tag{72}$$

In the limit of small θ, we recover the linear evolution equations from §3.1.

We can also solve explicitly for the parameters A and B from $r_{\mathrm{turn}} = r(\theta = \pi) = 2A$. From eq. (66), we have:

$$A = \frac{r_i(1+\delta_i)}{2[\delta_i - (\Omega_i^{-1} - 1)]} \tag{73}$$

substituting A and M into eq. (67)–(69) above to obtain B:

$$\begin{aligned}A^3 \ = \ \mathrm{G}MB^2 \ &= \ \mathrm{G}B^2\left(\frac{\Omega_i H_i^2 r_i^3(1+\delta_i)}{2\mathrm{G}}\right) \\ \frac{r_i^3(1+\delta_i)^3}{8[\delta_i - (\Omega_i^{-1}-1)]^3} \ &= \ \frac{B^2}{2}\Omega_i H_i^2 r_i^3(1+\delta_i) \\ B &= \frac{(1+\delta_i)}{2H_i\Omega_i^{1/2}[\delta_i - (\Omega_i^{-1}-1)]^{3/2}} \ .\end{aligned} \tag{74}$$

Assuming $\Omega_i = 1$ (MD), and keeping the leading order in $\delta_i \ll 1$, we obtain:

$$A \ = \ \frac{r_i(1+\delta_i)}{2\delta_i} = \frac{r_i}{2\delta_i} + \frac{r_i}{2} \approx \frac{r_i}{2\delta_i} \tag{75}$$
$$B \ = \ \frac{(1+\delta_i)}{2H_i\delta_i^{3/2}} = \frac{\delta_i^{-3/2}}{2H_i} + \frac{\delta_i^{-1/2}}{2H_i} \approx \frac{\delta_i^{-3/2}}{2H_i} \ . \tag{76}$$

Substituting A and B into eq. eq. (67)–(68), we obtain the evolution of the mass shell in a flat MD universe:

$$r = \frac{r_i}{2\delta_i}(1 - \cos\theta) \tag{77}$$
$$t = \frac{\delta_i^{-3/2}}{2H_i}(\theta - \sin\theta) \tag{78}$$

Equation 78 is more useful converting to redshift, again using $2/(3H_i) = t_i$ and $(t/t_i)^{2/3} = (1+z_i)/(1+z)$:

$$\begin{aligned} t &= \frac{\delta_i^{-3/2}}{2}\left(\frac{3}{2}t_i\right)(\theta - \sin\theta) \\ \frac{(1+z_i)^{3/2}}{(1+z)^{3/2}} &= \frac{3}{4}\delta_i^{-3/2}(\theta - \sin\theta) \\ (1+z) &= (1+z_i)\left(\frac{4}{3}\right)^{2/3}\delta_i(\theta - \sin\theta)^{-2/3} \ . \end{aligned} \tag{79}$$

If we define collapse at $\theta = 2\pi$ [i.e. $\delta(z_{\rm coll}) \to \infty$], we can solve for the redshift of collapse:

$$(1+z_{\rm coll}) = \delta_i(1+z_i)\left(\frac{4}{3}\right)^{2/3}(2\pi)^{-2/3} = 0.356\ \delta_i(1+z_i) \ . \tag{80}$$

Notice that the collapse occurs roughly on the time-scale of the Hubble time. So let's say a perturbation has $\delta_i = 0.1$ (already quasi-linear) at $z_i \approx 100$. It collapses at $z_{\rm coll} \approx 2.5$. This highlights that initial conditions for structure formation or N-body simulations must be started at very high redshifts.

What about a $\delta_i = 0.01$ perturbation at $z_i \approx 100$? It hasn't collapsed yet (and maybe never will, given that we are no longer in a matter-dominated regime). What about $\delta_i \sim 10^{-6}$ at $z_i \sim 1000$ (like we see in the CMB)? We now explicitly see that *without dark matter decoupling from the CMB much earlier than $z_i \sim 1000$, we would not be here!*

We would like to standardize an "initial redshift" when talking about perturbations, so we do not have to refer to both δ_i and $(1+z_i)$. Noting that in eq. (80) the terms δ_i and $(1+z_i)$ appear as a product, we can do so using the linear evolution result from the previous section. Specifically, in §3.1 we saw that the linear evolution of density perturbations follows:

$$\delta(z) \ = \ \frac{3}{5}\delta_i\left(\frac{1+z_i}{1+z}\right) \ + \ \frac{2}{5}\delta_i\left(\frac{1+z_i}{1+z}\right)^{-3/2} \ . \tag{81}$$

So 3/5 of the initial amplitude is in the 'growing mode'. It is common to write perturbations in terms of their $z = 0$ linear extrapolations of their growing mode:

$$\delta_0 \equiv \frac{3}{5}\delta_i(1+z_i) \ . \tag{82}$$

With this definition, we can re-write eq. (80) into the more common form:

$$\boxed{(1+z_{\rm coll}) \ = \ \frac{\delta_0}{1.686}} \tag{83}$$

Structures collapse in the spherical collapse model when their linearly extrapolated over-density is $\delta_0 = \delta_c \equiv 1.686$. In other words, we can associate with each linearly-extrapolated mode a collapse redshift according to eq. (83). We will see in the next section that this is an important ingredient in the analytic framework for modeling the abundance of DM halos.

What actually happens at "collapse"? According to the simple assumptions above that shells do not cross, the structure should collapse to a point. Actually, lacking pressure, shells would go through the center in a sort of damped harmonic oscillator. We call this period "virialization" as the inner structure settles into gravitational equilibrium. This non-linear evolution is difficult to model from first principles. However we can roughly predict when this occurs, again using simple energy conservation arguments, as well as the virial theorem:

$$2\ {\rm KE}_{\rm vir} + {\rm PE}_{\rm vir} = 0 \ . \tag{84}$$

Recalling that at the maximum extent of the shell (i.e. turnaround), we had $\mathrm{KE_{turn}} = 0$ and $\mathrm{E_{tot}} = \mathrm{PE_{turn}} = -\mathrm{G}M/r_{\mathrm{turn}}$, we can write at virialization:

$$\begin{aligned} \mathrm{KE_{vir}} + \mathrm{PE_{vir}} &= E_{\mathrm{tot}} \\ \frac{\mathrm{PE_{vir}}}{2} &= -\frac{\mathrm{G}M}{r_{\mathrm{turn}}} \\ -\frac{\mathrm{G}M}{2r_{\mathrm{vir}}} &= -\frac{\mathrm{G}M}{r_{\mathrm{turn}}} \\ r_{\mathrm{vir}} &= \frac{r_{\mathrm{turn}}}{2} \ . \end{aligned} \tag{85}$$

We can also ask what is the enclosed density at virialization. To answer this, let us first evaluate the other quantities at turnaround (r_{turn}). The over-density can be computed from eq. (72), evaluated at $\theta = \pi$:

$$\begin{aligned} (1+\delta_{\mathrm{turn}}) &= \frac{9}{2}\frac{(\theta-\sin\theta)^2}{(1-\cos\theta)^3} = \frac{9}{2}\frac{(\pi-0)^2}{2^3} \\ (1+\delta_{\mathrm{turn}}) &= \frac{9\pi^2}{16} \ , \end{aligned} \tag{86}$$

and the redshift of turnaround can be expressed in terms of the collapse redshift exploiting the scaling $(1+z) \propto (\theta - \sin\theta)^{-2/3}$ (eq. 79):

$$\begin{aligned} 1+z_{\mathrm{turn}} &= (1+z_{\mathrm{coll}})\frac{(\pi-\sin\pi)^{-2/3}}{(2\pi-\sin 2\pi)^{-2/3}} \ = \ (1+z_{\mathrm{coll}})\frac{\pi^{-2/3}}{(2\pi)^{-2/3}} \\ 1+z_{\mathrm{turn}} &= 2^{2/3}(1+z_{\mathrm{coll}}) \ . \end{aligned} \tag{87}$$

Using the relations (85)–(87), we can obtain the over-density at virialization:

$$\begin{aligned} \rho_{\mathrm{vir}} \ &= \ 2^3\rho_{\mathrm{turn}} \ = \ 8\left[\bar{\rho}_0(1+z_{\mathrm{turn}})^3\frac{9\pi^2}{16}\right] \ = \ 8\bar{\rho}_0\left[2^2(1+z_{\mathrm{coll}})^3\frac{9\pi^2}{16}\right] \\ \rho_{\mathrm{vir}} &= 18\pi^2\bar{\rho}_0(1+z_{\mathrm{coll}})^3 = 180\bar{\rho}_{z=z_{\mathrm{coll}}} \ . \end{aligned} \tag{88}$$

So we see that the mean density of virialized objects is roughly ~ 200 times the background density. This is an important prediction of the spherical collapse model, and is approximately confirmed by numerical simulations[7].

To summarize, we can see the (physical) evolution of a collapsing mass shell in the spherical collapse model in Fig. 8. Starting from some initial radius r_i, the shell expands initially with the Hubble flow, then slows down until it reaches a maximum radius at turnaround, $r_{\mathrm{turn}} = r_i(1+\delta_i)/[\delta_i - (\Omega_i^{-1}-1)] \approx r_i/\delta_i$ (where the final approximation assumes MD). Turnaround occurs at a scale factor which is $2^{-2/3}$ times the collapse scale factor. The collapse occurs when the growing mode has a linearly-extrapolated amplitude of 1.686. Shell crossing however occurs somewhat before that, and the approximations break down. The mean density of the virialized object is $18\pi^2$ times the background density.

This model is used in many applications, including estimating mass, radius, virial temperature, collapse redshift, and halo abundances (as we will see in the next section).

[7]It should be noted that in numerical simulations, it is somewhat ambiguous how to define a halo; an issue we will return to later in §. One of such 'halo finding' algorithms involves the mean density definition in eq. (88).

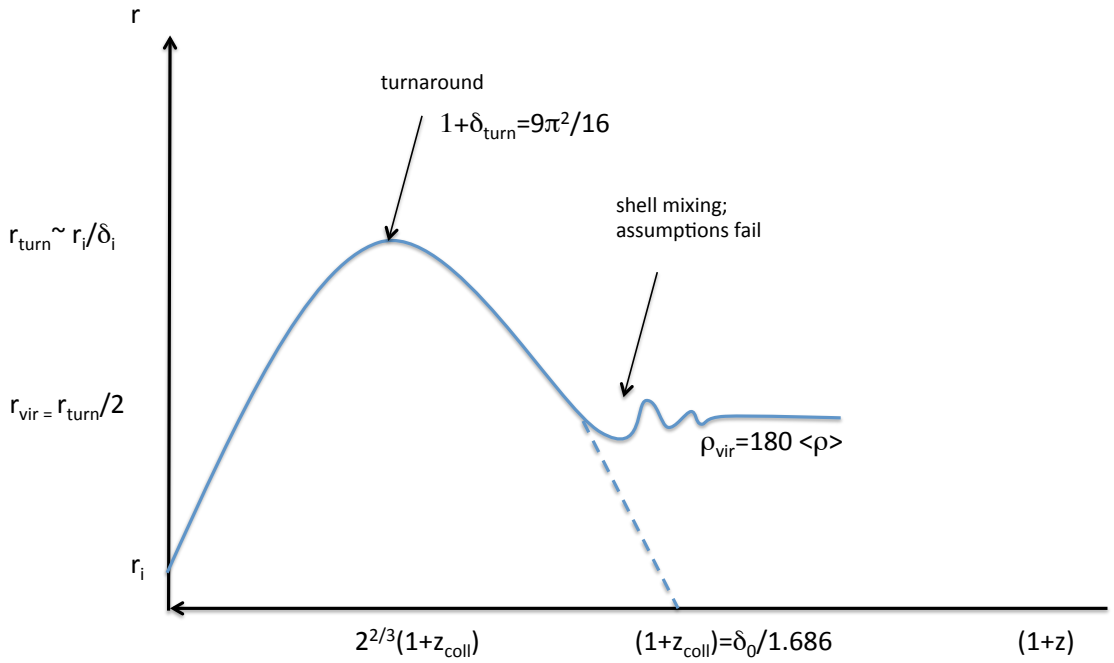

Fig. 8.— Evolution of a collapsing mass shell in the spherical collapse model, as described in the text.

3.3. Statistics of Perturbations: Excursion set theory and halo mass functions

Up to now, we have discussed the evolution of a single density perturbation. The Universe is comprised of many perturbations (i.e. structures). This means that we need to develop statistical tools to describe these structures, and compare them to theoretical predictions. This entails: (i) quantifying the initial conditions (ICs) which generated the perturbations; (ii) knowledge of how they evolve with time, and (iii) a way of comparing a given evolved realization[8] to a cosmological model. Step (ii) can be done either analytically (for example, using the spherical collapse model from the previous section), or with numerical simulations. Here we define the statistics needed for (i) and (iii). We closely follow Padmanabhan (1993), largely keeping the same notation.

Let us define a joint probability of a given realization, $P[\delta(\mathbf{x}); t]$. This is the *joint* probability that at time t, the density at position $\mathbf{x_1}$ is $\bar{\rho}(t)[1+\delta(\mathbf{x_1}, t)]$, *and* the density at position $\mathbf{x_2}$ is $\bar{\rho}(t)[1+\delta(\mathbf{x_2}, t)]$, and so forth... As time progresses, neighboring perturbations influence each others growth through gravity. However at very early times, one could approximate the perturbations as evolving independently of one another. This is more true in k-space[9] than in real space, allowing different k-modes to remain independent of one another for longer. The independence of modes at early times allows us to write out the joint probability as a product of the probabilities of individual modes:

$$P[\delta(\mathbf{k}); t] \equiv P[\delta(\mathbf{k_1}), \delta(\mathbf{k_2}), ...; t] \approx \prod g_k[\delta(\mathrm{k}); t] \ , \tag{89}$$

where g_k is the probability of a *single* mode having a specific amplitude and phase $\delta_k = r_k e^{i\phi_k}$. [10]

Note that g_k can be thought of being constructed as a sum over (infinitely) many δ_x (i.e. its Fourier conjugate). Since at early times, δ_x can be thought of as an uncorrelated (or weakly correlated) random variable, the central limit theorem motivates adopting a Gaussian distribution for g_k. A Gaussian random field has the property of being solely defined by its mean (which for our density contrast field is zero by definition) and its variance:

$$\langle \delta_k \rangle = 0 \qquad \langle |\delta_k|^2 \rangle \equiv \sigma_k^2 \qquad \langle \delta_k \delta_p \rangle = 0 \ \ \text{(for all } k \neq p) \tag{90}$$

In other words, the phases of g_k are uncorrelated, and the amplitudes are statistically specified by the variance.

Analogously, the real-space density contrast δ_x at any given point is a sum over many uncorrelated random variables δ_k, and can thus also be described as a Gaussian field according to the central limit theorem. The probability of a real-space location $\mathbf{x}$ having a density contrast with value $\delta_x = q$ at time t can be expressed as:

$$P[q] = \frac{1}{\sqrt{2\pi\Delta_x^2}} e^{-\frac{q^2}{2\Delta_x^2}} \ , \tag{91}$$

where the real-space variance, Δ_x^2, can be written out in terms of Fourier components as

$$\Delta_x^2 \equiv \langle \delta_x^2 \rangle = V^{-2} \sum_{k,p} \langle \delta_k \delta_p^* \rangle e^{i(\mathbf{k}-\mathbf{p})\cdot\mathbf{x}} = V^{-2} \sum_k \sigma_k^2 \tag{92}$$

---

[8] A realization refers to the outcome of an experiment. For example, obtaining the sequence *head, head, tails, head*, after four flips of a coin. Given a theoretical model for the initial conditions (a fair coin with a 50% chance of obtaining heads on any given toss), we need a way of testing if the outcome of the experiment (i.e. realization) is consistent with this model. In the case of the Universe, this is complicated by the fact that gravity "erases" some of these initial conditions as time progresses and structures virialize.

[9] Throughout we will use the standard Fourier conventions of $\delta(\mathbf{k}) \equiv \int \delta(\mathbf{x}) e^{-i\mathbf{k}\cdot\mathbf{x}} d\mathbf{x}$ and $\delta(\mathbf{x}) \equiv V^{-1} \int \delta(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}} d\mathbf{k}$, where V the real space volume.

[10] Note that for brevity, we will often write the dependent variables $\mathbf{x}$ and $\mathbf{k}$ as subscripts, e.g. $\delta(\mathbf{k}) \Longrightarrow \delta_k$. We will also drop the vector notation, assuming the principles of isotropy and homogeneity, for which the properties of a field only depend on the magnitude of the vector, $|\mathbf{k}| \equiv k$.

with the final expression taking advantage of the properties of a Gaussian field, eq. (90). We can convert the unwieldy discrete sum over modes into an integral by using the standard density of states relations. Assuming $V = L^3$ to be large enough to contain a representative volume of the Universe (beyond which we can assume homogeneity), the boundary conditions must be satisfied: $\mathbf{k} = \frac{2\pi}{L}\mathbf{n}$, requiring the three spatial components of $\mathbf{n} = (n_x, n_y, n_z)$ to be integers 0, 1, 2... From this boundary condition, we obtain the density of states, $d^3k = \frac{(2\pi)^3}{V} d^3n$, through which we can convert the sum into its continuum limit:

$$\sum_{\mathbf{k}} = \sum_{n} d^3n \Longrightarrow \frac{V}{(2\pi)^3} \int d^3k \ . \tag{93}$$

Using this we can rewrite the variance from eq. (92):

$$\begin{aligned} \Delta_x^2 \equiv \langle \delta_x^2 \rangle &= V^{-2} \sum_k \sigma_k^2 = \frac{V}{(2\pi)^3 V^2} \int \sigma_k^2 d^3k \\ &= V^{-1}(2\pi)^{-3} \int_0^\infty 4\pi k^2 \sigma_k^2 dk \equiv \int_0^\infty \Delta_k^2 \frac{dk}{k} \ , \end{aligned} \tag{94}$$

where the second line performs the volume integral in spherical coordinates assuming isotropy, and the commonly-used *power per ln k* is defined as

$$\Delta_k^2 \equiv \frac{V^{-1}}{2\pi^2} k^3 \sigma_k^2 \tag{95}$$

This definition is the most-commonly used way of showing power spectra.[11]

In the above, we have expressed δ_x as a Gaussian random field. If δ_x is a Gaussian random field, then so are linear functions of δ_x. Consider the excess mass in a sphere with radius R around a given point $\mathbf{x}$:

$$\delta M_R(\mathbf{x}) \equiv \bar{\rho} \int_{|r| \leq R} \delta(\mathbf{x} + \mathbf{r}) d^3r \tag{97}$$

Numerically evaluating this integral is more efficient if we replace the definite integral with an indefinite one, but introducing a "window function", with the following properties:

$$W(\mathbf{r}) = \begin{cases} \sim 1 & \text{if } |r| \leq R \\ \sim 0 & \text{if } |r| > R \end{cases} \tag{98}$$

The window function is also referred to as a "filter response", according to its common usage in signal processing, with the properties of eq. (169) defining a "low pass" filter as it smooths out small scale fluctuations and preserves only long-wavelength (low frequency) ones. Using the window function, eq. (97) can be written as an integral over all space:

$$\delta M_R(\mathbf{x}) \approx \bar{\rho} \int_V \delta(\mathbf{x} + \mathbf{r}) W(\mathbf{r}) d^3r \tag{99}$$

Equation (99) is a convolution in real space, which then becomes a multiplication in k-space:

$$\delta M_R(\mathbf{k}) = \bar{\rho} \delta_k W_k^* \ . \tag{100}$$

[11] The real-space equivalent of power spectra is the two point correlation function:

$$\xi(\mathbf{r}) \equiv \langle \delta(\mathbf{y} + \mathbf{r}) \delta(\mathbf{y}) \rangle_{\mathrm{y}} = V^{-2} \sum_{k,p} e^{i\mathbf{k}\cdot(\mathbf{r}+\mathbf{y}) - i(\mathbf{p}\cdot\mathbf{y})} = V^{-2} \sum_k \sigma_k^2 e^{i\mathbf{k}\cdot\mathbf{r}} \ , \tag{96}$$

where the final equality assumes a Gaussian random field. The correlation function is in certain cases more intuitive, as it can be defined as the excess probability (over random) of finding an object at a distance r from another object. Hence it is commonly used in galaxy surveys.

We can also define the average mass corresponding to this scale R,

$$\bar{M}_R \equiv \bar{\rho} \int W(\mathbf{r}) d^3r \equiv V_W \bar{\rho} \ , \tag{101}$$

where we have also defined an "effective volume", V_W, of the window function. Note that for a perfect low-pass filter (where the "$\sim$" in eq. 169 can be replaced with "$=$"), we have $V_W = 4\pi R^3/3$.

We can now compute the fractional excess, or mass perturbation, on scale R. In k-space we have:

$$\left(\frac{\delta M}{\bar{M}}\right)_R (\mathbf{k}) = \frac{\bar{\rho}\delta_k W_k^*}{V_W \bar{\rho}} = \frac{\delta_k W_k^*}{V_W} \ . \tag{102}$$

Fourier transforming back into real-space using the continuum limit:

$$\left(\frac{\delta M}{\bar{M}}\right)_R (\mathbf{x}) \equiv \delta_M = \frac{V}{(2\pi)^3} \int \left(\frac{\delta_k W_k^*}{V_W}\right) e^{i\mathbf{k}\cdot\mathbf{x}} d^3k \tag{103}$$

Finally, we can write the variance in the mass field on a scale R, which looks like the density contrast variance from eq. (94) but with an additional factor of W_k^2/V_W^2:

$$\boxed{\sigma_M^2(R) \equiv \langle \delta_M^2 \rangle = \frac{1}{V_W^2} \int_0^\infty \Delta_k^2 W_k^2 \frac{dk}{k}} \tag{104}$$

As we shall see below, *the mass variance from eq. (104) is a fundamental quantity for predicting the number density of collapsed structures (halos).* It is an integral over the matter power spectrum, weighted by the window function which defines the scale of interest. Since it is constructed from a linear function of a Gaussian distributed quantity, δ_x, it itself is Gaussian distributed, with the probability of finding a mass fluctuation with amplitude between δ_M and $\delta_M + d\delta_M$, at a given scale R, at redshift z being:

$$P(\delta_M, R, z)\ d\delta_M = \frac{1}{\sqrt{2\pi}\sigma_M} e^{-\frac{\delta_M^2}{2\sigma_M^2}}\ d\delta_M \ . \tag{105}$$

When computing the mass variance, there are three common choices for window functions: the spherical top hat, the sharp k-space, and a Gaussian. Their functional forms in real-space and k-space, as well as their effective volumes are as follows:

Spherical top hat:

$$W_r = \begin{cases} 1 & \text{if } |r| \le R \\ 0 & \text{if } |r| > R \end{cases} \qquad W_k = 4\pi R^3 \left[\frac{\sin(kR)}{(kR)^3} - \frac{\cos(kR)}{(kR)^2}\right] \qquad V_W = \frac{4\pi}{3} R^3 \tag{106}$$

Sharp k-space:

$$W_r = 3\left(\frac{r}{R}\right)^{-3} \left[\sin\left(\frac{r}{R}\right) - \frac{r}{R}\cos\left(\frac{r}{R}\right)\right] \qquad W_k = \begin{cases} 1 & \text{if } |kR| \le 1 \\ 0 & \text{if } |kR| > 1 \end{cases} \qquad V_W = 6\pi^2 R^3 \tag{107}$$

Gaussian:

$$W_r = e^{-\frac{r^2}{2R^2}} \qquad W_k = V_W e^{-\frac{(kR)^2}{2}} \qquad V_W = (2\pi)^{3/2} R^3 \tag{108}$$

Each window function has advantages and disadvantages. The spherical top hat is intuitive and clean in real space, but its Fourier transform has an infinite extent, and oscillates (so-called "ringing") which can result in some spurious signals when implemented on a discrete grid (so-called "aliasing"). The sharp k-space on the other hand is simple in k-space, which facilitates the mass function derivation below, but is prone to aliasing in real space. The Gaussian instead is smooth in both spaces, but is much more diffuse as a result, overly smoothing the density field.

The other component of the mass variance is the matter power spectrum, σ_k^2. This is a fundamental quantity in cosmology, and is usually expressed as (e.g. Eisenstein & Hu 1999):

$$\sigma_k^2 = Ak^n T^2(k) D^2(z,k) \tag{109}$$

Here A is a normalization constant, which by convention is normalized using *the present-day fluctuation on scales of R=8 h^{-1} Mpc, computed with a spherical top hat window function.* Currently, observations suggest a normalization resulting in $\sigma_M(z=0, R=8h^{-1}\text{Mpc}) = 0.82$ (Planck Collaboration XIII et al. 2016). The k^n term is the primordial spectrum set by inflation, which seems well represented by a power law. The T^2 term is the so-called transfer function, which accounts for modifications of the primordial power spectrum on "small scales" due to processes at Matter-Radiation equality and Recombination. Finally the $D(z,k)$ in the last term is the so-called growth factor, which describes the linear growth of perturbations, and is normalized to be unity at $z = 0$. As seen from eq. (57), during the MD regime the growth factor is simply proportional to the scale factor and is scale independent, i.e. $D(z,k) \propto a \propto (1+z)^{-1}$. *This scaling also implies that the linear matter variance (and power spectra) scales as $\sigma_k^2 \propto a^2$ in the MD regime.*

3.3.1. Halo Mass Functions

We now have the tools we need to evaluate a so-called halo mass function. Let's define M_h to be the mass corresponding to a collapsed structure, i.e. a halo, and $f_{\rm coll}(> M_h, z)$ to be the fraction of matter in the Universe contained inside halos of mass greater than M_h at time z (i.e. the collapsed fraction). Taking advantage of eq. (105),[12] we can express the collapse fraction as the integral over the δ_M distribution:

$$f_{\rm coll}(> M_h, z) = \int_{\delta_c}^{\infty} P(\delta_M, R, z)\, d\delta_M = \frac{1}{\sqrt{2\pi}\sigma_M(R,z)} \int_{\delta_c=1.686}^{\infty} e^{-\frac{\delta_M^2}{2\sigma_M^2(R,z)}}\, d\delta_M\ . \tag{110}$$

Here we only need to define the "critical density", δ_c, corresponding to the minimum overdensity of a collapsed halo. Since we are operating on the linear density field (i.e. using the linear power spectrum σ_M as we shall see below), we can just use the result from the spherical collapse model in §3.2, associating $\delta_c = 1.686$ with collapsed structures.

In Fig. 9 we illustrate this procedure with a schematic representing a realization of a Gaussian mass perturbation field. The top left panel shows δ_M at a redshift z_1, obtained by smoothing the matter field with a filter of width R_1. All of the mass above the threshold value $\delta_M \geq \delta_c = 1.686$ belongs to a collapsed structure, i.e. a halo. In this illustration, there are three halos with corresponding masses $M_h \gtrsim \bar{\rho} R_1^3$. These halos make up a fraction of all matter in the Universe, $f_{\rm coll}(> M_1, z_1)$, given by eq. (110).

If we ask how much mass is inside some larger halos, $M_2 > M_1$, at the same redshift, we can refer to the top right panel which is obtained by smoothing the same realization with a larger filter of corresponding width $R_2 > R_1$. We can see that the small-scale structure has been smoothed-out. In this case, there is only 1 halo remaining above the collapse threshold. If we change instead the redshift of interest, asking about the collapse fraction at some $z_2 > z_1$, we can refer to the bottom left panel. Since all modes of the Gaussian matter field grow in amplitude as $\delta \propto (1+z)^{-1}$, this is analogous to scaling the same realization from the top left panel by a factor of $(1+z_1)/(1+z_2)$. In this case, only one halo is found with mass $M_h > M_1$.

In practice, it is far more common to perform all calculations on the overdensity field linearly-extrapolated to $z = 0$. Since both the overdensity field and the critical threshold have the same redshift scaling, this is analogous to moving the redshift dependence from σ_M to $\delta_c(z) = 1.686/D(z)$ (c.f. the

[12] Note that M_h, R, and σ_M^2 are all used interchangeably to indicate a Lagrangian scale.

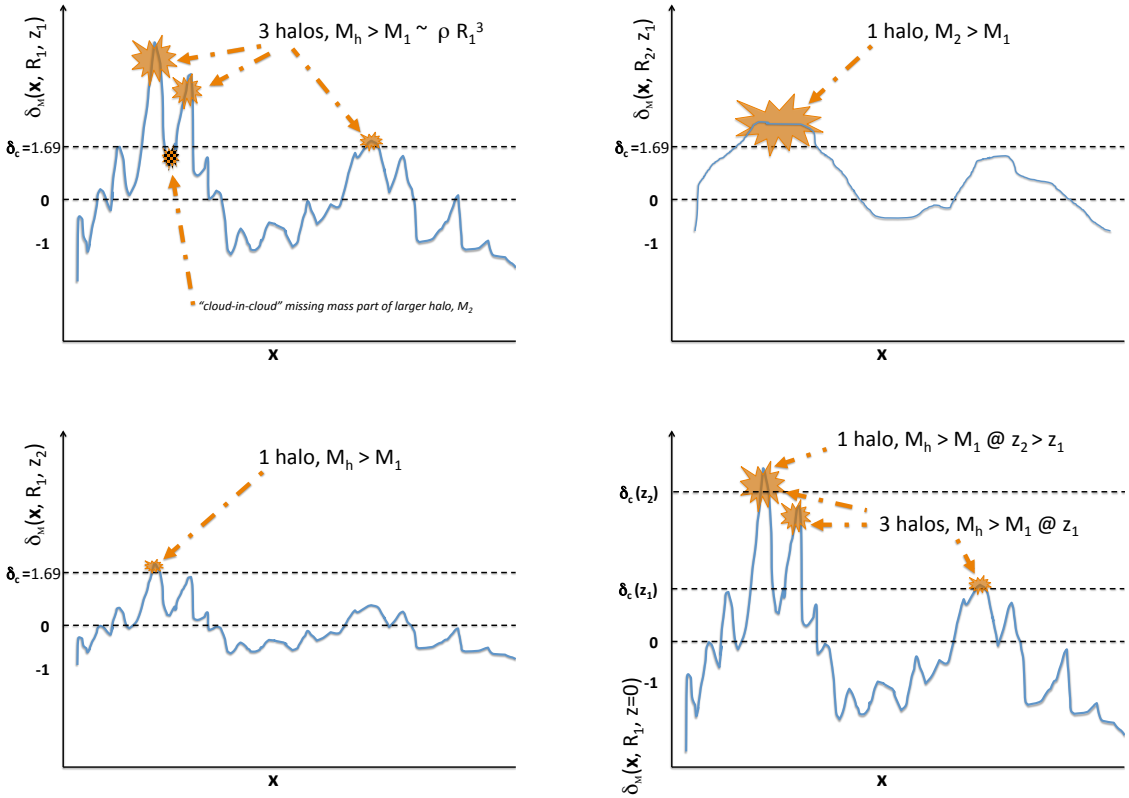

Fig. 9.— Illustration of a mass perturbation field, δ_M, along one spatial dimension, $\mathbf{x}$. The top left panel shows δ_M at a redshift z_1, obtained by smoothing the matter field with a filter of width R_1. The top right panel corresponds to the same realization, but smoothed by a larger filter. In the bottom left panel, we instead scale the realization from the top left to an earlier time, decreasing the amplitude of all modes by a factor of $D(z_2)/D(z_1) \approx (1+z_1)/(1+z_2)$. In the bottom right, the redshift dependence is moved to the critical collapse threshold, $\delta_c \to \delta_c(z) = 1.686/D(z)$, allowing us to operate only on the overdensity field linearly extrapolated to $z = 0$, as per convention.

bottom right panel of Fig. 9), making eq. (110):

$$f_{\rm coll}(> M_h, z) = \frac{1}{\sqrt{2\pi}\sigma_M(R)} \int_{\delta_c(z)=1.686/D(z)}^{\infty} e^{-\frac{\delta_M^2}{2\sigma_M^2(R)}}\, d\delta_M \ . \tag{111}$$

We can simplify this expression by performing a change of variables. Let $\tau \equiv \delta_M/\sqrt{2\sigma_M^2}$ and thus $d\tau = d\delta_M/\sqrt{2\sigma_M^2}$. We can then re-write the above equation as:

$$\begin{aligned} f_{\rm coll}(> M_h, z) &= \frac{1}{\sqrt{2\pi}\sigma_M} \int_{\delta_c}^{\infty} e^{-\frac{\delta_M^2}{2\sigma_M^2}}\, d\delta_M \\ &= \frac{1}{\sqrt{\pi}} \int_{\delta_c/\sqrt{2\sigma_M^2}}^{\infty} e^{-\tau^2} d\tau \\ &= \frac{1}{2}{\rm erfc}\left[\frac{\delta_c(z)}{\sqrt{2}\sigma_M(R)}\right] \ , \end{aligned} \tag{112}$$

where the final equality comes from the definition of the complimentary error function, erfc(x).

There is however a problem with this formalism. Since our matter field is discrete, we should recover the asymptotic limit $f_{\rm coll}(> M) \to 1$ as $M \to 0$. Instead, eq. (112) converges to 1/2 as the mass goes to zero. This missing factor of 2 was dubbed the "cloud-in-cloud" problem. Our mistake above is that we were not accounting for Lagrangian regions which, although too underdense to be counted as belonging

to a halo of mass M_1, still make up a part of a larger halo $M_2 > M_1$ (see the top left panel of Fig. 9). In other words, we should add a term to eq. (110):

$$f_{\rm coll}(> M_h, z) = \int_{\delta_c}^{\infty} P(\delta_M, R, z)\ d\delta_M + \int_{-\infty}^{\delta_c} C(\delta_M, \delta_c, R, z)\ d\delta_M\ . \tag{113}$$

This second term corresponds to the probability that a point with $\delta_M(R_1) < \delta_c$ will have $\delta_M(R_2) > \delta_c$.

The "cloud-in-cloud" problem motivates a different perspective of the calculation of the halo mass function. Instead of considering the mass perturbation field as a function of space, smoothed by a fixed mass scale as in the panels of Fig. (9), let us instead arbitrarily chose a given point in space, $\mathbf{x_1}$, and *vary the smoothing scale around that point.* At the largest scale, $M \to \infty$, the mass perturbation around $\mathbf{x_1}$ must by definition be zero. As the scale is decreased in small increments, the smoothed overdensity around this point can either increase or decrease, deviating from zero. Since our mass perturbation field is Gaussian, this process of decreasing the smoothing scale results in a *random walk* in $\delta_M(M)$: the *relative change* in δ_M with each step in scale, $\Delta\delta_M$, *does not depend on the current value of* δ_M *at that scale.* Moreover, if one uses a sharp k-space window function, $\Delta\delta_M$ is distributed as a zero-mean Gaussian with variance of $\sigma^2_{M+\Delta M} - \sigma^2_M$.[13]

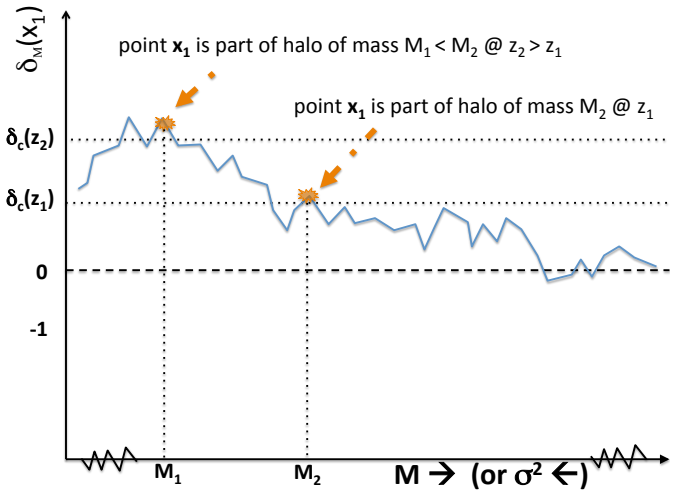

Fig. 10.— Realization of a mass perturbation field as a function of smoothing scale around a randomly chosen point in space, $\mathbf{x_1}$. Halo mass functions are constructed by sampling such realizations (which can be done analytically by solving the diffusion equation for the case that δ_c is scale-independent), and recording the distribution of the largest scales when the random walk trajectory crosses over the collapse threshold (so-called "first crossing distribution").

This random walk procedure is illustrated in Fig. 10, corresponding to a realization of a Gaussian mass perturbation field as a function of smoothing scale around some randomly chosen point in the universe, $\mathbf{x_1}$. Halo mass functions are constructed in the following manner:

1. Starting from the largest scales, the smoothing scale is decreased by small increments[14].
2. The random walk of $\delta_M(M)$ is followed, until the first time (largest scale) the trajectory goes above the $\delta_c(z)$ "barrier".

[13] Although this is only strictly true for a sharp k-space window function, the choice of window function does not have a large impact on the resulting halo mass functions (SHETH REF). Indeed, the spherical top-hat window function probably remains the most popular, helped by the fact that the power spectrum normalization σ_8 is defined with this filter and that it avoids the real-space "ringing" of the sharp k-space filter.

[14] "Small" is defined with respect to the probability of crossing the δ_c barrier in the next step, which should be modest.

