Mass and energy cascade in collisionless dark matter flow and relevant constraints on the dark matter particle mass

Zhijie (Jay) Xu, 1 *

1 Physical and Computational Sciences Directorate, Pacific Northwest National Laboratory; Richland, WA 99354, USA

ABSTRACT

The cascade phenomenon is fundamental to turbulence, a typical non-equilibrium system. In turbulence, energy is cascaded from large to small scales, where it is dissipated by fluid viscosity. Using Illustris and Virgo simulations, this paper presents the cascade in another non-equilibrium system, the self-gravitating collisionless dark matter flow. The cosmic energy of dark matter decreases with time, as if "dissipated" due to the expanding background. This is facilitated by the energy cascade from large to small haloes via halo merging and from large to small scales in individual haloes via particle migration in fluctuating non-uniform gravitational potential, such that the cosmic energy is "dissipated" on small scales. We identify an inverse mass cascade across haloes of different sizes, which leads to a random walk of haloes in halo mass space. The halo mass function (double-$\lambda$) is naturally given by the corresponding Fokker-Planck equation for halo random walk. Similarly, the random walk of particles in haloes leads to the distribution of particles. The halo density profile (double-$\gamma$) can be analytically derived from the corresponding Fokker-Planck equation for particle random walk. Universal scaling laws were identified that exhibit small-scale permanence for the halo density. The different inner density slopes of simulated haloes can be explained by the nonzero net mass and energy flux in individual haloes. The mass and energy cascade in dark matter flow establishes a statistically steady state to continuously release the system energy and maximize the system entropy. The key feature of this statistically steady state is scale-independent rates of cascade such that the statistical structures of the haloes are self-similar and scale-free, and there is no net accumulation of mass and energy on any intermediate scales. Since the waiting time and jumping length for particle random walk depends on the particle mass $m_p$, new mass constraints can be identified as $m_p \geq 10^{-15}\text{kg}$ or $10^{12}\text{GeV}$. This constraint excludes the standard WIMPs and suggests a heavy dark matter scenario (superheavy right-handed neutrinos, etc.).

Key words: Dark matter; N-body simulations; Theoretical models

Contents

1 Introduction  
2 N-body simulations and numerical data  
3 Cosmic energy evolution  
4 Mass and energy cascade in halo mass space  
4.1 Basic concepts  
4.2 Formulating the mass cascade in mass space  
4.3 Waiting time $\tau_g$ and universal scaling laws  
4.4 Halo random walk and halo mass function  
4.5 Formulating the energy cascade in mass space  
5 Energy cascade in spherical haloes  
5.1 Basic concepts  
5.2 Potential fluctuations and random walk in haloes  
5.3 Formulating the energy cascade in haloes  
5.4 Universal scaling laws and inner density of haloes  
5.5 Particle random walk and halo density profile  
6 Constraints on dark matter particle mass  
7 Conclusions  
A Mass cascade in halo mass space

1 INTRODUCTION

Collisionless systems often have properties suggesting common physical principles that control their motion and evolution. The self-gravitating collisionless fluid dynamics (SG-CFD) is the study of these principles for the motion of collisionless matter under the influence of gravity. Typical examples can be the large-scale gravitational collapse of collisionless dark matter, which is a non-equilibrium self-gravitating collisionless flow problem [1]. Gravitational instability leads to the self-organization of collisionless dark matter particles and the formation and evolution of large-scale structures. Within a CDM (cold dark matter) paradigm [2], the initial density fluctuation has a larger amplitude at smaller scales. The formation of structures starts from the gravitational collapse of small-scale density fluctuations and proceeds hierarchically such that small structures coalesce into large structures in a "bottom-up" fashion. Highly localized, over-dense, and virialized haloes are major manifestations of the nonlinear gravitational collapse [3, 4] and the building blocks of large-scale structures, whose abundance and internal structures have been extensively studied over the last several decades.

The abundance of haloes is often described by a halo mass function, one of the most fundamental quantities to probe large-scale structures and model the structure formation and evolution. The first landmark might be the Press-Schechter (PS) formalism [5, 6], which
allows one to predict the shape and evolution of mass function. The distribution of halo mass is determined by postulating that the probability of forming haloes is related to the amplitude of density fluctuations on that scale. Haloes will form at some mass scale once the smoothed linear density contrast on that scale exceeds a threshold value $\delta_c$. This threshold value must be analytically derived by examining the nonlinear collapse of a spherical top hat over-density \[7, 8\]. Alternative derivations using an excursion set approach (EPS) put the PS formalism on a firmer footing by removing the fudge factor in the original PS model \[6\]. This was further extended to the excursion set with correlated steps \[9, 10, 11\]. Although mathematically less rigorous, PS formalism and its extensions are still very useful and allow one to compute many different structural properties. Examples are the halo mass function, merging rates, and clustering properties.

The halo internal structures are usually characterized by the halo density profile \[12\], which can be analyzed both analytically and through numerical $N$-body simulations \[13, 14\]. Since the first study on the spherical collapse \[8\], the power-law density profile was proposed using the self-similar approximation \[15, 16\]. High-resolution $N$-body simulations have revealed a nearly universal profile exhibiting a cuspy density—less steep than the isothermal profile at smaller radii and becomes steeper at larger radii \[17, 18\]. However, there has yet to be a consensus on the exact slope value of the inner density slope $s$ obtained via N-body simulations. Since the first slope $s = -1.0$ described by the $N$FW profile \[17\], inner slopes of the simulated haloes have varied in a wide range, including values such as $s > -1.0$ \[19\], $s = -1.2$ \[20\], and $s \approx -1.3$ \[21, 22, 23\]. When compared with observations, the predicted cuspy inner density tend to be steeper \[24, 25, 26, 27\]. This gives rise to the cusp-core problem, one of the so-called small scale challenges.

To the author’s knowledge, the exact origin of the nearly universal density profile and the different inner density slopes of simulated haloes are not yet fully understood \[4\]. This paper treats the halo mass function and density profiles from a different perspective. By revisiting the cascade in turbulence, we reveal the existence of a mass and energy cascade in the collisionless dark matter flow. The mass cascade describes the mass transfer between haloes on different mass scales, which leads to the distribution of haloes with respect to the halo mass, i.e., the halo mass function. The energy cascade in haloes describes the energy flow along the halo radial direction, which suggests the distribution of particles in haloes, i.e., the halo density profile. This new perspective offers a theory for the near-universal halo mass functions and density profiles, as well as the widespread inner density slopes of simulated haloes.

Turbulence is a typical non-equilibrium system. Constant chaotic fluid motion and energy flowing through different scales prevent turbulent from reaching a stable equilibrium state. The energy cascade is fundamental to turbulence. At high Reynolds numbers, turbulence consists of a random collection of eddies (building blocks) at different length scales that interact with each other and dynamically change, which can be described by a famous poem \[28\]:

- Big whirls have little whirls, That feed on their velocity;
- And little whirls have lesser whirls, And so on to viscosity.

The poem describes a conceptual picture that large eddies feed smaller eddies, which feed even smaller eddies and then lead to viscous dissipation at the smallest scale, i.e., the concept of a direct energy cascade. There is a broad spectrum of eddy sizes within fully developed turbulence. Large eddies are usually created by the instability of large-scale mean flow at integral scale $L$. They rapidly break up and pass their kinetic energy to smaller eddies due to inertial force. Smaller eddies are transient and, in turn, pass their energy to even smaller eddies. The cascade continues down the scale and stops to operate at the smallest eddies (dissipation scale $\eta$), where the viscous force becomes dominant over the inertial force. At high Reynolds numbers, there exists a range of length scales where the viscous force is negligible and the inertial force is dominant. The rate $\varepsilon$ of energy passing down the cascade should be scale-independent in this range and match exactly the rate of energy dissipation at the smallest scale. The direct energy cascade is a dominant feature of three-dimensional turbulence. However, an inverse energy cascade was predicted for two-dimensional turbulence, where kinetic energy is transferred from small to large scales \[29\].

From this brief description, turbulence has two key features enabling the cascade phenomenon: i) A broad spectrum of eddies that mediate the energy cascade across different scales; ii) The viscous force operating on small scales to dissipate the system energy into heat (radiation) and maintain a steady energy cascade at a constant rate. Apparently, as another example of the non-equilibrium system, the self-gravitating collisionless dark matter flow (SG-CFD) shares some striking similarities with turbulence.

First, there also exists a broad spectrum of halo sizes to maximize the system entropy \[30\]. Small haloes are created by gravitational instability and interacting and merging with other haloes. Haloes pass their mass onto larger and larger haloes. Consequently, we expect a continuous cascade of mass from small to large mass scales. By simply replacing "whirls" in that poem with "haloes," a new poem now describes the inverse mass cascade from small to large scales that is consistent with the hierarchical structure formation:

- Little haloes have big haloes, That feed on their mass;
- And big haloes have greater haloes, And so on to growth.

Second, the total energy of the collisionless dark matter in an expanding background constantly decreases with time, as shown by the cosmic energy evolution (Section 3). Since there is no viscous force operating in collisionless dark matter, the decrease in total energy is caused by the Hubble expansion (Eq. (6)). In this regard, the Hubble parameter $H$ plays a "similar" role as the viscosity in turbulence that leads to the energy "dissipation" in collisionless dark matter flow. The quotation indicates the "dissipation" here is different.

Based on these similarities, we hypothesize the existence of cascade phenomena in the collisionless dark matter flow (SG-CFD). However, due to the collisionless nature and long-range gravity, we also expect the cascade in SG-CFD to be unique and distinct from turbulence. This paper will develop and formulate the ideas of mass and energy cascades in SG-CFD with the help of large-scale cosmological simulations. Despite the path already explored by many pioneers, these ideas provide a new understanding of the two important aspects of gravitational collapse: the abundance and internal structure of haloes. The rest of the paper is organized as follows: Section 2 introduces the simulation and numerical data. Section 3 presents the cosmic energy evolution that demonstrates the decrease of the total energy of dark matter in an expanding background. Section 4 presents the mass and energy cascade in halo mass space that leads to the halo random walk, halo mass functions, and scaling laws in mass space. Section 5 presents the energy cascade in individual haloes that lead to the particle random walk, halo density profiles, and universal scaling laws in haloes, followed by the Section 6 discussing the potential constraints on dark matter particle mass.

2 N-BODY SIMULATIONS AND NUMERICAL DATA

The large-scale cosmological Illustris simulation (Illustris-1-Dark) \[31\] was used to demonstrate and validate the concepts. Illustris is
a suite of large-volume cosmological dark matter only and hydro-dynamical simulations. The selected Illustris-1-Dark is dark matter only simulations that has a cosmological volume of a 106.5Mpc$^3$ and 1820$^3$ DM particles for a high mass resolution. Each DM particle has a mass around $m_p = 7.6 \times 10^6 M_{\odot}$. The gravitational softening length is around 1.4 kpc. The simulation has cosmological parameters of dark matter density $\Omega_{DM} = 0.2726$, dark energy density $\Omega_{DE} = 0.7274$ at $z = 0$, and Hubble constant $h = 0.704$. Some key parameters of N-body simulations are listed in Table 1.

For cross-validation, the cosmological N-body simulations carried out by the Virgo consortium were also used. A comprehensive description can be found in [32, 33]. The SCDM simulation of the Virgo consortium focuses on matter-dominant dark matter-only simulations with a standard CDM power spectrum (SCDM). The same set of simulation data has been widely used in a number of different studies, from clustering statistics [33] to the formation of halo clusters in large-scale environments [34], and testing models for halo abundance and mass functions [35]. This simulation has a lower mass resolution with particle mass $m_p = 2.27 \times 10^{11} M_{\odot}/h$. The simulation box is around 240 Mpc/h, where $h$ is the dimensionless Hubble constant in the unit of 100 km/s/Mpc.

The friends-of-friends algorithm (FOF) was used to identify all haloes in each simulation that depend only on a dimensionless parameter $b$, which defines the linking length $b (N/V)^{1/3}$, where $V$ is the volume of the simulation box. In this work, haloes were identified with a linking length parameter of $b = 0.2$. By identifying all haloes of different sizes, all dark matter particles were divided into halo particles with a total mass $M_h$ and out-of-halo particles that do not belong to any halo. Therefore, $M_h$ is the total mass of all haloes. We focus on the evolution of mass and energy in haloes of different mass $m_h$. All haloes were grouped into halo groups of various sizes according to the halo mass $m_h$ (or $n_p$, the number of particles in the halo), where $m_h = n_p m_p$. The total mass for a halo group of mass $m_h$ is $m_g = m_h n_k$, where $n_k$ is the number of haloes in that group. Two different cosmological simulations were used to demonstrate the fundamental concepts. The same approach can be easily extended to other cosmological simulations.

### Table 1. Virgo (SCDM) and Illustris simulation (Illustris-1-Dark) parameters

| Run | $\Omega_0$ | $h$ (Mpc/h) | $L$ (Kpc/h) | $m_p$ ($M_{\odot}/h$) | $t_{soft}$ (Gyr) |
|-----|------------|-------------|-------------|----------------------|-----------------|
| SCDM | 1.0        | 0.5         | 239.5       | 256$^3$              | 2.27 x 10$^{11}$ |
| Illustris | 0.24   | 0.704       | 75          | 1820$^3$            | 5.28 x 10$^9$   |

3 COSMIC ENERGY EVOLUTION

To better understand the cascade phenomenon, we first provide the energy evolution in self-gravitating collisionless dark matter flow. The equations of motion for $N$ collisionless particles in comoving coordinates $x$ and physical time $t$ read [36]:

$$\frac{d^2x_i}{dt^2} + 2H \frac{dx_i}{dt} = -\frac{Gm_p}{a^3} \sum_{j \neq i} \frac{x_i - x_j}{|x_i - x_j|^3},$$

where $N$ particles have equal mass $m_p$. The Hubble parameter $H(t) = \dot{a}/a$. Here, $H$ has a “damping” effect, which leads to the decrease in total energy of the N-body system (Eq. (6) and Fig. 1). In the matter-dominant era, the Hubble parameter satisfies $Ht = 2/3$.

Next, we will derive the energy evolution based on equations of motion (Eq. (1)). We first introduce a transformed time variable $s$ as $ds/dt = a\rho$, where $p$ is an arbitrary exponent. In terms of the new time variable $s$, the original Eq. (1) can be transformed to

$$\frac{d^2x_i}{ds^2} + \frac{dx_i}{ds} (p + 2) a^{-p} H \equiv a^{-(3+2p)} \mathbf{F}_i,$$

$$\frac{F_i}{m_p} = -Gm_p \sum_{j \neq i} \frac{x_i - x_j}{|x_i - x_j|^3} - \frac{\partial P_s}{\partial x_i},$$

where $\mathbf{F}_i$ is the resultant force on particle $i$ from all other particles, while $P_s$ is the total specific potential energy in comoving coordinates. Equation (2) reduces to the original Eq. (1) when $p = 0$. With $p = -2$, the first-order derivative vanishes in Eq. (2) and $s$ is the time variable for integration in N-body simulations. Setting $p = -1$, $s$ is the conformal time. By setting $p = -3/2$ along with $H_0^2 = H^2 a^3$ for the matter-dominant era, the equation of motion becomes

$$\frac{d^2x_i}{ds^2} + \frac{1}{2} H^2 \frac{dx_i}{ds} = \frac{F_i}{m_p}.$$

In this transformed equation, the scale factor $a$ is not explicitly involved. The time-dependent Hubble parameter $H$ is replaced by a Hubble constant $H_0$. This transformation offers significant convenience in analytically solving the cosmic energy evolution [37].

We first identify the transformation between velocity $v_i$ in time variable $s$ and the peculiar velocity $u_i$,

$$v_i = \frac{dx_i}{ds} = a^{-p} \frac{dx_i}{dt} = a^{-p-1} u_i, \quad u_i = a \frac{dx_i}{dt},$$

$$K_s = K_p a^{-2p-2}, \quad P_s = a P_y,$$

where the kinetic energy $K_s$ and the potential energy $P_s$ in the transformed equation can now be related to the peculiar kinetic energy $K_p$ and the potential energy $P_y$ in the physical coordinates.