3. The point $\mathbf{x_1}$ is flagged as belonging to a halo with a mass corresponding to the scale at which the trajectory first crossed $\delta_c(z)$. In the figure, this corresponds to a halo of mass M_2 for $z = z_1$, or a halo of some smaller mass M_1 at an earlier redshift $z = z_2$.[15]
4. A new realization of the mass perturbation field is created around some arbitrary point $\mathbf{x_2}$.
5. Steps (1) – (4) are repeated a statistically significant number of times.
6. The resulting distribution of first crossing masses corresponds to the halo mass function, $df_{\rm coll}(> M, z)/dM$.

Constructing halo mass functions from this random walk framework allows us to see the solution of the "cloud-in-cloud" problem. Since the δ_c barrier is a constant function of scale, the problem can just be cast in terms of the classical stochastic diffusion equation (Chandrasekhar 1943). For every random walk trajectory which goes above the δ_c barrier at M_1, there is an *equal probability* that it will continue going above or continue going below δ_c for $M < M_1$. In other words, the second term in equation 113 is *equal* to the first term which we already evaluated. Thus,

$$\begin{aligned} f_{\rm coll}(> M_h, z) &= \int_{\delta_c}^{\infty} P(\delta_M, R, z)\ d\delta_M + \int_{-\infty}^{\delta_c} C(\delta_M, \delta_c, R, z)\ d\delta_M \\ &= 2 \times \int_{\delta_c}^{\infty} P(\delta_M, R, z)\ d\delta_M \\ &= 2 \times \frac{1}{2}{\rm erfc}\left[\frac{\delta_c(z)}{\sqrt{2}\sigma_M(R)}\right] \\ &= {\rm erfc}\left[\frac{\delta_c(z)}{\sqrt{2}\sigma_M(R)}\right]\ . \end{aligned} \tag{114}$$

If we differentiate with respect to mass, we obtain the halo mass function (Press & Schechter 1974):

$$\boxed{\frac{df_{\rm coll}(> M, z)}{dM} = \sqrt{\frac{2}{\pi}}\frac{\delta_c(z)}{\sigma_M^2(M)}\left|\frac{d\sigma_M(M)}{dM}\right|\exp\left[-\frac{\delta_c^2(z)}{2\sigma_M^2(M)}\right]}\ . \tag{115}$$

Noting that the fraction of collapsed mass inside halos of mass around M is equal to their number density times their mass divided by the average density of the universe, $df_{\rm coll} = (dN/dV) \times (M/\bar{\rho}) \equiv dn \times (M/\bar{\rho})$, we can also write the halo mass function in terms of the comoving number density of halos:

$$\boxed{\frac{dn(> M)}{dM} = \frac{\bar{\rho}_0}{M}\frac{df_{\rm coll}(> M, z)}{dM}}\ . \tag{116}$$

We can go even further! In constructing the halo mass function, we are throwing away all information about the random walk trajectory except for the first crossing scale of a barrier at a given redshift. However, as seen in Fig. 10, the random walk (also known as the excursion-set) framework allows us to follow the whole growth/merger history of the structures enclosing our point $\mathbf{x_1}$. The growth history of this halo is obtained by looking at each step in the trajectory, backwards in time from M_2 at z_2. For example, we know the point can be identified as belonging to a halo of mass M_1 at a redshift $z = z_2 > z_1$.

This can be quantified with the so-called *conditional mass function* (e.g. Bond et al. 1991; Lacey & Cole 1993; Somerville & Kolatt 1999). If we are only interested in the relation of two points in the trajectory, we can solve for the conditional probability analytically. Say for example, we wish to know $f_{\rm coll}(> M, z \mid M_{\rm bias}, \delta_{\rm bias})$, i.e. the mass fraction of halos with mass $> M$ at z, *given* that they reside in regions which have a linear over density $\delta_{\rm bias}$ on scale $M_{\rm bias}$. Because the steps in the random walk are uncorrelated, this conditional collapse fraction or mass function can be evaluated directly from the above equations, simply by *translating the origin of the random walk* from $M =$ infinity, $\delta = 0$, to $M = M_{\rm bias}$,

[15] We again see that our framework corresponds to structure formation which is hierarchical: small structures form earlier than larger ones, as the spherical collapse barrier is a monotonic function of redshift.

$\delta = \delta_{\rm bias}$. With Gaussian fields, we add the variance in quadrature of the two points. For example, the conditional collapsed fraction is just:

$$f_{\rm coll}(> M, z \mid M_{\rm bias}, \delta_{\rm bias}) = {\rm erfc}\left[\frac{\delta_c(z) - \delta_{\rm bias}}{\sqrt{2(\sigma_M^2 - \sigma_{M_{\rm bias}}^2)}}\right] , \tag{117}$$

and the same substitution of $\delta_c \to \delta_c - \delta_{\rm bias}$, $\sigma_M^2 \to \sigma_M^2 - \sigma_{M_{\rm bias}}^2$ can be made to evaluate the conditional halo mass function from eq. (115).

It is worthwhile to take a step back and admire the power of this simple framework. Starting from some basic physics, we can evaluate the linear evolution of Gaussian overdensities, and the collapse threshold in the spherical collapse model. From these two simple relations, $\delta \propto (1+z)^{-1}$ and $\delta_c = 1.686$, we can construct detailed distributions of collapsed structures as a function of time, including their merger histories.

Surprisingly these analytic halo mass functions agree relatively well with results from N-body simulations, which follow the non-linear evolution of structure in detail. However, it was noted that they somewhat underestimate the abundance of massive halos, and overestimate the abundance of small halos. An alternative form was suggested by Sheth & Tormen (1999):

$$\frac{dn(> M, z)}{dM} = -\frac{\bar{\rho_0}}{M}\frac{\partial \ln \sigma_M}{\partial M}\sqrt{\frac{2}{\pi}} A \left(1 + \frac{1}{\hat{\nu}^{2p}}\right) \hat{\nu} \exp[-\frac{\hat{\nu}^2}{2}] , \tag{118}$$

with $\hat{\nu} \equiv \sqrt{a}\delta_c(z)/\sigma_M$. This analytic form was shown in Sheth et al. (2001) to result from a random walk procedure, but replacing the scale-independent barrier from spherical collapse, $\delta_c(z) = 1.686/D(z)$, with a scale-dependent barrier whose functional form is motivated by an ellipsoidal collapse model:

$$\delta_c(M, z) = \sqrt{a}\delta_c(z)\left[1 + b\left(\frac{\sigma_M^2(M)}{a\delta_c^2(z)}\right)^c\right] . \tag{119}$$

By fitting to the results of N-body simulations, we can evaluate the constants in this so-called Sheth-Torman mass function, obtaining (e.g. Jenkins et al. 2001): $A = 0.353$, $p = 0.175$, $a = 0.707$, $b = 0.34$, $c = 0.81$. These mass functions are still widely used today, forming the basis for all analytic studies of early structure formation.

3.4. Halo clustering and the non-linear density field

We have built an analytical framework to compute the abundance of halos of a given mass at a given redshift. We now turn to how these halos are spatially distributed, and their connection to the underlying non-linear matter field. Recall that the linear matter power spectrum is a direct and powerful probe of physical cosmology (c.f. eq. 109). However, we cannot directly measure σ_k from the evolved Universe for two reasons: (i) gravity acts to "couple modes", erasing the linear power on small scales; and (ii) we cannot observe the matter directly but instead must rely on baryonic tracers (e.g. gas, stars galaxies).

In this section we will introduce an analytic treatment to account for (i) and (ii) parametrically. This will allow us to infer the linear matter power spectrum from late-time tracers such as galaxies. We know galaxies must sit inside DM halos. If we can relate the galaxy field $\rightarrow$ halo field $\rightarrow$ matter field, then we could learn about both galaxy properties and cosmology by comparing models with observations.

3.4.1. Linear Lagrangian halo bias

So how do we relate the halo field to the underlying matter density field? We can get an approximation on large-scales using the same toolkit from the previous chapter. We can approximate the overdensity field, δ, as a sum of small-scale fluctuations, δ_{pk} sitting on top of a large-scale (linear) background, δ_{bg}:

$$\delta \approx \delta_{bg} + \delta_{pk} \tag{120}$$

This is commonly referred to as the "peak-background split" (Kaiser 1984; Cole & Kaiser 1989; Mo et al. 1997; Sheth et al. 2001), and shown in Fig. 11.

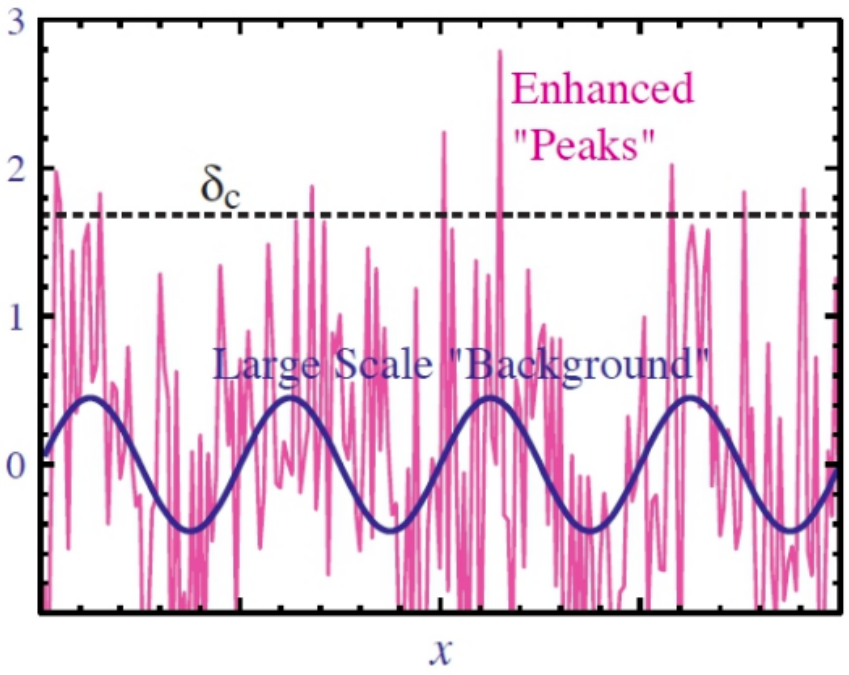

Fig. 11.— peak background schematic

Halos form preferentially where the large-scale background "helps" push the density field above the critical value. As discussed above for the conditional halo mass functions, we can rescale (i.e. move the

origin of the random walk):

$$\delta_{cpk} = \delta_c - \delta_{bg} \tag{121}$$
$$\sigma^2_{pk} = \sigma^2 - \sigma^2_{bg} \tag{122}$$
$$\nu \equiv \delta_{cpk}/\sigma_{pk} \tag{123}$$

We see from the second line that covariances between the large-scale background and small-scale fluctuations are ignored.

Since our halo mass functions scale as $n_M \equiv \mathrm{d}n_h/\mathrm{d}\ln M = C\nu\exp[-\nu^2/2]$, where C is a constant, and by definition $|\delta_{\rm bg}| \ll 1$, we can Taylor expand around the mean background, $\delta_{\rm bg} = 0$, to first order:

$$n_M(\delta_{\rm bg}) = n_M|_{\delta_{\rm bg}=0} + \left.\frac{dn_M}{d\nu}\frac{d\nu}{d\delta_{bg}}\right|_{\delta_{\rm bg}=0}\Delta\delta_{bg} \tag{124}$$
$$= \bar{n}_M + \left[C\exp[-\frac{\nu^2}{2}] - C\nu^2\exp[-\frac{\nu^2}{2}]\right]\left.\left(\frac{d(\delta_{cpk}/\sigma_{pk})}{d\delta_{bg}}\right)\right|_{\delta_{bg}=0}\Delta\delta_{bg} \tag{125}$$
$$= \bar{n}_M + \bar{n}_M(\nu^{-1} - \nu)(-\sigma^{-1}_{pk})\Delta\delta_{bg} \tag{126}$$
$$= \bar{n}_M\left[1 + \frac{\nu^2 - 1}{\nu\sigma_{pk}}\Delta\delta_{bg}\right] \tag{127}$$
$$= \bar{n}_M\left[1 + \frac{\nu^2 - 1}{\delta_{cpk}}\Delta\delta_{bg}\right] \tag{128}$$

Defining the Lagrangian halo overdensity, $\delta^L_h \equiv n_M/\bar{n}_M - 1$, and noting that by construction $\delta_{cpk} \approx \delta_c$ we have:

$$\delta^L_h = \frac{\nu^2 - 1}{\delta_{cpk}}\Delta\delta_{bg} \approx \frac{\nu^2 - 1}{\delta_c}\Delta\delta_{bg} \tag{129}$$
$$\delta^L_h \equiv b(M_h)\delta^{\rm lin}_\Delta \tag{130}$$

Here $b^L(M_h)$ is called the linear (Lagrangian) halo bias, since $\delta^{\rm lin}_\Delta \equiv \Delta\delta_{bg}$ is the large-scale (linear) matter overdensity.

This framework allows us then to relate the linear halo correlation function in Lagrangian space between halos of characteristic mass M_1 and M_2, and its Fourier dual, the halo power spectra, to the linear matter field (which can be computed analytically as shown in the previous section):

$$\xi^L_{hh}(r|M_1, M_2) = b^L(M_1)b^L(M_2)\xi^{\rm lin}_{\Delta\Delta}(r) \tag{131}$$
$$P^L_{hh}(k|M_1, M_2) = b^L(M_1)b^L(M_2)P^{\rm lin}_{\Delta\Delta}(k) \tag{132}$$
$$\tag{133}$$

As seen from the above equation, in the Press-Schechter formalism we have

$$b^L(M_h) = \frac{\nu^2 - 1}{\delta_c}\ . \tag{134}$$

Or plugging in the ellipsoidal barrier from Sheth-Tormen, we have

$$b^L(M_h) = \frac{a\nu^2 - 1}{\delta_c} + \frac{2p}{\delta_c[1 + (a\nu^2)^p]}\ , \tag{135}$$

Where the constants a and p are the same as from the last section obtained by calibrating to simulations (e.g. $a \approx 0.7$, $p \approx 0.2$ Jenkins et al. 2001).

The framework discussed here applies to the clustering of halos in Lagrangian space. One can account for non-linear evolution to Eularian space using the spherical collapse model, which results in $b = b^L + 1$ (e.g. Section 2.3 in Mo & White 1996). More accurately, the halo bias can be fit empirically to N-body simulations. The resulting *non-linear* halo biases, correlation flunctions and power spectra can be compared to galaxy observations using the Halo Model, discussed below.

3.5. Halo model

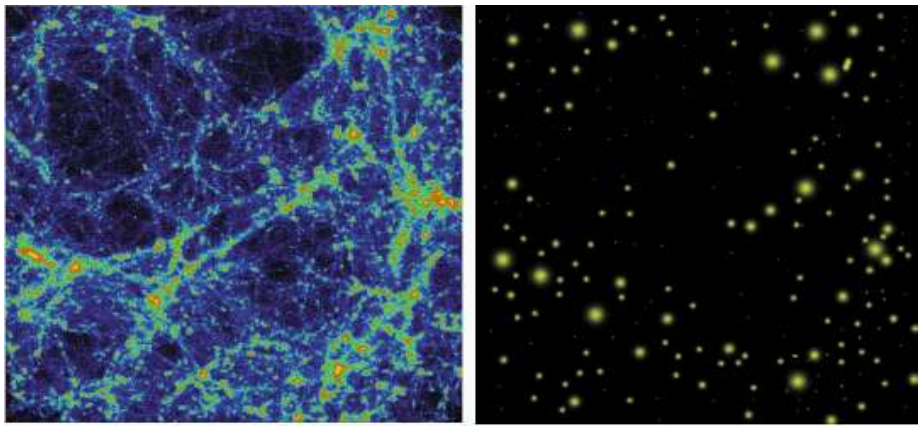

Fig. 12.— The non-linear matter field (*left*) can be reconstructed from the superposition of spherically symmetric halo profiles (*right*), up to second order. From Cooray & Sheth (2002).

In the above we effectively assumed halos are point objects, i.e. delta functions. While this is a decent approximation for scales much larger than the halo virial radius, going towards smaller scales requires accounting for the halo density profile. The resulting framework is called "the halo model" (e.g. Cooray & Sheth 2002), and is very widely used to construct two-point statistics of various matter field tracers such as galaxy and line intensity maps.

Let us assume that the matter field can be discretized into halos that have a spherically-symmetric density profile according to:

$$\rho_h(r|M_h) = M_h u(r|M_h) \ , \tag{136}$$

where $u(r|M_h)$ is the normalized profile (in units of inverse volume), $\int_0^\infty 4\pi r^2 u(r|M_h)dr = 1$. Note that $u(r|M_h)$ is effectively a window function that has a characteristic scale of $r = R_{\rm vir}$.

Now, let us partition the Universe (at a given redshift) into vanishingly small volumes ΔV, such that each ΔV_i contains at most 1 halo *center* of mass M_i. Thus the halo number occupancy of each cell is $N_i \in 0, 1$, and also $N_i = N_i^2 = N_i^3$.... We can write the matter density at any point in space, $\mathbf{x}$:

$$\rho(\mathbf{x}) = \sum_i N_i M_i u(\mathbf{x} - \mathbf{x_i}|M_i) \tag{137}$$

And the general (non-linear) matter correlation function:

$$\xi^{\rm nl}_{\Delta\Delta}(r) \equiv \langle \delta^{\rm nl}(\mathbf{x})\delta^{\rm nl}(\mathbf{x}+r)\rangle_V = \frac{1}{\bar\rho^2}\langle [\rho(\mathbf{x}) - \bar\rho]\,[\rho(\mathbf{x}+r) - \bar\rho]\rangle_V \tag{138}$$

$$= \frac{1}{\bar\rho^2}\left[\langle\rho(\mathbf{x})\rho(\mathbf{x}+r)\rangle_V - \cancelto{\bar\rho^2}{\langle\rho(\mathbf{x})\bar\rho\rangle_V} - \cancelto{\bar\rho^2}{\langle\rho(\mathbf{x}+r)\bar\rho\rangle_V} + \bar\rho^2\right] \tag{139}$$

$$= \frac{1}{\bar\rho^2}\langle\rho(\mathbf{x})\rho(\mathbf{x}+r)\rangle_V - 1 \tag{140}$$

We can re-write the volume averaged term plugging in the density from eq. (137), and defining $\mathbf{x_1} \equiv \mathbf{x}$ and $\mathbf{x_2} \equiv \mathbf{x_1} + r$:

$$\langle\rho(\mathbf{x})\rho(\mathbf{x}+r)\rangle_V = \langle\sum_i N_i M_i u(\mathbf{x_1} - \mathbf{x_i}|M_i)\sum_j N_j M_j u(\mathbf{x_2} - \mathbf{x_j}|M_j)\rangle_V \tag{141}$$

$$= \sum_i\sum_j \langle N_i N_j M_i M_j u(\mathbf{x_1} - \mathbf{x_i}|M_i)u(\mathbf{x_2} - \mathbf{x_j}|M_j)\rangle_V \tag{142}$$

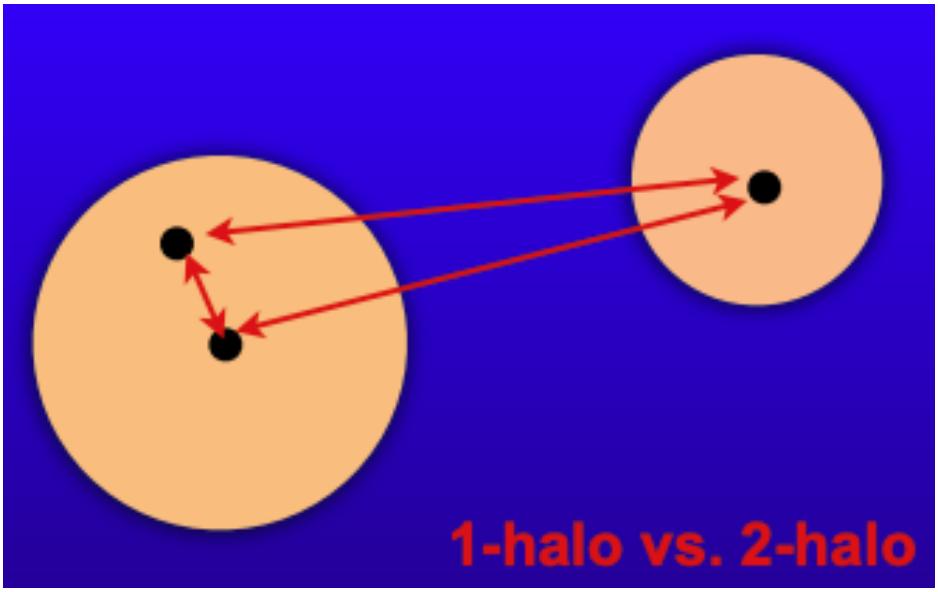

Fig. 13.— FIGURE: 1 halo and 2 halo schematic... split is roughly at scales of Rvir

This expression can be split into 1-halo and 2-halo terms (c.f. Figure 13):

$$\langle\rho(\mathbf{x})\rho(\mathbf{x}+r)\rangle_V = \langle\rho(\mathbf{x})\rho(\mathbf{x}+r)\rangle_{1h} + \langle\rho(\mathbf{x})\rho(\mathbf{x}+r)\rangle_{2h} \tag{143}$$

The 1-halo term corresponds to $i = j$, where we are correlating two different points inside the same halo:

$$\langle\rho(\mathbf{x})\rho(\mathbf{x}+r)\rangle_{1h} = \sum_i \langle N_i N_i M_i M_i u(\mathbf{x_1} - \mathbf{x_i}|M_i) u(\mathbf{x_2} - \mathbf{x_i}|M_i)\rangle_V \tag{144}$$

$$= \sum_i \langle n(M_i)\Delta V M_i^2 u(\mathbf{x_1} - \mathbf{x_i}|M_i) u(\mathbf{x_2} - \mathbf{x_i}|M_i)\rangle_V \ . \tag{145}$$

Here the last line makes the substitution $N_i^2 = N_i = n(M_i)\Delta V$. This sum is only non zero when it picks up the center of a single halo of mass M_i, correlating points with a separation $\mathbf{r} \equiv \mathbf{x_2} - \mathbf{x_1}$ inside this halo.

Recall that averaging over volume is analogous to a statistically-representative ensemble average of a population (aka ergodicity). Because we assumed the density field can be represented by a sum of spherically-symmetric halos, this ergodicity argument allows us to replace the volume average above with an integral over the halo mass function, $\langle\rangle_V \to \int d\ln M$:

$$\langle\rho(\mathbf{x})\rho(\mathbf{x}+r)\rangle_{1h} = \sum_i \Delta V \int d\ln M n(M) M^2 u(\mathbf{x_1} - \mathbf{x_i}|M) u(\mathbf{x_2} - \mathbf{x_i}|M) \tag{146}$$

$$= \int d\ln M n(M) M^2 \int d^3\mathbf{y}\ u(\mathbf{x_1} - \mathbf{y}|M) u(\mathbf{x_2} - \mathbf{y}|M) \tag{147}$$

with the last line replacing the discrete sum over volume with a volume integral.

The 2-halo term corresponds to $i \neq j$, where we are correlating points inside two different halos:

$$\langle\rho(\mathbf{x})\rho(\mathbf{x}+r)\rangle_{2h} = \sum_i \sum_{j\neq i} \langle N_i N_j M_i M_j u(\mathbf{x_1} - \mathbf{x_i}|M_i) u(\mathbf{x_2} - \mathbf{x_j}|M_j)\rangle_V \tag{148}$$

$$= \sum_i \sum_{j\neq i} \int d\ln M_1 M_1 n(M_1)$$

$$\int d\ln M_2 M_2 n(M_2) \Delta V_i \Delta V_j u(\mathbf{x_1} - \mathbf{x_i}|M_1) u(\mathbf{x_2} - \mathbf{x_j}|M_2) \left[1 + \xi_{hh}(\mathbf{x_i} - \mathbf{x_j}|M_1, M_2)\right] \ . \tag{149}$$

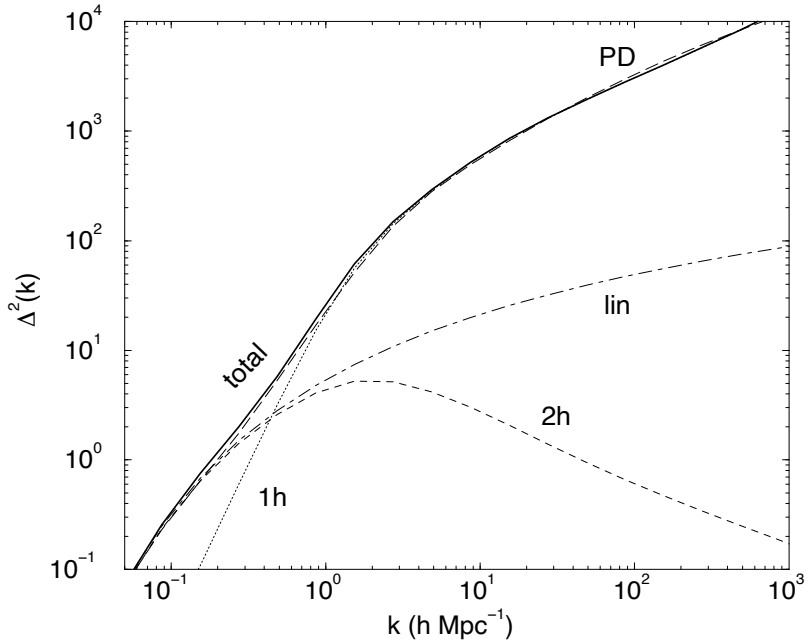

Fig. 14.— The non-linear matter power spectrum at $z = 0$, constructed as a combination of 1 and 2 halo terms. From Cooray & Sheth (2002).

Here when substituting the volume average with (halo) population averages, we have to account for the biased locations of halos, with their non-linear correlation function, $\xi_{hh}(r|M_1, M_2)$. Thus we pick up a factor of $[1+\xi_{hh}]$ in the second line, where the unity term corresponds to a random (uncorrelated) chance of finding halos at a separation of r, while the second is the enhancement of probability due to halo clustering (indeed this is the definition of a correlation function).

We can explicitly multiply through to get these two terms (random and in-excess of random):

$$\langle\rho(\mathbf{x})\rho(\mathbf{x}+r)\rangle_{2h} = \sum_i \sum_{j\neq i} \int d\ln M_1 M_1 n(M_1) \int d\ln M_2 M_2 n(M_2) \Delta V_i \Delta V_j u(\mathbf{x_1}-\mathbf{x_i}|M_1) u(\mathbf{x_2}-\mathbf{x_j}|M_2) \quad \bar{\rho}^2$$
$$+ \sum_i \sum_{j\neq i} \int d\ln M_1 M_1 n(M_1) \int d\ln M_2 M_2 n(M_2) \Delta V_i \Delta V_j u(\mathbf{x_1}-\mathbf{x_i}|M_1) u(\mathbf{x_2}-\mathbf{x_j}|M_2) \xi_{hh}(\mathbf{x_i}-\mathbf{x_j}|M_1, M_2) \tag{150}$$

$$= \bar{\rho}^2 + \int d\ln M_1 M_1 n(M_1) \int d\ln M_2 M_2 n(M_2)$$
$$\int d^3y_1 \int d^3y_2 \; u(\mathbf{x_1}-\mathbf{y_1}|M_1) u(\mathbf{x_2}-\mathbf{y_2}|M_2) b(M_1) b(M_2) \xi^{\rm lin}_{\Delta\Delta}(\mathbf{y_1}-\mathbf{y_2}) \; , \tag{151}$$

where in the final line we relate the halo correlation function to the matter correlation function with a non-linear bias term.

We now have all of the componenets to build our non-linear matter field from the linear field (allowing us to analytically constrain cosmology from observed, late-time 2-point statistics). Summarizing, in

configuration space we have the following relations:

$$\xi^{\rm nl}_{\Delta\Delta}(r) = \xi_{1h}(r) + \xi_{2h}(r) \tag{152}$$

$$\xi_{1h}(r) = \bar{\rho}^{-2} \int d\ln M M^2 n(M) \int d^3\mathbf{y}\ u(\mathbf{x}-\mathbf{y}|M)\ u(\mathbf{x}+\mathbf{r}-\mathbf{y}|M) \tag{153}$$

$$\xi_{2h}(r) = \bar{\rho}^{-2} \int d\ln M_1 M_1 b(M_1) n(M_1) \int d\ln M_2 M_2 b(M_2) n(M_2)$$
$$\times \int d^3\mathbf{y_1} \int d^3\mathbf{y_2}\ u(\mathbf{x}-\mathbf{y_1}|M_1)\ u(\mathbf{x}+\mathbf{r}-\mathbf{y_2}|M_2)\ \xi^{\rm lin}_{\Delta\Delta}(\mathbf{y_1}-\mathbf{y_2}) \tag{154}$$

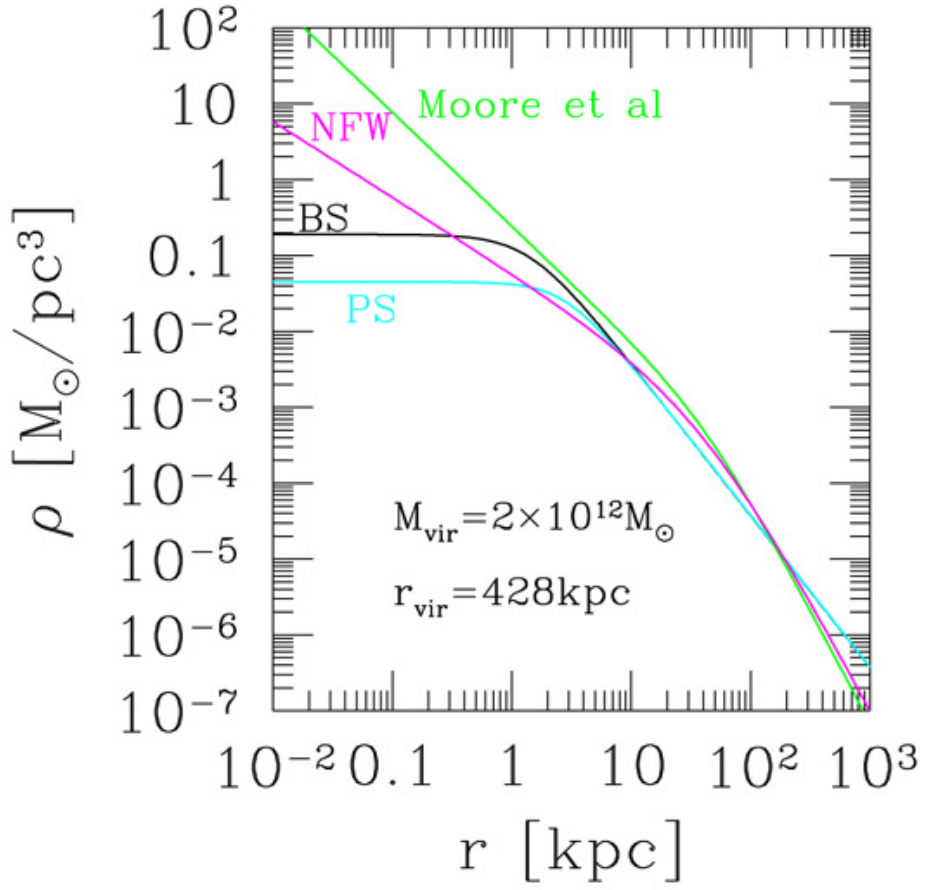

Fig. 15.— Examples of dark matter profiles, including the popular NWF profile from Navarro et al. (1996), an even steeper profile from Moore et al. (1998), and flatter profiles from Persic et al. (1996) and Bahcall & Soneira (1980). Figure from Brook & Di Cintio (2015).

The halo model formalism contains the following ingredients:

- The halo mass function, $n(M)$. This can be taken from analytic forms, such as Press-Schechter, Sheth-Torman, etc. or from N-body simulations.
- The bias, $b(M)$. In the simplest cases with a step function selection, this can be calculated analytically as shown above, or taken from simulations. However, invariably the observation has a more complicated selection function and we need to understand the astrophysics of galaxies to correctly compute the effective bais for each observation. This bias term is generally the largest source of uncertainty, encoding galaxy or IGM physics.
- The halo profile, $u(r|M)$. By definition, the halo profile is non-linear and influenced by both baryons and dark matter. This is taken from simulations. Navarro et al. (1996) suggested the following,

self-similar fit to their simulations: $u(r) \propto \left(\frac{r}{r_s}\right)^{-1}\left(1+\frac{r}{r_s}\right)$. They find that the profiles can be characterized with a single parameter, the so-called concentration parameter $c \equiv R_{\rm vir}/r_s$, which increases with redshift as halos get more concentrated. The outer profile is isothermal, $\rho \propto r^{-2}$, as we would expect from spherical collapse, while the inner one is flatter, $\rho \propto r^{-1}$, due to the influence of baryons (see Figure 15). The inner profile is highly influenced by both stellar feedback and dark matter properties. Motivated by observations of "cores" (instead of the expected "cusps" from pure CDM models), simulations have experimented with strong feedback as well as DM models with small-scale suppression of power (e.g. thermal relic warm dark matter; see Fig. 15).

The halo model is more often applied in Fourier space, where convolutions become multiplications and the formulas simplify to:

$$P^{\rm nl}_{\Delta\Delta}(k) = P_{1h}(k) + P_{2h}(r) \tag{155}$$

$$P_{1h}(r) = \bar{\rho}^{-2}\int d\ln M M^2 n(M)|u(k|M)|^2 \tag{156}$$

$$P_{2h}(r) = P^{\rm lin}_{\Delta\Delta}(k)\bar{\rho}^{-2}\left[\int d\ln M M b(M) n(M) u(k|M)\right]^2 \tag{157}$$

where $P^{\rm lin}_{\Delta\Delta}(k)$ is the linear matter power spectrum from eq. (95), and

$$u(k|M) = \int u(\mathbf{x}|M) e^{-\mathbf{k}\ \mathbf{x}} d^3\mathbf{x} = 4\pi\int_0^\infty u(r|M)\frac{\sin kr}{kr} r^2 dr \tag{158}$$

is the Fourier dual of the halo profile.

The halo model framework is fast and flexible and is widely used for two point statistics of galaxy and line intensity mapping. It can even be extended to non-matter fields, by changing the profile. For example, Mesinger & Furlanetto (2009) used a $1/r^2$ flux profile to model the ionizing background power spectra, and Schneider et al. (2021) used it to model the cosmic 21-cm power spectrum. There is a public halo model implementation in python at `https://github.com/halomod/halomod` (Murray et al. 2021).

3.6. Lagrangian Perturbation theory; the Zel'dovich approximation

Zel'Dovich (1970) proposed an alternative model for the evolution of collisionless matter, more appropriate for the frame of reference of a particle/fluid element. Instead of perturbing the density field (so-called Eulerian perturbation theory we explored in §3.1) by introducing δ (which is bound to fail as $\delta \sim 1$), he suggested to *perturb the displacement* of a particle/fluid element (so-called Lagrangian perturbation theory):

$$\mathbf{x}(\mathbf{q}, t) = \mathbf{q} + \mathbf{\Psi}(\mathbf{q}, t) \tag{159}$$

Here, $\mathbf{q}$ is the initial (so-called Lagrangian) position of a particle, $\mathbf{x}$ is the evolved (so-called Eulerian) position at time t, and $\mathbf{\Psi}(\mathbf{q}, t)$ is the displacement vector.

Recall from eq. (47) that to first order we can write:

$$\dot{\delta} \approx -a^{-1}\nabla\cdot\mathbf{v}_{\rm pec} \equiv -\nabla\cdot\dot{\mathbf{x}} \ . \tag{160}$$

Plugging-in eq. (159):

$$\dot{\delta} \approx -\nabla\cdot(\cancel{\dot{\mathbf{q}}} + \dot{\mathbf{\Psi}}) \ . \tag{161}$$

Integrating out the time, we can relate to first order:

$$\delta(\mathbf{q}, t) \approx -\nabla\cdot\mathbf{\Psi}(\mathbf{q}, t) \ . \tag{162}$$

We now take the ansaltz that $\mathbf{\Psi}(\mathbf{q}, t) \equiv g(t)\psi(\mathbf{q})$ is a separable function of time and space (corresponding to straight line, "ballistic" trajectories of particles). In that case, we can write eq. (162) as:

$$\delta(\mathbf{q}, t) \approx -g(t)\nabla\cdot\psi(\mathbf{q}) \ . \tag{163}$$

The spatial component is purely a function of the density field, and can be easily evaluated in k-space: $\psi(\mathbf{q}) = V^{-1}\sum_k \frac{i\mathbf{k}}{k^2}\delta_k e^{i\mathbf{k}\cdot\mathbf{q}}$. Moreover, since we know from linear theory that the overdensity scales with the growth factor, $\delta \propto D(t)$, we can immediately identify the temporal component of the displacement vector as $g(t) = D(t)$. Thus we have

$$\mathbf{\Psi}(\mathbf{q}, t) = D(t)\psi(\mathbf{q}) \tag{164}$$

$$\mathbf{v} = \dot{\mathbf{x}} = \dot{D}(t)\psi(\mathbf{q}) \tag{165}$$

This formalism therefore allows us to "move" particles from their initial positions along a straight line, with the direction purely specified by the initial density field, and the distance along the line traveled purely specified by the scale factor. This is a very quick and easy way to evolve the density field, and results in much better agreement with numerical gravity solvers than just the Eulerian perturbation theory. In modern cosmology, this so-called Zeldovich approximation (and its extension to higher-order) has two uses: (i) setting the initial conditions for N-body simulations (discussed in the next chapter); and (ii) completely replacing expensive gravity solvers in generating large-scale (quasi-linear) density fields (see Fig. 16).

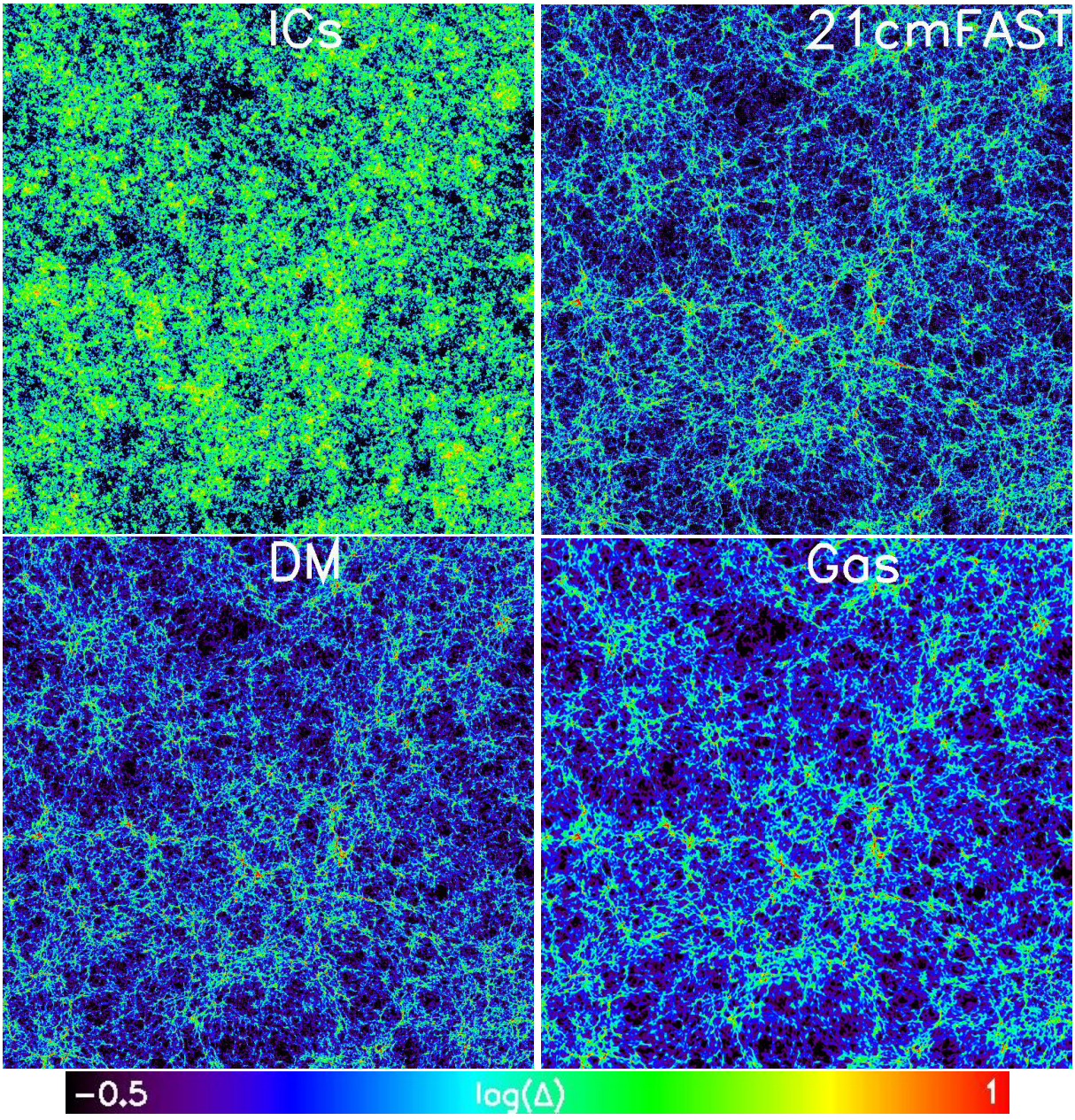

Fig. 16.— A 0.19 Mpc thick slice through a cosmological simulation on a periodic box with sides of length 143 Mpc (adapted from Trac, Cen, Loeb 2009). Panels show the log of the overdensity, $\Delta \equiv \rho/\bar{\rho}$, at $z = 7$. In the upper left (right), the initial conditions are evolved using first order Eularian (Lagrangian) perturbation theory. In the bottom left (right) are results from an N-body (hydrodynamic) simulation (taken from Mesinger et al. 2011).

3.7. Simulating the dark matter distribution with cosmological N-body codes

Cosmological N-body simulations discretize a volume of the Universe, evolving the particles within according to the laws of gravity. This physical set-up has several common features (c.f. the review of Bagla & Padmanabhan 1997):

- The simulated volume is not in isolation, and matter cannot be "created" or "lost" at the edges. This forces us to use periodic boundary conditions, which is most naturally achieved using a cubical or cuboid volume.
- The volume must be large enough so that the scale of the box remains linear. This avoids spurious tidal forces from neighboring copies.
- The particle mass must be small enough to resolve the structures of interest (for dark matter halos you need ∼1000 particles; e.g. cite)
- Each particle is actually a group of many matter particles. This means that dealing with collisions is complicated and the dynamics should generally be collisionless.

The work-flow of cosmological N-body codes is the following:

1. Generate initial conditions (usually specialized codes for this)
2. Compute the force
3. Move the particles
4. IF output is desired at this step (snapshot) print locations or just halo info by identifying the clustering of particles in real (or phase) space
5. Go to next time step and repeat steps (ii)–(iv)

We discuss each in turn. For nice reviews see Bagla & Padmanabhan (1997), Vogelsberger et al. (2020) and Springel (2016).

3.7.1. Initial conditions

How do you implement the initial Gaussian field? An obvious solution could be to sample your cosmological matter power spectrum at some early time on a grid and assign each grid particle the mass corresponding to $\bar{M}[1 + \delta(\mathbf{x}|z)]$, where $\bar{M}$ is the mean mass inside a cell. However, this approach means that all particles would have different masses, which translates into significant memory overheads and slows down algorithms for moving the particles.

Instead N-body codes use the same mass for all particles, but perturb their initial locations so as to recover the cosmological matter power spectrum (and associated force field). This is done using Lagrangian perturbation theory, discussed in the previous chapter §3.6. In recent years, the desire for higher accuracy has pushed many simulations to start at earlier redshifts ($z \sim$ few$\times$100 instead of the traditional choice of $z \approx 100$) when matter perturbations were more linear. Similarly now second order Lagrangian perturbation theory (2LPT; Scoccimarro & Sheth 2002) and even third order LPT are regularly used instead of the first order Zel'Dovich approximation.

The steps involved are the following:

1. randomly sample your matter power spectrum to obtain a density field realization

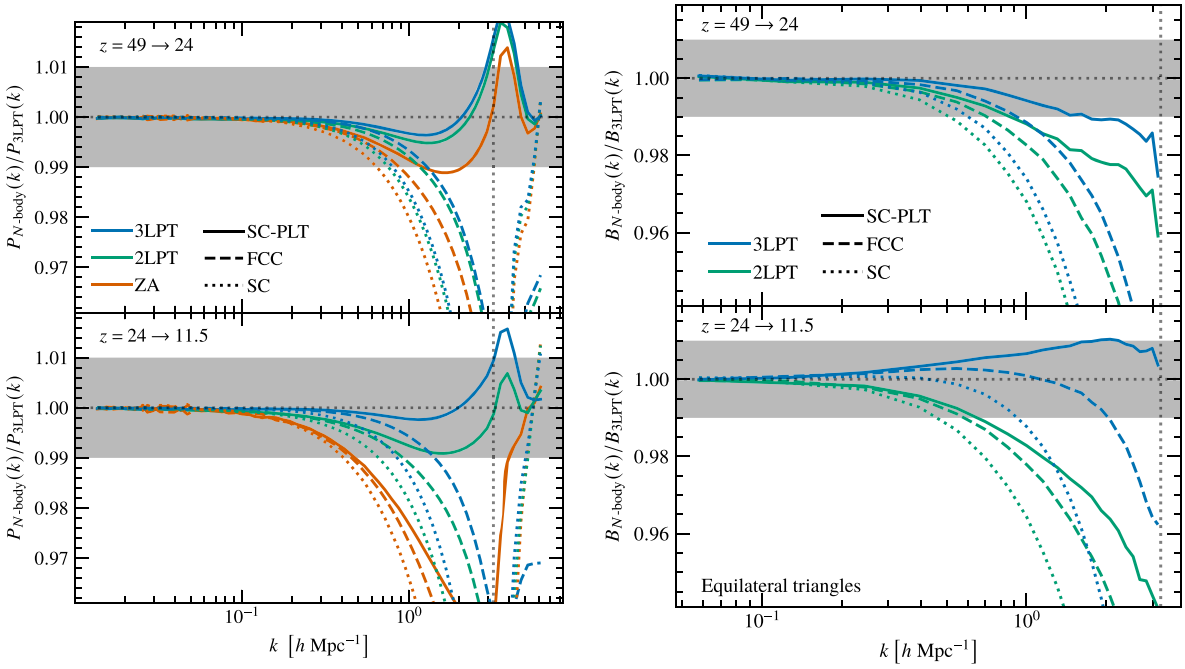

Fig. 17.— Impact of different particle initialization prescriptions on the resulting N-body field. The plots show the ratio between power spectra (*left panel*) and equilateral bispectra (*right panel*) computed using N-body simulations and 3LPT (taken to be the ground truth). Top (bottom) panels correspond to $z = 24$ (11.5). The N-body simulations were evolved from zstart = 49 (24) down to z = 24 (11.5) using initial conditions based on ZA, 2LPT and 3LPT (blue, green, and orange, respectively) and initialized from different perturbed lattices. Solid lines show the results of the simple cubic lattice initial conditions including corrections for particle discreteness and the dotted lines without. The vertical dotted line indicates the particle Nyquist wavenumber of the initial conditions. One per cent agreement is represented as a shaded area and a perfect agreement as a dotted horizontal line. These figures are taken from Michaux et al. 2020.

2. create a "uniform" grid of particles
3. displace particles from initial locations using PT computed on the density field from (1)
4. compute potential from the displaced particles in step (3)
5. assign initial velocities corresponding to this potential from (4)

The impact of some common choices of this procedure are shown in Fig. 17, taken from Michaux et al. (2020).

3.7.2. Force calculation

The main pieces of the N-body code are therefore computation of force and moving the particles. Computing the force is in principle an $O(N^2)$ calculation (every particle impacts every other particle), while moving the particles according to the infered velocity field is $O(N)$. Thus N-body codes spend most of the time in the force calculation. Direct force calculation is impractical for modern cosmological simulations, which can consist of upwards of billions of particles. Instead codes make shortcuts, and they can be divided into two main classes (although some codes apply a mix of these):

Particle mesh - the force is computed on a grid and then interpolated to particle locations

Tree - each particle has tree heirarchy that groups distant particles together, having them act as a single particle at their center of mass.

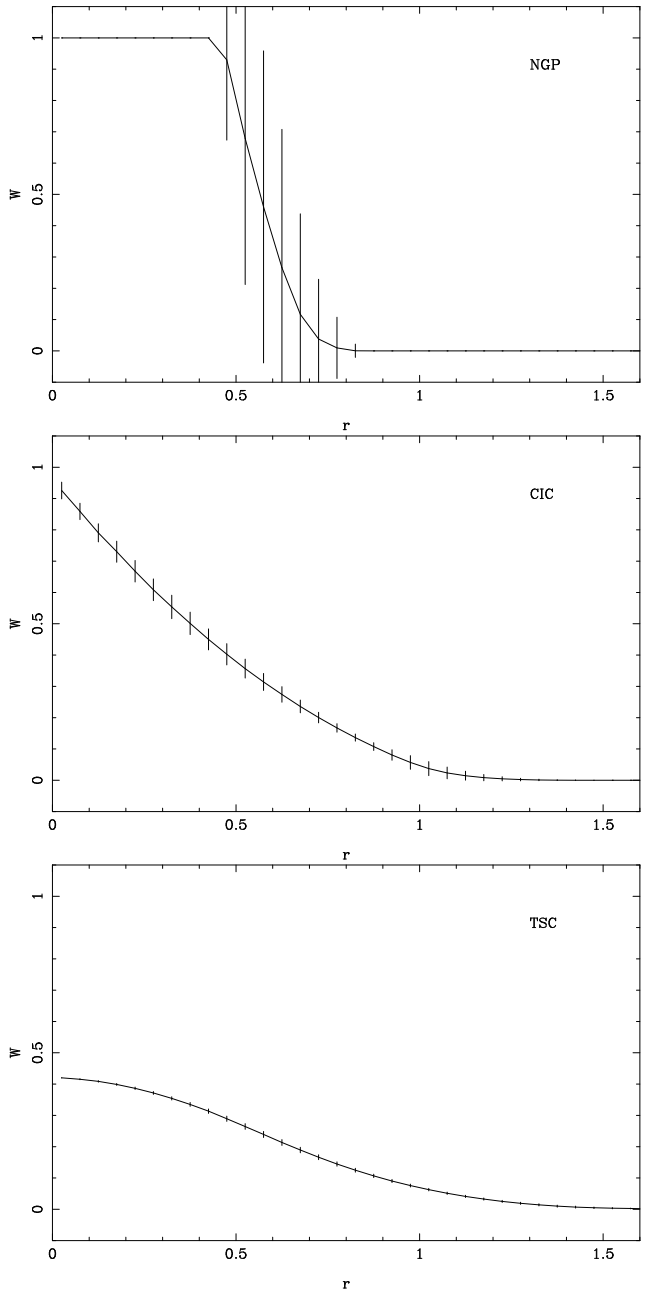

Fig. 18.— N-body window functions. Curves correspond to the average weight assigned to particles at a distance r from the center of a cell, while vertical bars indicate 1 σ dispersion around the mean. Panels correspond to *Nearest Grid Point*, *Cloud in Cell*, and *Triangular Shaped Cloud*, window functions, top to bottom. This figure is taken from Bagla & Padmanabhan (1997).

Here we take for reference particle mesh codes, which are the most common in cosmological simulations. Particle mesh codes group particles to create a density field, with Eularian overdensity at a cell located at $\mathbf{x}$:

$$\Delta(\mathbf{x}) \equiv \frac{M(\mathbf{x})}{\bar{M}}$$
$$M(\mathbf{x}) = \sum_{\text{particles,i}} M_i(\mathbf{x_i}) W(|\mathbf{x} - \mathbf{x_i}|) \tag{166}$$

Here we again have to use a window function, this time applied over all particles at locations $\mathbf{x_i}$. The window function, defined as above, has to sum to unity, be "local", and be reasonably isotropic so as to not create spurious torque. We again have to make tradeoffs, especially given that cells are usually cubic with side length L_{cell}, and thus do not naturally allow for isotropy. There are three common choices (though in practice the second of these is used the most frequently):

- *Nearest Grid Point* - this is the N-body analogy of the spherical top hat, discussed previously, assigning all of the particle's mass to the nearest cell center:

$$W(r) = \begin{cases} 1 & \text{if } r \leq L_{\text{cell}}/2 \\ 0 & \text{if } r > L_{\text{cell}}/2 \end{cases} \tag{167}$$

  This approach however becomes anisotropic as one approaches the grid scale (c.f. Fig. 18)

- *Cloud in Cell* - this window function linearly distributing a particle's mass between the closest two cells according to the distance to the cell's center:

$$W(r) = \begin{cases} 1 - r/L_{\text{cell}} & \text{if } r \leq L_{\text{cell}} \\ 0 & \text{if } r > L_{\text{cell}} \end{cases} \tag{168}$$

  Compared to the Nearest Grid Point, Cloud in Cell is more diffusive, but more isotropic.

- *Triangular Shaped Cloud* - this window function is a quadratic spline:

$$W(r) = \begin{cases} 3/4 - (r/L_{\text{cell}})^2 & \text{if } r \leq L_{\text{cell}}/2 \\ 9/8(1 - |2r/3L_{\text{cell}}|)^2 & \text{if } L_{\text{cell}}/2 \leq r \leq 3L_{\text{cell}}/2 \\ 0 & \text{if } r > 3L_{\text{cell}}/2 \end{cases} \tag{169}$$

  the most diffusive and isotropic. also computationally challenging.

As can be seen from Fig. 18 the cloud in cell is a compromise between the other two methods in terms of isotropy, diffusiveness and ease of computation. It is the most used window function.

Once we have the density field from eq. (166), we compute the Poisson equation to get the gravitational potential:

$$\nabla^2 \psi_{\mathbf{x}} = (\Delta_x - 1)/a \tag{170}$$

This is most conveniently done in Fourier space, since the Fourier dual of a gradient is simply multiplying by $-i\mathbf{k}$, and the inverse is multiplying by $-i\mathbf{k}/|\mathbf{k}|^2$:

$$\psi_{\mathbf{k}} = \frac{-i\mathbf{k}}{|\mathbf{k}|^2}(\Delta_k - 1)/a \tag{171}$$

Although FFTs have aliasing errors on discrete grids, these are far sub-dominant to errors resulting from numerical differentiation; c.f. Fig. 5 in Bagla & Padmanabhan 1997).

3.7.3. Moving the particles

Once the force is computed as discussed in the previous section, we need to move the particles. Writing the updated position and velocity of a particle i at time $t + \Delta t$ as a Taylor expansion of their values at time t:

$$x_i(t + \Delta t) = x_i(t) + v_i(t)\Delta t + O[(\Delta t)^2] \tag{172}$$
$$v_i(t + \Delta t) = v_i(t) + a_i(t)\Delta t + O[(\Delta t)^2] \tag{173}$$

, where the leading error if truncating to first order is $O[(\Delta t)^2]$. In practice however, we cannot chose an infinitely small time step. We therefore should reduce the error as much as possible for a finite time step Δt.

The most common approach is the so-called Leap-Frog method, which achieves $O[(\Delta t)^3]$ error by updating the position and velocity at alternate half time steps:

$$x_i(t + \Delta t) = x_i(t) + v_i(t + \Delta t/2)\Delta t + O[(\Delta t)^3] \tag{174}$$
$$v_i(t + \Delta t/2) = v_i(t - \Delta t/2) + a_i(t)\Delta t + O[(\Delta t)^3] \tag{175}$$

3.7.4. Finding halos

Friends of Friends in either real space or phase space.. spherical top hat

4. Baryonic structures: the formation of galaxies

For the second section of the course, we will focus on baryonic (i.e. "normal") matter. Although less mysterious than dark matter, baryons are in some ways a lot more interesting... Dynamically, baryons are able to do two main things which dark matter cannot[16]: (i) be pressure supported against gravitational contraction; and (ii) interact with radiation which allows it to cool and heat. Because of these effects, baryonic structure formation initially proceeds slower, but later is able to achieve much higher densities via radiative cooling. Highly collapsed objects, such as galaxies, stars, planets, people, etc. are all composed (primarily) of baryons.

We will first explore the life-cycle of baryons which end up in galaxies and stars. Afterwards we will study those baryons which remain in the intergalactic medium (IGM). The former, although a more interesting life-cycle, applies to only a few percent of all baryons. Moreover, as we shall see later, the fate of all of the baryons is interconnected through intergalactic radiation fields which permeate the Universe.

4.1. Linear evolution with pressure

Let us begin by returning to our fluid equations, including an additional pressure term for the baryons. The additional pressure force per unit proper volume can be written as:

$$\mathbf{F} = -a^{-1}\nabla p \ . \tag{176}$$

In most cases, the pressure can be expressed as just a function of the baryon density, ρ:

$$\begin{aligned}\mathbf{F} &= -a^{-1}\frac{\partial p}{\partial \rho}\nabla\rho \\ &= -a^{-1}c_s^2\bar{\rho}\nabla\delta \ ,\end{aligned} \tag{177}$$

where in the last equation we note that the adiabatic derivative of the pressure with respect to density is the square of the sound speed, c_s, we expressed the density as a linear perturbation over the mean, $\rho = \bar{\rho}(1+\delta)$, and the gradient is comoving while other terms are in proper units (as per convension).

If we add this additional term (dividing by $\bar{\rho}$ to get the force per unit mass) to our linear equation of motion, eq. (48) we obtain:

$$\dot{\mathbf{v}}_{\rm pec} + \frac{\dot{a}}{a}\mathbf{v}_{\rm pec} = -a^{-1}\nabla\phi_{DM+b} - a^{-1}c_s^2\nabla\delta \ . \tag{178}$$

Note that this equation describes the behavior of the gas, but the gravitational potential, ϕ, depends on both the dark matter and the baryons. Now, taking the divergence of this expression, $a^{-1}\nabla\cdot$:

$$a^{-1}\nabla\cdot\dot{\mathbf{v}}_{\rm pec} + \frac{\dot{a}}{a^2}\nabla\cdot\mathbf{v}_{\rm pec} = -a^{-2}\nabla^2\phi_{DM+b} - a^{-2}c_s^2\nabla^2\delta \ . \tag{179}$$

As we did in §3.1 we will subtract this from $\partial/\partial t$ of the linear order continuity equation:

$$\ddot{\delta} - \frac{\dot{a}}{a^2}\nabla\cdot\mathbf{v}_{\rm pec} + a^{-1}\nabla\cdot\dot{\mathbf{v}}_{\rm pec} = 0 \tag{180}$$

Subtracting eq. (179) from eq. (180), we obtain:

$$\ddot{\delta} - 2\frac{\dot{a}}{a^2}\nabla\cdot\mathbf{v}_{\rm pec} = a^{-2}\nabla^2\phi_{DM+b} + a^{-2}c_s^2\nabla^2\delta \ . \tag{181}$$

We again make the substitutions $\dot{\delta} = -a^{-1}\nabla\cdot\mathbf{v}_{\rm pec}$ (first order continuity) and $\nabla^2\phi_{DM+b} = 4\pi G\bar{\rho}_0 a^{-1}\delta_{DM+b}$ (Poisson):

$$\boxed{\ddot{\delta} + 2\frac{\dot{a}}{a}\dot{\delta} = 4\pi \mathrm{G}\bar{\rho}_0 a^{-3}\delta_{DM+b} + a^{-2}c_s^2\nabla^2\delta} \ . \tag{182}$$

[16]Strictly speaking, dark matter has a form of pressure support on small scales, due to the its residual velocity dispersion. Moreover, dark matter can annihilate and decay, releasing energy and particle showers. However, these effects are only important for structure formation in a small subset of dark matter models.

4.2. Cosmological Jeans mass

Equation (182) is analogous to eq. 53 for the linear evolution of an overdensity perturbation, but with the final additional term corresponding to "drag" from baryonic pressure. As is done in the classical Jeans mass derivation, we can compute the static solution in which pressure balances gravity. For simplicity, we assume $\delta_{DM+b} \approx \delta_b$, which is reasonable for times later than recombination and large scales (see Naoz & Barkana 2005 and references therein for a more general solution). Thus the static solution becomes:

$$0 = 4\pi \mathrm{G}\bar{\rho}_0 a^{-3}\delta + a^{-2}c_s^2\nabla^2\delta \ . \tag{183}$$

Intuition is better obtained by looking at the Fourier transform (noting that $\nabla \to -i\mathbf{k}$ and $\nabla^2 \to -|\mathbf{k}|^2$):

$$0 = \left[4\pi \mathrm{G}\bar{\rho}_0 a^{-3} - a^{-2}c_s^2|\mathbf{k}|^2\right]\delta_k \ . \tag{184}$$

Define k_j to be the magnitude of the (comoving) wavemode satisfying this stability criterion,

$$k_j = \sqrt{\frac{4\pi \mathrm{G}\bar{\rho}_0}{c_s^2 a}} \ , \tag{185}$$

and the corresponding (comoving) wavelength, $\lambda_j \equiv 2\pi/k_j$:

$$\boxed{\lambda_j = \sqrt{\frac{\pi c_s^2}{G\bar{\rho}_0 a^{-1}}}} \ . \tag{186}$$

Thus density perturbations with wavelengths of $\lambda > \lambda_j$ start to collapse due to gravity, while those with $\lambda < \lambda_j$ are pressure supported against gravitational collapse.

We can make an analogy with the classical expression by writing the Jeans wavelength in proper units, $\lambda_j^p = a\lambda_j = \sqrt{\frac{\pi c_s^2}{G\bar{\rho}_0 a^{-3}}} = \sqrt{\frac{\pi c_s^2}{G\bar{\rho}(z)}}$:

$$\frac{\lambda_j^p}{c_s} = \sqrt{\frac{\pi}{G\bar{\rho}}} \ . \tag{187}$$

The term on the LHS corresponds to the sound crossing time, $t_{\rm sc}$, while the term on the RHS corresponds to the free-fall or dynamical time, $t_{\rm dyn}$. Thus we recover the classical result that if $t_{\rm sc} > t_{\rm dyn}$ pressure is unimportant, while if $t_{\rm sc} < t_{\rm dyn}$ density perturbations will stabilize as acoustic waves.

Let's write the mass corresponding to the cosmic Jeans length:

$$\begin{aligned} M_J &\equiv \frac{4\pi}{3}\left(\frac{\lambda_j}{2}\right)^3 \bar{\rho}_0 = \frac{4\pi}{3}\frac{c_s^3}{8}\left(\frac{\pi a}{G\bar{\rho}_0}\right)^{3/2}\bar{\rho}_0 \\ &= \left(\frac{\pi^{5/2}}{6G^{3/2}}\bar{\rho}_0^{-1/2}\right)\ \left(c_s^2 a\right)^{3/2} \ . \end{aligned} \tag{188}$$

The first term in the above equation is a constant, while the second term evolves with time. Thus we see that $M_J \propto (c_s^2 a)^{3/2}$.

How does the cosmic Jeans mass evolve? Consider the following epochs:

- **After matter-radiation equality, before recombination** ($1100 \lesssim z \lesssim 3000$) – Here the baryons are a relativistic fluid, with $c_s^2 = c^2/3 =$ *constant*. Thus $M_J \propto a^{3/2}$. Evaluating at $z \sim 1000$, we have $M_J \sim 10^{18} M_\odot$. This is larger than a super-cluster. Thus before recombination, relevant baryonic perturbations are pressure supported to collapse. (thankfully, Dark Matter has no such restrictions).
- **After recombination, before thermal decoupling from the CMB** ($200 \lesssim z \lesssim 1100$) – The sound speed for a classical ideal gas is $c_s^2 = \frac{5k_B T}{3\mu m_p}$. But before thermal decoupling the gas temperature traces the CMB temperature (through Compton scattering with the residual electron fraction), $T = T_\gamma \propto a^{-1}$. Thus, $M_J \propto (a^{-1}a)^{3/2} =$ *constant*. Evaluating gives $M_J \sim 10^5 M_\odot$, which is the mass of globular cluster. Thus, the first baryonic perturbations to start gravitation collapse were on the scale of globular clusters.

- **After thermal decoupling, before the first stars** ($30 \lesssim z \lesssim 200$) – Here the cosmic gas expanded to low enough densities to make Compton scattering with the CMB inefficient; thus the gas cools mostly adiabatically, $T \propto a^{-2}$. Therefore $c_s^2 \propto T \propto a^{-2}$, and $M_J \propto (a^{-2}a)^{3/2} \propto a^{-3/2}$, decreasing with time. Evaluating, we get $M_J \sim 7 \times 10^3 \left(\frac{1+z}{10}\right)^{3/2} M_\odot$.

- **Reionization** ($6 \lesssim z \lesssim 10$) – We will later study the thermal evolution of cosmic gas when the first stars form in more detail, but it is important to note that radiation from early galaxies can dramatically raise the Jeans mass by ionizing the gas. As an ionization front passes over cosmic gas, the temperature is raised to $T \sim 10^4$ K and $\mu = 0.6$. If the IGM temperature remains roughly constant, with photo heating balancing radiative losses, the Jeans mass scales as $M_J \propto (a^0 a)^{3/2} \sim 10^8 \left(\frac{1+z}{10}\right)^{-3/2} M_\odot$. Thus the phase transition of Cosmic Reionization could have a profound impact on later structure formation by raising the Cosmological Jeans mass.

We end this discussion by noting that the Jeans mass is an instantaneous quantity. However pressure waves take a finite amount of time to respond to changes in the environment. Thus a more relevant quantity (albeit less intuitive) is the "time-averaged" Jeans wavenumber, i.e. the so-called "filtering" wavenumer, k_F, proposed by Hui & Gnedin (1997):

$$k_F^{-2} = \frac{3}{a} \int_0^a \frac{da'}{k_j^2} \left[1 - \sqrt{\frac{a'}{a}}\right] \ . \tag{189}$$

Since the Jeans mass during the early Universe was larger at earlier times, this raises somewhat the mass required for pressure support. However, non-linear evolution makes quantitative predictions based on either the Jeans or the filtering mass difficult.

4.3. Thermal evolution of collapsing gas

The linear theory argument from the previous section only specifies which perturbations in the gas density are unstable at a given time, and thus *start* to collapse. It is a reasonable approximation at large scales. It is however important to remember that dark matter has a head start with respect to the baryons. This means that on small-scales, the baryons "lag" behind the dark matter, and end up accreting onto already formed dark matter structures. In this stage of its evolution, the accreting gas can get gravitationally heated to the virial temperature of the dark matter halo (e.g. Barkana & Loeb 2004):

$$T_{\rm vir} \approx 10^4 {\rm K} \left(\frac{\mu}{0.6}\right) \left(\frac{M_h}{10^8 M_\odot}\right)^{2/3} \left(\frac{1+z}{10}\right) \left[\frac{\Omega_{m,0}}{0.3} \frac{1}{\Omega_m(z)} \frac{\Delta_c}{18\pi^2}\right] . \quad (190)$$

Here M_h is the total mass of the halo, $\Delta_c = \rho/\bar{\rho}$ the average density of the halo, and μ is the mean atomic weight with $\mu \approx 0.6$ (1.2) for ionized (neutral) primordial gas.

For the formation of galaxies, the virial temperature of the halo is a more important quantity than the IGM temperature. In order to continue collapsing inside the DM halo, the gas must be able to cool to compensate for the gravitational heating. *This is a fundamental difference between the DM and baryons: baryons can collapse to higher densities because they can radiate away thermal energy.*

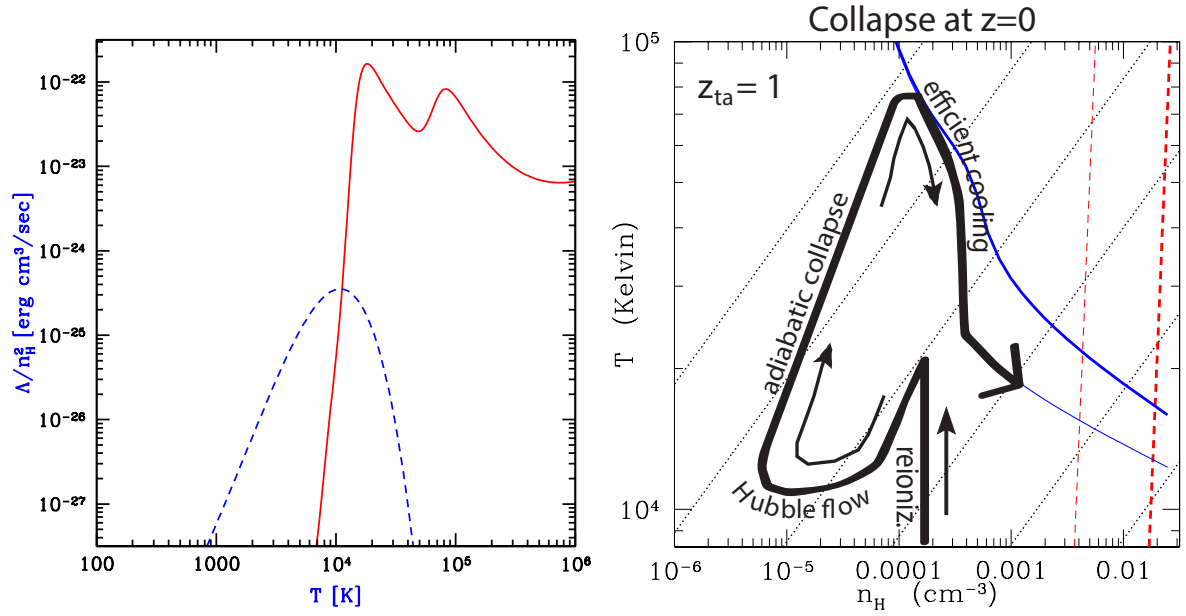

Fig. 19.— *Left:* the cooling function, taken from Barkana & Loeb (2001), computed using Abel online calculator, assuming primordial abundances. The solid red curve corresponds to atomic gas (H and He), while the dashed blue curve corresponds to molecular hydrogen. Note that in the literature the cooling function is sometimes also defined as the y-axis label, implicitly including the $n_{\rm H}^{-2}$ term in the definition. *Right:* the trajectory through phase space for gas collapsing onto a halo. This specific example is for a halo at $z = 0$, but most of the milestones are applicable to higher redshifts (taken from Noh & McQuinn 2014). The gas decouples from the Hubble flow at turn-around (see §3.2), and begins adiabatically contracting onto the enclosed halo. The gas either contracts or is shock heated (not shown in the figure) to the halo's virial temperature. Further collapse then proceeds relatively isothermally, staying at the peaks of of the cooling curves from the left panel (shown here with blue lines).

The efficiency of this cooling radiation as a function of the gas temperature can be seen in the left panel of Fig. 19, for a gas with primordial composition. The two peaks of the solid red curve correspond to cooling by H and He (at $\sim 10^4$ and $\sim 10^5$ K respectively). Indeed, most galaxies obtain their gas from the recombination and collisional excitation cooling of hydrogen. However, hydrogen cooling becomes very inefficient below $\sim 10^4$ K (corresponding to halos of mass $\sim 10^8 M_\odot$ at $z \sim 10$; see eq. 190). Since structure formation is hierarchical, smaller DM halos formed earlier. Thus the first galaxies in

our Universe likely accreted their gas through the rotational-vibrational transitions of the H_2 molecule (see the dashed blue curve in the left panel of Fig. 19). Molecular hydrogen allowed gas to condense onto even smaller halos, with virial temperatures of order $T_{\rm vir} \sim 10^3$ K (corresponding to halo masses of $\sim 10^6 - 10^7 M_\odot$; e.g. Haiman et al. 1996; Abel et al. 2002; Bromm et al. 2002). However, H_2 is fragile, and can be directly dissociated once the first galaxies form by radiation in the so-called Lyman-Werner (LW) band ($\sim$10.2–13.6 eV), or by disrupting H^- which is a catalyst in the chemical process by which H_2 abundances are enhanced over the primordial ones as the gas collapses to higher densities. How and when H_2 stops being an important coolant for gas accreting onto halos is a popular topic in modern research (e.g. Holzbauer & Furlanetto 2012; Fialkov et al. 2013; Wolcott-Green et al. 2017; Schauer et al. 2021; Muñoz et al. 2022).

Therefore, if the virial temperature of the halo is higher than required for efficient cooling: $\gtrsim 10^3$ K (10^4 K) for H_2 (H) cooling, the accreting gas can radiate away the energy obtained by gravitational heating, and keep going deeper into the potential well of the halo. This phase of the collapse is relatively isothermal, as the gas maintains the temperature of the peak of its cooling curve (c.f. the right panel of Fig. 19 taken from Noh & McQuinn 2014). The collapse stops being isothermal when the dynamical timescale of the gas becomes shorter than the cooling timescale: $t_{\rm dyn} < t_\Lambda$.