The energy evolution of the N-body system can be obtained by multiplying $v_i = d_x/ds$ on both sides of Eq. (2) and adding the equation of motion for all particles together [37]. An exact and simple equation (in time variable $s$) for the specific kinetic energy $K_s$ and the total potential energy $P_s$ can be obtained as

$$\frac{dK_s}{ds} + 2HK_s (p + 2)a^{-p} + a^{-(3+2p)} \frac{dP_s}{ds} = 0.$$  

By setting $p = 0$ and using the relations in Eq. (4), the exact cosmic energy equation for energy evolution of the N-body system reads

$$\frac{dE_x}{dt} + H (2K_p + P_y) = 0,$$

which describes the energy evolution in an expanding background. Here $K_p$ is the peculiar kinetic energy, $P_y$ is the potential energy in physical, coordinates and $E_x = K_p + P_y$ is the total specific energy. This is also known as the Layzer-Irvine equation [38, 39].

The cosmic energy evolution (Eq. (6)) admits a linear solution

$$K_p = -\varepsilon_x t, \quad P_y = \frac{7}{2} \varepsilon_x t, \quad E_y = -\frac{3}{2} \varepsilon_x t,$$

where $\varepsilon_x < 0$ is an important physical constant (unit: $m_{\odot}/h^3$) and the focus of this paper. It represents the rate of time variation of cosmic energy. The solution in Eq. (7) can be directly validated by N-body simulations. Figure 1 presents the energy evolution from Virgo simulation (SCDM). The (specific) kinetic energy $K_p$ and potential energy $P_y$ were calculated as the mean energy of all dark matter particles in the N-body system. Figure 1 confirms the solution in Eq. (7), i.e. a linear increase of $K_p$ with time and a negative potential energy $P_y = -1.4K_p$. Two key messages can be obtained:
Figure 1. The variation of specific kinetic energy $K_p$ (energy per unit mass), potential energy $P_y$, and total energy $E_y = K_p + P_y$ (unit: $km^2/s^2$) with time $t$ (normalized by $t_0$, the present cosmic time) from Virgo SCDM simulation. Solution in Eq. (7) is also presented for comparison with $\epsilon_u = -4.6 \times 10^{-7} m^2/s^3$. Simulation confirms a linear increase of $K_p = -\epsilon_u t$ with time and negative potential energy $P_y = -1.4K_p$. The total energy $E_y = -0.4K_p$, also decreases with time. The figure demonstrates that the total energy of the N-body system in an expanding background decreases with time as if the total energy is "dissipated" at a rate of $0.4\epsilon_u$. This continuous decrease in total energy leads to and maintains the energy cascade in collisionless dark matter. An "effective" potential exponent $n_e = 2K_p/P_y = 10^{-7}$ can be obtained.

i) According to the virial theorem, in virial equilibrium, $2K_p = n_eP_y$, where $n_e$ is an exponent of the interaction potential $r^{n_e}$. For standard gravity, exponent $n_e = -1$. Since gravity is the only interaction in the N-body system, we would expect $n_e = -1$ for an N-body system. However, the virial equilibrium can never be reached by the non-equilibrium N-body system (e.g., SG-CFD of dark matter flow). Instead of the virial equilibrium, a statistically steady state is established to continuously release system energy and maximize system entropy, which is manifested by the mass and energy cascade. In that particular state, the solution of Eq. (7) is valid such that an "effective" exponent $n_e = 10^{-7}$ is obtained, which deviates from -1;

ii) The total energy $E_y = -0.4K_p$ decreases with time. In analogy to hydrodynamic turbulence, the total energy of the N-body system in an expanding background decreases with time, as if "dissipated," even though there is no viscous force in the collisionless dark matter. What is the more fundamental nature of this "dissipation"? Will this energy "dissipation" lead to any dark radiation? These questions require further exploration [40]. Here, we admit that cosmic energy decreases with time and estimate the rate of energy "dissipation":

$$\epsilon_u = -\frac{K_p}{t} = -\frac{3}{\frac{2}{t} \approx -4.6 \times 10^{-7} m^2/s^3},$$

(8)

where $u_0 \equiv u (t = t_0) \approx 350 km/s$ is the one-dimensional velocity dispersion of all dark matter particles, and $t_0 \approx 13.7$ billion years is the physical time at present epoch or the age of the universe. Different simulations may have slightly different values of $u_0$ due to different cosmological parameters. However, the rate of energy "dissipation" $\epsilon_u \approx -10^{-7} m^2/s^3$ should be a good estimate.

Hereafter, the word "dissipation" stands for the decrease in total energy caused by the expanding background instead of the dissipation caused by the viscous force in turbulence. In the statistically steady state, the energy "dissipated" on small scales should balance the energy cascaded from large scales. Therefore, Equation (7) also suggests an inverse cascade of kinetic energy from small to large scales at a rate of $-1.4\epsilon_u$ and a direct cascade of potential energy at a rate of $-0.4\epsilon_u$. The total energy is directly cascaded to small scales at a rate of $-0.4\epsilon_u$ to provide the energy that is "dissipated" on small scales due to the Hubble expansion. In the next sections, we will illustrate how the mass and energy cascade initiate and proceed and their impacts on structure formation and evolution.

4 MASS AND ENERGY CASCADE IN HALO MASS SPACE

4.1 Basic concepts

For collisionless dark matter, long-range gravity requires the formation of a broad spectrum of haloes to maximize the entropy of the self-gravitating collisionless system [30]. Highly localized halo structures are a major manifestation of non-linear gravitational collapse [3, 4]. As the building blocks of dark matter flow (in contrast to the "eddies" in turbulence), haloes facilitate the mass and energy cascade across different halo masses. Figure 2 provides a schematic description of the mass and energy cascade in the halo mass space. This description is fully consistent with the hierarchical structure formation, where haloes grow via a series of merging events.

In Fig. 2, the merging of a halo (mass $M$) with a single merger of mass $m$ (depending on the mass resolution) leads to a mass and energy flux to a larger mass scale $M + m$, that is, the halo of mass $M$ moving into the next mass scale $M + m$ after merging. For a given halo of mass $m_h$, the merge event occurs with an average waiting time $\tau_g (m_h, z)$. On average, haloes of mass $m_h$ need to wait for a specific waiting time $\tau_g$ before merging with a single merger. That waiting time is dependent on the mass of haloes. Larger haloes tend to accrete mass faster with a shorter waiting time ($\tau_g \propto m_h^{-1}$ in Eq. (19)) with halo geometry parameter $\lambda = 2/3$. 

Vol. 000, 1–28 (2024)
For a group of \( n_h \) haloes of the same mass \( m_h \), the average waiting time for a merge event of any halo in that group can be found as \( \tau_h(m_{h1},z) = \tau_h/n_h \). A simple explanation is provided here. Let us randomly pick one halo in that group to merge in every round and repeat until the given halo is picked. The probability of picking that given halo in the first round is \( 1/n_h \), and the corresponding waiting time for that specific halo is \( \tau_h \). The probability of picking that given halo in the second round is \( (n_h - 1)/n_h + 1/n_h \) (product of probability not selected in the first round and the probability selected in the second round). The corresponding waiting time is \( 2\tau_h \). Therefore, the waiting time \( \tau_g \) for a given halo is

\[
\tau_g = \sum_{k=1}^{\infty} \frac{k \tau_h}{n_h} \left( \frac{n_h - 1}{n_h} \right)^{k-1} = \frac{\tau_h}{n_h} + \frac{\tau_h}{n_h} - \frac{(2\tau_h)}{n_h} + \ldots = n_h \tau_h. \tag{9}
\]

Based on this picture, the rate of mass cascade (or mass flux) in halo mass space can be conveniently defined as

\[
\Pi_m(m_h, z) = \frac{m_h}{\tau_{m}} = \frac{n_h m_h}{\tau_{g}} = -\frac{m_h}{\tau_{g}}, \tag{10}
\]

\( \epsilon_m(z) = \Pi_m(m_h, z) \) for \( m_h \leq m_h^* \), i.e., the mass of halo \( m_h \) is passed to a larger mass scale during a period of \( \tau_g \). Here, \( m_g = n_h m_h \) is the total mass for \( n_h \) haloes of the same mass \( m_h \) (or the mass of the halo group) such that \( \tau_g \) is the time required to pass the entire group of mass \( m_g \) to a larger mass scale. The negative sign represents the inverse cascade from small to large scales, in contrast to the direct cascade from large to small scales. The rate of the mass cascade is scale independent; that is, \( \epsilon_m \) is independent of the mass scale \( m_h \) in the mass range \( m_h < m_h^* \). The formulation of \( \epsilon_m \) in the mass cascade and the results of the N-body simulations are presented in Section 4.2 (Figs. 3 and 4).

Accompanied by the mass cascade, there exists also a simultaneous energy flux across haloes of different scales. To understand this, we will first demonstrate the energy "dissipation" that is intrinsically associated with every merging of haloes. This provides a small-scale mechanism for the "dissipation" of the total cosmic energy (Fig. 1). As energy continuously "dissipates" via halo merging, an energy cascade in halo mass space is required to balance the "dissipation". Here, we consider the two-body merging for haloes of mass \( M_1 \) and \( M_2 \), and size \( r_{h1} \) and \( r_{h2} \), respectively. After the merging, the new halo has a mass of \( M_{h1} = M_1 + M_2 \) and size \( r_{h3} \). We calculate the difference in the total energy before and after the merging. From the virial theorem, the halo kinetic energy and potential energy are related as \( KE = n_e K/2 \). Therefore, the change in the total energy upon the merging reads

\[
\Delta E = -\left(1 + \frac{n_e}{2}\right) \frac{GM_1^2}{r_{h1}} + \left(1 + \frac{n_e}{2}\right) \frac{GM_2^2}{r_{h2}} + \frac{GM_1^2 + GM_2^2}{r_{h1}} \tag{11}
\]

The density of haloes is related to the mean density \( \bar{\rho}_{DM} \) as

\[
M_{\frac{1}{3}} = \frac{M_2}{r_{h1}^3} = \frac{M_3}{r_{h2}^3} = \frac{4}{3}\pi \Delta c \bar{\rho}_{DM}, \tag{12}
\]

where \( \Delta_s = 18\pi^2 \) is the critical density ratio from the spherical collapse model. Substituting Eq. (12) into Eq. (11), the change in the total energy due to the halo merging reads

\[
\Delta E = \left(1 + \frac{n_e}{2}\right) \left(\frac{4}{3} \pi \Delta c \bar{\rho}_{DM}\right)^{\frac{1}{3}} \frac{GM_1^2}{r_{h1}} \left(1 + \gamma^2 - (1 + \gamma)^{\frac{1}{2}}\right), \tag{13}
\]

where \( \gamma = M_2/M_1 \) is the mass ratio between two haloes. Clearly, the change in total energy due to the merging is always negative or \( \Delta E < 0 \). Since a closed system tends to minimize its energy and maximize its entropy, the structure formation must proceed in a "bottom-up" fashion via halo merging to release system energy, not the "top-down" fashion. The halo merging provides a viable mechanism to "dissipate" the total cosmic energy in an expanding background. For two haloes with a large mass ratio or \( \gamma = M_2/M_1 \ll 1 \),

\[
\Delta E = -\frac{5}{3} \left(1 + \frac{n_e}{2}\right) \frac{GM_1^2}{r_{h1}} \frac{GM_2^2}{r_{h2}} < 0, \tag{14}
\]

where the energy change is proportional to \( M_1^2/M_2 \).

In halo mass space, the kinetic energy is cascaded from small to large mass scales at a rate of \( \dot{\epsilon}_u \) (the same direction as the mass cascade). In contrast, the potential energy is cascaded in the opposite direction at a rate of \( -\dot{\epsilon}_u \). Similarly to the mass flux in Eq. (10), the flux of the specific kinetic energy across haloes reads (also see Eq. (40))

\[
\Pi_{pv}(m_h, z) = -\frac{m_h}{\tau_{m}} \frac{K_{pv}}{\Lambda_m} = -\frac{\Pi_{m}}{\Lambda_m} K_{pv}, \tag{15}
\]

\( \epsilon_u = \Pi_{pv}(m_h, z) \) for \( m_h \leq m_h^* \), where \( K_{pv} \) is the mean specific kinetic energy in all haloes greater than \( m_h^* \) (defined in Eq. (38)), while \( \Lambda_m \) is the total mass in all haloes greater than \( m_h \) (defined in Eq. (17)). For every waiting time of \( \tau_p \), the energy cascade leads to an increase in the specific kinetic energy of \( m_h K_{pv}/\Lambda_m \) on all scales larger than \( m_h \). The formulation of \( \dot{\epsilon}_u \) in the energy cascade and the results of the N-body simulation are presented in Section 4.5 (Fig. 8).

In summary, the rate of the mass cascade \( \epsilon_m \), the rate of the energy cascade \( \epsilon_u \), and the waiting time \( \tau_g \) are three central quantities that we will focus on. The mass and energy cascade establishes a statistically steady state to continuously release the energy of the system and maximize the entropy [30]. When the dark matter flow reaches that statistically steady state, the rates of mass and energy cascade must be scale independent (i.e., \( \epsilon_m \) and \( \epsilon_u \) are independent of mass scale \( m_h \)). If this is not the case, there would be a net accumulation of mass and energy on some intermediate-scale mass below \( m_h^* \). The mathematical formulations of the mass and energy cascade are presented in Sections 4.2 and 4.5.

### 4.2 Formulating the mass cascade in mass space

In this section, we present a quantitative description of the mass cascade in the halo mass space. The energy cascade in the mass space can be formulated in a similar way (Section 4.5). First, we introduce the mass flux (\( \Pi_m \)) across haloes of different sizes

\[
\Pi_m(m_h, t) = -\int_{m_h}^{\infty} \frac{\partial}{\partial t} \left[ M_h(t) f_M(m, m_h^*) \right] dm, \tag{16}
\]

where \( M_h \) is the total mass in all haloes of all sizes. Here, \( f_M \) is the halo mass function. The cumulative mass function \( \Lambda_m(m_h, t) \) for the total mass in all haloes greater than \( m_h \) reads

\[
\Lambda_m(m_h, t) = \int_{m_h}^{\infty} M_h(t) f_M(m, m_h^*) dm. \tag{17}
\]

The total mass in all haloes can be obtained by setting \( m_h \rightarrow 0 \) in Eq. (17), i.e., \( M_h(t) = \Lambda_m(m_h \rightarrow 0, t) \). The time derivative of \( \Lambda_m \) describes the mass flux from all haloes below the scale \( m_h \) to all haloes above the scale \( m_h \), i.e., the rate of mass cascade \( \Pi_m \) in Eq. (16). In N-body simulations, we use the difference of \( \Lambda_m \) at
two different redshifts $z_1$ (or $t_1$) and $z_2$ (or $t_2$) to compute the time derivative and hence the mass flux $\Pi_m$. Figure 3 presents the variation of the mass flux $\Pi_m(m_h,z)$ from the Illustris simulation. The mass flux $\Pi_m(m,a)$ can be computed from the cumulative mass function $\Lambda_m$ at two different redshifts (see Eq. (16)). We propose a scale-independent rate of cascade $\epsilon_m(z)$ for $m_h < m_h^*$ in the propagation range since we require that the statistical structures of the haloes be self-similar and scale-free for haloes smaller than $m_h^*$. The simulation results confirm this concept. After reaching a statistically steady state around $z = 4$, a propagation range is formed with a scale-independent $\epsilon_m(a)$ for $m_h$ below a critical mass scale $m_h^*$. That rate $\epsilon_m(a) \propto a^{-1}$, that is, decreases with time and is about $0.02M_\odot/s$ at $z = 0$. Haloes in the propagation range pass their mass to larger haloes, while the group mass $m_g$ (total mass of all haloes with the same mass $m_h$) remains constant (see Fig. 5). The energy cascade is shown in Fig. 8.

4.3 Waiting time $\tau_g$ and universal scaling laws

The mass cascade involves a waiting time $\tau_g$ for haloes to merge and migrate from one mass scale to the neighboring mass scale (Eq. (10)). Combining $\tau_g$ with the scale-independent rate of the mass cascade $\epsilon_m$, some important universal scaling laws can be easily identified.