4.4. The first stars and black holes

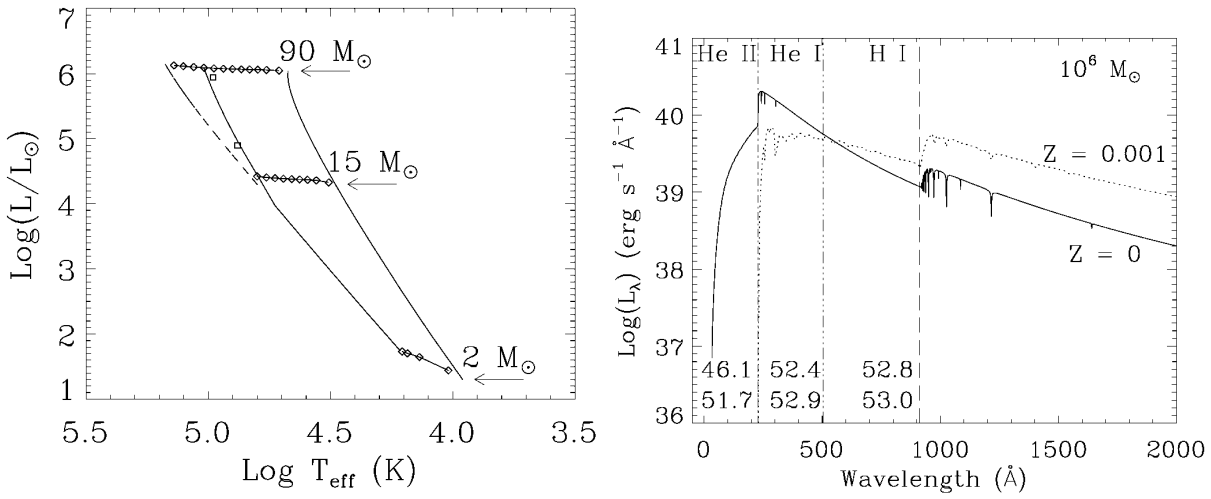

Fig. 20.— *Left panel:* Bolometric luminosity vs. effective temperatures for zero-age main sequence stars with masses spanning $2 \leq M/M_\odot \leq 90$. Population I stars with $Z \sim Z_\odot = 0.02$ are represented with the right curve, while Population III stars with $Z = 10^{-10}$ are shown with the left solid curve. The left dashed curve corresponds to $Z = 0$ for the $M > 15M_\odot$ models. The diamonds mark decreasing decades in metallicity from $Z = 0.02$. *Right panel:* Model spectra of a $10^6 M_\odot$ PopII (dotted curve) and PopIII (solid curve) star burst at an age of ~ 1 Myr. The numbers in the lower left represent the log of the number of ionizing photons in each band. Figures are taken from Tumlinson & Shull (2000).

What kind of objects (i.e. stars) result from this collapse? The final stages of collapse and the resulting distribution of the stellar masses (the so-called initial mass function; IMF) depends on the details of angular momentum transport, turbulence, and radiation pressure (see the review of Milosavljević & Safranek-Shrader 2016 and references therein). However, we can get a rough estimate by studying the fragmentation scale of the collapsing gas in the linear regime. Fragmentation occurs when the gas is Jeans unstable, i.e. the sound crossing time is longer than the dynamical time. Note that the Jeans mass scales as $M_J \propto \rho^{-1/2} c_s^3$. While the collapse is isothermal ($t_{\rm dyn} > t_\Lambda$), the temperature and sound speed are constant, and so we have $M_J \propto \rho^{-1/2}$. In other words, as the density increases during collapse, the Jeans mass decreases, resulting in the collapsing gas continuing to break up into smaller clumps. However, as the collapse transitions to being adiabatic ($t_{\rm dyn} < t_\Lambda$), we have $M_J \propto \rho^{-1/2} c_s^3 \propto \rho^{-1/2} T^{3/2}$. Given that $T \propto \rho^{2/3}$ for adiabatic contraction, $M_J \propto \rho^{1/2}$. Thus, during adiabatic collapse, fragmentation stops as the Jeans mass increases with further increase in density.

Thus the Jeans mass at the transition from isothermal to adiabatic collapse ($t_{\rm dyn} \sim t_\Lambda$) gives us an estimate of the final mass of the star. Cooling is very complicated in this regime, but we can parametrize it as a fraction of the black-body luminosity (i.e. the maximum allowed radiative efficiency):

$$\frac{dE_\Lambda}{dt} = f_{\rm eff} L_{\rm BB} = f_{\rm eff} 4\pi\sigma T^4 R_*^2 \ . \tag{191}$$

Here σ is the Stefan Boltzmann constant, and R_* is the scale of the clump. We can compare this with the gravitational heating rate[17]:

$$\begin{aligned} \frac{dE_g}{dt} &\approx \frac{E_g}{t_{\rm dyn}} \approx \frac{GM_*^2/R_*}{(G\rho)^{-1/2}} \\ &= \sqrt{\frac{3}{4\pi}} \frac{G^{3/2} M_*^{5/2}}{R_*^{5/2}} \ , \end{aligned} \tag{192}$$

[17]Note that we are discussing the final stages of the collapse inside the ISM of the galaxy. Here the gravitational potential is dominated by the gas itself, with the DM contributing a sub-dominant, fairly smooth component to the potential.

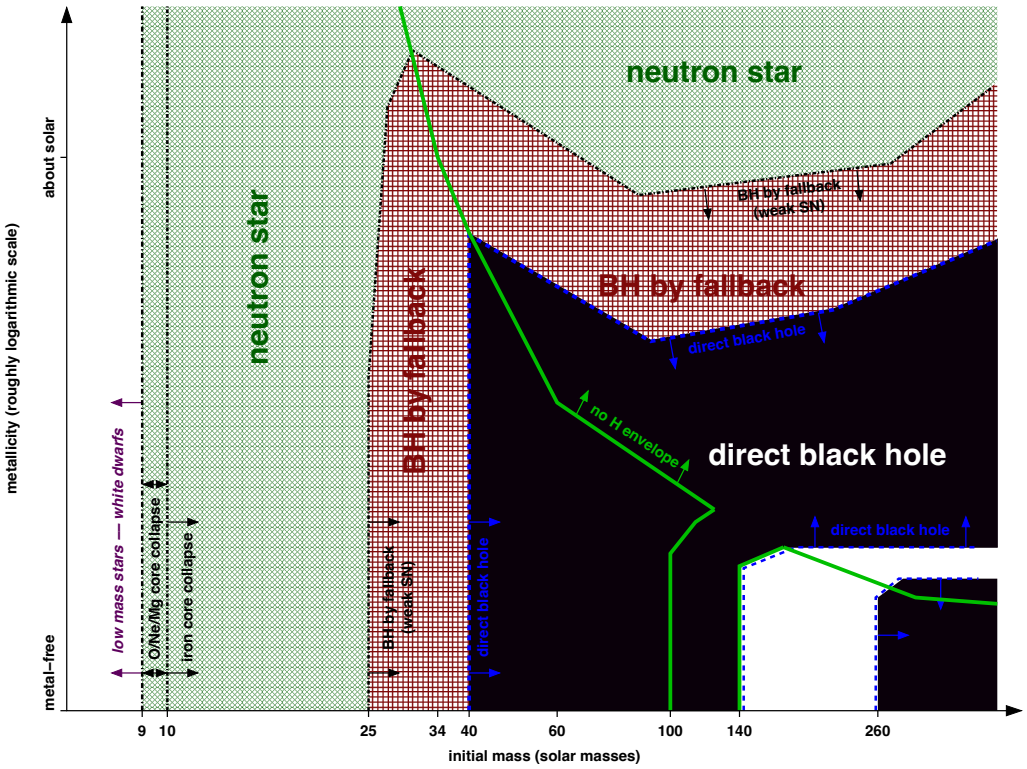

Fig. 21.— 1D, non-rotating models of stellar remnants from massive stars (taken from Heger et al. 2003). The initial metallicity is plotted along the vertical axis, while the initial mass is plotted along the horizontal axis. The white patch in the bottom right corresponds to pair-instability supernovae that leave no remnant.

where M_* is the mass of the clump (destined to be a star). Equating the radiative cooling rate (eq. 191) with the gravitational heating rate (eq. 192):

$$\begin{aligned}
\frac{dE_g}{dt} &\approx \frac{dE_\Lambda}{dt} \\
M_*^5 &\approx \frac{(4\pi)^3}{3G^3} f_{\rm eff}^2 \sigma^2 T^8 R_*^9 \approx \frac{(4\pi)^3}{3G^3} f_{\rm eff}^2 \sigma^2 T^8 \left(\frac{3}{4\pi}\right)^3 \rho^{-3} M_*^3 \\
M_*^2 &\approx \frac{9}{(G\rho)^3} f_{\rm eff}^2 \sigma^2 T^8 \\
M_* &\approx 3 f_{\rm eff} \sigma T^4 G^{-3/2} \rho^{-3/2}
\end{aligned} \tag{193}$$

We can simplify this further and relate the clump radius to its mass through the Jeans mass expression from eq. (188): $\rho^{-1/2} = \frac{6G^{3/2}}{\pi} \frac{M_*}{c_s^3}$. Substituting into the above:

$$\begin{aligned}
M_* &\approx \frac{3 f_{\rm eff} \sigma T^4}{G^{3/2}} \left(\frac{6G^{3/2}}{\pi} \frac{M_*}{c_s^3}\right)^3 \approx 20 f_{\rm eff} \sigma T^4 G^3 M_*^3 c_s^{-9} \\
M_*^2 &\approx 0.05 G^{-3} f_{\rm eff}^{-1} \sigma^{-1} T^{-4} c_s^9 \\
M_* &\approx 0.22 G^{-3/2} f_{\rm eff}^{-1/2} \sigma^{-1/2} T^{-2} \left(\frac{5 k_B T}{3 m_p}\right)^{9/4}
\end{aligned}$$

Simplifing and substituting in constants, we obtain:

$$\boxed{M_* \sim 0.1 M_\odot \left(\frac{f_{\rm eff}}{0.01}\right)^{-1/2} \left(\frac{T}{10^4 {\rm K}}\right)^{1/4}}, \tag{194}$$

which is typical of a dwarf star. Here, the fiducial choice of $f_{\rm eff}$ is loosely motivated by CO cooling in local molecular clouds.

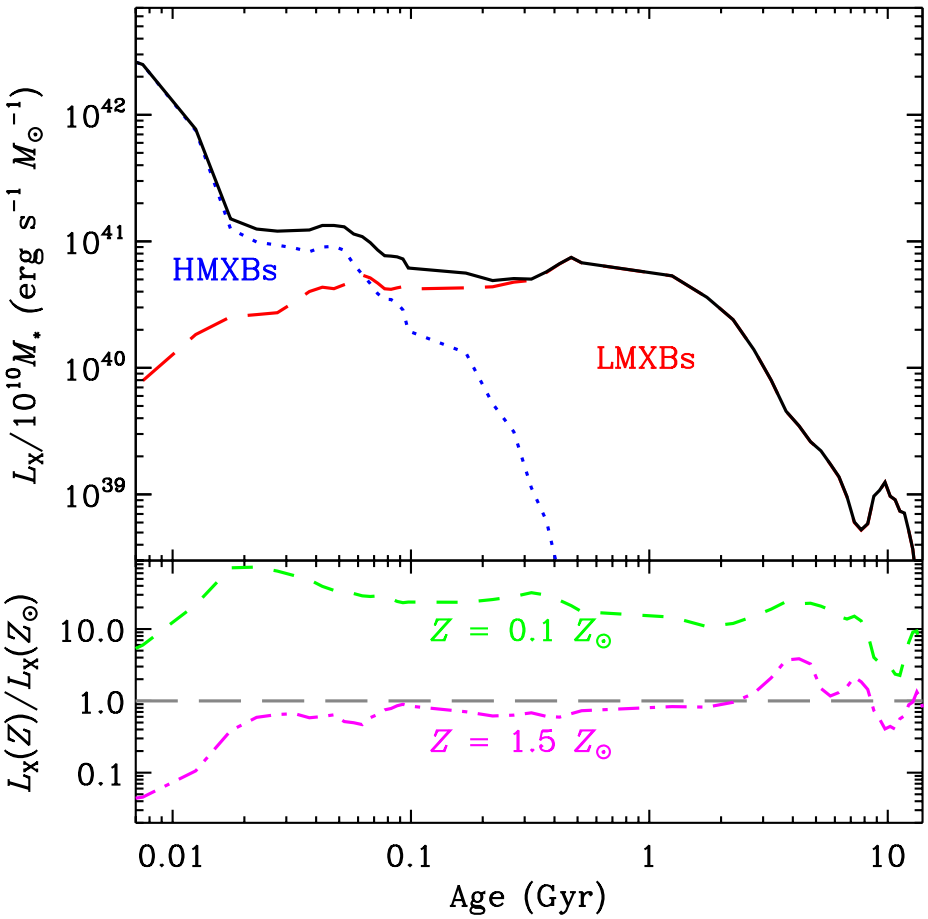

Fig. 22.— Evolution of the X-ray luminosity from a single star burst. The top panel shows the bolometric luminosity, initially dominated by high mass X-ray binaries, and subsequently by low mass X-ray binaries. The bottom panel shows the dependence of the X-ray luminosity on the assumed initial metallicity, normalized to solar. Binary stars forming in low metallicity environments are expected to be more luminous, primarily because metal-driven winds are less efficient and more mass is available to fuel accretion onto the compact object. The figure is taken from Fragos et al. 2013.

The above mass scale is the correct order of magnitude for the locally-observed stellar population. However, the primordial collapsing gas can cool less efficiently than metal-enriched gas. If cooling is less efficient (i.e. a lower $f_{\rm eff}$), the fragmentation scale is larger. Thus, qualitatively we expect the very first stars born from primordial gas (so-called Population III stars) to be more massive than later-generation PopII stars.

Nuclear burning also proceeds differently in PopIII stars. Lacking ingredients for the CNO cycle, PopIII stars are powered by the PP chain. The relative inefficiency of the PP chain means that core temperatures must be much higher than normal to provide sufficient energy output for pressure support. Subsequently helium starts burning via the triple-α process, creating trace amounts of heavy elements which facilitates the beginning of the CNO cycle.

As a result, the effective temperature of PopIII stars is expected to be much higher than present-day stars. In Fig. 20, we see predictions for PopIII temperatures (*left panel*) and spectral energy distributions (SEDs; *right panel*), computed using 1D stellar evolution models (Tumlinson & Shull 2000). The left panel shows that the absence of metals can result in an effective temperature up to a factor of few higher than PopII stars of equivalent mass. The resulting spectra (*right panel*) are much harder, with a factor of few–ten more HI and HeI ionizing photons.

PopIII stars also likely have different end products. This is shown in Fig. 21, taken from the 1D, non-rotating models of (Heger et al. 2003). The initial metallicity is shown along the vertical axis, while the

initial mass is plotted along the horizontal axis. Comparing the end products for $M_* \gtrsim 40 M_\odot$ stars which have "about solar" metallicity to those which are "metal-free", we see the added possibilities of direct collapse black holes as well as pair-instability supernovae (SNe). Pair-instability supernova result from the production of free electrons and positrons reducing the thermal pressure inside the core, driving an accelerated burning whose thermonuclear explosions blows apart the star completely, leaving no remnant. It is simple to argue that PopIII stars cannot all result in pair-instability supernovae, or there would be virtually no metal pollution which is required for the stellar populations to transition to those seen today and in high-redshift galaxies.

Finally, it is important to also note that the primordial gas of the first galaxies likely resulted in different properties of binary stars. As we shall see more below, the X-rays produced as gas is accreted from a donor star to a compact object in binary systems are thought to be the dominant sources of heat for the IGM, prior to reionization. Due to the higher fragmentation scale, we expect a somewhat higher binary fraction in metal-free ISM. Moreover, the lack of metals means decreased stellar winds, allowing more mass to be available to fuel accretion onto the compact object (e.g. Belczynski et al. 2010). This can result in a factor of ~ 10 higher bolometric X-ray luminosities from high mass X-ray binary systems (see the bottom panel of Fig. 22 which assumes the same IMF, but follows the evolution of binaries of different metalicities; taken from Fragos et al. 2013).

4.5. Analytic models of star formation inside the first galaxies

Star formation is highly complicated. It is an advanced area of research, requiring state-of-the-art numerical simulations capturing very small-scales (see the review of Milosavljević & Safranek-Shrader 2016 and references therein). However, in cosmology we are often interested in general trends and empirical averages. Therefore it is often convenient to describe star formation with simple analytic recipes, if possible, which would capture general trends of the galaxy population. Here we discuss some general properties of basic, semi-analytic models of galaxy formation (SAMs).

The star formation rate of a given galaxy can be expressed as:

$$\frac{dM_*}{dt} = \int_{t_0}^{t} dt' \frac{dM_{\rm acc}}{dt'} f_* P_*(t - t') \ . \tag{195}$$

Here t_0 corresponds to the time at which the host DM halo was sufficiently massive to allow gas to efficiently cool. For example, $T_{\rm vir} \sim 10^4$ K corresponds to the peak of the H-cooling function, which would be relevant assuming that star formation inside molecularly-cooled minihalos was sterilized very early. $dM_{\rm acc}/dt' \times dt'$ is the mass of fresh cosmic gas which cooled and accreted onto the galaxy at a time between t' and $t' + dt'$. f_* corresponds to the fraction of this gas which ends up in stars (over a relevant time-scale, e.g. t_H), and $P_*(\tau)$ is the probability per unit time that a star formed from a gas element with "age" $\tau \equiv t - t'$ since being accreted onto the DM halo. This probability should be normalized to unity over the same time-scale used to define f_*, e.g. $\int_0^{t_H} d\tau P_*(\tau) \approx 1$.

4.5.1. Gas accretion

New gas is brought into the galaxy via a combination of accretion and mergers (which can be viewed as a form of "clumpy" accretion).The gas needs to cool to get into the galaxy and potentially form stars. For a spherically symmetric profile, the cooling time as a function of distance from the halo center, r, can be written as:

$$t_\Lambda(r) = \frac{3}{2} \frac{\mu m_p k_B T}{\rho(r) \Lambda(T)} \ , \tag{196}$$

where $\Lambda(T)$ is the cooling rate in erg cm^3 s^{-1} at temperature T (see Fig. 19, though note the different naming convention), ignoring metals since we assume primordial enrichment in the infalling gas[18].

SAMs usually compute a cooling radius, $r_{\rm cool}$ at which the cooling time is equal to the dynamical time, i.e. $t_\Lambda(r_{\rm cool}) = t_{\rm dyn}(r_{\rm cool}) \approx [G\rho(r_{\rm cool})]^{-1/2}$. Note that during the initial isothermal collapse, we have $t_\Lambda \propto \rho(r)^{-1}$ while $t_{\rm dyn} \propto \rho(r)^{-1/2}$, so at radii smaller than $r_{\rm cool}$ the cooling time becomes shorter and gas accretion onto the galaxy is limited just by the dynamic time. Once can define two different cooling regimes:

- $r_{\rm cool} > R_{\rm vir}$: infalling gas can efficiently radiate energy and cool onto the galaxy. In this regime, the accretion rate onto the galaxy is

$$\frac{dM_{\rm acc}}{dt} \sim \frac{M_{\rm g}}{t_{\rm dyn}} \sim \frac{M_h}{t_{\rm dyn}} \left(\frac{\Omega_b}{\Omega_m}\right) f_b \sim \frac{dM_h}{dt} \left(\frac{\Omega_b}{\Omega_m}\right) f_b \tag{197}$$

  where $M_{\rm g}$ is the total gas mass inside a DM halo of mass $M_{\rm h}$, f_b is the baryon fraction with respect to the cosmic mean value, Ω_b/Ω_m, and the final equality makes use of the fact that the dynamical time is less then or comparable to the halo growth time scale at high-z (i.e. $t_{\rm dyn} \lesssim t_H$). The halo efficiently brings in gas along with the dark matter as it grows.

[18] Although mergers can bring in polluted gas, the impact on the cooling function at the relevant virial temperatures is small.

- $r_{\rm cool} < R_{\rm vir}$: infalling gas settles onto a quasi static hot halo, and is slowly accreted along the boundary according to:

$$\frac{dM_{\rm acc}}{dt} \sim 4\pi\rho_{\rm cool} r_{\rm cool}^2 \frac{dr_{\rm cool}}{dt} \ , \tag{198}$$

  where $\rho_{\rm cool}$ is the gas density evaluated at the cooling radius.

Which regime is more appropriate for our early galaxies? Let's evaluate the cooling time at the virial radius. If we assume an isothermal gas profile (roughly matching galaxy rotation curves), we can write $\rho(r) = \rho_{\rm vir}(R_{\rm vir}/r)^2$, where $\rho_{\rm vir}$ is the gas density at the virial radius of the halo. Integrating out to the virial radius to get the enclosed mass:

$$\begin{aligned} \int_0^{R_{\rm vir}} 4\pi r^2 \rho(r) dr &= M_h \left(\frac{\Omega_b}{\Omega_m}\right) f_b \\ 4\pi R_{\rm vir}^2 \rho_{\rm vir} \int_0^{R_{\rm vir}} r^2 r^{-2} dr &= M_h \left(\frac{\Omega_b}{\Omega_m}\right) f_b \\ \rho_{\rm vir} &= \frac{M_h}{4\pi R_{\rm vir}^3} \left(\frac{\Omega_b}{\Omega_m}\right) f_b \approx 60\bar{\rho}(z) \ , \end{aligned} \tag{199}$$

where in the final approximation, we take $f_b = 1$ and use the result from the spherical collapse model that the average density inside the halo is 180 times the average density of the universe at that time $180\bar{\rho}(z) = \frac{M_h \Omega_b/\Omega_m}{\frac{4}{3}\pi R_{\rm vir}^3}$. We can then evaluate the cooling time at the virial radius:

$$\begin{aligned} t_\Lambda(R_{\rm vir}) &= \frac{3}{2}\frac{\mu m_p k_B T_{\rm vir}}{\rho_{\rm vir}\Lambda(T_{\rm vir})} = \frac{3}{2}\frac{k_B T_{\rm vir}}{60\bar{n}(z)\Lambda(T_{\rm vir})} = \frac{3}{2}\frac{k_B T_{\rm vir}}{60 n_0 (1+z)^3 \Lambda(T_{\rm vir})} \\ &\approx 0.1{\rm Myr}\left(\frac{T_{\rm vir}}{10^4{\rm K}}\right)\left(\frac{10^{-22}{\rm erg\ cm^3\ s^{-1}}}{\Lambda}\right)\left(\frac{10}{1+z}\right)^3 \end{aligned} \tag{200}$$

So we see that the cooling time is much much smaller than the Hubble time for the first galaxies with $T_{\rm vir} \sim 10^4$–10^5 K. Thus it is common to approximate accretion onto early galaxies with eq. 197. From the above scallings, we see that as time progresses (lower redshifts and higher halo masses), the situation reverses. Taking for example, $z = 0$ and $T_{\rm vir} \sim 10^8$ K, we see that the cooling time can impede the gas accretion.

4.5.2. Star formation probability

We now return to the star-formation probability from eq. (195), i.e. $P_*(\tau)$ which is the probability per unit time that a star formed from a gas element with "age" τ since being accreted onto the DM halo. This is a fundamental quantity of any model, depending on the details of energy transfer in the ISM of early galaxies. It is very difficult to get far from first principles; however, it is useful in practice to look at some parametric forms, which can hopefully be tested against detailed galaxy observations.

We list three, useful parametric forms for $P_*(\tau)$, in order of increasing complexity:

- *Instantaneous* – in the limit that the stars form immediately after the gas is accreted onto the halo, $P_*(\tau)$ approaches a Dirac delta function:

$$P_*(\tau) \approx \delta(\tau) \tag{201}$$

- *Extended with a characteristic time-scale* – if star formation occurs on a characteristic (e-folding) time-scale, t_*, following gas accretion: (e.g. Cen & Ostriker 1992):

$$P_*(\tau) \approx \frac{\tau}{t_*^2} e^{-\tau/t_*} \tag{202}$$

Here t_* could be the dynamical time-scale[19] of the galactic disk or DM halo ($\sim$ few-100 Myr at high-z), approaching the Hubble time for continuous star formation. This parametrization is commonly used.

- *Sequence of bursts with a characteristic duration and recovery time* – this is a more general formulation of the above, allowing for non-monotonic star formation. It is motivated by the idea that feedback (radiative and SNe ejecta) from a star formation episode, can suppress subsequent star formation for some characteristic recovery time-scale, $t_{\rm reco}$. This results in a periodic Gaussian:

$$P_*(\tau) \propto \sum_i e^{-\frac{[\tau - i(t_{\rm reco}+t_*)]^2}{2t_*^2}} \tag{203}$$

with a "duty-cycle" of order $\sim t_*/(t_*+t_{\rm reco})$ corresponding to the fraction of the time a given galaxy is actively star forming (or analogously when averaging over many galaxies, it corresponds to the fraction of galaxies which are star-forming at any given time).

4.5.3. f_* and feedback-regulated star formation

The simplest choice for the stellar fraction is a constant. Locally, we see roughly $f_* \sim 0.01 - 0.1$. However a constant value does not match observations. When considering only galaxies inside small halos (i.e. those unaffected by AGN feedback, as is relevant at high redshifts), we see that star formation becomes less efficient at smaller masses. This scaling can be naturally explained if we relate the efficiency of star formation to the halo potential. Feedback via SNe driven winds provides such a scaling, as we discuss below.

By ejecting energy (thermal and mechanical) into the surrounding ISM, SNe explosions can temporarily suppresses further star formation episodes. This SNe feedback should be especially powerful in the small-mass galaxies we expect in the early Universe, since they reside in shallower potential wells, facilitating gas ejection.

We can get a few simple estimates of this process, starting with the total energy in SNe from a star-formation episode. Stars with $M_* \gtrsim 8M_\odot$ produce SNe after $\sim$ 10s Myr, with a typical energy of $E_{\rm SN} \sim 10^{51}$ erg (e.g. Portinari et al. 1998). Therefore, a star-burst with a total mass of ΔM_* will on average result in the following number of supernovae:

$$N_{\rm SNe} = \Delta M_* \int_{8M_\odot}^{200?M_\odot} \phi(M_*) dM_* \ , \tag{204}$$

where $\phi(M_*)dM_*$ is the initial mass function (IMF), i.e. the number of stars born with masses between M_* and M_*+dM_* per total stellar mass, normalized such that: $\int_{0.1M_\odot}^{200?M_\odot} M_*\phi(M_*)dM_* = 1$. A commonly-used Salpeter IMF Salpeter (1955) has a power-law form: $\phi(M_*) \propto M_*^{-2.35}$. Note that the poorly-known upper mass limit for star formation has a negligible impact on the total number of supernovae, due to the steepness of this power-law. Then the total SNe energy released from the star-burst is:

$$\Delta E_{\rm SNe} = N_{\rm SNe} E_{\rm SNe} \approx 7 \times 10^{53} {\rm erg} \left(\frac{\Delta M_*}{10^5 M_\odot}\right) \ , \tag{205}$$

where the fiducial choice for $\Delta M_* \sim 10^5 M_\odot$ is motivated by the typical size of globular clusters.

How does this energy get transferred to the ISM?. If the energy was instantly converted to thermal energy at the typical ISM density, it would radiate away almost instantly: $t_\Lambda(T \sim 10^7{\rm K}, n \sim 1{\rm cm}^{-3}) \sim$

[19]Note that the mean dynamical time scale should be halo mass independent and scale as the Hubble time, during matter domination: $t_{\rm dyn} \propto \rho^{-1/2} \propto (1+z)^{-3/2} \propto t_H$. Here the mean in the halo is assumed to be proportional to the mean density of the Universe, following the spherical collapse model.

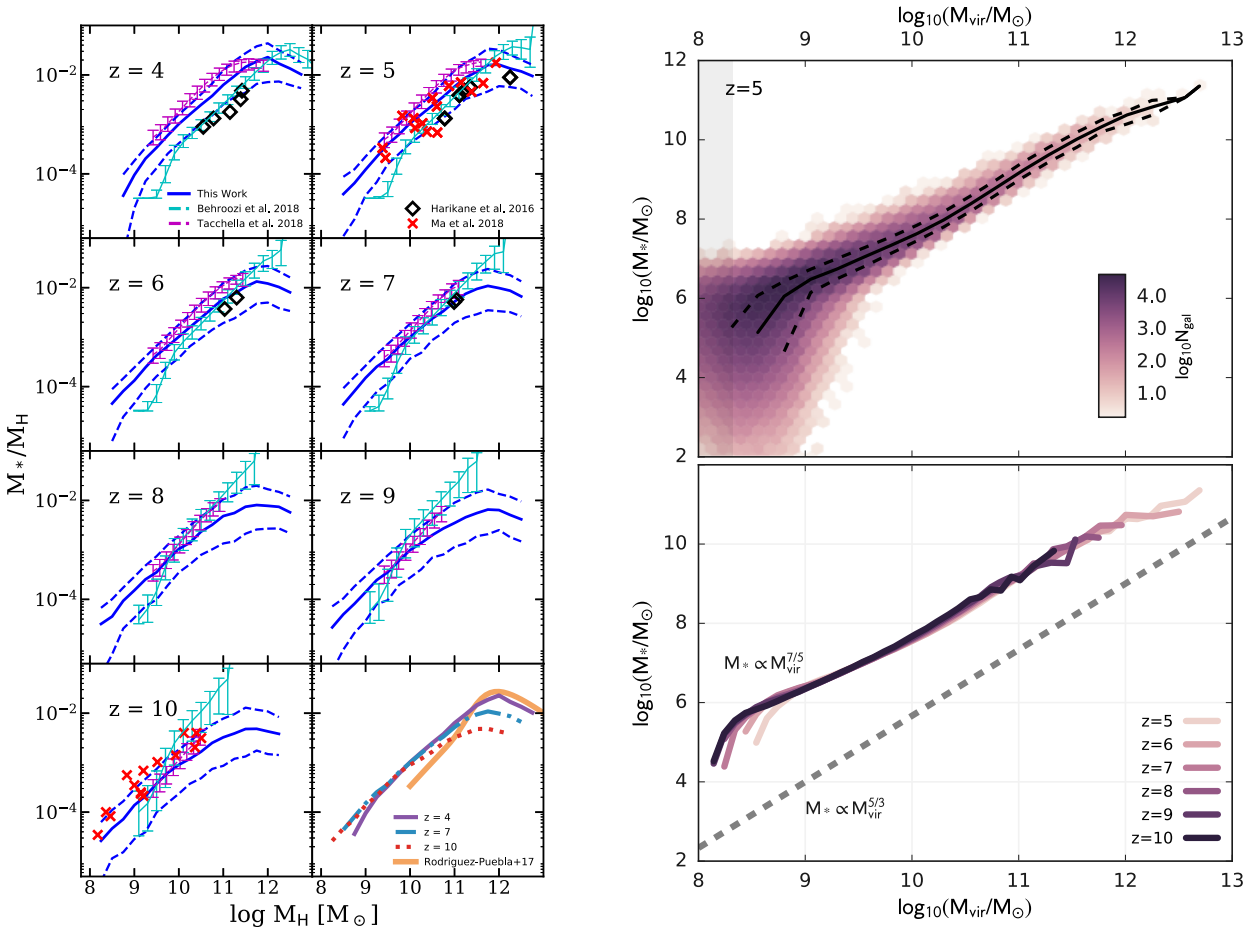

Fig. 23.— Estimates of f_* for high-redshift galaxies. The left panel shows a compilation of the stellar mass to halo mass relations for high redshift galaxies, taken from Yung et al. (2019). These include SAMs (Yung et al. 2019), semi-empirical scalings (Tacchella et al. 2018; Behroozi et al. 2019), hydro simulations (Ma et al. 2018), observational estimates (Harikane et al. 2019), and abundance matching estimates (Rodríguez-Puebla et al. 2016). The right panel corresponds to the stellar to halo mass relation from the SAM of Mutch et al. (2016). Most SAMs predict a remarkably constant power law index for this scaling for the dominant galaxies residing in small mass halos. This is consistent with the galaxy main sequence seen at lower redshifts.

Myr. This was realized early in the development of cosmological simulations, and dubbed the "over-cooling problem": i.e. if you treat SNe feedback just by increasing the gas temperature in hydrodynamic simulations, you generally end up with inefficient feedback and overproduce the total number of stars at low redshifts. The details of energy transfer are very complicated, requiring highly-specialized numerical simulations. Yet somehow, this energy should drive winds which regulate the SFR. We can parametrize this in terms of an efficiency, $\epsilon_{\rm wind}$, which is the fraction of energy released which couples as kinetic energy in the ISM. Using this efficiency, we can estimate the mass of the ISM which gets ejected out of the halo, $m_{\rm eject}$, with:

$$\Delta E_{\rm SNe}\epsilon_{\rm wind} > \frac{GM_h}{R_{\rm vir}} m_{\rm eject} \tag{206}$$

We can now ask what is the maximum allowed burst which would evacuate the halo, depriving it of gas. In other words, if we set $m_{\rm eject} = M_g \approx M_h f_b \Omega_b/\Omega_m$, we have:

$$\Delta E_{\rm SNe}\epsilon_{\rm wind} \approx \left(\frac{GM_h^2}{R_{\rm vir}}\right)\left(\frac{\Omega_b}{\Omega_m} f_b\right) \tag{207}$$

We can identify the first term on the RHS as twice the binding energy of the halo, (c.f. Barkana & Loeb

2001):

$$\frac{GM_h^2}{R_{\rm vir}} \approx 10^{54}{\rm erg}\left(\frac{M_h}{10^8 M_\odot}\right)^{5/3}\left(\frac{1+z}{10}\right) \ . \tag{208}$$

Substituting into eq. (207), we have:

$$7\times 10^{53}{\rm erg}\left(\frac{\Delta M_*}{10^5 M_\odot}\right)\epsilon_{\rm wind} \approx 10^{54}{\rm erg}\left(\frac{M_h}{10^8 M_\odot}\right)^{5/3}\left(\frac{1+z}{10}\right)\left(\frac{\Omega_b}{\Omega_m}f_b\right) \ . \tag{209}$$

Let's now return to the stellar fraction, f_*. For a burst of star formation we have, $f_* = \Delta M_*/M_g \sim \Delta M_*(M_h f_b \Omega_b/\Omega_m)^{-1}$. If we now define $f_{*,\rm max}$ to be the maximum value of the star fraction so as not to completely evacuate all of the gas from the halo, we can use ΔM_* from eq. (209) to derive:

$$\boxed{f_{*,\rm max} \approx 10^{-3}\epsilon_{\rm wind}^{-1}\left(\frac{M_h}{10^8 M_\odot}\right)^{2/3}\left(\frac{1+z}{10}\right) \ .} \tag{210}$$

From the above, we see explicitly that even if a tiny fraction ($\epsilon_{\rm wind} \gtrsim 0.001$) of the SNe energy is able to couple dynamically to the ISM, the gas will be evacuated. The resulting dearth of cool gas inside the halo would suppress future star formation. The above result also implies that if star formation is maximally efficient, limited only by SNe feedback, we expect to find the stellar fraction to scale with the halo mass as $f_* \sim f_{*,\rm max} \propto M_h^{2/3}$, and analogously the total stellar mass to scale as $M_* \propto M_h f_* \propto M_h^{5/3}$. These simple scalings are roughly consistent with high-redshift observations and more sophisticated models (see e.g. Fig. 23).

4.6. Empirical trends of galaxy formation

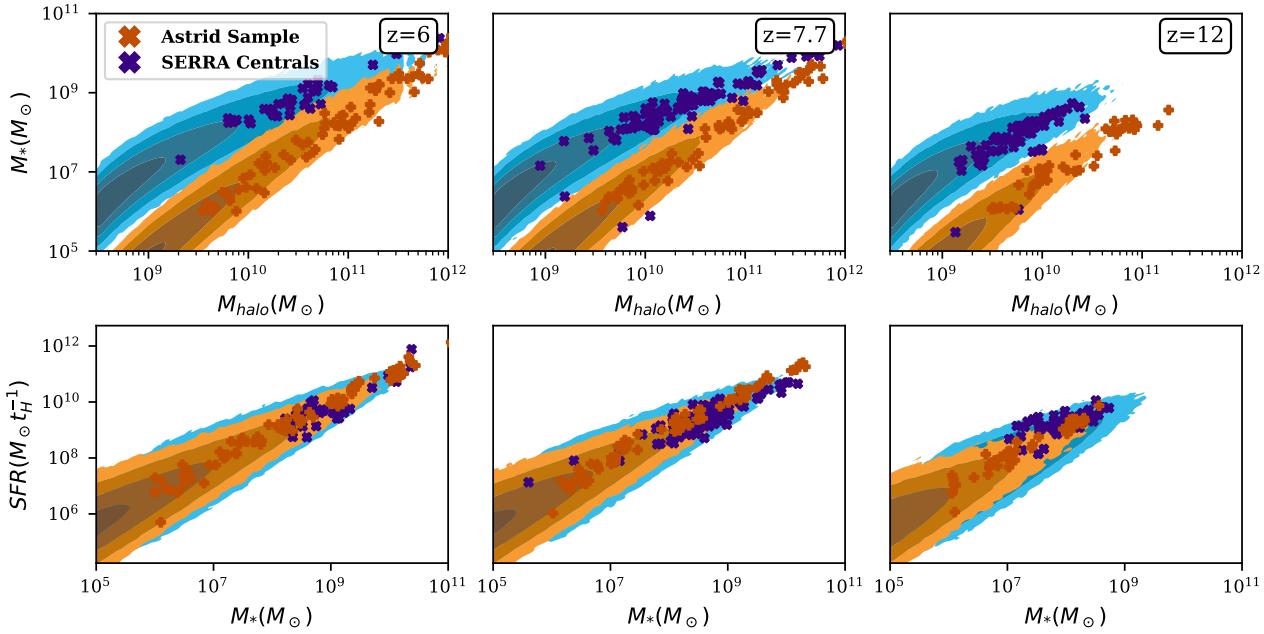

Fig. 24.— Distributions of stellar to halo mass (*top panels*) and star formation rate to stellar mass (*bottom panels*) at $z = (6.0, 7.7, 12)$. Orange and blue points correspond to galaxies taken from two hydrodynamical simulations: the zoom-in simulation suite SERRA (Pallottini et al. 2022) and the large cosmological simulation Astrid (Bird et al. 2022) For the SERRA sample every central galaxy is plotted, while for Astrid a subsample randomly selected in fixed logarithmic mass bins is plotted. Both codes have been calibrated to reproduce observable data at $M_h \gtrsim 10^{12} M_\odot$, but predict very different SHMRs for the unseen, faint galaxies that dominate during the first billion years. The orange and blue contours correspond to two different parameter combinations in `21cmFASTv4` whose conditional distributions can characterize the correspondingly-coloured galaxy populations from the different hydro codes. These contours correspond to 2–5 σ of the *joint* distributions, $P(M_*, M_h)$ and $P(SFR, M_*)$, highlighting how the vast majority of galaxies are expected to be far below the resolution limits for large-scale cosmological simulations. This figure is taken from Davies et al. in prep.

At low redshifts where there is more data (and complexity), popular SAMs (e.g. De Lucia & Blaizot 2007; Bower et al. 2006; Croton 2009) have dozens of free parameters, and thus have become rather opaque in terms of physical insight. Simple scalings as derived in the previous sections, while oversimplifications, do give physical insight and might be more relevant at higher redshifts where structure formation is simpler.

Another alternative to SAMs are empirical scaling relations. These lack direct physical insight, but allow us to characterize observations in useful parametric forms. Like for SAMs, our goal is to connect galaxy properties to the mass of the host dark matter halo (whose number densities and spatial distribution we understand much better).

One approach is to use observations and simulations to motivate *conditional probability distributions* that relate various galaxy properties. There are well-known relations that emerge from observations and simulations, such as the stellar-to-halo mass relation (SHMR) and the star forming main sequence (SFMS). These can motivate the corresponding conditional probability distributions, $P(M_* \mid M_h)$, and $P(\mathrm{SFR}|M_*)$. For a given property, x, such conditional distributionscan be characterized by log-normal functional forms:

$$P(\ln(x)) = \mathcal{N}(\mu_{\ln x}, \sigma_{\ln x}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\ln(x) - \mu_{\ln x}}{\sigma_{\ln x}}\right)^2\right] \tag{211}$$

where the mean and sigma of the distribution can be motivated by observations/simulations, and then

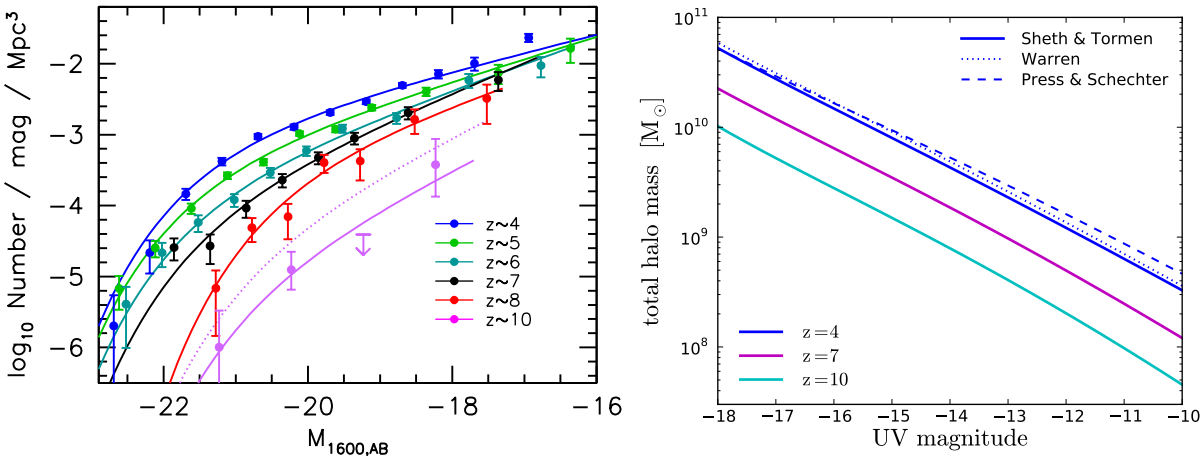

Fig. 25.— *Left:* Observed high-z luminosity functions (from Bouwens et al 2015). *Right:* luminosity to halo mass relation, assuming a constant duty cycle of unity (from Kuhlen & Faucher-Giguere 2012). *The vast majority of galaxies are unobserved!*

extrapolated to lower masses or higher redshifts, where we do not have data (see Fig. 24). Alternatively, we can *infer* the mean and sigma from intergalactic medium data, as we will discuss further below.

The most robust observations of high-redshift galaxies are luminosity functions: the number density of galaxies in a given UV magnitude bin, usually centered around 1500 – 1600 Å (see the left panel of Fig. 25 taken from Bouwens et al. 2015).[20] These (non-ionizing) UV luminosity functions (LFs) can be tied to the halo masses which host the observed galaxies (the theoretical quantity which is most robust) with:

$$\frac{dn(>M_{\rm uv},z)}{dM_{\rm uv}} = \int_0^\infty \frac{dn(>M_h,z)}{dM_h} P(>M_{\rm uv}|M_h) dM_h \ . \tag{212}$$

Here, $P(>M_{\rm uv}|M_h)dM_{\rm uv}$ is the conditional probability that a galaxy hosted by a halo with mass M_h has a UV luminosity between $M_{\rm uv}$ and $M_{\rm uv}+dM_{\rm uv}$. If we ignore the spread in $P(>M_{\rm uv}|M_h)$, and assume it is monotonically increasing (i.e. brighter galaxies must reside in more massive halos), we can then relate $M_{\rm uv} \leftrightarrow M_h$ by matching the abundances of galaxies and halos (so-called "abundance matching"; Vale & Ostriker 2006):

$$\int_{-\infty}^{M_{\rm uv}} \frac{dn(>M'_{\rm uv},z)}{dM'_{\rm uv}} dM'_{\rm uv} = \int_\infty^{M_h} \frac{dn(>M'_h,z)}{dM'_h} f_{\rm duty}(M'_h) dM'_h \ , \tag{213}$$

where we have introduced a duty cycle, $f_{\rm duty}(M_h)$, signifying what fraction of halos of mass M_h are actively star-forming at z and are thus observable. The right panel of Fig. 25 shows this $M_{\rm uv} \leftrightarrow M_h$ relation, obtained by further simplifying $f_{\rm duty}(M_h)=1$ (taken from Kuhlen & Faucher-Giguere 2012).

When using any empirical relations fit to data, it is also important not to neglect *galaxy to galaxy scatter* in their properties. Empirical fits usually fit a line in log-log space, which can cause biases in various quantities if stochasticity is ignored (e.g. the mean of a quantity increases if a log-normal distribution is widened; **??**). For example, when evaluating a mean global quantity like the star formation rate density, SFRD, using the SHMR and SFMS, one could write the following equation if these relations are deterministic:

$$\mathrm{SFRD}(>M_{\rm uv},z) = \int_{M_h(M_{\rm uv})}^\infty dM_h \frac{dn(>M_h,z)}{dM_h} \ \mathrm{SFR}[M_*(M_h),z] \tag{214}$$

[20] It is important to keep in mind that these are galaxy *candidates*. They are usually selected using the Lyman break technique (e.g. Steidel et al. 1996; Madau et al. 1996) with broad-band photometry. Spectroscopic follow-up is required to clearly identify these candidates from low-redshift interlopers having similar broad-band colors.

where the SFR dependencies on halo and stellar masses are taken from empirical fits. However, since these relations all have scatter, we should actually write the following:

$$\mathrm{SFRD}(>M_{\rm uv},z)\int_{M_h(M_{\rm uv})}^{\infty}dM_h\frac{dn(>M_h,z)}{dM_h}\times\int dM_*\ p(M_*|M_h)\times\int d\mathrm{SFR}\ p(\mathrm{SFR}|M_*,z) \tag{215}$$

We end this section by highlighting an important result seen in Figures 25 and 24: *current observations of high-redshift galaxies likely only capture the brightest few.* The LFs are very steeply increasing going down towards the faintest observed magnitudes. And from theoretical considerations, we would expect to have star-formation occurring in halos with masses as low as the atomic cooling threshold: $\sim 10^8 M_\odot$ at $z\sim 10$. The simple abundance matching method shown in the right panel suggests that these would have $M_{\rm uv}\sim -10$, orders of magnitude fainter than accessible even with infrared telescopes like the *James Webb Space Telescope (JWST)*. We are therefore unlikely to directly see the bulk of the galaxy population at the highest redshifts. However, their radiation ionizes and heats the intergalactic medium (IGM). This makes the IGM (see next section) a democratic, though indirect, probe of the combined galaxy population.

4.7. Radiative Transfer

Thus far, we have discussed radiation in the context of gas cooling; without cooling radiation, we would not have dense structures like galaxies, stars, chickens... However radiation also impacts the surrounding environment by heating and/or ionizing gas. Moreover, almost everything we know about the Universe is obtained by observing and analyzing the light we receive. Here we discuss the fundamentals of how radiation passes through a medium, so-called radiative transfer (RT).

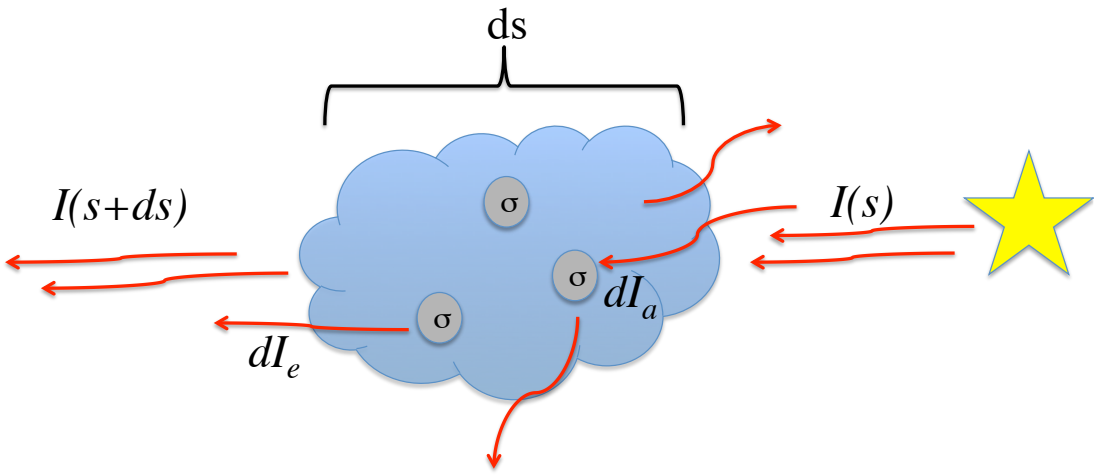

Fig. 26.— Schematic of radiative transfer through a gas cloud. The incoming radiation is affected by emission, absorption, and scattering.

The basic set-up is shown in Fig. 26. Radiation of specific intensity, I (in erg s^{-1} cm^{-2} sr^{-1} Hz^{-1}), from some background source passes through a path length of ds through gas with given properties, and is affected by three processes: (i) emission, (ii) absorption, and (iii) scattering.

1. *Emission* - assuming the emission is isotropic, the increase in intensity contributed by the gas can be characterized as
$$dI_e = \frac{\epsilon n}{4\pi} ds \equiv j\ ds\ . \tag{216}$$
Here ϵ is the specific emissivity of the gas (in erg baryon^{-1} s^{-1} Hz^{-1}), n is its number density, and j is the so-called emission coefficient. For any $2 \rightarrow 1$ atomic transition, the emission coefficient can be expressed as $4\pi j = h\nu_{21} n_2 A_{21} \phi(\nu)$, where $h\nu_{21}$ is the energy corresponding to the transition, A_{21} is the associated Einstein coefficient denoting the probability per unit time for a spontaneous $2 \rightarrow 1$ emission, and $\phi(\nu)$ is the corresponding line profile which is peaked at ν_{21} (see the left panel of Fig. 32 and the associated discussion).

2. *Absorption* - if each particle has an absorption cross section of $\sigma(\nu)$, then the total absorbing area of an infinately thin slice through the gas cloud of dimensions $\sim ds \times dA$ can be expressed as $n\ \sigma\ ds\ dA$. The absorbing area per unit area (i.e. the fraction subtended by absorbers) is then $n\ \sigma\ ds$. Thus, the decrease in intensity absorbed by the gas is
$$-dI_a = n\sigma I ds \equiv \alpha_{\rm abs} I\ ds\ . \tag{217}$$
Here $\alpha_{\rm abs} = n\sigma$ is the absorption coefficient, which we can again express in terms of the Einstein coefficient B_{12} for a $1 \rightarrow 2$ line transition, $4\pi\alpha_{\rm abs} = h\nu_{21} n_1 B_{12} \phi(\nu)\{1 - \exp[-h\nu/(kT)]\}$, where the $\exp[-h\nu/(kT)]$ term accounts for the likelihood of stimulated $2 \rightarrow 1$ emission.

3. *Scattering* - For most cosmological purposes, scattering out of the line of sight can be treated exactly like absorption following eq. (217), as the the light which gets scattered back into the line of sight is

negligible. In some cases it might be useful to keep track of scattered radiation, in order to estimate the surface brightness or polarization profiles of galaxies, for instruments that are sensitive to detect these properties. Radiative transfer of scattered radiation can be done with Monte Carlo methods which statistically sample the the distance, direction, and polarization from a scattering event (for more details, see e.g. Lee et al. 1994; Loeb & Rybicki 1999; Gronke & Dijkstra 2014).

Combining the above, we can write the differential change in intensity of light passing through gas as:

$$\frac{dI}{ds} = -\alpha_{\rm abs} I + j \tag{218}$$

It is useful to define an *optical depth*, $\tau = \int \alpha_{\rm abs} ds$, so $d\tau = \alpha_{\rm abs} ds = n\sigma ds$. Using this change of variables, we can write the radiative transfer equation above as:

$$\frac{dI}{d\tau} = -I + S \ , \tag{219}$$

where $S \equiv j/\alpha_{\rm abs}$ is the so called source function. To solve eq. (219), we can apply a change of variables. Defining $\mathcal{I} \equiv I \exp(\tau)$ and $\mathcal{S} \equiv S \exp(\tau)$, equation 219 becomes

$$\begin{aligned} \frac{d(\mathcal{I}e^{-\tau})}{d\tau} &= -\mathcal{I}e^{-\tau} + \mathcal{S}e^{-\tau} \\ \frac{d\mathcal{I}}{d\tau}\cancel{e^{-\tau}} - \cancel{\mathcal{I}e^{-\tau}} &= -\cancel{\mathcal{I}e^{-\tau}} + \cancel{\mathcal{S}e^{-\tau}} \\ \frac{d\mathcal{I}}{d\tau} &= \mathcal{S} \ , \end{aligned} \tag{220}$$

and integrating we obtain

$$\mathcal{I}(\tau) = \mathcal{I}(0) + \int_0^\tau \mathcal{S}(\tau')d\tau' \ . \tag{221}$$

Going back to our original physical variables,

$$I(\tau)e^{\tau} = I(0) + \int_0^\tau S e^{\tau'} d\tau', \tag{222}$$

and dividing through by $\exp(\tau)$ we obtain the standard equation of 1D radiative transfer,

$$\boxed{I(\tau) = e^{-\tau}\left[I(0) + \int_0^\tau S e^{\tau'} d\tau'\right]}. \tag{223}$$

We see from the above equation that the optical depth corresponds to an e-folding of the absorption. In other words, in the absence of emission, radiation passing through gas will be attenuated by a factor of e, after one optical depth.

If we can take the gas properties to be constant over the length of interest, S is a constant which can be removed from the integral above. Then eq. (223) becomes:

$$I(\tau) = I(0)e^{-\tau} + S\left(1 - e^{-\tau}\right) \tag{224}$$

$$= S + e^{-\tau}\left[I(0) - S\right] \ . \tag{225}$$

From the above, we can clearly see the asymptotic trends that for an optically-thick medium, with $\tau \to \infty$, we have the intensity approaching the appropriately-named source function, $I \to S$. Similarly, for an optically-thin medium, with $\tau \to 0$, the intensity remains unchanged from the incoming background radiation, $I \to I(0)$.

5. The Intergalactic Medium

Thus far in our study of baryons, we have focused on those residing inside dark matter halos, i.e. galaxies. One can argue that they have the most interesting fates. However, the fraction of baryons which reside in galaxies is actually very small: atomic cooling halos host at most a few percent of the baryons at $z \gtrsim 6$. The vast majority of matter lies in the diffuse web stretching between the galaxies, the so-called intergalactic medium (IGM).

The IGM can be characterized by the following fundamental properties: (i) density; (ii) ionization state; (iii) temperature. We discuss each of these in the following sections.

5.1. Ionization evolution: the Epoch of Reionization (EoR)

The Epoch of Reionization (EoR) is the last major phase change of the IGM. Light from the first stars and galaxies, discussed in the previous section, spread out throughout the Universe, ionizing and heating the IGM. It is a complex process, encoding the physics of the first structures and how they impacted their surroundings. It is challenging also to model, as the epoch involves a huge range of scales, with the small-scale physics of star formation driving ionization structures which are inhomogeneous on cosmological scales. Here we will review a basic analytic framework, and encourage readers to delve deeper in the field with reviews such as Mesinger (2016).

Let's begin with an early, star-forming galaxy surrounded by the neutral IGM. Ionizing radiation from its stars can escape the galaxy into the IGM, driving a local, expanding HII region[21] with comoving volume, $V_{\rm HII}$. The evolution of this HII region can be written as:

$$\langle n_{\rm H}\rangle \frac{dV_{\rm HII}}{dt} = \frac{dN_\gamma}{dt} - \alpha_{\rm HII}\langle n_{\rm H}^2\rangle V_{\rm HII} a^{-3} \ . \tag{226}$$

Here the LHS is the rate at which new HI is ionized as the HII region expands. The first term on the RHS corresponds to the rate at which ionizing photons are escaping the galaxy into the IGM, while the second term corresponds to the number of recombinations per time inside the HII region (note the final a^{-3} term converts the recombination coefficient, $\alpha_{\rm HII}$, [22] to comoving units). For the cosmic HII region

[21] Note that the width of the ionization fronts roughly correspond to the mean free path of the typical ionizing photons. For any UV source, this mean free path in the IGM is very small, of order $\sim$ kpc. Therefore the EoR is an inhomogeneous process with almost fully ionized HII regions around the first galaxies expanding into almost fully neutral HI regions. Here we assume a completely bimodal IGM: either fully neutral or fully ionized. Therefore, we have no ionized fraction terms in eq. (226). In §5.3.1, we shall relax this assumption, which primarily impacts the recombination rate inside the cosmic HII regions.

[22] The recombination coefficient for a given species (hydrogen or helium) is usually written as being either "case A", $\alpha_{\rm A}$, or "case B", $\alpha_{\rm B}$. The case A coefficient includes the sum of probabilities of a recombination to *any* state (including directly to the ground state), while the case B excludes recombinations directly to the ground state (which result in the emission of an ionizing photon). For hydrogen at a temperature of 10^4 K, we have $\alpha_{\rm A} = 4.2 \times 10^{-13}$ cm^3 s^{-1}, and $\alpha_{\rm B} = 2.6 \times 10^{-13}$ cm^3 s^{-1} (e.g. Osterbrock 1989). When computing the ionization balance of the IGM, it is more appropriate to use case A if the recombinations are taking place in optically-thick systems (at low redshifts referred to as Lyman limit systems; LSSs). The reasoning behind this is that the photons resulting from ground state recombinations are likely to be absorbed locally, inside the LLS). After some number of ionizations/absorptions, the recombination happens into an excited state, and there is no more ionizing photon. Thus the ionizing photons resulting from ground state recombinations do not escape the LLS, and so do not contribute to the ionization balance in the diffuse IGM (Miralda-Escudé 2003). The case B recombination coefficient is more appropriate when recombinations are happening in more diffuse, optically thin systems. In this case, the ground state photon can travel in the IGM, and result in another IGM ionization. As a result this photon is "ionization neutral" when computing the IGM ionization state, and so is not counted in the rate equations. While it is clear that for the post-reionization IGM the case A is more appropriate, it is really not clear what is better at high redshifts, as it depends on knowing the properties of the systems which are dominating the recombinations (are they occurring mostly in dense systems or in the actual diffuse IGM which one is modeling). In the next chapter, we develop the framework for studying recombining systems, but current uncertainties in the strength of the ionizing background prevent us from knowing which recombination coefficient is more appropriate. In this chapter therefore, we use a general notation, $\alpha_{\rm HII}$, to indicate that the appropriate coefficient is somewhere between case A and case B.

to grow, the emission rate of ionizing photons has to be larger than the recombination rate.

We can expand the emission term as a product of the following:

$$N_\gamma = f_{\rm esc} N_{\gamma/b} f_* N_b^{\rm halo} \tag{227}$$

Here the number of baryons in the galaxy is $N_b^{\rm halo}$, f_* is the fraction of those baryons inside stars (c.f. §4.5.3), $N_{\gamma/b}$ is the number of ionizing photons produced per stellar baryon, and $f_{\rm esc}$ the fraction of these ionizing photons which manage to escape into the IGM.

For the absorption term, it is convenient to define a *clumping factor*, $C \equiv \langle n_{\rm H}^2 \rangle / \langle n_{\rm H} \rangle^2$. The clumping factor is a measure of substructure, and should *only be computed inside the ionized regions* which contribute to recombinations. With these definitions, we can rewrite eq. (226) as:

$$\frac{dV_{\rm HII}}{dt} = \frac{1}{\langle n_{\rm H} \rangle} \frac{d[f_{\rm esc} N_{\gamma/b} f_* N_b^{\rm halo}]}{dt} - \alpha_{\rm HII} \langle n_{\rm H} \rangle C V_{\rm HII} a^{-3} \ . \tag{228}$$

To simplify this further, we can define the terms that depend on ISM properties ($f_{\rm esc}$, $N_{\gamma/b}$, f_*) in terms of their steady-state or ensemble-averaged values (see the discussions in the previous chapter), keeping only the time derivative of the the growth of the galaxy ($N_b^{\rm halo}$) in the first term on the RHS. Then if we divide by some "total" (large enough to be representative) volume $V_{\rm tot}$, we obtain the evolution of the filling factor (fraction of total volume) of this particular cosmic HII region:

$$V_{\rm tot}^{-1} \frac{dV_{\rm HII}}{dt} = \frac{f_{\rm esc} N_{\gamma/b} f_*}{V_{\rm tot} \langle n_{\rm H} \rangle} \frac{dN_b^{\rm halo}}{dt} - \alpha_{\rm HII} \langle n_{\rm H} \rangle C \frac{V_{\rm HII}}{V_{\rm tot}} a^{-3} \ . \tag{229}$$

So far we discussed a single HII region. The Universe during the EoR contains many HII regions. We are now in the position to perform an ensemble average over various individual $V_{\rm HII}$. The total ionized volume, summing over all cosmic HII regions is $\sum_i V_{\rm HII}^i$. Analogously, the filling factor of HII regions is $Q_{\rm HII} \equiv V_{\rm tot}^{-1} \sum_i V_{\rm HII}^i$. Finally, we note that $[V_{\rm tot} n_{\rm H}]^{-1} \sum_i N_b^{\rm halo,i} = [N_{\rm tot}]^{-1} \sum_i N_b^{\rm halo,i}$ is the fraction of baryons inside star-forming galaxies. If we assume that star-forming galaxies are hosted by halos with masses above some critical threshold mass (set by cooling or feedback), $M_{\rm min}$, then the fraction of baryons inside star-forming galaxies is just the collapsed fraction, $f_{\rm coll}(> M_{\rm min})$ from §3.3. With this ensemble averaging, we arrive at the evolution of the HII filling factor:

$$\boxed{\frac{dQ_{\rm HII}}{dt} = f_{\rm esc} N_{\gamma/b} f_* \frac{df_{coll}(> M_{\rm min}, z)}{dt} - \alpha_{\rm HII} \langle n_{\rm H} \rangle C a^{-3} Q_{\rm HI}} \ . \tag{230}$$

Each of the parameters in the above equation is the subject of topical research.

- *The fraction of galactic baryons inside stars, f_*,* depends on the efficiency of star formation, as discussed in §4.5.3. Simple estimates which scale the halo mass function by a constant amount to fit the LF suggest values of order $f_* \sim$ per cent, for the bulk of the high-redshift galaxy population (e.g. Vale & Ostriker 2006; Dijkstra et al. 2014; Dayal et al. 2014; Mutch et al. 2016; Park et al. 2019).

- *The typical number of ionizing photons produced per stellar baryon, $N_{\gamma/b}$,* depends on the IMF of the stars. Population II stars should produce roughly 5000 ionizing photons over their lifetime, while more top-heavy IMFs or metal-free PopIII stars could increase this number by an order of magnitude (see e.g. Fig. 20 and Tumlinson & Shull 2000; Schaerer 2002).

- *The halo mass threshold for star-formation, $M_{\rm min}$,* depends on cooling efficiency or feedback, and can take on values ranging between $M_{\rm min} \sim 10^6 M_\odot$ for the first, molecularly-cooled halos (e.g. Bromm et al. 2002; Abel et al. 2002; Yoshida et al. 2008), $M_{\rm min} \sim 10^8 M_\odot$ for atomically-cooled halos. If feedback was efficient in quenching star formation in these small-mass halos the threshold could be as high as $M_{\rm min} \sim 10^{10} M_\odot$, corresponding the faintest high-redshift galaxies observed today (see Fig. 25).

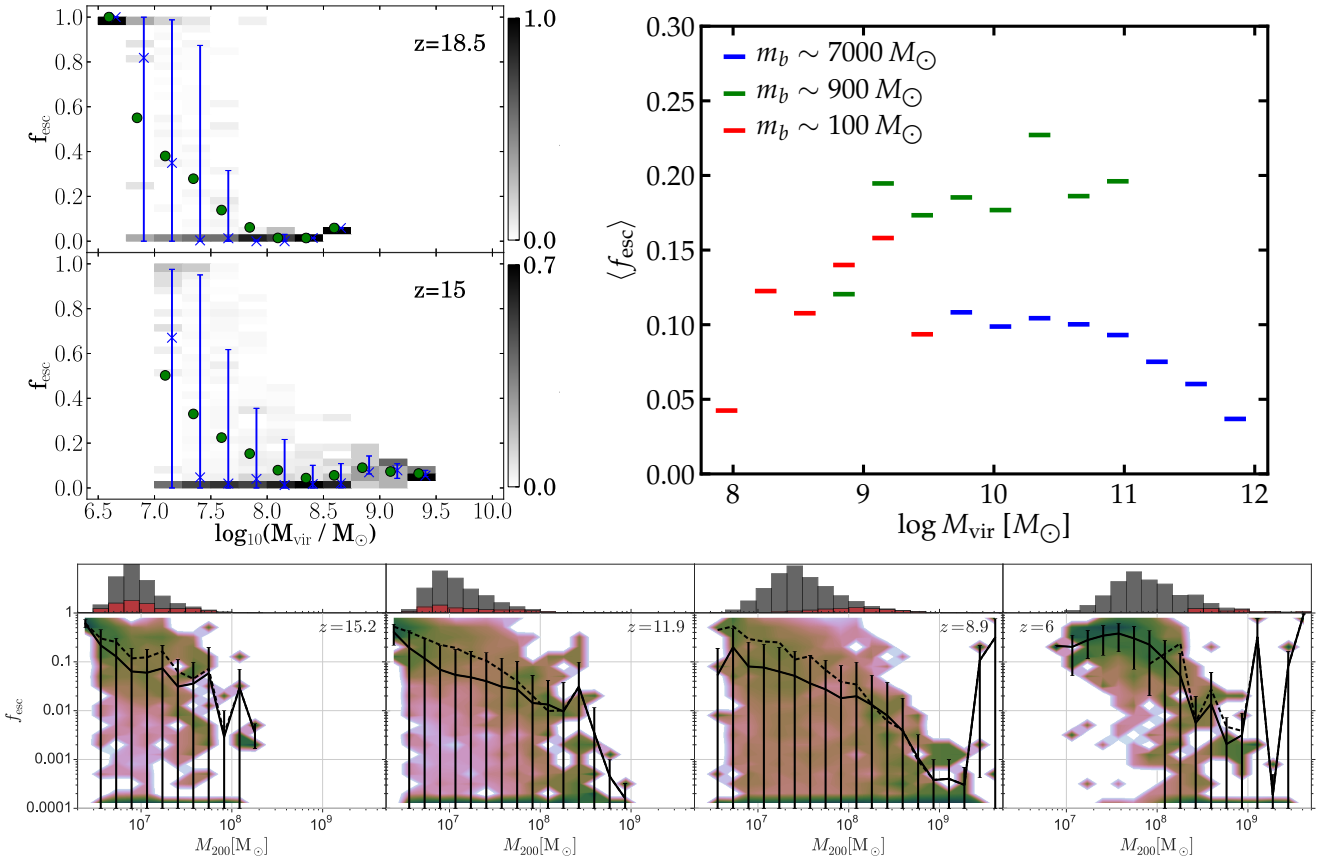

Fig. 27.— Predictions from hydrodynamic simulations of the dependence on f_{esc} on halo mass at high redshifts. Panels are taken from Xu et al. (2016), Ma et al. (2020), Paardekooper et al. (2015), clockwise from upper right. There is currently no consensus on the normalization and scaling of the escape fraction. Moreover, these simulations still do not resolve the density and velocity subscricture in birth clouds, and thus would require additional calibration.

- *The fraction of ionizing photons which escape the galaxy,* f_{esc}, depends on the galactic morphologies and the corresponding distribution of column densities. These in turn are likely set by a combination of dynamical and thermal evolution, with strong SNe feedback episodes likely clearing away the surrounding medium, facilitating the escape of ionizing photons. Direct observations of Lyman continuum emission are impossible at high redshifts given current technology. Stacks of Lyman break galaxies at lower redshifts, $z \sim 3$–4, motivate typical values of $f_{esc} \sim$ per cent (e.g. Steidel et al. 2001; Shapley et al. 2006; Siana et al. 2007; Marchi et al. 2017); however fainter galaxies at high redshifts are expected to have higher escape fractions as low column density sightlines are easier to be created by SNe explosions in shallower potential wells (e.g. Kimm & Cen 2014; Ferrara & Loeb 2013; Paardekooper et al. 2015; Xu et al. 2016; though see Ma et al. 2020). If M_{min} is much larger the atomic cooling threshold, so that only rare bright galaxies drive the EoR, we would need to have escape fractions of order tens of per cent to have the Universe reionize by $z \sim 5$–6 (e.g. Mitra et al. 2013; Kuhlen & Faucher-Giguere 2012; Robertson et al. 2013; Greig & Mesinger 2017). There is currently no consensus on even the qualitative trends for f_{esc} (see Fig. 27).

- *The clumping factor inside the ionized IGM, C,* is expected to be of order unity – few for the bulk of reionization, but could be much larger in the initial EoR stages if the ionized gas is heated and smoothed as its Jeans mass increases (e.g. Emberson et al. 2013; Pawlik et al. 2017), or rise rapidly in the later stages of the EoR as the ionization fronts penetrate into increasingly dense clumps thus allowing higher densities to contribute to the recombination rate (e.g. Furlanetto & Oh 2005; Sobacchi & Mesinger 2014; see §5.3.1).

We can simplify eq. (230) even further if we assume that these astrophysical parameters are redshift-independent. In this case, we can integrate over cosmic time:

$$Q_{\rm HII}(z) = f_{\rm esc} N_{\gamma/b} f_* \int_\infty^{z(t)} \frac{df_{\rm coll}(>M_{\rm min}, z)}{dt'} dt' - \int_\infty^{z(t)} \frac{dn_{\rm rec}}{dt'} dt' \tag{231}$$

$$= f_{\rm esc} N_{\gamma/b} f_* f_{\rm coll}(>M_{\rm min}, z) - n_{\rm rec}(z) \ , \tag{232}$$

where the number of recombinations per baryon is explicitly denoted as $n_{\rm rec}$. To first order, we can take the recombinations to be linearly distributed in $Q_{\rm HII}$ (i.e. assuming a weaker dependence on the other terms). This allows us to write:

$$Q_{\rm HII}(z) \approx f_{\rm esc} N_{\gamma/b} f_* f_{\rm coll}(>M_{\rm min}, z) - \bar{n}_{\rm rec} Q_{\rm HII}(z) \ , \tag{233}$$

where $\bar{n}_{\rm rec}$ is the total number of recombinations per baryon during the EoR. Finally, we have:

$$\boxed{Q_{\rm HII}(z) \approx \frac{f_{\rm esc} N_{\gamma/b} f_*}{(1+\bar{n}_{\rm rec})} f_{\rm coll}(>M_{\rm min}, z)} \tag{234}$$

As seen from eq. (234), the EoR only depends on the product of the aforementioned astrophysical quantities. Therefore, it is common to define this product as an "ionizing efficiency", ζ:

$$\zeta \equiv 20 \left(\frac{f_{\rm esc}}{0.1}\right) \left(\frac{f_*}{0.03}\right) \left(\frac{N_{\gamma/b}}{5000}\right) \left(\frac{1.5}{1+\bar{n}_{\rm rec}}\right) \tag{235}$$

with eq. (234) becoming simply:

$$\boxed{Q_{\rm HII} = \zeta f_{\rm coll}} \ . \tag{236}$$

5.1.1. Patchy Reionization

Finally, it is important to remember that reionization by UV photons is a very inhomogeneous process, with a fraction $\sim Q_{\rm HII}$ of the Universe virtually fully ionized, while the remaining $1 - Q_{\rm HII}$ is virtually fully neutral. The topology of this process thus tells us how the star-forming galaxies are spatially distributed (e.g. McQuinn et al. 2007). We can simulate this patchy reionization with large radiative transfer simulations; however the results are uncertain as we do not know the ionizing efficiencies of galaxies. Luckily, we can build some intuition analytically. As was noted by Furlanetto et al. (2004), we can use the same excursion-set tools we used to build the halo mass functions.