The waiting time $\tau_g$ is inversely proportional to the halo potential $\tau_g \propto \Phi_h^{-1} = (Gm_h/r_h)^{-1}$ [41], where $r_h$ is the size of that halo. This is also explained in Section 5.2 (Eq. (55)), where the halo merging and accretion can be related to the particle migration in haloes. Larger haloes with a higher gravitational potential tend to accumulate mass more quickly with a shorter waiting time $\tau_g$. Putting all of these together, we can write

$$\frac{dm_h}{dt} = \frac{m_p}{\tau_g},$$

$$\tau_g \propto \Phi_h^{-1} = (Gm_h/r_h)^{-1},$$

$$\rho_h = \frac{m_h}{\frac{4}{3} \pi r_h^3} = \Delta_c \rho_{DM} \propto a^{-3},$$

where $m_p$ is the mass of a single merger (or dark matter particle). The first equation describes that, on average, the halo accretes a mass of $m_p$ during every waiting time $\tau_g$. This equation describes the average mass accretion. The second equation quantifies the waiting time. The third equation relates the average halo density to the mean density $\rho_{DM}$, where $\Delta_c = 18\pi^2$ is the critical density ratio. Solving these equations leads to:

$$\tau_g \propto m_h^{-2/3} \quad \text{and} \quad r_h \propto m_h^{1/3}.$$  

Therefore, for a halo with a given mass $m_h$, the waiting time can be written as $\tau_g \propto a m_h^{-\lambda}$, where $\lambda$ is a key halo parameter and $\lambda = 2/3$ is expected for large haloes.

From the definitions of the rate of mass cascade $\epsilon_m$ and the halo group mass $m_g$ (Eq. (10)), we have

$$-\epsilon_m(z) = \frac{m_g(m_h)}{\tau_g(a,m_h)} = \frac{d}{dt}M_h(t)$$

$m_g(m_h) = M_h(t) f_M(m_h, z)m_p = n_h m_h$.

Here, $n_h$ is the number of haloes with the same mass $m_h$. The mass of the halo group $m_g$ is independent of the redshift $\epsilon$ due to the scale-independent mass cascade. That is, the mass flux into a halo group is equal to the mass flux out of the same group, so the mass of the group $m_g$ is a function of $m_h$ only and is constant over time. In addition, with mass injected from the smallest scale $m_p$ at a scale-independent
rate of $\varepsilon_m$, the rate of the mass cascade also equals the rate of change in the total mass $M_h(t)$ in all haloes. Combining Eq. (20) with Eq. (19) for $\tau_g$, universal scaling laws can be identified when the system reaches the statistically steady state with a scale-independent rate $\varepsilon_m$ for mass cascade. First,

$$m_g(m_h) = -\varepsilon_m(z)\tau_g \propto -\varepsilon_m(z) a m_h^{-2/3},$$

leads to

$$\varepsilon_m(z) \propto a^{-1}, \quad M_h \propto a^{1/2}, \quad m_g \propto m_h^{-2/3}, \quad n_h \propto m_h^{-5/3}.$$  (21)

The halo group mass follows a simple power law $m_g \propto m_h^{-2/3}$ that is redshift independent. This is the so-called small-scale permanence (see Fig. 5), i.e., the group mass $m_g$ at different redshifts $z$ collapses into the same scaling law $m_g \propto m_h^{-2/3}$.

Using Eq. (20), the halo mass function $f_M$ satisfies

$$f_M(m_h, z) = \frac{m_g(m_h)}{M_h(t) m_p} \propto m_h^{-2/3} a^{-1/2},$$

(22)

where the halo mass function follows a simple power law in propagation range $m_h < m_h^*$. By absorbing the redshift dependence into the characteristic mass $m_h^*$, we may express the halo mass function in a very simple power-law form (also see Eq. (30)). Combined with Eq. (22), the scaling for the evolution of $m_h^*$ can be obtained,

$$f_M(m_h^*, m_h) \propto \frac{1}{m_h} \left( \frac{m_h}{m_h^*} \right)^{-\lambda} \text{ leads to } m_h^* \propto a^{1/2}. $$

(23)

Finally, using Eqs. (19) ($\lambda = 2/3$) and Eq. (23), for haloes with the characteristic mass $m_h^*$ and size $r_h^*$, we found the scaling

$$r_h^* \propto t, \quad m_h^* \propto t \quad \text{and} \quad \tau_g^* \propto t^{\lambda}.$$  (24)

4.4 Halo random walk and halo mass function

The power law mass function in the propagation range can be derived from universal scaling laws (Eq. (22)), while the complete halo mass function in the entire range of mass scales can be derived from the random walk of haloes in mass space.

First, the inverse mass cascade leads to a random walk of haloes in the mass space that mimics the random work of particles for diffusion. As shown in Fig. 2, haloes continuously migrate in mass space from one scale ($m_h$) to the neighboring scale ($m_h + m$) by merging with single mergers of mass $m$. There is a waiting time $\tau_g$ associated with the halo random walk. The waiting is position-dependent, i.e., dependent on the position $m_h$ in mass space (Eq. (19)).

Second, the particle random walk for diffusion leads to an evolving particle distribution according to the Fokker-Planck equation. Similarly, the random walk of haloes in mass space leads to a probability distribution to find a halo at a given mass. The distribution of haloes with respect to the halo mass, i.e., the halo mass function, is naturally obtained from the corresponding Fokker-Planck equation.

This approach is mathematically convenient and reveals some fundamental aspects of the halo mass function. The evolving halo mass function continuously maximizes the system entropy contained in the distribution of the halo mass, just as the evolving particle distribution maximizes the entropy contained in the particle distribution. Mathematically, the random walk of haloes in mass space can be
described by the stochastic Langevin equation [41]
\[
\frac{\partial m_h(t)}{\partial t} = \sqrt{2D_p} (m_h) \zeta(t) \propto \frac{m_p}{\tau_g} \propto \frac{m_p}{m_H^\lambda}.
\] (25)

Unlike the standard particle random walk for diffusion, the waiting time \( \tau_g \) for the halo random walk depends on the position in the mass space (or the halo mass \( m_h \)). For a power-law waiting time \( \tau_g \propto m_h^{-\lambda} \) (\( \lambda = 2/3 \) from Eq. (19)), the position-dependent diffusivity should take the form of (comparing Eq. (25) to Eq. (18))
\[
D_p (m_h) = D_{p0} (t) m_h^{2\lambda}.
\] (26)

Here, \( D_{p0} (t) \) is a proportional constant for diffusivity \( D_p \). The white Gaussian noise \( \zeta(t) \) satisfies the covariance \( \langle \zeta(t) \zeta(t') \rangle = \delta(t - t') \) with zero mean \( \langle \zeta(t) \rangle = 0 \).

Next, in Stratonovich’s interpretation [42, 43], the Langevin equation (Eq. (25)) yields the Fokker-Planck equation for the halo mass function \( f_M (m_h, t) \) (in analogue to the particle diffusion):
\[
\frac{\partial f_M (m_h, t)}{\partial t} = D_{p0} \frac{\partial}{\partial m_h} \left[ \frac{m_p}{\tau_g} \frac{\partial}{\partial m_h} \left( m_h^2 f_M (m_h, t) \right) \right].
\] (27)

Finally, the halo mass function \( f_M (m_h, t) \) can be analytically solved from Eq. (27), which is a stretched Gaussian with an exponential cut-off at large \( m_h \) and a power-law at small \( m_h \),
\[
f_M (m_h, t) = \frac{m_h^{-\lambda}}{\sqrt{\pi D_{p0} t}} \exp \left[ -\frac{m_h^{2-2\lambda}}{4 (1 - \lambda)^2 D_{p0} t} \right].
\] (28)

The characteristic mass \( m_h^* \) can be related to the mean square displacement in mass space as
\[
\left\langle m_h^2 \right\rangle = \int_0^\infty f_M (m_h, t) m_h^2 dm_h
= \frac{1}{\sqrt{\pi}} \Gamma \left[ \frac{3 - \lambda}{2 - 2\lambda} \right] \frac{1}{\left( 1 - \lambda \right)^2 D_{p0} t} \gamma_0 m_h^{2\lambda}.
\] (29)

where \( \gamma_0 \) is a proportional constant. With \( \lambda < 1 \) and the exponent \( 1/(1 - \lambda) \geq 1 \) in Eq. (29), the random walk of haloes in mass space is essentially a super-diffusion.

Now, the halo mass function can be rewritten as
\[
f_M (m_h, t) = \frac{1 - \lambda}{m_h^{\sqrt{\eta_0}}/m_h^{\lambda}} \exp \left[ -\frac{1}{4 \eta_0} \left( \frac{m_h}{m_h^*} \right)^{2-2\lambda} \right],
\] (30)

where the dimensionless constant
\[
\eta_0 = \frac{1}{4} \left[ \frac{\gamma_0}{\Gamma ((3 - \lambda)/(2 - 2\lambda))} \right]^{-1-\lambda}.
\] (31)

The time dependence of \( f_M \) is absorbed into \( m_h^* \). Clearly, Eq. (30) reduces to the Press-Schechter (PS) mass function if \( \lambda = 2/3 \) and \( \eta_0 = 1/2 \) or \( \gamma_0 = 15 \). Note that this approach for deriving the halo mass function does not rely on any collapse model (either spherical or ellipsoidal). By introducing the dimensionless variable \( \nu = (m_h/m_h^*)^{3/2} \), mass function can be expressed in terms of \( \nu \)
\[
\nu = (m_h/m_h^*)^{3/2},
\]
and the dimensionless double-\( \lambda \) mass function reads,
\[
\nu = \frac{p}{2} \frac{1}{\sqrt{\eta_0}} \exp \left( \frac{\nu^p}{4 \eta_0} \right),
\] (32)

where parameter \( p = 3 \) \((1 - \lambda)\) is on the order of unity.

In principle, two different \( \lambda \) (that is, double-\( \lambda \)) are required for two different ranges (propagation and deposition range). The halo mass function in Eq. (30) can be naturally generalized to a double-\( \lambda \) halo mass function with \( \lambda_1 = 2/3 \) for propagation and deposition ranges, respectively [41]. Based on this idea, a double-\( \lambda \) mass function can be proposed based on the single \( \lambda \) mass function in Eq. (30),
\[
f_M (m_h, z) = \frac{(2 \sqrt{\eta_0})^{-q} 2}{q!(q/2)} \frac{(1 - \lambda_1)}{\lambda_2} \frac{m_h^{\lambda_1}}{m_h^{\lambda_2}} \exp \left[ -\frac{1}{4 \eta_0} \left( \frac{m_h}{m_h^*} \right)^{2-2\lambda_2} \right],
\] (33)

where \( q = (1 - \lambda_1)/(1 - \lambda_2) \) is the ratio between two \( \lambda \) values to satisfy the normalization constraint. The 4th order moment of the double-\( \lambda \) mass function is
\[
\int_0^\infty f_M (m_h, z) (m_h^*)^k dm_h
\]
\[
= \frac{q/2}{\Gamma ((1 + q/2)/(2 - 2\lambda_2))} \Gamma \left( \frac{q}{2} + \frac{k}{2 - 2\lambda_2} \right) (m_h^*)^k.
\] (34)

With the dimensionless variable defined as \( \nu = (m_h/m_h^*)^{2-2\lambda_2} \), the dimensionless double-\( \lambda \) mass function reads,
\[
f_{DA} (\nu) = \frac{(2 \sqrt{\eta_0})^{-q}}{q!(q/2)} \frac{1}{\nu q/2} \nu^{q/2-1} \exp \left( -\frac{\nu}{4 \eta_0} \right).
\] (35)

Figure 7 plots different dimensionless mass functions (log(f(\nu))) (Appendix A6)) compared to the mass function obtained from the Virgo simulation (solid blue). The PS mass function overestimates the mass in large haloes. The fitted JK mass function matches the simulation for a given range of halo size but not the entire range. The double-\( \lambda \) mass function (Eq. (35)) matches both the simulation and ST mass function for the entire range.
4.5 Formulating the energy cascade in mass space

This section presents a quantitative description of the energy cascade in the halo mass space. To better describe the energy cascade, we first decompose the kinetic energy of the halo particles into two parts of different nature. In N-body simulations, every halo particle can be characterized by a mass $m_p$ and a velocity vector $v_p$. The particle velocity $v_p$ can be decomposed as [30]

$$v_p = v_h + v_p', \quad (36)$$

namely, the halo mean velocity, $v_h = \langle v_p \rangle_h$, and the velocity fluctuation, $v_p'$. Here, $\langle \rangle_h$ represents the average of all particles in the same halo, and $v_h$ represents the velocity of that halo. Consequently, the total kinetic energy $K_p$ of a given halo particle can be divided into $K_p = K_{ph} + K_{pv}$. Here, $K_{ph} = \frac{1}{2} m_h v_h^2$ (halo kinetic energy) is the contribution from the motion of entire haloes $v_h$ due to the inter-halo interaction of that particle with all other particles in different haloes and all out-of-halo particles. This part of the kinetic energy is related to interactions on large scales that are still in the nonlinear regime.

The other part, $K_{pv} = \frac{1}{2} v_p'^2 / 2$ (virial kinetic energy), is the contribution of the velocity fluctuation $v_p'$ due to the intra-halo interaction of that particle with all other particles in the same halo. This part of the kinetic energy is from halo virialization and is due to interactions on a shorter distance and smaller scales in the non-linear regime. Similarly to the energy cascade associated with nonlinear interactions in turbulence, the energy cascade in dark matter flow focused on the cascade of the virial kinetic energy $K_{pv}$ due to the nonlinear interactions on small scales.

Similar to the cumulative mass function $\Lambda_m$ in Eq. (17), the cumulative kinetic energies ($\Lambda_{ph}$ and $\Lambda_{pv}$) represent the total kinetic energies $K_{ph}$ and $K_{pv}$ in all haloes greater than $m_h$, such that

$$\Lambda_{ph}(m_h, t) = \int_{m_h}^{\infty} M_h(t) f_M(m, m_h^*) K_{ph} dm,$$

$$\Lambda_{pv}(m_h, t) = \int_{m_h}^{\infty} M_h(t) f_M(m, m_h^*) K_{pv} dm, \quad (37)$$

$$K_{ph} = \frac{3}{2} \sigma_h^2 \langle m \rangle \quad \text{and} \quad K_{pv} = \frac{3}{2} \sigma_{v'}^2 \langle m \rangle. \quad (38)$$

Here, $\sigma_h^2$ and $\sigma_{v'}^2$ are the dispersion of the one-dimensional velocity for the halo velocity $v_h$ and the velocity fluctuation $v_p'$.

Next, we will use the cumulative kinetic energy and cumulative mass $\Lambda_m$ to formulate the mean specific halo kinetic energy (energy per unit mass) $K_{ph}$ and the virial kinetic energy $K_{pv}$ in all haloes above any mass scale $m_h$, that is,

$$\overline{K_{ph}}(m_h, t) = \frac{\Lambda_{ph}}{\Lambda_m} \quad \text{and} \quad \overline{K_{pv}}(m_h, t) = \frac{\Lambda_{pv}}{\Lambda_m}. \quad (38)$$

We will focus on the energy cascade of the specific virial kinetic energy $K_{pv}$ due to nonlinear interactions on small scales. For the inverse mass cascade in Eq. (16), the change in total halo mass above the scale $m_h$, that is, the cumulative mass function $\Lambda_m(m_h, a)$, comes entirely from the mass cascade or the interactions between all haloes below the scale $m_h$ and all haloes above $m_h$. Similarly, the change in the (specific) virial kinetic energy $K_{pv}$ for all haloes above the scale $m_h$ comes entirely from the energy cascade due to interactions between haloes below and above the scale $m_h$. Therefore, similar to the mass cascade $\Pi_m$ in Eq. (16), the rate of cascade for the virial kinetic energy $K_{pv}$ reads

$$\Pi_{pv}(m_h, t) = -\frac{\partial}{\partial t} \left( \overline{K_{pv}} \right) = -\frac{\partial}{\partial t} \left( \frac{\Lambda_{pv}}{\Lambda_m} \right), \quad (39)$$

where $\overline{K_{pv}}$ is defined in Eq. (38), i.e. the specific virial kinetic energy in all haloes greater than $m_h$. Similar to Eq. (16), Eq. (39) describes the rate of transfer of specific virial kinetic energy ($K_{pv}$) from haloes below the scale $m_h$ to haloes above the scale $m_h$ at a rate of $\Pi_{pv}$.