We can rephrase the global evolution in eq. (236), by realizing that each sub-region of the Universe is itself ionized if:

$$\zeta f_{\rm coll}(>M_{\rm min}, z|M_{\rm HII}, \delta_{\rm HII}) \geq 1 \ . \tag{237}$$

Here we have replaced the global collapsed fraction, with the conditional one: the fraction of matter inside collapsed structures above $M_{\rm min}$ at z, *given* that they reside in a large-scale region that has a matter overdensity $\delta_{\rm HII}$ on a scale $M_{\rm HII}$. We can express this conditional collapsed fraction as (§3.3):

$$f_{\rm coll}(>M_{\rm min}, z|M_{\rm HII}, \delta_{\rm HII}) = {\rm erfc}\left[\frac{\delta_{\rm crit}(z) - \delta_{\rm HII}}{\sqrt{2[\sigma^2(M_{\rm min}) - \sigma^2(M_{\rm HII})]}}\right] \ . \tag{238}$$

Plugging this into eq. (237), and inverting the complimentary error function:

$$\frac{\delta_{\rm crit}(z) - \delta_{\rm HII}}{\sqrt{2[\sigma^2(M_{\rm min}) - \sigma^2(M_{\rm HII})]}} \leq {\rm erfc}^{-1}(\zeta^{-1}) \ . \tag{239}$$

Therefore, we can stipulate that a region of scale $M_{\rm HII}$ is ionized at redshift z, if it has an overdensity of:

$$\boxed{\delta_{\rm HII} \geq \delta_{\rm crit}(z) - {\rm erfc}^{-1}(\zeta^{-1})\sqrt{2[\sigma^2(M_{\rm min}) - \sigma^2(M_{\rm HII})]}} \tag{240}$$

This overdensity is analogous to the "critical overdensity" for the collapse of dark matter halos. In §3.3, we constructed a halo mass function from the distribution of first upcrossings of the barrier $\delta_{\rm crit}$. Analogously, here we can construct the "HII region mass function" from the distribution of first upcrossings of the barrier $\delta_{\rm HII}$. This can be done either numerically, or analytically by linearizing the function in σ^2: $\delta_{\rm HII} \equiv B_0 + B_1\sigma^2(M)$. The constants B_0 and B_1 can be obtained by considering the asymptotic limit on large-scales, $\sigma^2(M_{\rm HII}) \to 0$. In this large-scale limit, the barier becomes $\delta_{\rm HII} \to B_0 = \delta_{\rm crit} - {\rm erfc}^{-1}(\zeta^{-1})\sqrt{2[\sigma^2(M_{\rm min}) - \cancelto{0}{\sigma^2(M_{\rm HII})}]} = \delta_{\rm crit} - \sqrt{2}\sigma(M_{\rm min}){\rm erfc}^{-1}(\zeta^{-1})$. Moreover the slope of the barrier becomes $\partial\delta_{\rm HII}/\partial\sigma^2 \to B_1 = {\rm erfc}^{-1}(\zeta^{-1})/\sqrt{2[\sigma^2(M_{\rm min}) - \cancelto{0}{\sigma^2(M_{\rm HII})}]} = {\rm erfc}^{-1}(\zeta^{-1})/\sqrt{2\sigma^2(M_{\rm min})}$. Having a linear barrier allows us to use the ellipsoidal functional form derivation of the halo mass function by Sheth et al. (2001). In analogy to the linear ellipsoidal barrier used for the ST mass functions, we can write the HII bubble mass fuction, i.e. the comoving number density of HII regions of mass scale $M_{\rm HII} \sim (4/3)\pi R_{\rm HII}^3\bar{\rho}$, as (Furlanetto et al. 2004):

$$\boxed{\frac{dn}{d\ln M_{\rm HII}} = \sqrt{\frac{2}{\pi}}\frac{\bar{\rho}}{M_{\rm HII}}\left|\frac{d\ln\sigma}{d\ln M_{\rm HII}}\right|\frac{B_0}{\sigma}\exp\left[\frac{(B_0+B_1\sigma^2)^2}{2\sigma^2}\right]} \tag{241}$$

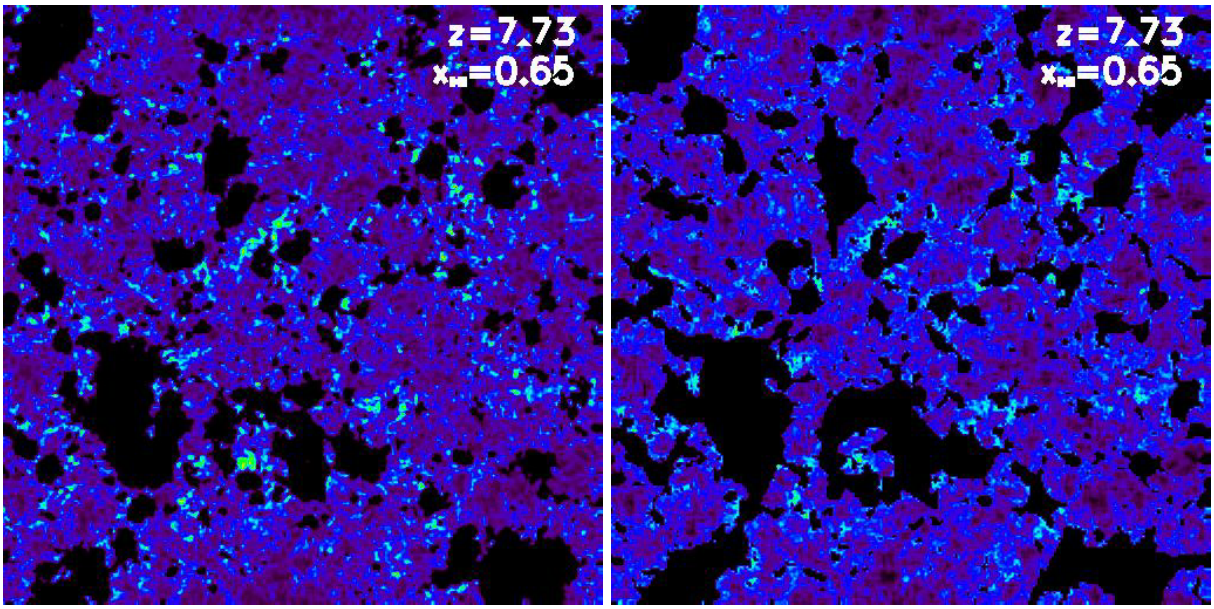

Fig. 28.— Slices through a simulated 21-cm signal during the EoR, with black corresponding to cosmic ionized patches (from Mesinger et al. 2011). The left panel was generated from a hydrodynamic radiative-transfer simulation (with a computation time of $\sim 10^7$ CPU hours), while the right panel was generated using an analytic excursion-set procedure applied to density fields which were evolved with the ZA (with a computation time of ~ 0.1 CPU hours). Both share the same initial conditions. All slices are 143 Mpc on a side and 0.56 Mpc thick.

In addition to the analytic "HII mass function", the excursion-set approach discussed above has been applied directly to 3D realizations. This is computationally very efficient, since smoothing the 3D density field to obtain $f_{\rm coll}(> M_{\rm min}, z|M_{\rm HII}, \delta_{\rm HII})$ just involves doing an FFT on the scale $M_{\rm HII}$. Starting from some maximum scale corresponding to a horizon for ionizing photons, the criterion from eq. (237) is evaluated at each cell of the simulation. Cells which reside in sufficiently large overdensities smoothed on that scale are marked as ionized. Then the smoothing scale is decreased and the procedure is iterated.

Ionization fields obtained with this procedure are in a good agreement with computationally-intensive radiative transfer methods, on moderate to large scales ($\gtrsim 1$ Mpc; e.g. Zahn et al. 2011; see also Fig. 28). The conditional collapsed fraction from eq. (237) can be computed using (i) the halo field directly from N-body simulations (Zahn et al. 2007); (ii) the halo field from perturbation theory (Mesinger & Furlanetto 2007); (iii) the evolved density field (Mesinger et al. 2011). The later, although a little more

approximate, has the advantage of facilitating a nearly unlimited dynamical range. This is important when modeling the signal on very large scales, such as is required for 21-cm observations (see §5.5).

5.2. Current EoR probes

Our current knowledge about the EoR stems from two classes of probes: (i) integral constraints from the CMB in the form of the Thompson scattering optical depth to the lass scattering surface (LSS)[23]; and (ii) astrophysical "flashlights" which illuminate the intervening IGM. We briefly discuss each in turn.

5.2.1. *Optical depth to the CMB*

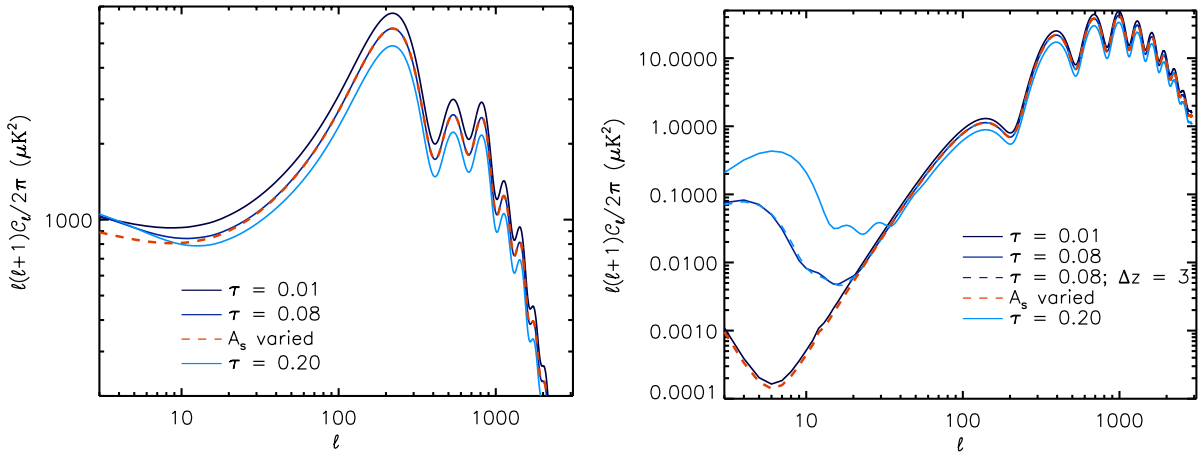

Fig. 29.— CMB temperature (*left*) and E-mode (zero curl) polarization (*right*) power spectra, for several different values of the mean Thompson scattering optical depth, τ_e. Increasing τ_e dampens the temperature fluctuations; however, this effect is strongly degenerate with reducing the primordial amplitude, A_s. Although a far weaker signal, the large-scale polarization fluctuations do not suffer from this degeneracy. These figures are taken from Reichardt (2016).

As light from the last scattering surface (LSS; i.e. the CMB) passes through the Universe, it interacts with free electrons through Thompson scattering. Thompson scattering is gray scattering, thus the dominant effect is to dampen the CMB temperature fluctuations, as light from hot spots gets scatter into lines of sight towards cold spots, and visa versa. The more free electrons (corresponding to an earlier EoR), the stronger is the distortion of the primordial CMB.

This imprint of the EoR can be characterized through the mean Thompson scattering optical depth,

$$\tau_e = \langle \int_0^{z_{\rm LSS}} n_e \sigma_T \left| \frac{cdt}{dz} \right| dz \rangle_{\rm LOS} \tag{242}$$

Here, n_e is the electron number density, σ_T the Thompson scattering cross-section, $c\ dt$ the line element to the LSS, and the averaging is performed over all lines of sight (LOSs). Thus the higher the τ_e, the more the CMB temperature fluctuations are damped (see the left panel of Fig. 29). This damping is easy to detect. Unfortunately, it is also strongly degenerate with the primordial power spectrum amplitude, A_s, as shown in the left panel of Fig. 29.

Luckily, the CMB has a large-scale quadrupole anisotropy. This means the EoR creates a linear polarization signal in the CMB, which peaks on scales larger than the horizon during the EoR. Unlike for

[23] Alternative probes such as E-mode polarization as a function of angular scale (e.g. Mortonson & Hu 2008), the patchiness of τ_e (e.g. Dvorkin & Smith 2009), the kinetic Sunyaev-Zel'dovich signal from patchy reionization (e.g. Mesinger et al. 2012), could yield interesting results in the future provided systematics can be controlled (see the review of Reichardt 2016).

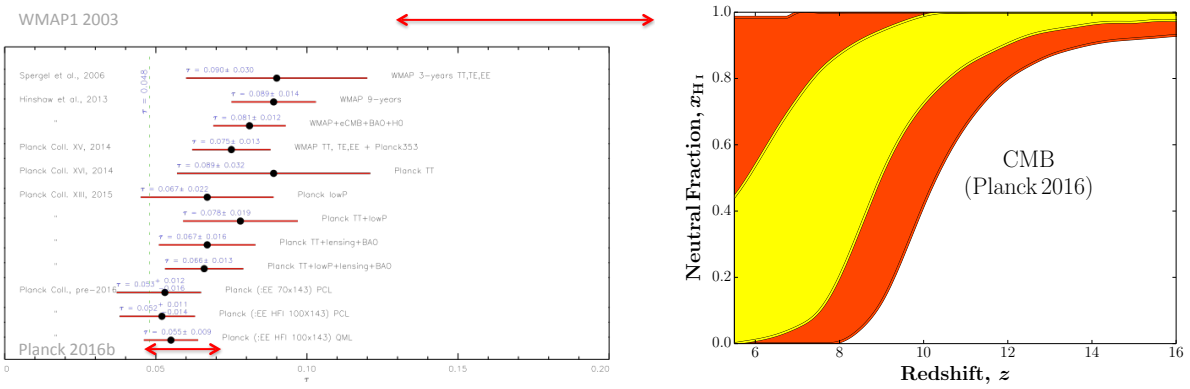

Fig. 30.— *Left*: Historical trend of the 1 σ constraint on the mean Thompson scattering optical depth to the CMB, τ_e. Figure is adapted from Planck Collaboration XLVI (2016), with the addition of the 1-yr *WMAP* result of $\tau_e = 0.17 \pm 0.04$ using only the temperature power spectrum at the top (Kogut et al. 2003) and the alternate HFI estimate from Planck Collaboration XLVII (2016) $\tau_e = 0.058 \pm 0.012$ at the bottom. *Right*: Constraints on the evolution of the average neutral fraction, $\bar{x}_{\rm HI} = 1 - Q_{\rm HII}$, corresponding to the latest *Planck* estimate of $\tau_e = 0.058 \pm 0.12$ (Planck Collaboration XLVII 2016). 68% C.L. are shown in yellow, while 95% C.L. are shown in red. $\bar{x}_{\rm HI}(z)$ was sampled from physically-motivated EoR models, based on eq. (234), with the optical depth used to compute a χ^2 likelihood. Taken from Greig & Mesinger (2017).

the temperature power spectra, the impact of τ_e on the polarization power spectra is *not* degenerate with cosmology (see the right panel of Fig. 29). Unfortunately, this signal is much weaker and more difficult to detect, compared to the temperature fluctuations.

In the left panel of Fig. 30, we show the historical trend of τ_e estimates. Starting with the *WMAP* satellite, the first estimate using only the temperature power-spectra was $\tau_e = 0.17 \pm 0.04$ (1σ) (Kogut et al. 2003). This unexpectedly-high optical depth implied there were *abundant* ionizing sources in the very Universe ($z > 15$), at a time when the furthest objects were at $z \sim 6$. The resulting implications on structure formation caused much excitement/confusion in the community.

However, in subsequent years the value of τ_e decreased, with the errors shrinking. This was driven mainly by the addition of polarization data, first through the temperature-polarization cross-power spectra and then through the detection of the polarization auto power spectra with the *Planck* satellite. The current (2017) conservative estimate is $\tau_e = 0.058 \pm 0.012$ (Planck Collaboration XLVII 2016), obtained using *Planck*'s high frequency instrument (HFI).

How does this constrain the reionization history? CMB anisotropies result from the cumulative contribution of free electrons to the last scattering surface: thus it is an integral measurement. Translating such an observation to a reionization history requires assuming a functional form for $Q_{\rm HII}(z)$ and then marginalizing over the parameters that regulate it. In the right panel of Fig. 30 we show the 1-σ (*yellow*) and 2-σ (*red*) constraints on the reionization history created by sampling EoR models based on eq. (234), using $\tau_e = 0.058 \pm 0.012$ to compute a χ^2 likelihood, and marginalizing over the free parameters in the model. We see that the mean reionization redshift implied by Planck Collaboration XLVII (2016) is $z = 7.64^{+1.34}_{-1.82}$. We caution however that the exact shape of these EoR history constraints are model-dependent, depending on the $Q_{\rm HII}(z)$ functionals and their corresponding priors (e.g. Mitra et al. 2015; Planck Collaboration XLVII 2016; Heinrich et al. 2017).

5.2.2. Lyα damping wing absorption

The Lyα line of hydrogen has emerged as a powerful probe of the EoR. To understand its utility, let's consider the schematic shown in Fig. 31. Sources during the EoR (galaxies and QSOs) emit an intrisic

Lyα flux (*bottom right panel*), whose profile is set by local and interstellar properties of the source. These photons emerge from the galaxy/QSO into some local patch of the IGM, which has already been ionized by the contribution from neighboring sources; the residual HI inside these local ionized patches (*top right panel*) is determined by the local density and ionizing radiation, as we shall see in the next section. The photons pass through the local HII region, redshifting along the way. Those which are not scattered out of the line of sight by the residual HI inside the local ionized patch then pass through the large-scale EoR topology of cosmic HI and HII regions (*left panel*), redshifting as they travel towards us.

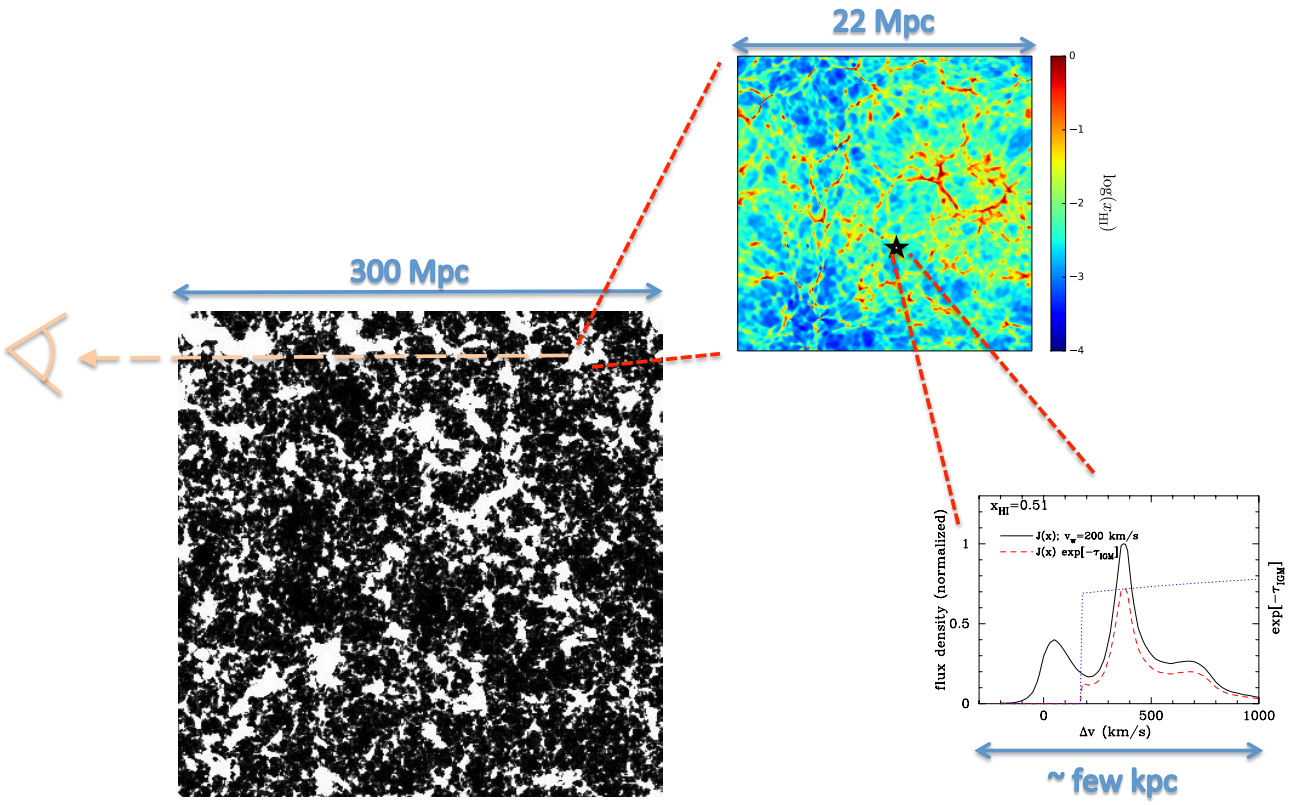

Fig. 31.— Schematic showing the various components determing the observed Lyα line from high redshift QSOs and galaxies druring the EoR. From left to right we show: (i) a 0.75 Mpc thick slice through large-scale reionization simulation at $Q_{\rm HII} \sim 0.5$ (Sobacchi & Mesinger 2014); (ii) a 21 kpc slice through hydro simulation of the ionized IGM surrounding high-z galaxies (Mesinger et al. 2015); (iii) the intrinsic Lyα line emerging from a galaxy including RT through local outflows (Dijkstra et al. 2011).

The observed flux at a wavelength, $\lambda_{\rm obs}$, for a source at redshift $z_{\rm s}$ can be expressed as:

$$F_{\rm obs}(\lambda_{\rm obs}) = F_0 \left(\frac{\lambda_{\rm obs}}{1+z}\right) \, e^{-\tau(\lambda_{\rm obs})} \, , \tag{243}$$

where F_0 is the intrinsic (i.e. emerging from the galaxy/QSO into the IGM) spectrum, evaluated at a rest frame wavelength $\lambda_{\rm obs}/(1+z)$, and the total IGM optical depth due to Lyα absorption, τ, is given by (neglecting peculiar velocities):

$$\tau(\lambda_{\rm obs}) = \int_0^{z_{\rm s}} dz \, \frac{c \, dt}{dz} \, n_H \, x_{\rm HI} \, \sigma \tag{244}$$

where $c(dt/dz)$ is the proper line element in a given cosmology, $n_H(z)$ is the hydrogen number density at redshift z, $x_H(z)$ is the hydrogen neutral fraction at redshift z, and $\sigma[\lambda_{\rm obs}/(1+z)]$ is the Lyα absorption cross section.

As described above, each source sits inside a local HII region, allowing the total optical depth to be separated into a component sourced by the resonant absorption, τ_R, and that from the damping wing of the cross section, τ_D. The common practice is to use the size of the local HII region, R_S, to separate the

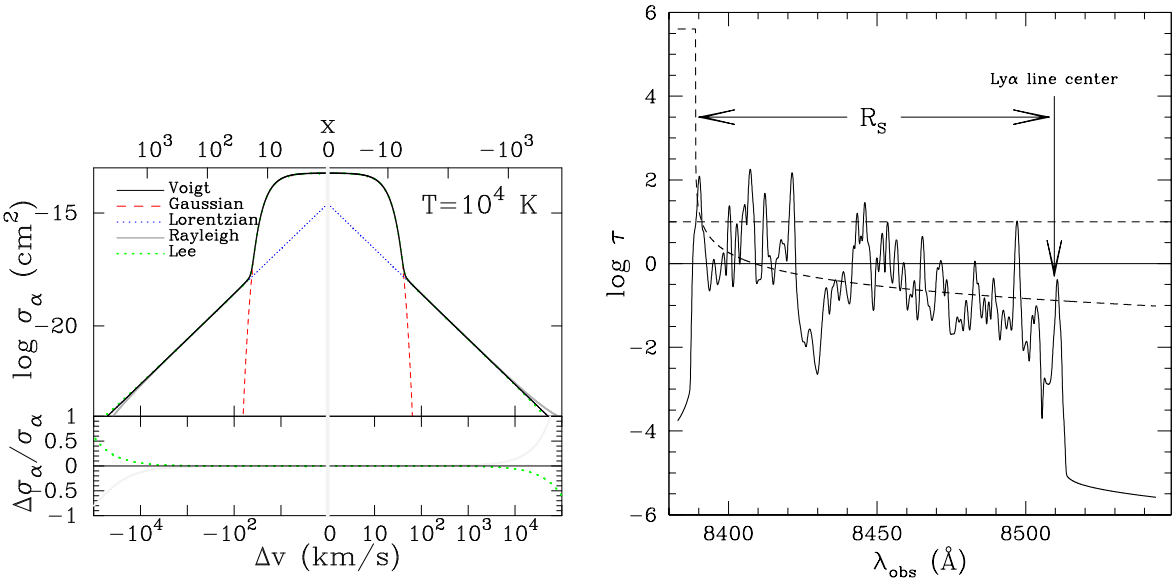

Fig. 32.— *Left:* Lyα cross section. Like all line transitions, the Lyα cross section consists of a relatively narrow core, whose width is set by a combination of turbulent motions and thermal Doppler broadening, and Lorenzian tails extending far from the core of the line (e.g. Rybicki & Lightman 1979). The figure is taken from Dijkstra (2014). *Right:* Optical depth contributions from within (τ_R) and from outside (τ_D) the local HII region for a typical line of sight towards a $z_s = 6$ quasar embedded in a fully neutral IGM. The *dashed line* corresponds to τ_D, and the *solid line* corresponds to τ_R. In this example, the damping wing of the IGM, τ_D, contributes significantly to the total optical depth at $\lambda_{\rm obs} \sim 8430$ Å and $\lambda_{\rm obs} \gtrsim 8470$ Å. The figure is taken from Mesinger & Haiman (2007).

terms:

$$\tau = \tau_R + \tau_D \qquad (245)$$
$$= \int_{z_{\rm HII}}^{z_s} d\tau_R + \int_{z_{\rm end}}^{z_{\rm HII}} d\tau_D \ .$$

Here $z_{\rm HII}$ corresponds to the redshift of the edge of the local HII region, and z_{end} denotes the redshift by which HI absorption is insignificant along the line of sight to the source (of order a hundred Mpc from the source).

The two components in eq. (245) are qualitatively different, as can be seen from the right panel of Fig. 32. Due to the relatively narrow core, τ_R pics up density and residual HI fluctuations inside the local HII region; thus it is a rapidly fluctuating quantity resulting in the so-called Lyα forest in QSO spectra. On the other hand, the damping wing is a smooth function of wavelength, averaging over opacity fluctuations over relatively large scales.

The strength of the damping wing absorption depends directly on the neutral fraction of the IGM. Studies looking for the imprint of the damping wing in galaxy and QSO spectra either focus on its spectral smoothness (e.g. Mesinger & Haiman 2004, 2007; Schroeder et al. 2013) or on the absolute absorption on the red side of the Lyα line where resonant absorption is negligible (e.g. Miralda-Escude 1998; Haiman & Spaans 1999; Santos et al. 2004; Bolton et al. 2011; Mesinger et al. 2015). In fact the later approach was used by Greig et al. (2017) to obtain the first detection (2σ) of ongoing reionization from the spectrum of a bright $z = 7.1$ quasar (see Fig. 33).

5.2.3. Combining current probes

Fig. 34 summarizes the current state of knowledge on the history of reionization (pre-2017; taken from Greig & Mesinger (2017); see also similar results by Mitra et al. (2015); Price et al. (2016)). Fitting a physically-motivated basis set of $\bar{x}_{\rm HI}(z)$ to current observations, these authors constrain the epochs

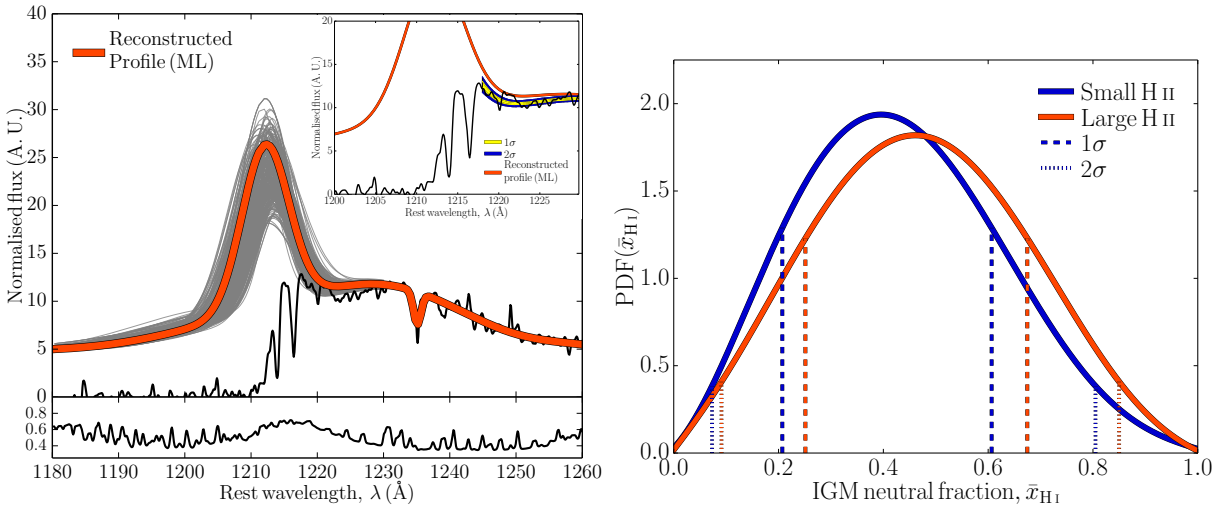

Fig. 33.— *Left:* FIRE spectrum of the $z = 7.08$ QSO, ULASJ1120+0641 is shown in black (Simcoe et al. 2012). The intrisic emission, F_0, of the QSO (before it passes through the intervening IGM) is shown in red (maximum likelihood) and gray (sampling the posterior), obtained by using the reconstruction procedure of Greig et al. (2017). The zoom-in inset also shows the 1 and 2 σ uncertainty on the total observed spectrum, $F_0 e^{-\tau_D}$, with τ_D computed from the simulations of Mesinger et al. (2016). The fact that the total observed spectrum is systematically higher than the intrinsic one is evidence of a non-zero τ_D from ongoing reionization. *Right:* The PDFs of $\bar{x}_{\rm HI} = 1 - Q_{\rm HII}$, quantifying the imprint of the damping wing shown in the right panel. The two curves correspond to opposite extreme assumptions about the topology of the EoR. *Figures are taken from Greig et al. (2017).*

corresponding to an average neutral fraction of (75, 50, 25) per cent, to $z = (8.52^{+0.96}_{-0.87}, 7.57^{+0.78}_{-0.73}, 6.82^{+0.78}_{-0.71})$, (1-$\sigma$). The strongest constraints here come from the *first detection of ongoing reionization*, obtained from the spectra of the $z = 7.1$ QSOs ULASJ1120+0641: $\bar{x}_{\rm HI}(z = 7.1) = 0.4^{+0.41}_{-0.32}$ (2-σ); see also the recent work by Mason et al. (2017) who obtain comparable limits from the disappearance of Lyman alpha emitting galaxies beyond $z \gtrsim 6$ (not shown in the figure).

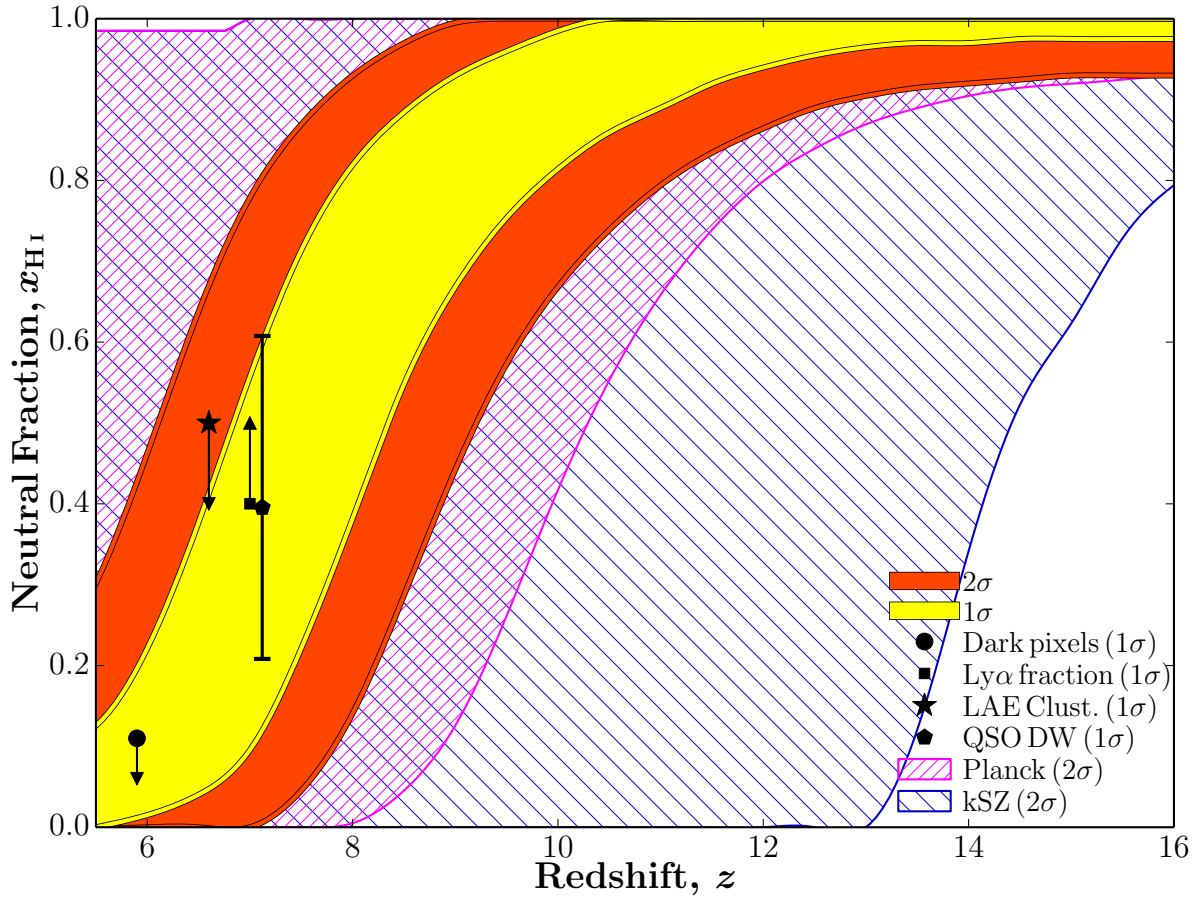

Fig. 34.— Constraints on the evolution of the average neutral fraction, $\bar{x}_{\rm HI} = 1 - Q_{\rm HII}$, from various probes (pre-2017). A physically-motivated EoR model was sampled, with the likelihood of each resulting $\bar{x}_{\rm HI}(z)$ curve provided by current observations. The figure is taken from Greig & Mesinger (2017).

.

5.3. Density evolution

Neglecting the impact of radiation, the density distribution of the IGM can be obtained by evolving the continuity equations from §4.1. The linear evolution of gas was already discussed above, when discussing the initial stages of collapse. However the IGM is only quasi-linear; thus hydrodynamic simulations are also used to obtain its density field. Fig. 35 shows the gas distribution from such a simulation by Viel et al. (2010) (*top left panel*), together with the corresponding DM field (*bottom left panel*). On large scales, the gas and dark matter trace each other very well, while on small scales the baryons are more diffuse owing to pressure support (note that the Jeans length in the mean density, ionized IGM is $\sim 0.6\sqrt{(1+z)/10}$ cMpc). On sub-galactic scales this trend is reversed, as radiative cooling allows baryons to collapse to much higher densities, creating stars and black holes.

For many applications, it would be very useful to have an analytic or parametric model of the IGM density distribution. In the linear regime, the density PDF is a Gaussian centered on $\Delta \equiv \rho/\bar{\rho} = 1$. We would expect structure formation to result in an extended tail towards large Δ, thus shifting the median of the distribution to $\Delta < 1$ (i.e. the under dense, so-called "voids" take up most of the volume of the Universe). This behavior is evident in Fig. 36. We could also expect the width of the distribution to be related to the Jeans scale. Using these guiding principles, Miralda-Escudé et al. (2000) (hereafter MHR00) proposed the following parametric form for the volume-weighted density PDF:

$$P(\Delta, z) = A\Delta^{-\beta} \exp\left[-\frac{(\Delta^{-2/3} - C_0)^2}{2(2\delta_0/3)^2}\right] . \tag{246}$$

where A and C_0 are constants set by volume and mass normalization, at each redshift:

$$\int_0^\infty P(\Delta) d\Delta = 1 \; ; \tag{247}$$

$$\int_0^\infty \Delta P(\Delta) d\Delta = 1 \; . \tag{248}$$

In the limit of $\Delta \to 1$ (perturbation around the mean density) and $C_0 \to 1$, we would recover the linear density field behavior, with the distribution approaching a Gaussian in $\Delta - 1$, with a standard deviation $\propto \delta_0$. Thus we expect $\delta_0 \propto (1+z)^{-1}$ following the evolution of the linear Growth factor in matter-dominated cosmologies. MHR00 take $\delta_0 \propto 7.61(1+z)^{-1}$, with the proportionality constant fit to match hydrodynamic simulations. The final constant, $\beta(z) \sim 2.2$–2.5, is also fit to simulation outputs at $z = 2$–4, though again we can "guesstimate" its value by noting that in the $\Delta \gg 1$ tail of the distribution which probes collapsed structures, the exponential factor in eq. (246) approaches unity. Thus the total distribution approaches $P(\Delta) \propto \Delta^{-\beta}$. If we assume collapsed structures, i.e. halos, follow an isothermal density profile: $\Delta(r) \propto r^{-2}$, then the fraction of the halo volume with density greater than $> \Delta$ is $V(>\Delta) \propto r^3 \propto \Delta^{-3/2}$, making the volume-averaged probability density scale as $P(\Delta) = dV(>\Delta)/d\Delta \propto \Delta^{-5/2}$. Thus isothermal structures result in $\beta = 2.5$, close to the fit found by MHR00.

How well does eq. (246) reproduce simulations? This can be seen from Fig. 37. Although there are some physically-motivated trends in eq. (246), *it is still an empirical fit to simulations* and therefore the agreement in the left panel (the original work from MHR00) is understandable. Bolton & Becker (2009) subsequently revisited this functional form and tested its agreement against larger simulations, over a more extended redshift range out to $z = 6$. Their results are shown in the right panel of Fig. 37. They find that the MHR00 form is accurate to withing a few percent over two decades around $\Delta = 1$, becoming increasingly inaccurate for large values. Note however that the high value tail is not known even in simulations, since the density distribution of gas in and around galaxies is very sensitive to SNe feedback (e.g. McQuinn et al. 2011).

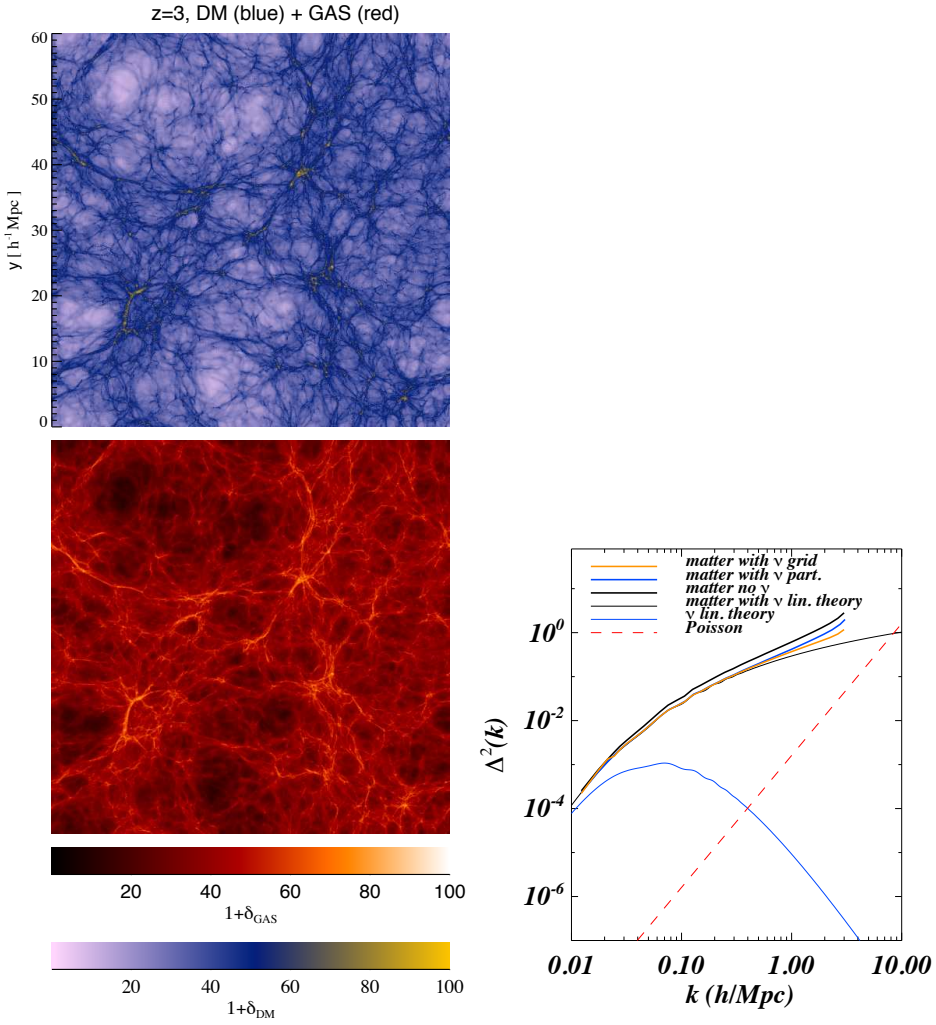

Fig. 35.— The simulated intergalactic medium at $z = 3$. The top left shows a 6 h^{-1} Mpc slice through the dark matter distribution in a 512^3 simulation, while the panel below it shows the correspond baryon field. Note that on large scales, the gas and dark matter trace each other very well, while on small scales the baryons are more diffuse owing to pressure support. On sub-galactic scales this trend is reversed, as radiative cooling allows baryons to collapse to much higher densities, creating stars and black holes. On the right, there is the corresponding dimensionless power spectra, including some models with massive neutrinos. Neutrino free streaming results in a suppression of small scale structure. The figures are taken from Viel et al. (2010).

5.3.1. Corresponding HI structure

As we saw in the previous section, observables generally do not depend only on the IGM density, but on the combination of the density and neutral fraction. In the (ionized) Universe, we can expect low density regions to be optically thin to ionizing radiation, while dense clumps are optically thick, capable of self-shielding against the ionizing background radiation. MHR00 also proposed a bi-modal model for self-shielding, shown in Fig. 38, with a critical density threshold separating the optically thin regime from the optically thick regime.

In this model, the gas sees a local ionizing background, $\Gamma(\Delta)$, which instantaneously transitions from the impinging (usually taken to be the mean) ionizing background, $\Gamma_{\rm bg}$, to zero at the critical self-shielding density:

$$\Gamma(\Delta) = \begin{cases} \Gamma_{\rm bg} & \text{if } \Delta < \Delta_{\rm ss} \\ 0 & \text{if } \Delta > \Delta_{\rm ss} \end{cases} \tag{249}$$

The above step function is a reasonable approximation, resulting in recombination rates similar to what

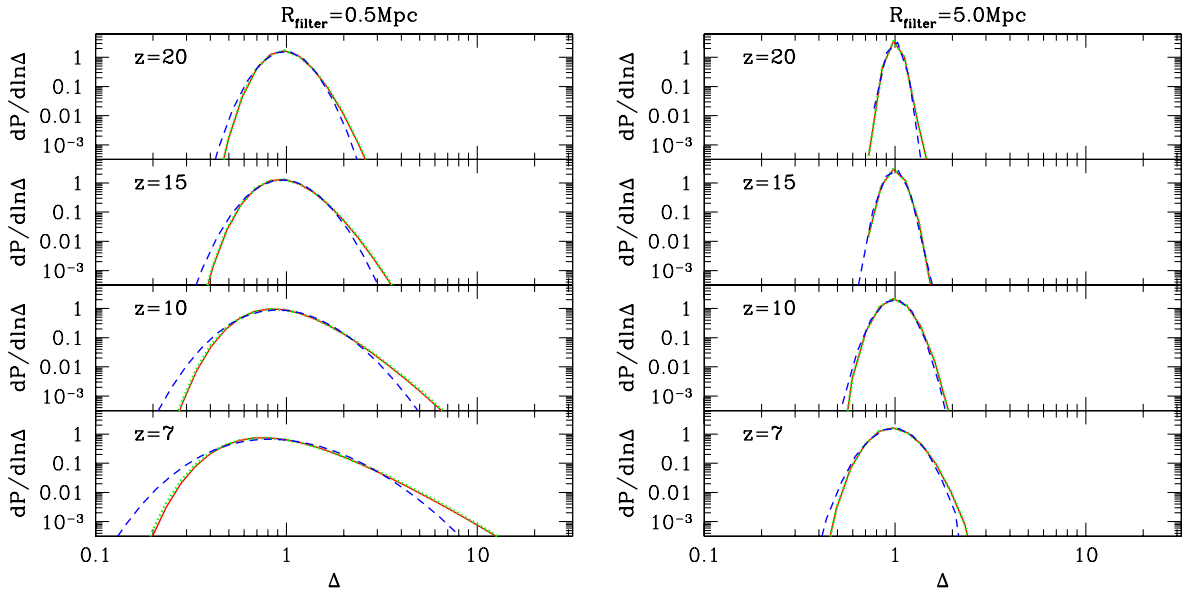

Fig. 36.— Volume-averaged PDFs of the density fields ($\Delta \equiv \rho/\bar{\rho}$) smoothed on scale $R_{\rm filter}$, computed from a 143 Mpc on a side, 1024^3 simulation of the gas (*solid red curves*), DM (*dotted green curves*), and linear perturbation theory (ZA; *dashed blue curves*) fields, at $z = 20$, 15, 10, 7 (top to bottom). The left panel corresponds to $R_{\rm filter} = 0.5$ Mpc; the right panel corresponds to $R_{\rm filter} = 5$ Mpc. At early times and large scales, the density field is still fairly linear (approaching a Gaussian in $\Delta - 1$, with a dispersion of σ_M). Non-linearity as one approaches late times and small scales is evident in the appearance of the high value tail, with the median of the distribution shifting to under-densities (matching the visual trends seen in the previous figure. All smoothing was performed with a real-space, top-hat filter. The figures are taken from Mesinger et al. (2011).

is seen in simulations (see below) but over-estimating the neutral hydrogen content of $\Delta \gg \Delta_{\rm ss}$ systems (e.g. McQuinn et al. 2011; Rahmati et al. 2013; Keating et al. 2014; Mesinger et al. 2015). The transition from optically thin to optically thick is more gradual, and depends somewhat on the spectral shape of the ionizing background (harder photons can penetrate deeper into clumps). Using a Haardt & Madau (2001) UV background, Rahmati et al. (2013) provide an empirical fit to their hydrodynamic simulations which show a more gradual self-shielding, in better agreement with observations of HI column densities:

$$\Gamma(\Delta) = \Gamma_{\rm bg} \times \left\{ 0.98 \left[1 + \left(\frac{\Delta}{\Delta_{\rm ss}} \right)^{1.64} \right]^{-2.28} + 0.02 \left[1 + \frac{\Delta}{\Delta_{\rm ss}} \right]^{-0.84} \right\} \tag{250}$$

Can we compute the characteristic self-shielding overdensity $\Delta_{\rm ss}$*?* Let's begin by first discritizing the density field into self-gravitating clouds on the *local* (i.e. Δ-dependent) Jeans scale, i.e. the distance a pressure wave can travel in a free-fall time (Schaye 2001):[24]

$$L_{\rm J} \equiv \frac{c_s}{\sqrt{G\rho}} \ . \tag{251}$$

Taking $c_s^2 = \frac{\gamma k_B T}{\mu m_p}$ and $\rho = \left(\frac{1}{f_g}\right) \rho_{\rm gas} = \left[\frac{1}{f_g(1-Y_{\rm He})}\right] \rho_{\rm H}$, where f_g is the fraction of the total mass in gas, $f_g(1-Y_{\rm He})$ is the fraction of the total mass in hydrogen, and $\rho_{\rm H} = m_p n_{\rm H}$ is the mass density of hydrogen,

[24]Note that the 1D distance is a more robust quantity than the Jeans mass, which assumes a 3D geometry for the conversion from scale to mass.

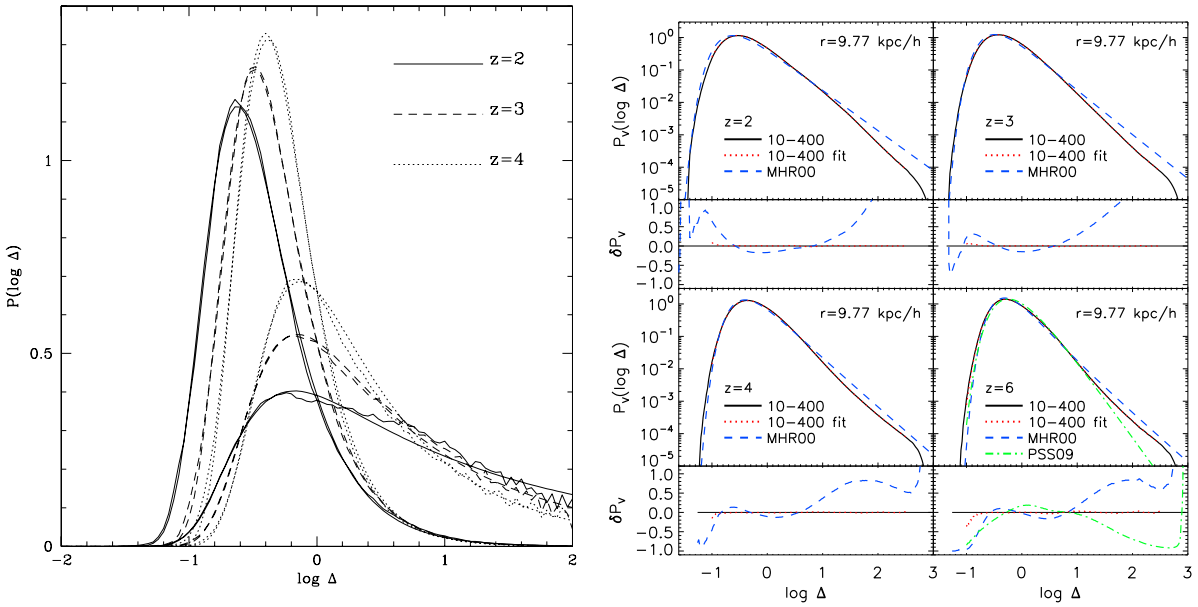

Fig. 37.— Density PDFs. On the left, we have the original results from Miralda-Escudé et al. (2000); the lines with noise correspond to a hydrodynamic simulation while the smooth curves correspond to their analytic fit. Narrow distributions are volume weighted, i.e. $P(\Delta, z)$, while broad distributions are mass weighted, i.e. $\Delta P(\Delta, z)$. On the right we have an analogous study by Bolton & Becker (2009) showing how the MHR00 distribution, although accurate at the percent level for two decades around the mean, becomes inaccurate at high densities for this model. It is worth noting that the true density distribution in the high value tail is sensitive to SNe feedback, and is poorly understood.

we can write:

$$L_{\rm J} = \sqrt{\frac{\gamma k_B T}{\mu m_p}} \sqrt{\frac{f_g(1-Y_{\rm He})}{G m_p n_{\rm H}}} = \sqrt{\frac{\gamma k_B}{G m_p^2}} \sqrt{\frac{f_g}{\mu}} \sqrt{\frac{T}{\Delta \bar{n}_{\rm H}}} \ . \tag{252}$$

The first factor is just made of constants, while the second factor can change somewhat due to ionization ($\mu = 0.6/1.2$ for an ionized/neutral medium, and we expect the gas fraction to be close to the cosmic mean value Ω_b/Ω_m). The final factor can vary significantly. For reference, the typical Jeans length inside galaxies can be $\sim$ pc, while the typical Jeans length inside the (ionized) IGM can be $\sim$ Mpc at high redshifts.

Given the Jeans length of our gas element, we can compute the corresponding HI column density, $N_{\rm HI}$, to see if it is sufficient to self-sheild:

$$N_{\rm HI} = x_{\rm HI} \Delta \bar{n}_{\rm H} L_{\rm J} = x_{\rm HI} \sqrt{T\Delta} \sqrt{\frac{\gamma k_B f_g \bar{n}_H}{G \mu m_p^2}} = 1.5 \times 10^{20} {\rm cm}^{-2} \ x_{\rm HI} \sqrt{T_4 \Delta_{100}} Z_7^{3/2} \ , \tag{253}$$

where we take $\gamma = 5/3$ for a monatomic gas, $f_g = \Omega_b/\Omega_m \approx 0.17$, $\mu = 0.6$, and subscripts in the last line denote the values used for normalization: $T_4 \equiv T/10^4{\rm K}$, $\Delta_{100} \equiv \Delta/100$, $Z_7 \equiv (1+z)/7$. The corresponding optical depth is obtained by multiplying with the ionization cross section, $\sigma_{\rm LL} \sim 6 \times 10^{-18}$ cm^2, resulting in: $\tau = N_{\rm HI}\sigma_{\rm LL} \approx 10^3 x_{\rm HI} \sqrt{T_4 \Delta_{100}} Z_7^{3/2}$. *Thus, even gas with a modest fraction of neutral hydrogen (our default values correspond to $x_{\rm HI} > 10^{-3}$), can self-shield against ionizing radiation.*

Since highly ionized gas can self shield, in computing $\Delta_{\rm ss}$ we can simplify the equation of ionization equilibrium, i.e.

$$x_{\rm HI} n_{\rm H} \Gamma = (1 + f_{\rm He}) \left[n_{\rm H} (1 - x_{\rm HI}) \right]^2 \alpha_{\rm HII} \ , \tag{254}$$

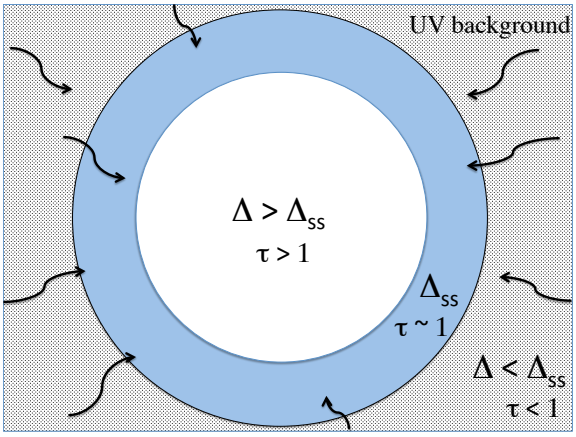

Fig. 38.— Simple model for the ionization structure of gas illuminated by a pervasive ionizing background. Roughly spherical, self-gravitating clumps have a thin shell with overdensity, $\Delta_{\rm ss}$, corresponding to an optical depth of unity. Gas inside this shell is self-shielded from the ionizing background, and remains largely neutral.

to its $x_{\rm HI} \ll 1$ limit:

$$x_{\rm HI} = \frac{(1+f_{\rm He})n_{\rm H}\alpha_{\rm HII}}{\Gamma} \ , \tag{255}$$

where $f_{\rm He} = (4/Y_{\rm He} - 3)^{-1} \approx 0.077$ is the helium number fraction, and we assume helium is singly-ionized together with hydrogen (an accurate assumption given their comparable energy thresholds). Taking the empirical fit for the recomination coefficient from Cen (1992): $\alpha_{\rm A}(T) \approx 4.2 \times 10^{-13}\ T_4^{-0.7}$ cm^3 s^{-1}, and defining $\Gamma_{12} = \Gamma/(10^{-12}{\rm s}^{-1})$, we have:

$$x_{\rm HI} \approx 3 \times 10^{-3}\ \Gamma_{12}^{-1} T_4^{-0.7} \Delta_{100} Z_7^3 \ . \tag{256}$$

Putting this into eq. (253) we have:

$$N_{\rm HI, x_{HI} \ll 1} \approx 4.5 \times 10^{17} {\rm cm}^{-2}\ \Gamma_{12}^{-1} T_4^{-0.2} \Delta_{100}^{3/2} Z_7^{9/2} \ . \tag{257}$$

If we define $\Delta_{\rm ss}$ to be the value for which $\tau = N_{\rm HI, x_{HI} \ll 1} \sigma_{\rm LL} = 1$, we obtain:

$$\boxed{\Delta_{\rm ss} \approx 50\ \Gamma_{12}^{2/3} T_4^{0.13} Z_7^{-3}} \tag{258}$$

We now have the framework to calculate one of the fundamental quantities for the IGM: the recombination rate. The recombination rate per hydrogen atom is computed by integrating over the density distribution:

$$\frac{dn_{\rm rec}}{dt} = \int_0^\infty \Delta^2 \bar{n}_{\rm H}\ \alpha_{\rm HII}\ [1 - x_{\rm HI}]^2\ P\ d\Delta \ . \tag{259}$$

Here the density PDF, P, can be computed from eq. (246), and the neutral fraction, $x_{\rm HI}$, from eq. (254). In Fig. 39 we show the evolution of the mean emissivity (*thick curves*) and recombination rate (*thin curves*) for several simulations in which the above framework was included via a sub-grid prescription (taken from Sobacchi & Mesinger 2014). The most complete model is denoted as "FULL". Towards the end stages of reionzation ($\bar{x}_{\rm HI} \lesssim 0.1$), the recombination rate balances the emission rate of ionizing photons (see also Furlanetto & Oh 2005), resulting in a so-called "photon-starved" end to reionization (e.g. Bolton & Haehnelt 2007).

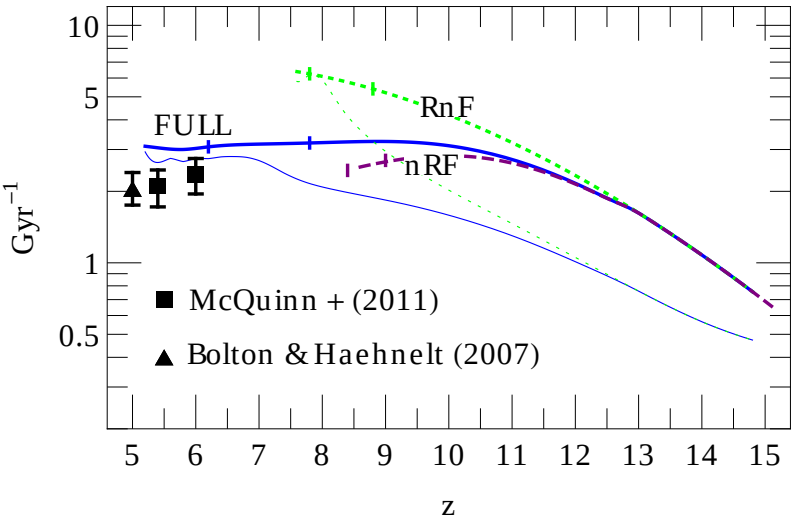

Fig. 39.— Evolution of the average emissivity (*thick*) and recombination rate per baryon (*thin*) with different models in Sobacchi & Mesinger (2014). The most complete model is denoted as "FULL". The vertical ticks correspond to $\bar{x}_{\rm HI} = 0.2$ and $\bar{x}_{\rm HI} = 10^{-2}$. For comparison we show the emissivity constraints inferred from the Ly α forest at $z \lesssim 6$ (Bolton & Haehnelt 2007; McQuinn et al. 2011).

.

5.4. Thermal evolution

Gas in the IGM behaves as a classical ideal gas, which quickly reaches local thermal equilibrium (LTE). In LTE, the temperature of any gas component with number density n_i and internal energy U_i can be written as:

$$\begin{aligned} T &= \frac{2U_{\rm i}}{3k_B n_{\rm i}} \\ &= \frac{2U_{\rm tot}}{3k_B n_{\rm tot}} \end{aligned} \tag{260}$$

The second line follows from equipartition of energy in an ideal gas in LTE. $U_{\rm tot}$ is the total internal energy per unit volume, and $n_{\rm tot}$ is the total number density of the gas, composed primarily of hydrogen and helium:

$$\begin{aligned} n_{\rm tot} = \sum_i n_i &\approx n_e + n_{\rm HI} + n_{\rm HII} + n_{\rm HeI} + n_{\rm HeII} \\ &\approx x_i(n_{\rm H} + n_{\rm He}) + (1-x_i)n_{\rm H} + x_i n_{\rm H} + (1-x_i)n_{\rm He} + x_i n_{\rm He} \\ &= x_i(n_{\rm H} + n_{\rm He}) + n_{\rm H} + n_{\rm He} \\ &= n_b(1+x_i) \end{aligned} \tag{261}$$

As in the previous section, here we assume that hydrogen and helium are singly ionized with an ionization fraction of x_i; thus $n_e = x_i(n_{\rm H} + n_{\rm He})$. Moreover, we ignore doubly-ionized Helium (assuming $n_{\rm HeIII}$=0), whose high ionization threshold requires very hard sources, and is thus expected to be ionized with the advent of QSOs at lower redshifts, $z \lesssim 4$.

We can explicitly solve for the evolution of the IGM temperature of an IGM gas element, starting with the time derivative of eq. (260):

$$\frac{dT}{dt} = \frac{2}{3k_B}\left[\frac{1}{n_{\rm tot}}\frac{dU_{\rm tot}}{dt} - \frac{U_{\rm tot}}{n_{\rm tot}^2}\frac{dn_{\rm tot}}{dt}\right]$$

Substituting in eq. (261), we get:

$$\begin{aligned}\frac{dT}{dt} &= \frac{2}{3k_B}\left[\frac{1}{n_b(1+x_i)}\frac{dU_{\rm tot}}{dt} - \frac{U_{\rm tot}}{n_b^2(1+x_i)}\frac{dn_b}{dt} - \frac{U_{\rm tot}}{n_b(1+x_i)^2}\frac{dx_i}{dt}\right] \\ &= \frac{2}{3k_B n_b(1+x_i)}\frac{dU_{\rm tot}}{dt} - \frac{T}{n_b}\frac{dn_b}{dt} - \frac{T}{(1+x_i)}\frac{dx_i}{dt}\end{aligned} \tag{262}$$

It is useful to explicitly denote the sources of heating:

$$\frac{dU_{\rm tot}}{dt} = \frac{dQ}{dt} + \frac{dU_{\rm adia}}{dt}\ , \tag{263}$$

where dQ/dt is the total, non-adiabatic heating rate per volume (we return to this below), and the adiabatic heating rate per volume can be written as:

$$\begin{aligned}\frac{dU_{\rm adia}}{dt} &= \frac{3}{2}\frac{dP}{dt} = \frac{3}{2}A\frac{dn_b^\gamma}{dt} = \frac{3}{2}A\gamma n_b^{\gamma-1}\frac{dn_b}{dt} \\ &= \frac{5}{2}(1+x_i)k_{\rm B}T\frac{dn_b}{dt}\end{aligned} \tag{264}$$

Here A is a constant of proportionality, and we have assumed the adiabatic equation of state: $P \propto n^\gamma$, with $\gamma = 5/3$ for a monatomic ideal gas.[25]

Substituting eq. (263-264) into eq. (262):

$$\begin{aligned}\frac{dT}{dt} &= \frac{2}{3k_B n_b(1+x_i)}\left[\frac{dQ}{dt} + \frac{5}{2}(1+x_i)k_{\rm B}T\frac{dn_b}{dt}\right] - \frac{T}{n_b}\frac{dn_b}{dt} - \frac{T}{(1+x_i)}\frac{dx_i}{dt} \\ &= \frac{2}{3k_B n_b(1+x_i)}\frac{dQ}{dt} + \frac{5T}{3n_b}\frac{dn_b}{dt} - \frac{T}{n_b}\frac{dn_b}{dt} - \frac{T}{(1+x_i)}\frac{dx_i}{dt} \\ &= \frac{2}{3k_B n_b(1+x_i)}\frac{dQ}{dt} + \frac{2T}{3n_b}\frac{dn_b}{dt} - \frac{T}{(1+x_i)}\frac{dx_i}{dt}\end{aligned} \tag{265}$$

Here the first term corresponds to the non-adiabatic sources of heating (discussed below), the second term corresponds to adiabatic heating and cooling, while the third term is the energy contribution from changing the number of species.

To see explicitly the contribution of the average expansion history and the growth of structure, it is also common to see this expression with the adiabatic term expanded using the local overdensity, $\Delta \equiv n_b/\bar{n}_b$, such that $n_b = \bar{n}_0 a^{-3}\Delta$ and thus $n_b^{-1}dn_b/dt = a^3\Delta^{-1}\left[a^{-3}\frac{d\Delta}{dt} - 3\Delta a^{-4}\frac{da}{dt}\right] = \frac{1}{\Delta}\frac{d\Delta}{dt} - 3H$. Putting this expression into the second term of eq. (265) and rearranging the terms we obtain:

$$\boxed{\frac{dT}{dt} = -2TH + \frac{2T}{3\Delta}\frac{d\Delta}{dt} - \frac{T}{(1+x_i)}\frac{dx_i}{dt} + \frac{2}{3k_{\rm B}n_b(1+x_i)}\frac{dQ}{dt}} \tag{266}$$

Here the first term on the RHS shows that in the absence of additional sources of heating, the adiabatic cooling due to the expansion of the Universe cools the mean densities at a rate of $-2TH$, i.e. $\frac{dT}{T} = -2\frac{da}{a} \rightarrow T \propto (1+z)^2$, something we derived already in §2.2. The second term of the RHS shows the additional adiabatic heating/cooling from the evolution of over/under densities. The third term is relevant during reionization, and the fourth term corresponds to non-adiabatic heating.

In the high redshift IGM, the fourth term is dominated by radiative cooling and heating, and can be expanded to (e.g. Upton Sanderbeck et al. 2016):

$$\frac{dQ}{dt} = \frac{dQ_{\rm Comp}}{dt} + \sum_i \frac{dQ_{\rm photo,i}}{dt} + \sum_p\sum_i R_{p,i}n_e n_i\ , \tag{267}$$

[25] More explicitly, using A as a constant of proportionality which accounts for all of our plasma species (whose number densities are all proportional to n_b): $P = An_b^\gamma = \frac{2}{3}U_{\rm adia} = n_{\rm tot}k_{\rm B}T$, and $k_{\rm B}T = \frac{An_b^\gamma}{n_b(1+x_i)} = An_b^{\gamma-1}/(1+x_i)$.

where $dQ_{\rm Comp}/dt$ is the Compton heating/cooling off of the CMB, $dQ_{\rm photo,i}/dt$ is the photo-heating rate of species i, and $R_{p,i}$ is the cooling rate coefficient for species i and cooling process p. At the redshifts and densities of interest, the third term is sub-dominant (e.g. McQuinn & Upton Sanderbeck 2015, but can include recombination, free-free and collisional cooling (e.g. Rybicki & Lightman 1979; Hui & Gnedin 1997). Note that we also ignore shock heating which has been shown to be inefficient in the bulk of the IGM (e.g. McQuinn & O'Leary 2012).

Compton cooling/heating can be expressed as (e.g. Seager et al. 2000):

$$\frac{2}{3k_{\rm B}n_b(1+x_i)}\frac{dQ_{\rm Comp}}{dt} = \frac{x_i}{1+x_i}\frac{8\sigma_{\rm T}u_\gamma}{3m_ec}(T_\gamma - T) \ , \tag{268}$$

where $\sigma_{\rm T}$ is the Thompson scattering cross-section, and u_γ is the energy density of the CMB.

Photo-heating of hydrogen and helium can be expressed as

$$\frac{dQ_{\rm photo,i}}{dt} = \int d\nu n_b f_i \sigma_i \frac{4\pi J}{h\nu}(h\nu - E_i^{\rm th}) f_{\rm heat} \tag{269}$$

where f_i is the number fraction of species i = HI, HeI, or HeII, σ_i the ionization cross-section, and $E_i^{\rm th}$ is the ionization threshold energy of species i. Furthermore, $f_{\rm heat}(h\nu - E_i^{\rm th}, x_i)$ is the fraction of the photo-ionized electron's energy going into heat (e.g. Shull & van Steenberg 1985; Furlanetto & Stoever 2010; Valdés et al. 2010), which depends on both the primary electron's energy, $(h\nu - E_i^{\rm th})$, and the ambient ionized fraction, x_i (which determines the partitioning of the primary electron's energy deposition into heating vs ionizing).

The angle-averaged specific intensity, $J(\nu, z)$, (in erg s^{-1} Hz^{-1} pcm^{-2} sr^{-1}; a prefix of 'p' in 'pcm' denotes proper units) can be computed integrating the comoving specific emissivity (energy per unit time per unit frequency per unit comoving volume), $\epsilon_{h\nu}(\nu_e, z')$ back along the light-cone (c.f. Haardt & Madau 1996):

$$\begin{aligned} J(\nu, z) &= \frac{1}{4\pi}\int_z^\infty dz' \frac{dV_{\rm com}}{dz'}\epsilon_{h\nu}e^{-\tau}\frac{(1+z)^2}{4\pi R_{z,z'}^2}\frac{(1+z)}{(1+z')}\frac{(1+z)}{(1+z')}\frac{(1+z')}{(1+z)} \\ &= \frac{(1+z)^2}{4\pi}\int_z^\infty dz' \cancel{4\pi R_{z,z'}^2}\frac{dR}{dz'}\frac{\epsilon_{h\nu}e^{-\tau}}{\cancel{4\pi R_{z,z'}^2}}\frac{1+z}{1+z'} \\ &= \frac{(1+z)^3}{4\pi}\int_z^\infty dz' \frac{cdt}{dz'}\epsilon_{h\nu}e^{-\tau} \ , \end{aligned} \tag{270}$$

where $R_{z,z'}$ is the comoving distance from z' to z, $(1+z)^2/(4\pi R_{z,z'}^2)$ is the proper surface area of the corresponding sphere, $e^{-\tau}$ is the probability a photon emitted at z' survives to reach z, $dV_{\rm com}/dz'$ is the differential of the comoving volume, and the three redshift factors at the end of the first line account for time dilation, redshifting of the photon energy and redshifting the frequency interval $(d\nu)$, respectively.

The comoving specific emissivity, $\epsilon_{h\nu}(\nu_e, z')$ is evaluated in the emitted frame, $\nu_e = \nu(1+z')/(1+z)$, and the IGM optical depth (dominated by photo-ionizations of H and He) between z and z' can then be written as $\tau(\nu, z, z') = \int_{z'}^z dz'(cdt/dz')(1 - Q_{\rm HII})\bar{n}_b\tilde{\sigma}$, where the photo-ionization cross-section is weighted over species, $\tilde{\sigma}(z, \nu_e) \equiv f_{\rm H}(1-\bar{x}_i)\sigma_{\rm H} + f_{\rm He}(1-\bar{x}_i)\sigma_{\rm HeI} + f_{\rm He}\bar{x}_i\sigma_{\rm HeII}$ and is also evaluated at $\nu_e = \nu(1+z')/(1+z)$ (Mesinger et al. 2011).

We can relate the specific emissivity to the collapse of structures as we did when discussing reionization. Assuming that the source luminosities, L in erg s^{-1} Hz^{-1} per solar baryon, follow a power law with spectral index α beyond some threshold ν_0 corresponding to absorption within the galaxy, $L = A(\nu_e/\nu_0)^{-\alpha}$, and adopting a normalization of $\dot{n}_\gamma$ photons produced per stellar baryon per time:

$$\dot{n}_\gamma = \int_{\nu_0}^\infty \frac{L}{h\nu}d\nu = A\nu_0^\alpha\int_{\nu_0}^\infty \frac{\nu^{-\alpha}}{h\nu}d\nu = \frac{A}{h\alpha} \ , \tag{271}$$

thus $L = N_\gamma\alpha h(\nu_e/\nu_0)^{-\alpha}$, we can write:

$$\epsilon_{h\nu}(\nu_e, z') = \frac{\alpha h}{\mu m_p}\dot{n}_\gamma\left(\frac{\nu_e}{\nu_0}\right)^{-\alpha} f_{\rm esc}\left[f_*\rho_{\rm crit,0}\Omega_b\frac{df_{\rm coll}(z')}{dt}\right] \ , \tag{272}$$

μm_p is the mean baryon mass, $\rho_{\rm crit,0}$ is the current critical density, f_* is fraction of baryons converted into stars. The quantity in the square brackets is the star formation rate density at z' (mass of stars formed per unit time per unit comoving volume).

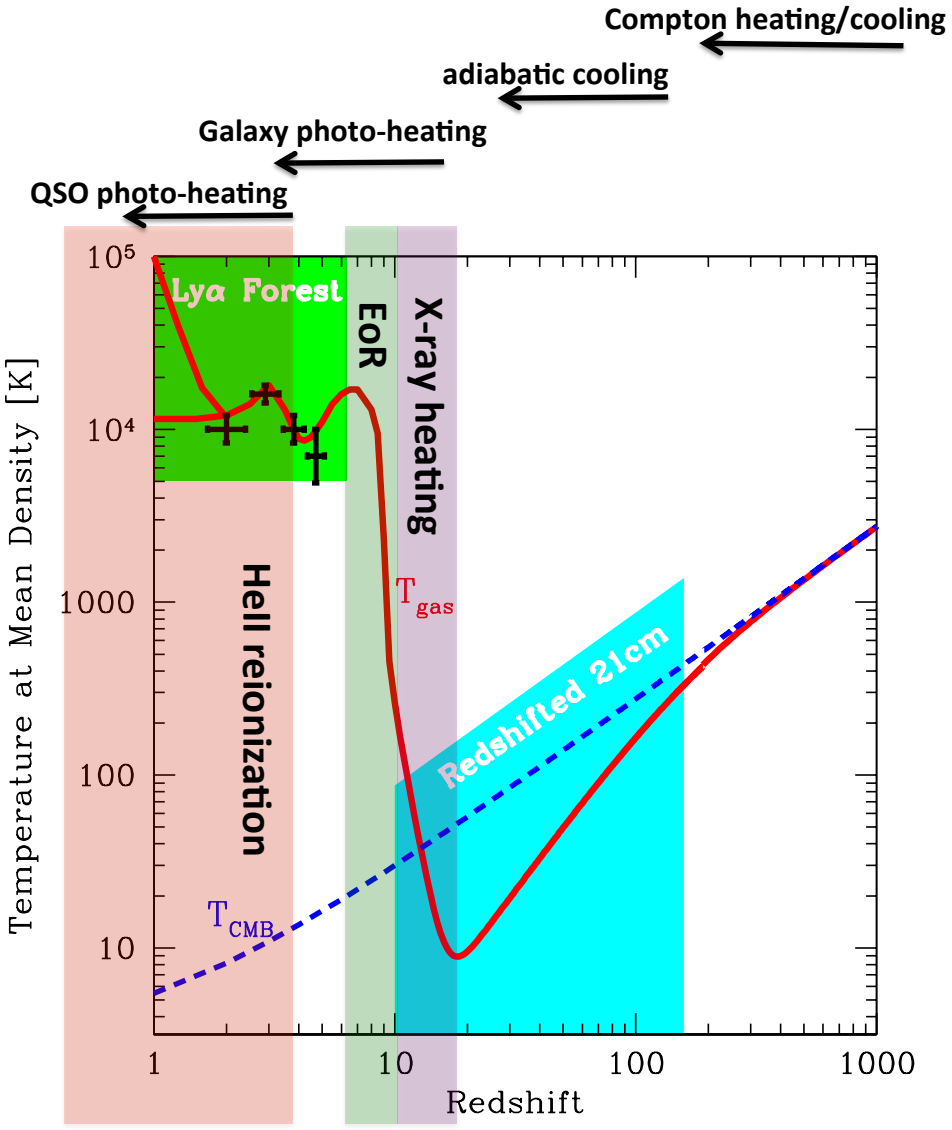

Fig. 40.— Schematic showing the evolution of the temperature of cosmic gas at mean density (adapted from McQuinn 2015).

In summary, our current understanding of the thermal evolution of the average-density IGM is shown in Fig. 40, adapted from McQuinn (2015). We note several milestones at $z > 6$, in order of decreasing redshift:

- **Thermal coupling to the CMB** ($200 \lesssim z \lesssim 1100$) – Following recombination, the baryon temperature was still tied to the CMB temperature through Compton scattering off of the $x_i \sim 10^{-4}$ residual electron fraction. Thus $\bar{T} = T_\gamma \propto (1+z)$.