Figure 8 plots the variation of $\Pi_{pv}$ with the halo mass $m_h$ and the redshifts $z$. The mean (specific) virial kinetic energy $K_{pv}$ at two different redshifts $z_1$ and $z_2$ was used to calculate $\Pi_{pv}$ in this figure. Similarly to the mass cascade in Fig. 3, if the statistical structures of the haloes are self-similar and scale-free for haloes smaller than $m_h^*$, the rate of the energy cascade $\epsilon_u$ should also be independent of the scale $m_h$ for $m_p < m_h < m_h^*$. The simulation results confirm a scale- and time-independent rate of cascade $\epsilon_u \approx -10^{-7} m^2/s^3$. Therefore, in the propagation range ($m_h < m_h^*$),

$$\epsilon_u = \Pi_{pv}(m_h, t) \approx -\frac{\partial}{\partial t} \left( \overline{K_{pv}} \right) \equiv \frac{\langle K_{pv} \rangle}{M_h}, \quad (40)$$

where the rate of energy cascade $\epsilon_u$ can be directly related to the rate of mass cascade $\epsilon_m$ (see Eq. (15)). Here, $\langle K_{pv} \rangle = \overline{K_{pv}}(0, t)$ is the mean specific kinetic energy for all halo particles. Similarly to Eq. (20), since $\epsilon_u$ is scale independent, with continuous injection of virial kinetic energy $K_{pv}$, at a constant rate of $\epsilon_u$ on the smallest mass scale, the rate of the energy cascade equals the rate of change of $\langle K_{pv} \rangle$ in all haloes. Therefore, we expect the total virial kinetic energy of all halo particles to be proportional to time $t$ or $\langle K_{pv} \rangle \propto t$. This is consistent with the cosmic energy evolution in Eq. (7).
Figure 9. Schematic plot for the mass and energy flow along the radial direction in dark matter haloes. Both kinetic and potential energy increase with scale $r$. The mass accretion and self-gravity of haloes give rise to the inflow of an infinitesimal mass $dm_r$ ("A" (green color)). The outflow of an infinitesimal mass $dm_r$ ("B" (red color)) is due to the Hubble expansion. On a sufficiently small scale $r$, the inflow should balance the outflow such that the net mass flux vanishes on scale $r$. However, the energy flux can be finite. The outflow of "B" leads to a decrease in the total energy of "B", while the inflow of "A" leads to an increase in the total energy of "A". Kinetic energy flows from small to large scales ("inverse") at a rate of $\varepsilon$, accompanied by potential energy flowing from large to small scales ("direct") at a rate of $2/n_e \varepsilon$. The total energy flows from large to small scales at a rate of $(1 + 2/n_e) \varepsilon$.

5 ENERGY CASCADE IN SPHERICAL HALOES

5.1 Basic concepts

The mass and energy cascade in halo mass space describes the interactions between haloes of different masses. At this level, a statistically steady state is established to continuously maximize system entropy and release system energy while the N-body system evolves toward the limiting equilibrium. A key feature of that statistically steady state is the scale-independent rates of the mass and energy cascade ($\varepsilon_m$ and $\varepsilon_e$), as we demonstrate in Figs. 3 and 8.

Next, on the level of individual haloes, there exists a radial flow and radial velocity dispersion in dark matter, as shown by various simulations [44, 45]. Similarly, mass and energy flow also exist along the radial direction, which is associated with the radial flow. In analogy to the cascade in the halo mass space, this section describes the mass and energy flow in haloes that is particularly important for the internal structures of haloes (or density profile) [41], the universal scaling laws of dark matter haloes [46], and the properties of dark matter particles [40]. The same theory can also be applied to the evolution and dynamics of the galaxy and associated scaling laws for the mass and size of bulges [47].

To understand the mass and energy flow in haloes, we start from the stable clustering hypothesis (SCH), a fundamental assumption for the nonlinear gravitational collapse at small scales. In principle, Hubble expansion leads to an outward radial flow and mass flux, while the halo mass accretion and self-gravity of haloes lead to an inward radial flow and mass flux. The SCH assumes that on sufficiently small scales, clusters of mass particles are bound and stable with a fixed mean physical separation between particles [48, 49]. There is no stream motion between particles in physical coordinates. In this sense, the peculiar motion cancels out the Hubble flow such that the inflow balances the outflow, and the net mass flux vanishes. However, the energy flux can be finite due to the non-uniform gravitational potential in haloes.

Figure 9 describes the mass and energy flow across a small scale $r$, where $v_r^2$ is the typical velocity dispersion on that scale. According to the stable clustering hypothesis, the structures below scale $r$ are bound and stable with constant mass and energy. During an infinitesimal time $dt_r$, the inflow of an infinitesimal mass $dm_r$ across scale $r$ due to the halo gravity (denoted by "1" in green color) balances the outflow due to the Hubble expansion (denoted by "2" in red color). The two infinitesimal masses can be far from each other without direct interactions to exchange energy or momentum. The mass flux into the scale $r$ should equal the mass flux out of the scale $r$, such that the net mass flux vanishes. Mathematically, the mass flux $\dot{m}_r$ and the infinitesimal time $dt_r$ read

$$\dot{m}_r = \frac{dm_r}{dt_r} = m_{r_{out}} - m_{r_{in}} \quad \text{and} \quad \Delta v_r = \frac{dr}{v_r}$$

The net mass flux vanishes, i.e., $m_{r_{out}} = m_{r_{in}} = 0$.

Next, let us consider the energy change on a scale of $r$. There are kinetic and potential energies associated with infinitesimal mass $dm_r$, both of which increase with $r$ and are related to each other by the virial theorem ($2KE = n_r PE$). Here, $n_e$ is an effective exponent and $n_e = -1$ for standard gravity and $n_e = -10/7$ from cosmic energy evolution (Eq. (7)). The potential and kinetic energies read

$$KE = \frac{1}{2} dm_r v_r^2 \quad \text{and} \quad PE = \frac{1}{n_e} dm_r v_r^2.$$  

(42)

Note that these energies are from the intra-halo interactions between particles in the same halo. Only these energies are relevant to the energy flow in haloes. This is why we need to decompose the total particle energy into intra-halo and inter-halo contributions (Eq. (36)).

In Fig. 9, for an infinitesimal mass $dm_r$ (denoted as "A") at location $r + dr/2$ moves inward to $r - dr/2$ during time $dt_r$, the inflow (green arrow) due to the halo self-gravity leads to an increase in total energy (PE+KE) of infinitesimal mass $dm_r$,

$$\Delta E_1 = -\frac{1}{2} \left(1 + \frac{2}{n_e}\right) \frac{dm_r}{dt_r} v_r^2.$$  

(43)

Now we can introduce a key parameter $\varepsilon^*$ to represent the rate of change in the specific energy of the infinitesimal mass $A$

$$\varepsilon^* = \frac{\Delta E_1}{dm_r dt_r} = -\frac{1}{2} \left(1 + \frac{2}{n_e}\right) \frac{v_r^2}{dt_r}.$$  

(44)

This is the key parameter describing the flow of total energy to small scales. Since the total energy is the sum of kinetic and potential energy, kinetic energy flows from small to large scales (inverse cascade with $\varepsilon^* < 0$), while the potential energy flows from large to small scales (direct cascade with $\varepsilon_{PE} > 0$),

$$\varepsilon_r = \varepsilon_{KE} = -\frac{1}{2} \frac{d v_r^2}{dt_r} = \frac{n_e}{2 + n_e} \varepsilon^* \quad \text{and} \quad \varepsilon_{PE} = \frac{2}{2 + n_e} \varepsilon^*.$$  

(45)

In the same figure, for an infinitesimal mass $dm_r$ (denoted as "B") at $r - dr/2$ moving outward to $r + dr/2$, the outflow (red arrow) due to the Hubble expansion leads to a decrease in the total energy of the infinitesimal mass $B$

$$\Delta E_2 = \frac{1}{2} \left(1 + \frac{2}{n_e}\right) \frac{dm_r}{dt_r} v_r^2.$$  

(46)

That decrease in the energy of mass $dm_r$ is "dissipated" due the Hubble expansion, as we demonstrate in the cosmic energy evolution (Fig. 1). The rate of decrease in specific energy (the rate of energy "dissipation") of that $dm_r$ reads

$$\varepsilon_a = \frac{\Delta E_2}{dm_r dt_r} = -\frac{1}{2} \left(1 + \frac{2}{n_e}\right) \frac{v_r^2}{dt_r} = -\varepsilon^*,$$

(47)

which must balance the energy decrease on a cosmic scale (Eq. (8)).
Since the energy decrease in mass $B$ is due to the Hubble expansion (leaving the system as if "dissipated"), this leads to a continuous energy flow from large to small scales (the inflow of mass $A$). Clearly, the energy decrease ($\varepsilon_A$) of mass $B$ is compensated by the energy flux ($\varepsilon^*$) due to the inflow of mass $A$ such that the structures below scale $r$ maintain constant mass and energy.

Next, an important hypothesis we need to make is that the key parameter $\varepsilon_r$ for the flow of kinetic energy is independent of the scale $r$. While these haloes evolve toward limiting equilibrium, a statistically steady state is established. A key feature of this statistically steady state is a scale-independent rate for the energy flow in these haloes. Of course, this hypothesis must be tested and confirmed by simulations, as we demonstrate in Sections 5.3 (Fig. 11). At this time, we take this hypothesis to further illustrate the meaning of $\varepsilon_r$. Since $\varepsilon_r$ is independent of $r$, we can write $\varepsilon(m_h, z) = -\varepsilon_r$ that can be a function of halo mass and redshift only. Solution of Eq. (45) gives the two-thirds law (with $dt_r$ from Eq. (41))

$$v_r^2 = (3\varepsilon r)^{2/3}. \quad (48)$$

Integrating $dt_r = dr/v_r$ in Eq. (41), the typical time $t_r$ reads

$$t_r = \int_0^r \frac{dr}{v_r} = \frac{3}{2} \frac{r^{2/3}}{(3\varepsilon r)^{1/3}} = \frac{3}{2} \frac{r}{v_r} = \frac{v_r^2}{2\varepsilon}. \quad (49)$$

Now, the rate of energy flow in halo can be rewritten as

$$\varepsilon = \frac{v_r^2}{2t_r} = \frac{v_r^2}{3r}. \quad (50)$$

With changes in the specific energy from $v_r^2/2$ to zero during time $t_r$, the parameter $\varepsilon$ represents the rate of change in the specific energy of dark matter particles. Since the velocity dispersion $v_r^2$ increases with mass $m_h$, larger haloes tend to have a larger rate of energy flow.

Combined Eq. (48) with the virial theorem $v_r^2 = \gamma_r Gm_r/r$ ($\gamma_r \approx 0.5$ is a numerical factor), we have

$$m_r = \frac{1}{3\gamma_r} G^{-1} \varepsilon_r^{-1} v_r^4 \text{ and } \frac{dv_r^2}{dm_r} = \frac{2}{5} \frac{v_r^2}{m_r}. \quad (51)$$

where $m_r$ is the total mass enclosed in all scales below $r$. Finally, we can express $\varepsilon_r$ as (from Eqs. (48) and (51))

$$\varepsilon = \frac{1}{2} \frac{dv_r^2}{dt_r} = \frac{1}{5} \frac{v_r^2}{m_r} \frac{dm_r}{dt_r} \text{ or } \varepsilon m_r = \frac{1}{5} \frac{dm_r}{dt_r} v_r^2 = \frac{1}{5} \frac{dm_r}{dt_r} \varepsilon_r, \quad (52)$$

where $m_r$ is the mass flux at scale $r$. From Eq. (52), the product $\varepsilon m_r$ describes the energy flux associated with the mass flux $m_r$, that is, the infinitesimal energy of $(dm_r, v_r^2)$ is transferred across the scale $r$ during infinitesimal time $dt_r$. The mass flux $m_r$ and the energy flux $\varepsilon m_r$ are clearly related to the key parameter $\varepsilon$.

### 5.2 Potential fluctuations and random walk in haloes

In the previous section, we illustrate the mass and energy flow in haloes and identify a key parameter $\varepsilon$ to quantify the rate of flow. This section discusses the mass and energy flow on the particle level via random walk of particles in haloes. The random walk of haloes in halo mass space facilitates the mass cascade in mass space and leads to the halo mass function (Section 4). Similarly, the random walk of dark matter particles in haloes facilitates the energy cascade in haloes and leads to the halo density profile.

Figure 10 presents a schematic diagram for the random walk of particles in haloes with a fluctuating gravitational potential. Due to the discrete particle nature of haloes, the gravitational potential $\Phi$ in the haloes can never be smooth, as shown by N-body simulations.

The total potential at scale $r$ can be decomposed into the mean and fluctuations. Here, $\Phi(r) = \Phi_r(r) + \Phi_r^0(r)$, where $\Phi_r$ is the mean potential and $\Phi_r^0$ is potential fluctuation. There is an energy flux associated with every elementary jump of a particle from small to large $r$ (position "1" to "2"), where the particle kinetic energy $v_r^2/2$ is cascaded to a larger scale $r$ after every migration. For a given particle, migration occurs with an average waiting time $\tau_{gr}(r, m_p, z)$. Because particles have finite kinetic and potential energy, continuous migration facilitates an inverse kinetic energy cascade from small to large scales.

The total potential at scale $r$ can be decomposed into the mean and fluctuations. Here, $\Phi(r) = \Phi_r(r) + \Phi_r^0(r)$, where $\Phi_r$ is the mean potential at scale $r$, and $\Phi_r^0$ is the fluctuation of the potential. The mean potential, potential fluctuation, and density at scale $r$ read

$$\Phi_r(r) = \frac{G m_r}{r}, \quad \Phi_r^0(r) = \frac{G m_p}{d_r}, \quad \rho_r(r) = \frac{m_p}{d_r^3}. \quad (53)$$

where $m_r$ is the mass enclosed by $r$, $d_r$ is the mean spacing between particles at scale $r$, and $m_p$ is the mass of a single particle.

Next, we focus on the average waiting time for the random walk of particles in a fluctuating potential, for example, from position "1" to "2" in Fig. 10. The local temperature $T_r$ and the probability $P_r$ for a successful attempt reads

$$k_B T_r = m_p v_r^2 = m_p \gamma_r \frac{G m_r}{r} \text{ and } P_r = \exp \left( -\frac{m_p E_r}{k_B T_r} \right). \quad (54)$$

where $v_r$ is the typical velocity at scale $r$ and is related to the local potential $\Phi_r$ by virial theorem. Hereafter, we adopt the subscript ‘r’ to represent quantities on scale $r$, the subscript ‘p’ for these quantities on the smallest scale in haloes that is determined by the particle mass and dependent on the resolution of the N-BODY simulations, and the subscript ‘h’ for these quantities on the scale of halo size $r_h$.

Since the potential fluctuations are usually much smaller than the mean potential ($m_p E_r \ll k_B T_r$), the probability for a successful attempt and the waiting time $\tau_{gr}$ for particle random walk read

$$P_r \approx 1 - \left( \frac{m_p E_r}{k_B T_r} \right), \quad \tau_{gr}(r) = \frac{1}{\Gamma_r} (1 - P_r) = \frac{1}{\Gamma_r} \frac{m_p}{d_r} \gamma_r m_r = \frac{1}{\Gamma_r d_r} \gamma_r |\Phi_r|. \quad (55)$$

where $\tau_{gr}(r)$ is a position-dependent waiting time for the random walk of a given particle at location $r$. Here, $\Gamma_r$ is the attempt frequency at $r$ (similarly to the attempt frequency in solid diffusion), i.e., the number of attempts per unit of time for a particle to jump to a neighboring position. The waiting time $\tau_{gr}$ can be related to the
waiting time $\tau_g$ for the random walk of haloes in mass space (Eq. (18)), i.e., $\tau_g = \tau_{gr} (r = r_h)$. If we follow the growth of haloes of characteristic mass $m^*_h$ over time, we can write

$$\frac{dm^*_h}{dt} = \frac{m_p}{\tau'_g},$$

$$\tau'_g = \tau_{gr} (r = r'_h) = \frac{m_p}{r'_h \Gamma_{r'}} m^*_h. \tag{56}$$

These two equations lead to

$$u'_h = \Gamma_r d_r \frac{dm^*_h}{dt} \frac{r^*_h}{t} = \frac{dr^*_h}{dt}, \tag{57}$$

where $m^*_h$ and $r^*_h$ are the mass and size of haloes and $u'_h = \Gamma_r d_r$ represents the speed of growth of haloes. Here, we use the fact that $m^*_h \propto r, r'_h \propto t, \text{and} r^*_h \propto t^0$ (Eq. (24)). For a typical characteristic mass $m^*_h \approx 3 \times 10^{13} M_\odot$, the size $r^*_h \approx 1 \text{Mpc}$ at $z = 0$, and the age of the universe of 13.7 Billion years, the velocity $u'_h$ is around 80km/s. Since the speed of halo growth $u'_h$ remains as a constant while the density of characteristic haloes $\rho'_h \propto a^{-3}$ decreases with cosmic time, we estimate that the $u'_h$ is independent of halo density and should also be roughly independent of the scale $r$. The larger the mean particle distance $d_r$, the smaller the attempt frequency $\Gamma_r$. Therefore, the waiting time of the particles is inversely proportional to the local potential, that is, $\tau_{gr} \propto \Phi^{-1}_r$ from Eq. (55). Similarly, the waiting time of haloes in mass space is inversely proportional to the halo potential, i.e., $\tau_g \propto \Phi^{-1}_h$ in Eq. (18).

Next, consider a thin spherical shell of thickness $s_r$ located at $r$ that contains $N_r$ particles. We will identify the waiting time $\tau_{hr}(r)$ for a successful jump for any one particle of these $N_r$ particles in that spherical shell. In analogy to Eq. (9), we should have $\tau_{gr}(r) = \frac{N_r \tau_{hr}(r)}{s_r}$ and the waiting time

$$\tau_{hr}(r) = \frac{\tau_{gr}}{N_r} = \frac{\tau_{gr}}{4 \pi r^2 \rho r \langle s_r / m_p \rangle} = \frac{m^*_h}{\Gamma_r d_r \pi r^2 \rho r m_r s_r} \times \frac{1}{\pi r^2 \rho r m_r s_r}, \tag{58}$$

where $s_r$ is the thickness of the spherical shell or the jumping length of particle random walk. Similarly to Eq. (10) in mass space, the corresponding rate of mass flow due to the outward migration of particles should read (outflow in Fig. 9)

$$m_r = \frac{m_p}{\tau_{hr}} = 4 \pi r^2 \rho r \langle s_r / m_p \rangle m_r, \tag{59}$$

i.e. a mass of $m_p$ is transferred to larger scales during the time $\tau_{hr}$ via particle migration. Compared to Eq. (41), the particle mass $m_p$ is equivalent to the infinitesimal mass $dm_r$, while the waiting time $\tau_{hr}$ is equivalent to the infinitesimal time $dt_r$.

In analogy to Eqs. (49) and (50), we introduce a key parameter $\varepsilon$ (unit: $s^3/m^2$) to describe the rate of the energy flow along the halo radial direction, which takes the form of

$$\varepsilon = \frac{5}{2} \frac{d}{dt} \frac{m_r}{m_p} = \frac{5}{2} \frac{\tau_{hr}}{\pi r^2} \frac{m_r}{m_p} \text{ and } \varepsilon = \frac{v^2}{2} = \frac{v^2}{5 \pi r^2} \frac{m_p}{m_r}. \tag{60}$$

where $\tau_{hr}$ is the characteristic time on scale $r$. Here, we use the scaling $v \propto v^2$ and $m_p \propto v^2$ (Eq. (51)). With a particle on scale $r$ migrating from position “1” to ”2” (Fig. 10) during the waiting time $\tau_{hr}$, the specific energy of $m_p v^2 / m_r$ is transferred across scale $r$. Similar to Eq (52), the product $\varepsilon m_r$ in Eq. (60) represents the corresponding energy flux associated with the particle migration at scale $r$.

So far, we have only considered the outward mass flow as a result of the particle migration caused by the Hubble expansion, which leads to an increase in halo size $r_h$, a decreasing halo density with time, and energy "dissipation" (Eq. (46)). At the same time, the halo mass accretion (particles accreting onto the outskirts of haloes) and self-gravity lead to an increase in halo mass and introduce an inward radial flow as a result of the gravitational collapse (Fig. 10). At any radius $r$, the inflow is of the order $v_r$, the typical velocity at $r$ (Eq. (54)) such that the mass flow caused by the inflow reads

$$m_r = 4 \pi r^2 \rho_r v_r. \tag{61}$$

For individual haloes, the competition between the migration-induced outward flow (Eq. (59)) and the accretion-induced inward flow (Eq. (61)) might lead to a non-vanishing net mass flux. However, a vanishing met mass flux is expected for the average radial flow for all haloes of the same mass, as required by the stable clustering hypothesis. In that case, we equate the out-flow flux in Eq. (59) and in-flow flux in Eq. (61), the thickness $s_r$ and waiting time $\tau_{hr}$ read

$$s_r = v_r \tau_{gr} = \frac{m_p r_r v_r}{u'_h m_r} \text{ and } \tau_{hr} = \frac{m_p}{m_r} \frac{r_r}{4 \pi r^2 \rho_r v_r}. \tag{62}$$

In this section, in analog to the halo migration in mass space, we introduce the particle migration in the fluctuating halo potential field and the associated mass and energy flux. The rate of energy flux $\varepsilon$ and the waiting time $\tau_{gr}$ and $\tau_{hr}$ are the central quantities for the energy cascade in the haloes that determine the internal structure, density profile, and many other properties of haloes. The energy cascade in haloes establishes a statistically steady state to continuously release the halo energy and maximize the halo entropy, where the rates of the energy cascade must be scale independent ($\varepsilon$ is independent of $r$). This leads to a scale-independent cascade between scales $r < r < r_s$, where the smallest scale $r$ depends on the particle mass or simulation resolution, while $r_s$ is the usual scale radius. The mathematical formulations are presented in Sections 5.3.

5.3 Formulating the energy cascade in haloes

This section quantifies the energy cascade in haloes from N-body simulations. The same theory can also be applied to the energy flow in galaxy bulges and the associated scaling laws for bulge mass, size, dynamics, and the evolution of the supermassive black holes [47].

In analogy to the formulation of the energy cascade in halo mass space (Section 4.5), we examine the energy cascade in all haloes of a given mass $m_h$ such that the averaged mass flux vanishes. We start the formulation by introducing the cumulative functions along the halo radial direction $r$. The cumulative mass function $\Lambda^h_m(m_h, r, z)$ (similar to $\Lambda_m$ in Eq. (17)) represents the total mass above $r$ that is averaged for all haloes of the same mass $m_h$

$$\Lambda^h_m(m_h, r, z) = \int_{r}^{\infty} \rho_r (m_h, r', z) 4 \pi r'^2 \, dr'. \tag{63}$$

where $\rho_r$ is the average mass density for all haloes of mass $m_h$.

Next, similarly to the formulation in Section 4.5, we also decompose the velocity of the halo particles $v_p$ into the halo mean velocity, $v_h = \langle v_p \rangle_h$, and the velocity fluctuation, $v'_p$, that is, $v_p = v_h + v'_p$ (Eq. (36)). Here, $v_h$ represents the velocity of that halo, the average velocity of all particles in the same halo. Consequently, the total kinetic energy $K_P$ of a given halo particle can be divided into $K_P = K_P + K_{PV}$. The virial kinetic energy, $K_{PV} = v'_p^2/2$, is the contribution from the velocity fluctuation due to the intra-halo interactions on small scales that are in the nonlinear regime. Only this part of the kinetic energy is relevant for the energy cascade in haloes.

We introduce a cumulative function $\Lambda^h_{PV}(K_P)$ for $K_{PV}$

$$\Lambda^h_{PV}(m_h, r, z) = \int_{r}^{\infty} K_{PV} \rho_r (m_h, r', z) 4 \pi r'^2 \, dr'. \tag{64}$$

Vol. 000, 1–28 (2024)
Next, similarly to the energy cascade in the mass space (Eq. (38)), we introduce the specific kinetic energy on scale $r$ for all haloes of the same mass $m_h$ that reads

$$K_{PV}^h(m_h, r, z) = \frac{\Pi_{PV}^h(m_h, r, z)}{\rho_h(m_h, r, z)} = \int_0^r K_{PV}^h(m_h, r', z) 4\pi r' dr'. \tag{65}$$

Here, $K_{PV}^h$ is the specific energy (energy per unit mass) contained on scales above $r$. With $K_{PV}^h$ increasing with time, the energy flux $\Pi_{PV}^h$ along the halo radial direction is defined as (similarly to Eq. (39))

$$\Pi_{PV}^h(m_h, r, z) = -\frac{\partial}{\partial t} K_{PV}^h = -\frac{\partial}{\partial t} \left( \frac{\Lambda_{PV}^h}{\rho_h^2} \right). \tag{66}$$

As shown in Fig. 9, the inflow canceling out the outflow leads to a net vanishing mass flux. However, there exists a flux of specific kinetic energy $E$ from small to large scales that leads to the increase in the particular energy on all scales greater than $r$, i.e., $E = \Pi_{PV}^h$ (Fig. 9).

In a certain range of scales $r < r_s$ (inner haloes), the characteristic time on small scales is very small compared to the time on large scales. The small-scale motion does not feel the large-scale motion directly, except through the energy flux $E$ across scales. When a statistical equilibrium is established, similarly to the energy cascade in the halo mass space, we expect the rate of energy flow $\varepsilon(m_h, z)$ to be independent of the scale $r$, i.e.

$$\varepsilon(m_h, z) = \Pi_{PV}^h(m_h, r, z) = -\frac{\partial}{\partial t} E_{PV}^h \quad \text{for} \quad r < r_s. \tag{67}$$

This hypothesis can be tested by N-body simulations.

Figure 11 plots the variation of the energy flux $\Pi_{PV}^h(m_h, r, z)$ using Eq. (66) and the kinetic energy $K_{PV}^h$ in Eq. (65) at two different redshifts from the Illustris simulations. In this figure, the rate of energy flow $\varepsilon$ is clearly independent of the scale $r$ below a characteristic scale $r_s$, usually the scale radius of haloes. The key parameter $\varepsilon$ increases with the halo mass $m_h$ and the redshift $z$. For haloes with characteristic mass $m^*_h$, $\varepsilon(m^*_h, z) \equiv \varepsilon_u \approx 10^{-7} m^2/s^3$.

Similarly, we can calculate the rate of the energy cascade for haloes of different masses. Figure 12 plots the variation of $\varepsilon(m_h, z)$ with the halo mass $m_h$ and the redshift $z$. The figure shows that $\varepsilon \propto m_h^{2/3}$ and increases with redshift $z$.

$$\varepsilon(m_h, z) = \varepsilon_u \nu = \varepsilon_u \left( m_h/m_h^* \right)^{2/3} \propto m_h^{2/3} a^{-1}, \tag{68}$$

where $\varepsilon_u \equiv \varepsilon(m^*_h, z)$ is the rate of energy flow in haloes of characteristic mass $m^*_h$. The dimensionless parameter $\nu$ is defined as $\nu = (m_h/m^*_h)^{2/3}$ [41]. The rate of the energy cascade varies with the halo mass and time as $\varepsilon \propto m_h^{2/3} a^{-1}$, while the haloes evolve towards the limiting equilibrium. We only have a vanishing energy cascade or $\varepsilon = 0$ in fully virialized haloes with $a \to \infty$ (usually small and old haloes), a limiting state that real haloes can never reach.

In the picture of energy flow in individual haloes, particle migration leads to an outward mass flow from small to large $r$. Halo mass accretion and self-gravity lead to an inward radial flow from large to small $r$. The two mass fluxes cancel such that the net mass flux vanishes. In addition, there is a continuous flow of kinetic energy from small to large scales with an energy flux $E$ that is independent of the scale $r$ in a certain range of scales. A simultaneous flux of potential energy exists in the opposite direction. Again, according to the virial theorem (2KE + PE = 0), the rate of a cascade of potential energy from large to small scales is approximately twice the rate of kinetic energy or $-2\varepsilon$. The more accurate calculation from cosmic energy evolution shows a rate of $-7/5\varepsilon$ for the potential energy with $n_c = -10/7$ in Fig. 9 (Eq. (7)).

### 5.4 Universal scaling laws and inner density of haloes

Scaling laws are well known to be associated with the energy cascade in turbulence [28, 50]. In this section, we focus on the scaling laws associated with the energy cascade in spherical haloes. The starting...
point is the expression of \( \varepsilon \) (Eq. (50))

\[
\varepsilon = \frac{v_r^2}{2 \rho_r} = \frac{v_r^3}{3r} \tag{69}
\]

Since the rate of the energy cascade \( \varepsilon \) is independent of \( r \) in a certain range of scales, the 2/3 law can be easily obtained for kinetic energy \( v_r^2 \propto (-r \varepsilon r)^{2/3} \). With virial theorem \( v_r^2 \propto GM_r/r \), the 5/3 scaling law can be recovered for the mass enclosed within \( r \), i.e., \( m_r \propto r^{5/3} \). We can write the scaling laws for mass \( m_r \) (mass enclosed within \( r \)), density \( \rho_r \) (mean density of the halo enclosed within \( r \)), velocity \( v_r \), time \( t_r \), and kinetic energy \( v_r^2 \), and mass flux \( m_t \) on scale \( r \), all determined by three quantities \( \varepsilon, G \) and scale \( r \):

\[
m_r = \alpha_r \varepsilon^{2/3} G^{-1} r^{5/3},
\]

\[
\rho_r = \beta_r \varepsilon^{2/3} G^{-1} r^{-4/3},
\]

\[
v_r^2 = \gamma_r \varepsilon^{2/3} G^{-1} r^{1/3},
\]

\[
\tau_r = 0.5 \beta_\rho G r^{-1/3} r^{2/3},
\]

\[
v_r^2 = \beta_r G^{-2/3} \alpha_r \varepsilon^{2/5} (eGm_r)^{2/5},
\]

\[
m_t = 4 \pi \beta_\varepsilon^{1/3} \beta_r e G^{-1} r
\]

where \( \alpha_r \) and \( \beta_r \) are two numerical factors that can be determined by data fitting (Fig. 13). The evolution of these quantities is absorbed into \( \varepsilon(m_h, z) \). The relations between numerical factors are

\[
\beta = \frac{1}{3}, \quad \gamma_r \alpha_r = \beta_r^{2/3}, \quad \beta_r = \frac{5}{4 \pi} \alpha_r \beta
\]

Other relevant quantities include the mean particle distance \( d_r \), the attempt frequency \( \Gamma_r \), the waiting time \( \tau_{gr} \) and \( \tau_n \), and the shell thickness \( s_r \). These quantities depend on the particle mass \( m_p \) and the speed of growth for haloes of characteristic mass \( m_h^{*} \), i.e.,

\[
u_h^{*} = \Gamma_r d_r = 80 \text{ km/s.}
\]

From Eqs. (55) to (62),

\[
d_r = \beta_r^{-1} m_p^{1/3} e^{-2/3} G^{-1/3} r^{1/3},
\]

\[
\Gamma_r = \frac{1}{3} \beta_r u_h^{*} m_p^{1/3} e^{2/3} G^{-1/3} r^{-4/3},
\]

\[
\tau_{gr} = \frac{m_p}{\alpha_r \Gamma_r},
\]

\[
u_h^{*} = \beta_r^{3/2} m_p^{1/2} e^{3/2} G^{-1/2} r^{-2/3},
\]

\[
\tau_n = \frac{3}{5} \Gamma_r m_p e^{-1/2} G^{-1/2},
\]

\[
N_r = \frac{5 u_h^{* 3/2}}{3 \alpha_r \beta^{3/2}},
\]

\[
s_r = \frac{u_h^{* 1/2}}{\alpha_r \beta^{1/2}} m_p e^{-1/2} G^{-1/2}.
\]

Since these quantities involve the particle mass \( m_p \), information on these quantities may provide useful insights into the dark matter particle mass and properties.