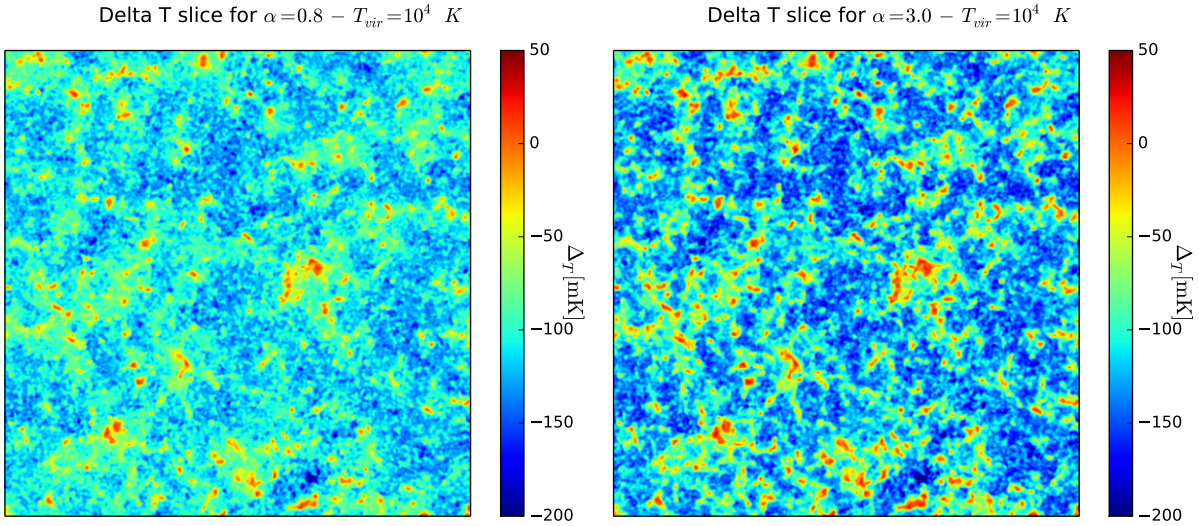

Fig. 41.— Slices through a 750 Mpc simulation of the 21-cm brigthness temeprature field when the inhomogeneity in the gas temperature is driving the peak signal. The left panel assumes the typical X-ray emission of the first galaxies is dominated by HMXB binaries, while the right panel assumes domination by the hot ISM. Since the HMXB spectra are much harder, with the typical X-ray photons having longer mean free paths, the heating is much more uniform in the left panel. The figures are taken from (Pacucci et al. 2014).

- **Adiabatic expansion during the Dark Ages** ($20 \lesssim z \lesssim 200$) – As the Universe expanded, the adiabatic cooling from the expansion dominated over the Compton heating. Thus the cosmic gas began cooling more rapidly, $\bar{T} \propto (1+z)^2$.

- **X-ray heating during the Cosmic Dawn** ($10 \lesssim z \lesssim 20$) – When the first stars and accreting black holes formed, their radiation began spreading through the Universe, heating and ionizing the IGM. Empirical relations of the X-ray luminosity to the SFR of nearby galaxies (e.g. Mineo et al. 2012a,b) suggest that X-rays from high mass X-ray binaries (HMXBs) and the hot ISM began heating the IGM well before reionization.[26] As x_i surpasses a few per cent, the bulk of the secondary electron's energy becomes deposited as heat through free-free interactions (as $x_i \to 1$, $f_{\rm heat} \to 1$). The associated rise of $\bar{T}$ traces the birth of the first galaxies and their typical X-ray luminosity (Mesinger et al. 2014). This epoch of heating (EoH) is inhomogeneous, since the mean free path of these X-rays through the IGM is a strong function of their energies, $\lambda_{\rm X} \approx 20\ \bar{x}_{\rm HI}^{-1} \left(\frac{E_{\rm X}}{300{\rm eV}}\right)^{2.6} \left(\frac{1+z}{10}\right)^{-2}$ cMpc. Therefore the hardness of the typical X-ray spectra could have strong imprint in the large-scale patchiness of the IGM temperature during these times (see Fig. 41).

- **Reionization** ($6 \lesssim z \lesssim 10$) – Cosmic reionization as we have seen is an inhomogeneous and likely extended process. After an ionization front passes through a patch of the IGM, it becomes impulsively heated to temperatures of $\sim$ 1–3 $\times 10^4$ K (with the exact value depending on the speed of the ionization front and the spectrum of the ionizing background, e.g. Hui & Gnedin 1997). Thus during reionization, some IGM patches should have temperatures of $\sim 10^4$ K, while the neutral ones remain at $\lesssim 100$ – 1000 K (with the exact value depending on the level of X-ray heating). Post reion-

[26] One can confirm this claim explicitly from the above equations by using the corresponding values of $N_\gamma \sim 0.1$ X-ray photons per stellar baryon and an $f_{\rm esc} \sim 1$ above $h\nu_0 \gtrsim 0.5$ keV (McQuinn & O'Leary 2012; Mesinger et al. 2013; Das et al. 2017).

ization, the temperature of the gas is governed by the optically-thin photo-heating expression in eq. (269). It approaches a well-studied temperature-density relation: $T \propto \Delta^{0.6}$, losing knowledge of its reionization history within $\Delta z \sim$2–3 (e.g. Hui & Gnedin 1997; McQuinn & Upton Sanderbeck 2015).

5.5. Cosmological 21-cm signal

Finally, we end with the most powerful future probe of the early Universe: the cosmic 21-cm signal. The 21-cm line corresponds to the spin-flip transition of the ground state of neutral hydrogen. There is an energy difference of 0.068 K between the ground states when the spin of the electron is aligned with that of the proton vs. when they are anti-aligned. The transition from the aligned to anti-aligned states thus releases a photon of energy 0.068 K, corresponding to a (radio) wavelength of λ_{21} =21 cm (ν_{21} = 1420 MHz).

This transition was predicted in 1944 by Hendrik van den Hulst, as a way of using radio lines to study the structure of our galaxy. It was subsequently detected in 1951 by Ewan and Purcell at Harvard University. The 21-cm line has proven to be very useful in mapping the neutral gas content of our galaxy as well as nearby galaxies. Since it is a line transition, it allows us to study the kinematics of neutral (cold) gas.

However, the 21-cm line has unmatched potential for cosmology. Before reionization, *the entire Universe was full of neutral hydrogen.* If we can observe this radio signal with interferometers, *we could map out the structure of the first billion years of our Universe!*

For this cosmological application, the neutral hydrogen would be seen in contrast against a radio background, most likely the CMB (though see, e.g. Ewall-Wice et al. 2018; Fialkov & Barkana 2019, for some extreme models in which an astrophysical background dominates over the CMB, motivated by the recent controversial EDGES detection Bowman et al. 2018).[27] At these low energies, we can use the Rayleigh-Jeans limit, with the intensity being proportional to the temperature. As per radio astronomy convention, we can thus re-write eq. (224) using temperature instead of intensity to predict the so-called brightness temperature, T_b, at an observed frequency $\nu_{\rm obs}$ corresponding to a patch of the IGM at $(1+z) = \nu_{21}/\nu_{\rm obs}$:

$$T_b(\nu_{\rm obs}) = T_\gamma e^{-\tau} + T_s\left(1 - e^{-\tau}\right) \ . \tag{273}$$

Here the background radiation is the CMB temperature, $T_\gamma = 2.73(1+z)$ K, and the source function is the so-called spin temperature of the gas. The spin temperature is defined according to the relative level populations of hydrogen atoms occupying the aligned (n_1) and anti-aligned (n_0) states:

$$\frac{n_1}{n_0} \equiv 3\exp\left[-\frac{0.068K}{T_s}\right] \ , \tag{274}$$

with the factor of 3 corresponding to the statistical weights of the two states, and 0.068 K the energy difference between them.

The astrophysics is encoded in the optical depth through the cosmic gas cloud:

$$\tau(\nu_{\rm obs}) \approx \int dr\ n_{\rm H} x_{\rm HI} \sigma(\nu_{\rm rest}) \ . \tag{275}$$

Here $dr = cdt$ is the proper cosmological line element, and the absorption cross section is evaluated in the rest frame of the gas:

$$\nu_{\rm rest} = \nu_{\rm obs}(1+z)\left(1 - \frac{v_r}{c}\right) = \nu_{\rm obs}\left[(1+z) - (1+z)\frac{v_r}{c}\right] \ , \tag{276}$$

with v_r denoting the peculiar radial (line of sight) velocity of the gas, and the $(1 - v_r/c)$ accounting for the Doppler shifting (to first order). We can simplify eq. (275) by converting the line of sight integral to

[27] There is an alternative use of the cosmic 21-cm line: to see the IGM as absorption features in the background radio spectra of radio-loud AGN. This is the so-called 21-cm forest, in analogy to the Lyα forest. Unfortunately, this requires bright, radio-loud AGN at redshifts before the Universe was ionized and heated. The existence of such sources at such high redshifts is very uncertain. Even if some sources are found, the 21-cm forest would only result in a handful of sight-lines towards them; this pales in comparison with using the CMB as a background which allows us to map the Universe.

a frequency integral. To do so, we take the differential of the rest frame frequency,

$$\begin{aligned}\frac{d\nu_{\rm rest}}{dr} &= \nu_{\rm obs}\left[\frac{dz}{dr} - \frac{dz}{dr}\frac{v_r}{c} - \frac{1+z}{c}\frac{dv_r}{dr}\right] \\ &= \nu_{\rm obs}\frac{dz}{dr}\left[1 - \frac{v_r}{c} - \frac{1+z}{c}\frac{dv_r}{dr}\frac{dr}{dz}\right] \ . \end{aligned} \tag{277}$$

Using the relation for the proper line element $\frac{dr}{dz} = -\frac{c}{H(1+z)}$:

$$\begin{aligned}\frac{d\nu_{\rm rest}}{dr} &= -\nu_{\rm obs}\frac{H(1+z)}{c}\left[1 - \frac{v_r}{c} + \frac{dv_r}{Hdr}\right] \\ \frac{d\nu_{\rm rest}}{dr} &= -\frac{\nu_{\rm rest}}{\cancel{(1+z)}(1-\frac{v_r}{c})}\frac{H\cancel{(1+z)}}{c}\left[1 - \frac{v_r}{c} + \frac{dv_r}{Hdr}\right] \\ \frac{d\nu_{\rm rest}}{dr} &= -\frac{\nu_{\rm rest}H}{c}\left[1 + \frac{dv_r}{dr}\frac{1}{H(1-\frac{v_r}{c})}\right] \\ dr &= d\nu_{\rm rest}\frac{-c}{H\nu_{\rm rest}}\left[1 + \frac{dv_r}{dr}\frac{1}{H(1-\frac{v_r}{c})}\right]^{-1}\end{aligned} \tag{278}$$

We can now replace the integrand in eq. (275), writing instead

$$\tau(\nu_{\rm obs}) \approx \int d\nu_{\rm rest}\ n_{\rm H}x_{\rm HI}\sigma(\nu_{\rm rest})\frac{c}{H\nu_{\rm rest}}\left[1 + \frac{dv_r}{dr}\frac{1}{H(1-\frac{v_r}{c})}\right]^{-1} \tag{279}$$

We can further simplify this expression by noting that the proper IGM velocities are non-relativistic, $v_r/c \ll 1$, and also approximating the line profile as a delta function around resonance, $\sigma(\nu_{\rm rest}) \propto \delta_{\rm Dirac}(\nu_{21})$.[28] The later approximation treats the gas properties as constants over the line profile, allowing us to evaluate the entire integrand at the line center (i.e. 21-cm resonance):

$$\tau(\nu_{\rm obs}) \propto n_{\rm H}x_{\rm HI}\frac{c}{\nu_{21}}\frac{\sigma(\nu_{21})}{(H + dv_r/dr)} \ . \tag{280}$$

If we explicitly evaluate all of the constants and simplify, we can write:

$$\tau(\nu_{\rm obs}) \sim 0.003\ x_{\rm HI}\ \Delta\left(\frac{1+z}{10}\right)^{1/2}\left(\frac{10{\rm K}}{T_s}\right)\left[\frac{H}{H + dv_r/dr}\right] \tag{281}$$

Note that the value of the optical depth is relatively reasonable (the Universe would have been nicer to make it a factor of 10–100 larger, but we take what we can get...). The IGM is then seen in contrast against the CMB, expressed as the brightness temperature offset:

$$\begin{aligned}\delta T_b &= T_b - T_\gamma = T_\gamma e^{-\tau} + T_s\left(1 - e^{-\tau}\right) - T_\gamma \\ &= (T_s - T_\gamma)\left(1 - e^{-\tau}\right) \approx (T_s - T_\gamma)\tau \\ &\approx 30\ x_{\rm HI}\ \Delta\left(\frac{H}{dv_r/dr + H}\right)\left(1 - \frac{T_\gamma}{T_{\rm S}}\right)\left(\frac{1+z}{10}\frac{0.15}{\Omega_{\rm M}h^2}\right)^{1/2}\left(\frac{\Omega_b h^2}{0.023}\right){\rm mK}.\end{aligned}$$

The second line assumes $\tau \ll 1$, which we saw was a safe approximation for most of the IGM.

The spin temperature, $T_{\rm S}$, interpolates between the CMB temperature, T_γ, and the gas kinetic temperature, T_K. Since the observation uses the CMB as a backlight, a signal is only obtained if $T_S \to T_K$. This coupling is achieved through either: (i) collisions, which are effective in the IGM at high redshifts, $z \gtrsim 50$; or (ii) a Lyman alpha background [so-called Wouthuysen-Field (WF) coupling; Wouthuysen (1952); Field (1958)], effective soon after the first sources turn on.

[28] This common approximation, although valid for the vast majority of IGM gas, does create an unphysical divergence at $dv_r/dr \to H$. For a through discussion of this, see Mao (2012).

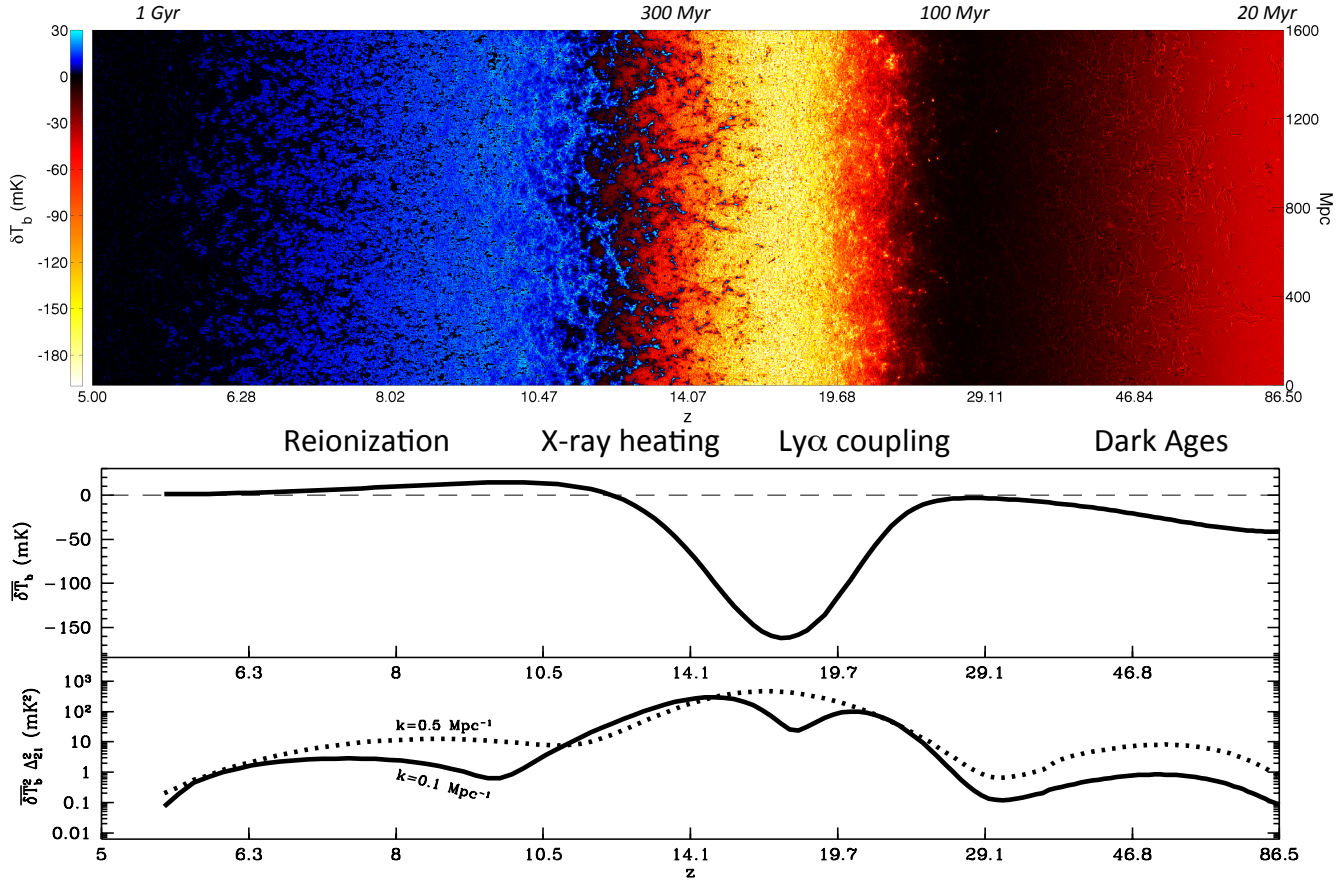

Fig. 42.— *Top:* light-cone strip of the 21-cm brightness temperature from the "Faint Galaxies" model of Mesinger et al. (2016). The main cosmic milestones can be seen from right to left: (i) first stars (Wouthuysen-Field coupling in black→yellow), (ii) first black holes (X-ray heating in yellow→blue), (iii) first galaxies (cosmic reionization in blue→black). The timing and the 'patterns' of the signal encode information about the first galaxies and the structure of the IGM. *Middle:* corresponding evolution of the global brightness temperature contrast. *Bottom:* corresponding evolution of the power spectrum amplitude at $k = 0.1$ Mpc^{-1} (*solid curve*) and $k = 0.5$ Mpc^{-1} (*dotted curve*).

In Fig. 42, we show a slice through the δT_b field in a "fiducial" model (for more details, see Mesinger et al. 2016). It is immediately obvious that the 21cm signal is a physics-rich probe, encoding information on various processes during and before reionization. Although the exact timing of the cosmic epochs is uncertain, the relative order is robustly predicted (c.f. Furlanetto 2006; §2.1 in McQuinn & O'Leary 2012):

1. **Collisional coupling**: The IGM is dense at high redshifts, so the spin temperature is uniformly collisionally coupled to the gas kinetic temperature, $T_K = T_S \lesssim T_\gamma$. Following thermal decoupling from the CMB ($z \lesssim 300$), the IGM cools adiabatically as $T_K \propto (1+z)^2$, faster than the CMB $T_\gamma \propto (1+z)$. Thus δT_b is negative. This epoch, serving as a *clean probe of the mater power spectrum* at $z \gtrsim 100$, is not shown in Fig. 42.
2. **Collisional decoupling**: The IGM becomes less dense as the Universe expands. The spin temperature starts to decouple from the kinetic temperature, and begins to approach the CMB temperature again, $T_K < T_S \lesssim T_\gamma$. Thus δT_b starts rising towards zero. Decoupling from $T_{\rm K}$ occurs as a function of the local gas density, with underdense regions decoupling first. Fluctuations are sourced by the density field, and again *offer a direct probe of cosmology.* Eventually ($z \sim 25$), all of the IGM is decoupled and there is little or no signal. This epoch corresponds to the red→black transition on the right edge of Fig. 42.
3. **WF coupling (i.e. Lyα pumping)**: The first astrophysical sources turn on, and begin coupling

T_S and T_K, this time through the Lyα background. δT_b becomes more negative, reaching values as low as $\delta T_b \sim 100$–200 (depending on the offset of the WF coupling and X-ray heating epochs). This epoch, offering a window on the *very first stars in our Universe*, corresponds to the black→yellow transition in Fig. 42.

4. **IGM heating**: The IGM is heated, with the spin temperature now coupled to the gas temperature, $T_K = T_S$. As the gas temperature surpasses T_γ, the 21cm signal changes from absorption to emission, becoming insensitive to the actual value of T_S (see eq. **??**). This epoch probes all processes which heat the IGM, *both astrophysical and cosmological*. The dominant source of heating is likely the X-rays from early accreting black holes (e.g. Furlanetto (2006)); however in some models more exotic processes dominate, such as the evaporation of cosmic strings, or DM annihilation (e.g. Valdés et al. (2013)). This epoch (assuming X-ray dominated heating) corresponds to the yellow→blue transition in Fig. 42.
5. **Reionization**: as the abundance of *early galaxies* increases, the IGM gradually becomes ionized, a process which is inside-out on large scales. The tomography of this process is sensitive to the nature and clustering of the dominant UV sources (e.g. McQuinn et al. (2007)). The cosmic 21cm signal decreases, approaching zero. This epoch corresponds to the blue→black transition in Fig. 42.

We see that the last three stages are sensitive to early astrophysical sources (and sinks) of cosmic radiation fields, while the first two (the Dark Ages) allow us to probe cosmology at redshifts much lower than recombination.

Current instruments....

5.5.1. What can we learn from the 21-cm signal?

REFERENCES

Abel T., Bryan G. L., Norman M. L., 2002, Science, 295, 93

Bagla J. S., Padmanabhan T., 1997, Pramana, 49, 161

Bahcall J. N., Soneira R. M., 1980, ApJS, 44, 73

Barkana R., Loeb A., 2001, Phys. Rep., 349, 125

Barkana R., Loeb A., 2004, ApJ, 609, 474

Behroozi P., Wechsler R. H., Hearin A. P., Conroy C., 2019, MNRAS, 488, 3143

Belczynski K., Bulik T., Fryer C. L., Ruiter A., Valsecchi F., Vink J. S., Hurley J. R., 2010, ApJ, 714, 1217

Bird S., Ni Y., Di Matteo T., Croft R., Feng Y., Chen N., 2022, MNRAS, 512, 3703

Bolton J. S., Becker G. D., 2009, MNRAS, 398, L26

Bolton J. S., Haehnelt M. G., 2007, MNRAS, 382, 325

Bolton J. S., Haehnelt M. G., Warren S. J., Hewett P. C., Mortlock D. J., Venemans B. P., McMahon R. G., Simpson C., 2011, MNRAS, 416, L70

Bond J. R., Cole S., Efstathiou G., Kaiser N., 1991, ApJ, 379, 440

Bouwens R. J., et al., 2015, ApJ, 803, 34

Bower R. G., Benson A. J., Malbon R., Helly J. C., Frenk C. S., Baugh C. M., Cole S., Lacey C. G., 2006, MNRAS, 370, 645

Bowman J. D., Rogers A. E. E., Monsalve R. A., Mozdzen T. J., Mahesh N., 2018, Nature, 555, 67

Bromm V., Coppi P. S., Larson R. B., 2002, ApJ, 564, 23

Brook C. B., Di Cintio A., 2015, MNRAS, 453, 2133

Cen R., 1992, ApJS, 78, 341

Cen R., Ostriker J. P., 1992, ApJ, 399, L113

Chandrasekhar S., 1943, Reviews of Modern Physics, 15, 1

Cole S., Kaiser N., 1989, MNRAS, 237, 1127

Cooray A., Sheth R., 2002, Phys. Rep., 372, 1

Croton D. J., 2009, MNRAS, 394, 1109

Das A., Mesinger A., Pallottini A., Ferrara A., Wise J. H., 2017, MNRAS, 469, 1166

Dayal P., Ferrara A., Dunlop J. S., Pacucci F., 2014, MNRAS, 445, 2545

De Lucia G., Blaizot J., 2007, MNRAS, 375, 2

Dijkstra M., 2014, ArXiv e-prints:1406.7292

Dijkstra M., Ferrara A., Mesinger A., 2014, MNRAS, 442, 2036

Dijkstra M., Mesinger A., Wyithe J. S. B., 2011, MNRAS, 414, 2139

Dvorkin C., Smith K. M., 2009, Phys. Rev. D, 79, 043003

Eisenstein D. J., Hu W., 1999, ApJ, 511, 5

Emberson J. D., Thomas R. M., Alvarez M. A., 2013, ApJ, 763, 146

Ewall-Wice A., Chang T. C., Lazio J., Doré O., Seiffert M., Monsalve R. A., 2018, ApJ, 868, 63

Ferrara A., Loeb A., 2013, MNRAS, 431, 2826

Fialkov A., Barkana R., 2019, arXiv e-prints, p. arXiv:1902.02438

Fialkov A., Barkana R., Visbal E., Tseliakhovich D., Hirata C. M., 2013, MNRAS, 432, 2909

Field G. B., 1958, Proceedings of the Institute of Radio Engineers, 46, 240

Fragos T., et al., 2013, ApJ, 764, 41

Furlanetto S. R., 2006, MNRAS, 371, 867

Furlanetto S. R., Hernquist L., Zaldarriaga M., 2004, MNRAS, 354, 695

Furlanetto S. R., Oh S. P., 2005, MNRAS, 363, 1031

Furlanetto S. R., Stoever S. J., 2010, MNRAS, 404, 1869

Furlanetto S. R., Zaldarriaga M., Hernquist L., 2004, ApJ, 613, 1

Greig B., Mesinger A., 2017, MNRAS, 465, 4838
Greig B., Mesinger A., Haiman Z., Simcoe R. A., 2017, MNRAS, 466, 4239
Greig B., Mesinger A., McGreer I. D., Gallerani S., Haiman Z., 2017, MNRAS, 466, 1814
Gronke M., Dijkstra M., 2014, MNRAS, 444, 1095
Haardt F., Madau P., 1996, ApJ, 461, 20
Haardt F., Madau P., 2001, in Neumann D. M., Tran J. T. V., eds, Clusters of Galaxies and the High Redshift Universe Observed in X-rays Modelling the UV/X-ray cosmic background with CUBA
Haiman Z., Spaans M., 1999, ApJ, 518, 138
Haiman Z., Thoul A. A., Loeb A., 1996, ApJ, 464, 523
Harikane Y., et al., 2019, ApJ, 883, 142
Heger A., Fryer C. L., Woosley S. E., Langer N., Hartmann D. H., 2003, ApJ, 591, 288
Heinrich C. H., Miranda V., Hu W., 2017, Phys. Rev. D, 95, 023513
Holzbauer L. N., Furlanetto S. R., 2012, MNRAS, 419, 718
Hui L., Gnedin N. Y., 1997, MNRAS, 292, 27
Jenkins A., Frenk C. S., White S. D. M., Colberg J. M., Cole S., Evrard A. E., Couchman H. M. P., Yoshida N., 2001, MNRAS, 321, 372
Kaiser N., 1984, ApJ, 284, L9
Keating L. C., Haehnelt M. G., Becker G. D., Bolton J. S., 2014, MNRAS, 438, 1820
Kimm T., Cen R., 2014, The Astrophysical Journal, 788
Kogut A., et al., 2003, ApJS, 148, 161
Komatsu E., et al., 2011, ApJS, 192, 18
Kuhlen M., Faucher-Giguere C.-A., 2012, MNRAS, 423, 862
Lacey C., Cole S., 1993, MNRAS, 262, 627
Lee H.-W., Blandford R. D., Western L., 1994, MNRAS, 267, 303
Loeb A., Rybicki G. B., 1999, ApJ, 524, 527
Ma X., Hopkins P. F., Garrison-Kimmel S., Faucher-Giguère C.-A., Quataert E., Boylan-Kolchin M., Hayward C. C., Feldmann R., Kereš D., 2018, MNRAS, 478, 1694
Ma X., Quataert E., Wetzel A., Hopkins P. F., Faucher-Giguère C.-A., Kereš D., 2020, arXiv e-prints, p. arXiv:2003.05945
Madau P., Ferguson H. C., Dickinson M. E., Giavalisco M., Steidel C. C., Fruchter A., 1996, MNRAS, 283, 1388
Mao X.-C., 2012, ApJ, 744, 29
Marchi F., et al., 2017, A&A, 601, A73
Mason C. A., Treu T., Dijkstra M., Mesinger A., Trenti M., Pentericci L., de Barros S., Vanzella E., 2017, ArXiv e-prints:1709.05356
McQuinn M., 2015, ArXiv e-prints:1512.00086
McQuinn M., Lidz A., Zahn O., Dutta S., Hernquist L., Zaldarriaga M., 2007, MNRAS, 377, 1043
McQuinn M., Oh S. P., Faucher-Giguère C.-A., 2011, ApJ, 743, 82
McQuinn M., O'Leary R. M., 2012, ApJ, 760, 3
McQuinn M., Upton Sanderbeck P., 2015, ArXiv e-prints:1505.07875
Mesinger A., 2016, Understanding the Epoch of Cosmic Reionization: Challenges and Progress, 423
Mesinger A., Aykutalp A., Vanzella E., Pentericci L., Ferrara A., Dijkstra M., 2015, MNRAS, 446, 566
Mesinger A., Ewall-Wice A., Hewitt J., 2014, MNRAS, 439, 3262
Mesinger A., Ferrara A., Spiegel D. S., 2013, MNRAS, 431, 621
Mesinger A., Furlanetto S., 2007, ApJ, 669, 663
Mesinger A., Furlanetto S., 2009, MNRAS, 400, 1461
Mesinger A., Furlanetto S., Cen R., 2011, MNRAS, 411, 955
Mesinger A., Greig B., Sobacchi E., 2016, MNRAS, 459, 2342
Mesinger A., Haiman Z., 2004, ApJ, 611, L69
Mesinger A., Haiman Z., 2007, ApJ, 660, 923
Mesinger A., McQuinn M., Spergel D. N., 2012, MNRAS, 422, 1403

Michaux Michae Hahn O., Rampf C., Angulo R. E., 2020, Monthly Notices of the Royal Astronomical Society, 500, 663

Milosavljević M., Safranek-Shrader C., 2016, in Mesinger A., ed., Astrophysics and Space Science Library Vol. 423 of Astrophysics and Space Science Library, Star Formation for Predictive Primordial Galaxy Formation. p. 65

Mineo S., Gilfanov M., Sunyaev R., 2012a, MNRAS, 419, 2095

Mineo S., Gilfanov M., Sunyaev R., 2012b, MNRAS, 426, 1870

Miralda-Escude J., 1998, ApJ, 501, 15

Miralda-Escudé J., 2003, ApJ, 597, 66

Miralda-Escudé J., Haehnelt M., Rees M. J., 2000, ApJ, 530, 1

Mitra S., Choudhury T. R., Ferrara A., 2015, MNRAS, 454, L76

Mitra S., Ferrara A., Choudhury T. R., 2013, MNRAS, 428, L1

Mo H. J., Jing Y. P., White S. D. M., 1997, MNRAS, 284, 189

Mo H. J., White S. D. M., 1996, MNRAS, 282, 347

Moore B., Governato F., Quinn T., Stadel J., Lake G., 1998, ApJ, 499, L5

Mortonson M. J., Hu W., 2008, ApJ, 672, 737

Muñoz J. B., Qin Y., Mesinger A., Murray S. G., Greig B., Mason C., 2022, MNRAS, 511, 3657

Murray S. G., Diemer B., Chen Z., Neuhold A. G., Schnapp M. A., Peruzzi T., Blevins D., Engelman T., 2021, Astronomy and Computing, 36, 100487

Mutch S. J., Geil P. M., Poole G. B., Angel P. W., Duffy A. R., Mesinger A., Wyithe J. S. B., 2016, MNRAS, 462, 250

Naoz S., Barkana R., 2005, MNRAS, 362, 1047

Navarro J. F., Frenk C. S., White S. D. M., 1996, ApJ, 462, 563

Noh Y., McQuinn M., 2014, MNRAS, 444, 503

Osterbrock D. E., 1989, Astrophysics of gaseous nebulae and active galactic nuclei. Research supported by the University of California, John Simon Guggenheim Memorial Foundation, University of Minnesota, et al. Mill Valley, CA, University Science Books, 1989, 422 p.

Paardekooper J.-P., Khochfar S., Dalla Vecchia C., 2015, MNRAS, 451, 2544

Pacucci F., Mesinger A., Mineo S., Ferrara A., 2014, MNRAS, 443, 678

Padmanabhan T., 1993, Structure Formation in the Universe

Pallottini A., Ferrara A., Gallerani S., Behrens C., Kohandel M., Carniani S., Vallini L., Salvadori S., Gelli V., Sommovigo L., D'Odorico V., Di Mascia F., Pizzati E., 2022, MNRAS, 513, 5621

Park J., Mesinger A., Greig B., Gillet N., 2019, MNRAS

Pawlik A. H., Rahmati A., Schaye J., Jeon M., Dalla Vecchia C., 2017, MNRAS, 466, 960

Persic M., Salucci P., Stel F., 1996, MNRAS, 281, 27

Planck Collaboration 2018, arXiv e-prints:1807.06209, p. arXiv:1807.06209

Planck Collaboration XIII et al., 2016, A&A, 594, A13

Planck Collaboration XLVI 2016, A&A, 596, A107

Planck Collaboration XLVII 2016, A&A, 596, A108

Portinari L., Chiosi C., Bressan A., 1998, A&A, 334, 505

Press W. H., Schechter P., 1974, ApJ, 187, 425

Price L. C., Trac H., Cen R., 2016, ArXiv e-prints:1605.03970

Rahmati A., Pawlik A. H., Raičević M., Schaye J., 2013, MNRAS, 430, 2427

Reichardt C. L., 2016, in Mesinger A., ed., Astrophysics and Space Science Library Vol. 423 of Astrophysics and Space Science Library, Observing the Epoch of Reionization with the Cosmic Microwave Background. p. 227

Robertson B. E., et al., 2013, ApJ, 768, 71

Rodríguez-Puebla A., Behroozi P., Primack J., Klypin A., Lee C., Hellinger D., 2016, MNRAS, 462, 893

Rybicki G. B., Lightman A. P., 1979, Radiative processes in astrophysics. New York, Wiley-Interscience, 1979. 393 p.

Salpeter E. E., 1955, ApJ, 121, 161

Santos M. R., Ellis R. S., Kneib J.-P., Richard J., Kuijken K., 2004, ApJ, 606, 683

Schaerer D., 2002, A&A, 382, 28
Schauer A. T. P., Glover S. C. O., Klessen R. S., Clark P., 2021, MNRAS, 507, 1775
Schaye J., 2001, ApJ, 559, 507
Schneider A., Giri S. K., Mirocha J., 2021, Phys. Rev. D, 103, 083025
Schroeder J., Mesinger A., Haiman Z., 2013, MNRAS, 428, 3058
Scoccimarro R., Sheth R. K., 2002, MNRAS, 329, 629
Seager S., Sasselov D. D., Scott D., 2000, ApJS, 128, 407
Shapley A. E., Steidel C. C., Pettini M., Adelberger K. L., Erb D. K., 2006, ApJ, 651, 688
Sheth R. K., Mo H. J., Tormen G., 2001, MNRAS, 323, 1
Sheth R. K., Tormen G., 1999, MNRAS, 308, 119
Shull J. M., van Steenberg M. E., 1985, ApJ, 298, 268
Siana B., et al., 2007, ApJ, 668, 62
Simcoe R. A., Sullivan P. W., Cooksey K. L., Kao M. M., Matejek M. S., Burgasser A. J., 2012, Nature, 492, 79
Sobacchi E., Mesinger A., 2014, MNRAS, 440, 1662
Somerville R. S., Kolatt T. S., 1999, MNRAS, 305, 1
Springel V., 2016, in Revaz Y., Jablonka P., Teyssier R., Mayer L., eds, Saas-Fee Advanced Course Vol. 43 of Saas-Fee Advanced Course, High Performance Computing and Numerical Modelling. p. 251
Steidel C. C., Giavalisco M., Dickinson M., Adelberger K. L., 1996, AJ, 112, 352
Steidel C. C., Pettini M., Adelberger K. L., 2001, ApJ, 546, 665
Tacchella S., Bose S., Conroy C., Eisenstein D. J., Johnson B. D., 2018, ApJ, 868, 92
Tseliakhovich D., Hirata C., 2010, Phys. Rev. D, 82, 083520
Tumlinson J., Shull J. M., 2000, ApJ, 528, L65
Upton Sanderbeck P. R., D'Aloisio A., McQuinn M. J., 2016, MNRAS, 460, 1885
Valdés M., Evoli C., Ferrara A., 2010, MNRAS, 404, 1569
Valdés M., Evoli C., Mesinger A., Ferrara A., Yoshida N., 2013, MNRAS, 429, 1705
Vale A., Ostriker J. P., 2006, MNRAS, 371, 1173
Viel M., Haehnelt M. G., Springel V., 2010, JCAP, 6, 015
Vogelsberger M., Marinacci F., Torrey P., Puchwein E., 2020, Nature Reviews Physics, 2, 42
Wolcott-Green J., Haiman Z., Bryan G. L., 2017, MNRAS
Wouthuysen S. A., 1952, AJ, 57, 31
Xu H., Wise J. H., Norman M. L., Ahn K., O'Shea B. W., 2016, ApJ, 833, 84
Yoshida N., Omukai K., Hernquist L., 2008, Science, 321, 669
Yung L. Y. A., Somerville R. S., Popping G., Finkelstein S. L., Ferguson H. C., Davé R., 2019, MNRAS, 490, 2855
Zahn O., Lidz A., McQuinn M., Dutta S., Hernquist L., Zaldarriaga M., Furlanetto S. R., 2007, ApJ, 654, 12
Zahn O., Mesinger A., McQuinn M., Trac H., Cen R., Hernquist L. E., 2011, MNRAS, 414, 727
Zel'Dovich Y. B., 1970, A&A, 5, 84