Another approach to derive these scaling laws using the dimensional argument is relatively simpler but more heuristic. When a statistically stable state is established in inner haloes \( (r < r_s) \), the fast motion on small scales does not feel the slow motion on large scales directly, except through the scale-independent rate of the energy cascade \( \varepsilon \). The flow fields on these scales are statistically similar so that all relevant physical quantities can be determined by and only by three quantities: \( \varepsilon, G \), and scale \( r \). Therefore, the scaling laws for any quantity \( Q = e^{x} G^{y} r^{z} \) can be determined from the dimensional analysis to obtain the exponents \( x, y, \) and \( z \). This approach leads to the same scaling laws in Eq. (70).

To validate the predicted scaling laws, we present the N-body simulation results. For example, the halo inner density should be

\[
\rho_r(r, m_h, z) = \beta_r G^{-1} r^{-4/3} = \beta_r e^{2/3} G^{-1} r^{-4/3}\left(\frac{m_h}{m_h^{*}(z)}\right)^{2/3},
\]

where scaling laws predict a limiting density slope of -4/3 and inner density \( \rho_\rho(r) \propto m_h^{4/3} a^{-2/3} \) for haloes of mass \( m_h \). Clearly, halo mass accretion leads to an increase in \( m_h \) and density \( \rho_r \), while particle migration leads to a decrease in \( \rho_r \) over time. The two competing effects are equal for haloes with a characteristic mass \( m_h^{*} \propto a^{3/2} \) such that the inner density \( \rho_r \) for haloes of characteristic mass \( m_h^{*} \) remains constant over time. This is the so-called small-scale permanence for halo density. The halo density at different redshifts collapses into the same -4/3 law. For old and small haloes with a mass less than \( m_h^{*} \), the mass accretion is slower, leading to a slow decrease in density \( \rho_r \) over the cosmic time \( t \), as shown in Eq. (73).

Figure 13 demonstrates the small-scale permanence. The figure plots the time evolution of the average density profiles for all haloes with characteristic masses \( m_h^{*}(z) \) from Illustris simulations at different redshifts. The energy cascade in haloes gives rise to a limiting density slope of -4/3 (Eq. (73)). The density profiles of all dark matter haloes of mass \( m_h^{*} \) converge to the predicted time-unvarying scaling (solid red line from Eq. (73)) on small scales, i.e., the small-scale permanence. The -4/3 scaling law in Fig. 13 should extend to smaller and smaller scales until the scale of the smallest halo structure. That scale can be the free streaming scale, which is dependent on the nature and mass of dark matter particles. From the scaling laws developed, the properties of dark matter particles may be obtained by extending these scaling laws to the smallest scale [40].

### 5.5 Particle random walk and halo density profile

The energy flow in haloes with a scale-independent rate leads to a power-law halo inner density that follows the -4/3 scaling on small scales. This section derives the complete halo density profiles in the entire range of scales. Similarly to the halo mass function obtained from the random walk of haloes in mass space, the halo density profile (or the distribution of halo particles) can be analytically derived based on a similar idea, given by the solution of the Fokker-Planck equations from the random walk of particles in haloes [41].

To find the halo density profile, we need to derive the particle distribution function due to the random walk in the three-dimensional halo space. The 3D particle random walk can be described by a
Langevin equation for the particle position $X_t$ (similar to Eq. (25)),
\[
\frac{dX_t}{dt} = \sqrt{2D_p(X_t)\xi(t)}.
\]
(74)
For the random walk of particles in haloes, without loss of generality, a power law $\tau_{gr}(r) \propto r^{-\gamma}$ (see Eq. (72) for $\gamma = 2/3$) can be assumed, i.e. a position-dependent waiting time. Here, $r \equiv |X_t|$ is the distance of the particle to the center of the halo. From this, the position-dependent diffusivity reads
\[
D_P(X_t) = D_0(t) r^{2\gamma},
\]
(75)
where $D_0(t)$ is a proportional constant. The smaller the distance $r$, the smaller the diffusivity or the longer waiting time, and the higher particle density. In 1D convention, the 3D Fokker-Planck equation in Cartesian coordinate can be directly obtained for particle distribution function $P_r(X, t)$ ($i = 1, 2, 3$ for Cartesian coordinates),
\[
\frac{\partial P_r(X, t)}{\partial t} = D_0 \frac{\partial}{\partial X_i} \left( r^{2\gamma} \frac{\partial}{\partial X_i} P_r(X, t) \right).
\]
(76)
The corresponding solution of Eq. (76) in spherical coordinate is
\[
P_r(r, t) = \frac{(2 - 2\gamma)^{\frac{3}{2} - 2\gamma} r^{-2\gamma}}{4\pi (D_0 t)^{\frac{3}{2}} (2 - 2\gamma)^{\frac{3}{2}} \Gamma \left( \frac{3}{2} - 2\gamma \right)} \exp \left( \frac{-r^{2-2\gamma}}{4(1 - \gamma^2)D_0 t} \right).
\]
(77)
Since the distribution function $P_r(r, t)$ is equivalent to the halo density, we find that halo density $\rho_r \propto r^{2\gamma}$. From this insight, assuming that $y$ is unknown, we can predict the value of $y$ and recover the -4/3 law. This can be an independent check as an alternative approach other than deriving the -4/3 law from the scale-invariant energy cascade $e$ (see Eqs. (69) and (70)).

Since the waiting time $\tau_{gr} \propto \Phi(r)^{-1} \propto r^{-\gamma}$ (Eq. (55)), the halo density should scale as $\rho_r \propto r^{2-2\gamma}$ from Eq. (77). The halo mass encased in $r$ scales as $M_r \propto \rho_r r^3 \propto r^{3-2\gamma}$. The local potential at $r$ should scale as $\Phi(r) \propto \rho_{\infty} r^{-3+2\gamma}$ and the inner density follows $\rho_{r} \propto r^{2\gamma}$. The well-known core-slope problem describes the discrepancy between cuspy halo density predicted by cosmological CDM-only simulations and the core density from observations. The predicted halo density exhibits a cuspy inner density $\rho_c \propto r^s$ with a wide range of slope parameter $s$ between -1.0 and -1.5 (see Introduction 1). There appears to be no consensus on the exact value of the slope $s$ and no solid theory for the density slope $s$. In this paper, we propose that the internal structure of haloes is primarily determined by the radial mass and energy flow [41, 46]. In this regard, we have the following considerations:

(i) Individual haloes with both non-zero mass flux and non-vanishing energy flow ($m_{\text{net}}^{\text{tot}} \neq 0$ in Eq. (41), $\epsilon \neq 0$). The outward radial flow can be caused by the migration of particles along the radial direction. Inward radial flow can be caused by halo mass accretion. The stable clustering hypothesis requires that the outflow cancels out the inflow such that the net mass flux vanishes (Fig. 9). However, most simulated haloes are non-equilibrium dynamic objects that might have both a nonzero mass flux and non-vanishing energy flow. The density slopes of these haloes depend both on the radial flow and the mass accretion of each halo [41]. This can be the reason for the wide variety of density slopes $s$ for simulated haloes. Such simulated haloes can generally be modeled by a double-$\gamma$ density profile
\[
\rho_r(r, t) = \rho_s(t) \left( \frac{r}{r_s} \right)^{\gamma - 2} \exp \left( \frac{1}{\beta} \left( 1 - \left( \frac{r}{r_s} \right)^{4\alpha} \right) \right),
\]
(84)
where $\rho_s(t)$ is the density at scale radius $r_s(t)$. This four-parameter double-$\gamma$ density profile ($\rho_s, r_s, \alpha, \beta$ in Eq. (84)) reduces to the standard three-parameter Einasto profile ($\rho_s, r_s, \alpha$) with $\alpha = 2\beta$.

(ii) Haloes with vanishing net mass flux and non-vanishing energy flow ($m_{\text{net}}^{\text{tot}} = 0$ and $\epsilon \neq 0$). Individual haloes may not satisfy the stable clustering hypothesis exactly and can have random mass flux with $\dot{m}_{\text{net}}^{\text{tot}} \neq 0$. However, in N-body simulations, we can average out the net mass flux by constructing composite haloes from all haloes of the same scale mass $m_h$. This is how we extract the halo density profiles from N-body simulations (Figs. 13 and 14), where relevant quantities are averaged for all haloes with the same mass. For these composite haloes, the net mass flux vanishes, and we only have the energy flow determining the internal structure of the halo such that the inner density follows the -4/3 law. The corresponding double-density profile for haloes of any mass $m_h$ (let $\alpha/\beta = 2/3$ in Eq. (84))
or \( \gamma_1 = 2/3 \) reads

\[
\rho_r(r, m^*_h, z) = \beta_r \varepsilon^{2/3} G^{-1} r^{-4/3} \left( \frac{r}{r_s} \right)^{-4} \exp \left[ -\frac{1}{\beta} \left( \frac{r}{r_s} \right)^{2\beta} \right],
\]

(Eq. 85)

\( \varepsilon(m^*_h, z) = (m^*_h / m^*_h(z))^{2/3} \varepsilon_u, \)

where \( \beta \approx 1 \) is an amplitude parameter, \( \beta \) is a parameter, and \( r_s \) is the scale radius. Here, \( \varepsilon_u \approx 10^{-7} \text{m}^2 / \text{s}^3 \) is the rate of the energy flow for haloes with characteristic mass \( m^*_h(z) \). While \( \varepsilon(m^*_h, z) \) is the rate of the energy flow in haloes of any mass \( m^*_h \) (Eq. (68)). For small \( r \), the inner density reduces to

\[
\rho_r(r, m^*_h, z) = \beta_r \varepsilon^{2/3} G^{-1} r^{-4/3} \left( \frac{m^*_h}{m^*_h(z)} \right)^{4/3} a^{-2/3} \text{ for } r \to 0.
\]

(Eq. 86)

The redshift dependence is absorbed into the parameter \( \varepsilon \). For haloes of characteristic mass \( m^*_h(z) \), there exists a small-scale permanence for the halo density, i.e., the density profiles at different redshifts \( z \) converge to a time-unvarying scaling,

\[
\rho_r(r, m^*_h, z) \equiv \rho_r(r) = \beta_r \varepsilon^{2/3} G^{-1} r^{-4/3} \text{ for } r \to 0.
\]

(Eq. 87)

The small-scale permanence is shown and discussed in Fig. 13.

Figure 14 presents the density profiles from Illustris simulations for (composite) haloes of different masses \( m^*_h \) at \( z = 0 \). The solid lines show the average halo density profiles for all haloes with a mass between \( 10^{10.0} \text{M}_\odot \) from the Illustris simulation, where halo mass \( m^*_h \) is between \( 10^{9} \text{M}_\odot \) and \( 10^{14} \text{M}_\odot \). The thick dashed (straight) line represents the -4/3 scaling law for halo density \( \varepsilon \) that involve the rate of cascade \( \varepsilon \approx m^*_h(z) \) (Eq. (68)). Here, \( \varepsilon = \varepsilon_u \) for haloes of characteristic mass \( m^*_h \) (Eq. (68)). The figure demonstrates the -4/3 law that we obtained from the energy cascade in haloes. That scaling was in agreement with the density profiles of haloes of more than six orders of magnitude. The thin dashed lines in both Figs. 14 and 13 present the double-\( \lambda \) density profile from Eq. (85) that has good agreement with simulation data.

(iii) The last scenario is completely virialized haloes with both vanishing net mass flux and vanishing energy flow \( (m^*_h = 0 \text{ and } \varepsilon = 0) \).

This is the limiting equilibrium state haloes evolve toward but can never reach. The -4/3 density slope should still be good in this limiting state with \( \varepsilon \to 0 \).

6 CONSTRAINTS ON DARK MATTER PARTICLE MASS

Previous sections discuss the mass and energy cascade in halo mass space and in individual haloes, which can be relevant to dark matter particle mass and properties. This section focuses on some potential constraints that can be established. More efforts are required for a better understanding. Ideas are only briefly presented here in the hope of leading to other better suggestions from the community.

The general ideas are: i) if gravity is the only interaction between dark matter particles, the propagation range with a scale-independent rate of the cascade may extend to the smallest scale (free streaming in Fig. 13). If the free stream scale is comparable to the particle mass, the scaling laws identified may provide useful insights into the properties of dark matter particles [40]; ii) the relevant physical quantities (wait time, etc.) for the particle random walk and migration involves the particle mass \( m_p \) (Eq. (72)). New constraints on particle mass can be identified if we can identify any constraints on these quantities. We will illustrate the second idea in this section:

(i) First, from the scaling for waiting time \( \tau_{hr} \) in Eq. (72), we have the expression for particle mass \( m_p \),

\[
m_p = \frac{5}{7} \varepsilon G^{-1} r \tau_{hr}.
\]

(Eq. 88)

For haloes with characteristic mass \( m^*_h \), the rate of energy flow \( \varepsilon \) = \( \varepsilon_u \approx 10^{-7} \text{m}^2 / \text{s}^3 \), and the halo size \( r_h \approx 1 \text{Mpc} \), we should have

\[
m_p = \frac{5}{7} \varepsilon_u G^{-1} r_h \tau^*_h.
\]

(Eq. 89)

The waiting time \( \tau^*_h \equiv \tau_{hr}(r = r_h^*) \) represents a time scale for a real physical process, i.e., the waiting time for the migration of particles on the halo boundary (Eq. (58)). It is reasonable to assume that the waiting time \( \tau^*_h \approx \tau_p \), where \( \tau_p \approx 5.4 \times 10^{-44} \text{s} \) is the Planck time, the smallest possible unit of time for any physical process. With this and \( \tau_r \approx 0.5 \) (Eq. (51)), the particle mass should satisfy

\[
m_p \geq \frac{5}{7} \varepsilon_u G^{-1} r_h \tau_{p} = 2.5 \times 10^{-18} \text{kg} = 10^3 \text{GeV}.
\]

(Eq. 90)

(ii) Second, we consider the growth of haloes of characteristic mass \( m^*_h \propto t \) and size \( r_h \propto t \) (Eq. (24)). Since the waiting time \( \tau_{hr} \propto m^{-2/3}_h \) (Eq. (19)) in mass space and \( \tau_{hr} \propto r^{-2/3} \) (Eq. (72)) in individual haloes, haloes with a characteristic mass \( m^*_h \) must have the fastest mass accretion and the shortest waiting time \( \tau^*_G \). The mass accretion of these characteristic haloes reads

\[
m^*_h = \frac{\text{dm}^*_h}{\text{dt}} = \frac{m_p}{\tau^*_G} = \frac{m^*_h(z = 0)}{t_0}.
\]

(Eq. 91)

where \( m^*_h \) is the rate of mass accretion. On average, characteristic haloes accrete one particle of mass \( m_p \) during an average waiting time \( \tau^*_G \). For a typical mass \( m^*_h(z = 0) \approx 3.2 \times 10^{13} \text{M}_\odot \) at the current
The halo density on the boundary \( \rho_h^* \) is in the same order as the critical density \( \rho_{\text{crit}} \). The halo boundary density \( \rho_h^* \) obtained from the \( -4/3 \) scaling is around \( 6 \times 10^{-25} \) kg/m\(^3\) (Eq. (70)). Relevant quantities are also listed in Table 2.

Figure 15. The schematic plot of relevant densities next to the boundary of haloes at \( \epsilon = 0 \). The mass-energy density of the universe is \( \rho_{\text{crit}} = 10^{-26} \) kg/m\(^3\). The man halo density \( \rho^*_h = 4 \times 10^{-25} \) kg/m\(^3\). Characteristic haloes have a mass of \( m^*_h = 3.2 \times 10^{13} M_\odot \) and a halo size \( r^*_h = 1 \) Mpc. The halo density on the boundary \( \rho_h^* \) is the same order as the universe density \( \rho_{\text{crit}} \). The halo boundary density \( \rho_h^* \) obtained from the \( -4/3 \) scaling is around \( 6 \times 10^{-25} \) kg/m\(^3\) (Eq. (70)). Relevant quantities are also listed in Table 2.

Epoch [4] and \( t_0 \approx 4.3 \times 10^{17} \) s (the age of the universe), the rate of mass accretion \( m^*_p \approx 1.5 \times 10^{56} \) kg/s. The particle mass \( m_p \) can be related to the waiting time as

\[
m_p = m^*_p \rho^*_g = 1.5 \times 10^{26} \frac{\text{kg}}{s} \frac{r^*_h}{r^*_g}.
\]

In N-body simulations, \( m_p \) is dependent on the mass resolution. For Illustris simulation with \( m_p \approx 6 \times 10^6 M_\odot \) (Table 1), the waiting time is on the order of \( t^*_g \approx 7 \times 10^{13} \) s.

The waiting time \( t^*_g \approx \tau_{GR} \) also represents a time scale for a real physical process, i.e., the waiting time for random walk in haloes. Therefore, it is also reasonable to assume that \( \tau_{GR} \geq t_p \).

With this consideration, the particle mass must satisfy

\[
m_p = 1.5 \times 10^{26} \frac{\text{kg}}{s} \frac{r^*_h}{r^*_g} \geq 1.5 \times 10^{26} \frac{\text{kg}}{s} \frac{t_p}{r^*_p} = 10^{-17} \text{kg}.
\]

This provides a similar lower limit for particle mass \( m_p \).

(iii) Third, a more stringent constraint estimates that the waiting time \( \tau^*_g > \Delta \tau_{SP} \), where \( \Delta \approx 200 \) is the contrast ratio of the halo density to the background dark matter density. To understand this, we start from equations for the growth of halo mass \( m^*_h \) and size \( r^*_h \),

\[
\frac{dm^*_h}{dt} = m^*_p \frac{m^*_h}{r^*_g} \quad \text{and} \quad \frac{dr^*_h}{dt} = \frac{r^*_h}{t} = u^*_h,
\]

\[
\frac{m^*_h}{r^*_h^3} = \rho_{\text{crit}} \Omega_{DM} \Delta c \quad \text{and} \quad \rho_{\text{crit}} = \frac{3H^2}{8\pi G},
\]

where \( H \) is the Hubble parameter, \( \rho_{\text{crit}} \approx 10^{-26} \) kg/m\(^3\) is the critical mass-energy density of the present universe, and \( \Omega_{DM} = 0.2449 \) is the mass fraction of dark matter. For a typical dark halo mass \( m^*_h = 3.2 \times 10^{13} M_\odot \) and halo size \( r^*_h = 1 \) Mpc (listed in Table 2), the speed of the halo growth is around \( u^*_h \approx 80 \) km/s (Eq. (57)). Figure 15 presents a schematic plot of the relevant densities next to the halo boundary. From Eq. (94), we can have the relation between the speed of halo growth \( u^*_h \) and the waiting time \( \tau^*_g \),

\[
u^*_h^2 = \frac{Gm_p}{3H^2} \Omega_{DM} \rho_{DM} \Delta c \frac{r^*_h}{r^*_g}.
\]

Without loss of generality, let the mass density on the halo boundary be \( \rho^*_h = \rho_{\text{crit}}/\alpha \). The mass-energy density in haloes is determined mostly by the dark matter density since dark matter is the dominant component in haloes. The mass-energy density of the universe is \( \rho_{\text{crit}} \), which includes all cosmic components, including dark matter, baryonic matter, and dark energy. For \( \rho^*_h \geq \rho_{\text{crit}} \), we expect \( \alpha \leq 1 \).

The halo mass now reads (from Eq. (94))

\[
m^*_h = \frac{4}{3} \pi r^*_h^3 \rho^*_h \Omega_{DM} \Delta c.
\]

With Eq. (94), we can further express the rate of mass accretion in terms of the boundary density \( \rho^*_h \),

\[
\frac{1}{3} \Omega_{DM} \Delta c \rho^*_h \frac{4}{3} \pi r^*_h^3 u^*_h = \frac{m^*_p}{\tau^*_g} \quad \text{or} \quad \frac{1}{3} \Omega_{DM} \Delta c \rho^*_h \frac{4}{3} \pi r^*_h^2 u^*_h = m^*_h.
\]

For a spherical shell on the halo boundary that contains one and only one particle (\( \tau_{GR} = \tau_{TR} \) for \( r = r^*_h \)), the thickness of that spherical shell \( s^*_h = s_r = r^*_h \) reads (from Eq. (97))

\[
s^*_h = \frac{m_p}{\rho^*_h 4 \pi r^*_h^2} = \frac{1}{3} \Omega_{DM} \Delta c a u^*_h r^*_h \rho^*_h \Delta c,
\]

where \( s^*_h \) is approximately the jumping length for the migration of particles on the halo boundary. Using Eq. (95), the speed of halo growth is related to that jumping length as

\[
u^*_h^2 = \frac{2a}{3H^2} \frac{Gm_p}{\rho^*_h 4 \pi r^*_h^2} \quad \text{and} \quad m_p = \frac{3H^2}{2a} u^*_h^2 s^*_h G^{-1}.
\]

For every waiting time \( \tau^*_g \), the particle migrates at a distance of \( s^*_h \). Similarly, since \( s^*_h \) represents a physical length scale for particle migration, it is reasonable to assume that the distance \( s^*_h \geq l_p \), where \( l_p = 1.6 \times 10^{-8} \) m is the Planck length, the smallest possible unit of length. With this constraint, the particle mass must satisfy

\[
m_p \geq \frac{3H^2}{2a} u^*_h^2 G l_p = 10^{-15} \text{kg}.
\]

(iv) Especially we estimate the speed of halo growth and the jumping frequency at scale \( r \) as

\[
u^*_h = \Gamma \frac{d_t}{d_l} \quad \text{or} \quad \Gamma \geq \frac{Gm_p}{d^2 l_p^2}.
\]

Since gravity is assumed to be the only force between dark matter particles, for particle migration in haloes, the smallest separation between two particles should be the Planck length scale \( l_p \). The typical acceleration \( Gm_p/d_t^2 \) on scale \( r \) changes its direction on the Planck length \( l_p \), which determines the jumping frequency \( \Gamma \).

Next, the relevant speed of that particle reads (using Eq. (99))

\[
s^*_h = \frac{1}{3} \Omega_{DM} \Delta c a u^*_h = \frac{1}{3} \Omega_{DM} \Delta c \sqrt{\frac{2a}{3H^2}} \frac{Gm_p}{s^*_h}.
\]

Finally, normalizing the jumping length \( s^*_p = \beta \) by Planck length \( l_p \) and waiting time \( \tau^*_g = \gamma l_p \) by Planck time \( l_p \), we obtain

\[
\left( \frac{\beta}{\alpha} \right)^{3/2} = \frac{1}{3} \Omega_{DM} \Delta c \sqrt{\frac{2a}{3H^2}} \frac{Gm_p}{s^*_h}.
\]

where \( c = 3 \times 10^8 \) m/s is the speed of light. Plugging in numbers into Eq. (103) and considering that \( \alpha \leq 1 \) and \( \beta \geq 1 \) \( m_p = 1.5 \times 10^{26} \) kg/s, \( \Omega_{DM} = 0.2449 \), and \( H(t=2) \), we should have

\[
\gamma \approx \Delta c \frac{\beta}{\alpha} \quad \text{or} \quad \tau^*_g = \Delta c \frac{\beta}{\alpha} s^*_p \geq \Delta c l_p.
\]

The density contrast ratio \( \Delta_c \) should influence the waiting time.
The relevant quantities in haloes of characteristic mass $m^*_h$ at $z=0$. The lower part of the table contains quantities dependent on the particle mass $m_p$.

| Quantity (m) | Symbol | Scaling | Value at $r_p$ | Value at $r^*_h$ |
|-------------|--------|---------|---------------|-----------------|
| Scale (m)   | $r$    | $r^{5/3}$ | $3 \times 10^{-13}$ | $3.2 \times 10^{22}$ |
| Mass (kg)   | $m_p$  | $r^{5/3}$ | $1.6 \times 10^{-15}$ | $6.4 \times 10^{23}$ |
| Density (kg/m$^3$) | $\rho_p$ | $r^{-4/3}$ | $5.3 \times 10^{22}$ | $1 \times 10^{-26}$ |
| Time (s)    | $\tau_p$ | $r^{2/3}$ | $1 \times 10^{-6}$ | $4 \times 10^{13}$ |
| Velocity (m/s) | $v_p$    | $r^{-1/3}$ | $4 \times 10^{-7}$ | $2.6 \times 10^{15}$ |
| Mass flow (kg/s) | $m_p$ | $r^{-1}$ | $1.6 \times 10^{-9}$ | $1.0 \times 10^{16}$ |
| Mean distance (m) | $d_p$ | $r^{4/3}$ | $3 \times 10^{-13}$ | $5.4 \times 10^{3}$ |
| Frequency (1/s) | $\Gamma_p$ | $r^{-4/3}$ | $2.8 \times 10^{17}$ | $14.8$ |
| Waiting time (s) | $\tau_{gr}$ | $r^{-2/3}$ | $4 \times 10^{-18}$ | $1.1 \times 10^{-41}$ |
| Waiting time (s) | $\tau_{hr}$ | $r^{-1}$ | $1 \times 10^{-6}$ | $1.1 \times 10^{-41}$ |

The ratio $\Delta_c$, the longer the waiting time $\tau^*_h$. An alternative interpretation of this can be based on the Eq. (58) for waiting time, $\tau_{hr} \propto (r \rho_p m_r s_r)^{-1}$. (105)

Let us consider a series of possible characteristic haloes of different size $r$, mass $m_p$, and density $\rho_p \propto \Delta_c \rho_{DM}$, where $\Delta_c$ is the density ratio. Since characteristic haloes have the fastest mass accretion such that the shell thickness on the halo boundary $s_r = 6p$ should be the same for all of these haloes. According to the scaling law in (105), the waiting time on the halo boundary should follow $\tau^*_h = \tau_{hr} \propto (r \rho_p m_r)^{-1} \propto (r^{-3} r^{5/3})^{-1} \propto \rho_p \propto \Delta_c \rho_{DM}$. (106)

Therefore, the waiting time is proportional to the density ratio $\Delta_c$. In the limiting situation $\Delta_c = 1$ or the halo has the same density as the background, the waiting time satisfies $\tau^*_h \geq t_p$ (Eq. (104)). For waiting time $\tau^*_h$ satisfying $\tau^*_h \geq \Delta_c t_p$, a more stringent constraint for particle mass reads (Eq. (92)) $m_p = 1.5 \times 10^{26} \frac{kg}{s} \tau^*_h \geq 1.6 \times 10^{-15} kg \approx 10^{12} GeV$. (107)

All these constraints (Eqs. (90), Eq. (93), (100), and (107)) exclude the standard WIMPs and strongly suggest a heavy dark matter scenario, where superheavy right-handed neutrinos might be a good candidate. Detailed discussion on the possible properties of dark matter particles from the mass and energy cascade is presented in a different article [40]. Let us take the particle mass $m_p = 10^{12} GeV$ in Eq. (107) as an example; all other quantities can be easily obtained from the scaling laws (Eq. (70)) and equations in this section. Table 2 summarizes these values of relevant physical quantities on the smallest and largest scale of $r$ in haloes of characteristic mass $m^*_h$ at the current epoch. For $m_p = 10^{12} GeV$, on the halo boundary, the waiting time $\tau_{gr} = \tau_{hr} = \Delta_c t_p$ and the jumping length $s^*_h = l_p$.

7 CONCLUSIONS

This paper presents a theory for the mass and energy cascade of dark matter on two different levels. On the system level, the cascade in halo mass space leads to the distribution of haloes with respect to halo mass, i.e., the halo mass function. For individual haloes, the cascade along the radial direction leads to the distribution of particles in haloes, i.e., the halo density profile. On both levels, the mass and energy cascade in dark matter flow establishes a statistically steady state to continuously release the system energy and maximize the system entropy. The key feature of the statistically steady state is scale-independent rates of cascade such that the statistical structures of the haloes are self-similar and scale-free, and there is no net accumulation of mass and energy on any intermediate scales.

For the mass cascade in halo mass space, the net mass transfer is upward in a bottom-up fashion, i.e., an inverse cascade that is consistent with the hierarchical structure formation. Two distinct ranges are identified, i.e., a propagation range with a scale-independent rate of mass transfer $\varepsilon_m$ below a characteristic mass $m^*_h$ and a deposition range to actively consume the mass cascaded from small scales to grow halo larger than $m^*_h$ (Fig. 3). The inverse mass cascade leads to a random walk of haloes in the mass space with a position-dependent waiting time $\tau_g$. The distribution of haloes in halo mass space, i.e., the halo mass function, is naturally given by the solution of the corresponding Fokker-Planck equation for halo random walk (double-$\lambda$ mass function in Eq. (33)). This approach for halo mass function is simple without resorting to a spherical or elliptical collapse model. The non-Gaussian features in the density field developed at small scales are highly expected to be an important signature of the inverse mass cascade. Since haloes have finite kinetic and potential energy, there also exists an inverse cascade of kinetic energy at a rate of $\varepsilon_u$ from small to large mass scales and a direct cascade of the potential energy at a rate of $1.4 \varepsilon_u$ from large to small mass scales (Fig. 8). Universal scaling laws in mass space can be developed with small-scale permanence for halo group mass $m_g$ (Fig. 5).

The N-body system in an expanding background is not energy conserved. The total energy continuously decreases with time (Fig. 1). This is facilitated by the energy cascade from large to small haloes in halo mass space and from large to small scales in individual haloes, such that the energy is “dissipated” on small scales. The energy cascade establishes a statistically steady state to continuously release halo energy and maximize halo entropy. In these haloes, the outflow of halo mass due to the Hubble expansion balances the inflow due to the halo mass accretion and self-gravity (Fig. 9). The net mass transfer vanishes due to the stable clustering hypothesis. The kinetic energy is inversely transferred from small to large scale $r$ at a rate-dependent rate $\varepsilon_u(m^*_h, z) \propto m^*_h^{-3/2} \dot{a}^{-1}$. In contrast, the potential energy is directly cascaded from large to small scales at a rate of $1.4 \varepsilon_u$ (Fig. 11). The outflow of mass can be described by the random walk of particles with a position-dependent waiting time $\tau_{gr}$. The distribution of particles in haloes, i.e., the halo density profile, can be analytically derived from the corresponding Fokker-Planck equation (double-$\lambda$ profile in Eq. (80)). Universal scaling laws (Eqs. (70)) can be identified with small-scale permanence for the halo density profile (Fig. 13). The different inner density slopes of simulated haloes can be explained by the nonzero net mass flux and energy flux in haloes (Eq. (84)).

Since the waiting time and jumping length for particle random walk in haloes depends on the particle mass $m_p$ (Eq. (72)), new constraints on the particle mass can be identified from the physical constraints on these quantities. Based on the assumption that the waiting time should be greater than the Planck time (the smallest unit of time) or the jumping length should be greater than the Planck length (the smallest unit of length), we propose a new constraint for particle mass $m_p \geq 10^{-15} kg$ or $10^{12} GeV$ (Eqs. (90), Eq. (93), (100), and (107)). These constraints exclude the standard WIMPs and strongly suggest a heavy dark matter scenario (superheavy right-handed neutrinos, etc.). This constraint is also consistent with the particle mass obtained by extending the established scaling laws to the smallest free-streaming scale [40].
ACKNOWLEDGMENTS

This research was supported by the Laboratory Directed Research and Development at Pacific Northwest National Laboratory (PNNL). PNNL is a multiprogram national laboratory operated for the U.S. Department of Energy (DOE) by Battelle Memorial Institute under contract no. DE-AC05-76RL01830.

DATA AVAILABILITY

Two datasets underlying this article, that is, halo-based and correlation-based statistics of dark matter flow, are available on Zenodo [51, 52], along with the accompanying presentation slides ‘A comparative study of dark matter flow & hydrodynamic turbulence and its applications’ [53]. All data are also available on GitHub [54].

References

[1] Z. Lukic, K. Heitmann, S. Habib, S. Bashinsky, and P. M. Ricker, Astrophysical Journal 671, 1160 (2007).
[2] G. R. Blumenthal, S. M. Faber, J. R. Primack, and M. J. Rees, Nature 311, 517 (1984).
[3] J. Neyman and E. L. Scott, Astrophysical Journal 116, 144 (1952).
[4] A. Cooray and R. Sheth, Physics Reports-Review Section of Physics Letters 372, 1 (2002).
[5] W. H. Press and P. Schechter, Physical Review 187, 425 (1974).
[6] J. R. Bond, S. Cole, G. Efstathiou, and N. Kaiser, Astrophysical Journal 379, 440 (1991).
[7] K. Tomita, Progress of Theoretical Physics 42, 9 (1969).
[8] J. E. Gunn and J. R. Gott, Astrophysical Journal 176, 1 (1972).
[9] M. Musso and R. K. Sheth, Monthly Notices of the Royal Astronomical Society: Letters 423, L102 (2012).
[10] A. Paranjape and R. K. Sheth, Monthly Notices of the Royal Astronomical Society 426, 2789 (2012).
[11] M. Maggiore and A. Riotto, The Astrophysical Journal 711, 907 (2010).
[12] A. Del Popolo and P. Kroupa, Astronomy & Astrophysics 502, 733 (2009).
[13] B. Moore, F. Governato, T. Quinn, J. Stadel, and G. Lake, Astrophysical Journal 499, L5 (1998).
[14] A. Klypin, A. V. Kravtsov, J. S. Bullock, and J. R. Primack, Astrophysical Journal 554, 903 (2001).
[15] E. Bertschinger, Astrophysical Journal Supplement Series 8, 39 (1985).
[16] J. A. Fillmore and P. Goldreich, Astrophysical Journal 281, 1 (1984).
[17] J. F. Navarro, C. S. Frenk, and S. D. M. White, Astrophysical Journal 490, 493 (1997).
[18] J. F. Navarro, E. Hayashi, C. Power, A. R. Jenkins, C. S. Frenk, S. D. M. White, V. Springel, J. Stade, and T. R. Quinn, Monthly Notices of the Royal Astronomical Society 349, 1039 (2004).
[19] J. F. Navarro, A. Ludlow, V. Springel, J. Wang, M. Vogelsberger, S. D. M. White, A. Jenkins, C. S. Frenk, and A. Helmi, Monthly Notices of the Royal Astronomical Society 402, 21 (2010), https://academic.oup.com/mnras/article-pdf/402/1/21/18573804/mnras0402-0021.pdf.
[20] J. Diemand and B. Moore, Advanced Science Letters 14, 297 (2011), arXiv:0906.4340 [astro-ph.CO].
[21] F. Governato, C. Brook, L. Mayer, A. Brooks, G. Rhee, J. Wadsley, P. Jonsson, B. Willman, G. Stinson, T. Quinn, and P. Madau, Nature 463, 203 (2010).
[22] D. McKeown, J. S. Bullock, F. J. Mercado, Z. Hafen, M. Boylan-Kolchin, A. Wetzl, L. Necib, P. F. Hopkins, and S. Yu, MNRAS 513, 55 (2022), arXiv:2111.03076 [astro-ph.GA].
[23] A. Lazar, J. S. Bullock, M. Boylan-Kolchin, T. K. Chan, P. F. Hopkins, A. S. Graus, A. Wetzl, K. El-Badry, C. Wheeler, M. C. Straight, D. Kereš, C.-A. Faucher-Giguère, A. Fitts, and S. Garrison-Kimmel, Monthly Notices of the Royal Astronomical Society 497, 2393 (2020).
[24] W. J. G. de Blok and A. Bosma, Astronomy & Astrophysics 385, 816 (2002), arXiv:astro-ph/0212076 [astro-ph].
[25] W. J. G. de Blok, A. Bosma, and S. McGaugh, Monthly Notices of the Royal Astronomical Society 340, 657 (2003), https://academic.oup.com/mnras/article-pdf/340/2/657/18649385/340-2-657.pdf.
[26] R. A. Swaters, B. F. Madore, F. C. van den Bosch, and M. Balcells, Astrophys. J. 583, 732 (2003), arXiv:astro-ph/0210152.
[27] R. Kuzio de Naray and T. Kaufmann, Monthly Notices of the Royal Astronomical Society 414, 3617 (2011), https://academic.oup.com/mnras/article-pdf/414/3617/18715139/mnras0414-3617.pdf.
[28] L. F. Richardson, Weather Prediction by Numerical Process (Cambridge University Press, Cambridge, UK, 1922).
[29] R. H. Kraichnan, Physics of Fluids 10, 1417 (1967).
[30] Z. Xu, A&A 675, A92 (2023), arXiv:2110.03126 [astro-ph].
[31] D. Nelson, A. Pillepich, S. Genel, M. Vogelsberger, V. Springel, P. Torrey, V. Rodriguez-Gomez, D. Sijacki, G. Snyder, B. Griffen, F. Marinacci, L. Blecha, L. Sales, D. Xu, and L. Hernquist, Astronomy and Computing 13, 12 (2015).
[32] C. S. Frenk, J. M. Colberg, H. M. P. Couchman, G. Efstathiou, A. E. Evrard, A. Jenkins, T. J. MacFarland, B. Moore, J. A. Peacock, F. R. Pearce, P. A. Thomas, S. D. M. White, and N. Yoshida, arXiv:astro-ph/0007362v1 (2000), 10.48550/arXiv:astro-ph/0007362.
[33] A. Jenkins, C. S. Frenk, F. R. Pearce, P. A. Thomas, J. M. Colberg, S. D. M. White, H. M. P. Couchman, J. A. Peacock, G. Efstathiou, and A. H. Nelson, Astrophysical Journal 499, 20 (1999).
[34] J. M. Colberg, S. D. M. White, A. Jenkins, and F. R. Pearce, Monthly Notices of the Royal Astronomical Society 308, 593 (1999).
[35] R. K. Sheth, H. J. Mo, and G. Tormen, Monthly Notices of the Royal Astronomical Society 323, 1 (2001).
[36] P. J. E. Peebles, The Large-Scale Structure of the Universe (Princeton University Press, Princeton, NJ, 1980).
[37] Z. Xu, arXiv e-prints , arXiv:2202.04054 (2022).
[38] W. M. Irvine, Local Irregularities in a Universe Satisfying the Cosmological Principle, Thesis, HARVARD UNIVERSITY (1961).
[39] D. Layzer, Astrophysical Journal 138, 174 (1963).
[40] Z. Xu, arXiv e-prints , arXiv:2202.07240 (2022).
[41] Z. Xu, Scientific Reports 13, 16531 (2023), arXiv:2210.01200 [astro-ph].
[42] R. L. Stratonovich, SIAM Journal on Control 4, 362–371 (1966).
[43] R. Metzler, J. H. Jeon, A. G. Cherstvy, and E. Barkai, Physical Chemistry Chemical Physics 16, 24128 (2014).
[44] A. Klypin, G. Yepes, S. Gottlober, F. Prada, and S. Hess
mass fraction of a halo group of size $\nu = 1$ at different redshifts to the concept of halo merger trees, the starting point is to define transfer (cascade) between halo groups of different sizes. Similarly the same mass of a halo group of size $\nu = 1$ that is inherited from the halo group of size $\nu = 0$ for halo group of size $n_p1 = 10$ at $z_1 = 0.1$ inherited from halo groups of different sizes at an earlier redshift $z_2 = 0.3$ (forward mass redistribution). Similarly, backward redistribution gives the fraction of mass passed to halo groups of different sizes at a later redshift $z_3 = 0.0$. Halo inherits and passes most of its mass from to and haloes of similar size (locality in mass scale).

mass of a given halo group. Figure A1 provides an example for two functions. For given $z_1$, $n_p1$, and $z_2$ or $z_3$, the normalization condition requires,

$$\sum_{n_p2} D_{FM}(z_1, n_p1, z_2, n_p2) = 1$$

and

$$\sum_{n_p3} D_{BM}(z_1, n_p1, z_3, n_p3) = 1.$$  

Figures. A2 and A3 plot the forward and backward redistribution functions as a function of the halo group size $n_p2$ or $n_p3$ for five different group sizes $n_p1 = 2, 5, 10, 20, 50$, respectively. Halo groups of different sizes $n_p1$ inherit and pass their mass to a distribution of halo group sizes. The interaction among halo groups is shown to be local on the mass scale. This demonstrates that smooth and minor merging is dominant over the major merging between two haloes of comparable mass. The halo group of size $n_p1$ inherits or passes most of its mass via merging/breaking between haloes of similar (neighboring) size to $n_p1$ and single mergers. Therefore, there are two peaks for forward/backward mass redistribution functions at around $n_p1 \approx n_p1$ (halo groups of similar size) and $n_p1 \approx 1$ (single mergers), where $i = 2$ or $3$ for forward or backward functions, respectively.

Groups of large haloes inherit/pass their mass from/to a wider distribution of halo sizes. In comparison, groups of small haloes inherit/pass their mass from/to a relatively narrower distribution of halo sizes. Both mass redistribution functions are not symmetric about the halo size $n_p1$, with more mass inherited from halo groups below the size $n_p1$ and more mass passed to halo groups above the size $n_p1$, i.e., a net transfer of mass to large scales. The sharp peaks for halo groups of smaller size and the widespread distribution for halo groups of larger size in Figs. A2 and A3 indicate that small haloes have relatively longer lifespans and can exist for a longer time such that most small haloes will remain in the same group at a later redshift. Large haloes tend to have a relatively shorter lifespan. Halo lifespan will be further discussed in Section 4.4.

**APPENDIX A: MASS CASCADE IN HALO MASS SPACE**

This section presents some additional details on the inverse mass cascade in halo mass space in matter dominant era.

### A1 Mass redistribution among halo groups

To study the mass transfer between different mass scales, we divide the entire system into two subsystems: 1) the out-of-halo subsystem with a total mass of $M_o$ includes all masses that do not belong to any haloes; 2) the halo subsystem with a total mass of $M_h$ includes all masses contained in all haloes. The inverse mass cascade involves continuous mass and energy exchange between two subsystems.

All haloes in the halo subsystem can be grouped into haloes with the same mass $m_h$ or particle number $n_p$. We focus on the mass transfer (cascade) between halo groups of different sizes. Similarly to the concept of halo merger trees, the starting point is to define functions that describe the mass redistribution among halo groups at different redshifts $z$. The forward mass redistribution function $D_{FM}(z_1, n_p1, z_2, n_p2)$ describes the mass fraction of a halo group of size $n_p1$ at the redshift $z_1$ that is inherited from the halo group of size $n_p2$ at an earlier redshift $z_2$. Similarly, the backward mass redistribution function $D_{BM}(z_1, n_p1, z_3, n_p3)$ can be defined as the mass fraction of a halo group of size $n_p1$ at redshift $z_1$ that will be passed to the halo group of size $n_p3$ at a later redshift $z_3$. These functions precisely quantify the progenitor and inheritor of the total

$$D_{FM}(z_1, n_p1, z_2, n_p2) = \frac{\sum_{n_p1} D_{FM}(z_1, n_p1, z_2, n_p2)}{\sum_{n_p2} D_{FM}(z_1, n_p1, z_2, n_p2)}$$

and

$$D_{BM}(z_1, n_p1, z_3, n_p3) = \frac{\sum_{n_p1} D_{BM}(z_1, n_p1, z_3, n_p3)}{\sum_{n_p3} D_{BM}(z_1, n_p1, z_3, n_p3)}.$$
Mass and energy cascade in dark matter flow

Figure A2. Forward mass redistribution function $D_{FM}(z_1, n_{p1}, z_2, n_{p2})$ from $z_2 = 0.3$ to $z_1 = 0.1$ for five different halo groups of sizes $n_{p1} = 2, 5, 10, 20, 50$, respectively. The figure shows the mass fraction of a halo group of size $n_{p1}$ at $z_1 = 0.1$ inherited from halo groups of various sizes at an earlier redshift $z_3 = 0.3$. The interaction among groups of haloes is shown to be local in the mass space, i.e., the halo group of size $n_{p1}$ inherits its mass mostly from the interaction between halo groups of similar (neighboring) sizes to $n_{p1}$ and single mergers. Note that there are two peaks at around $n_{p2} = n_{p1}$ (halo groups of similar size) and $n_{p2} = 1$ (single merger). Halo groups of larger size inherit their mass from a wider distribution in size $n_{p2}$.

Figure A3. Backward mass redistribution function $D_{BM}(z_1, n_{p1}, z_3, n_{p3})$ from $z_1 = 0.1$ to $z_3 = 0.0$ for five different halo groups of sizes $n_{p1} = 2, 5, 10, 20, 50$, respectively. The figure shows the mass fraction of a halo group of size $n_{p1}$ at $z_1 = 0.1$ passed to halo groups of various sizes at later redshift $z_3 = 0.0$. Again, the interaction among groups of haloes is shown to be local in the mass space, i.e., the halo group of size $n_{p1}$ passes most of its mass via merging/breaking to halo groups of similar (neighboring) size to $n_{p1}$ and single mergers. Hence, there are two peaks at around $n_{p3} = n_{p1}$ and $n_{p3} = 1$. Halo groups of larger size pass their mass to a wider distribution of halo group sizes at a later redshift. In comparison, halo groups of smaller size pass their mass to a relatively narrower distribution in size $n_{p3}$.

Figure A4. The net mass redistribution function $D_{NM}$ as a function of halo group size $n_{p2}$ for five different halo groups sizes $n_{p1}$, with $D_{NM} < 0$ indicating that the halo group of size $n_{p1}$ inherits more mass from the halo group of size $n_{p2}$ than the mass it passes to the group of the same size $n_{p2}$. The net effect of mass redistribution is that haloes are transferring mass from smaller scales to larger scales, i.e., an inverse mass cascade in mass space. In contrast, direct mass cascade refers to the mass transfer from larger to smaller scales.

To determine the direction of the mass cascade, we introduce a net mass redistribution function $D_{NM}$ as the difference between the backward and forward mass redistribution functions,

$$D_{NM}(z_1, n_{p1}, z_2, n_{p2}) = D_{BM}(z_1, n_{p1}, z_2, n_{p2}) - D_{FM}(z_1, n_{p1}, z_3, n_{p2})$$  \[A2\]

The net mass redistribution function $D_{NM}$ measures the net effect of the halo group size $n_{p1}$ at redshift $z_1$ on the mass cascade of halo group size $n_{p2}$ from redshift $z_2$ to $z_3$, with $D_{NM} < 0$ indicating that the halo group of size $n_{p1}$ inherits more mass from the halo group size $n_{p2}$ than the mass it passes to the halo group of the same size $n_{p2}$. Obviously, from Eq. (A1),

$$\sum_{n_{p2}} D_{NM}(z_1, n_{p1}, z_2, n_{p2}) = 0.$$  \[A3\]

Figure A4 plots the net mass redistribution function $D_{NM}$ for five halo group sizes $n_{p1}$, with $D_{NM} < 0$ for halo groups $n_{p2}$ smaller than size $n_{p1}$ and $D_{NM} > 0$ for halo groups $n_{p2}$ larger than size $n_{p1}$. The net effect is that haloes are transferring mass from small mass scales to large mass scales, i.e., an inverse mass cascade in the halo mass space. In contrast, the direct mass cascade refers to the transferring of mass from large to small mass scales. In short, three distinct features of inverse mass cascade can be identified:

(i) Locality: the transferring of mass is local in mass space. Haloes inherit/pass their mass mostly from/to haloes of the same or similar size. The interaction among haloes is shown to be local on a mass scale. For any finite time interval $\Delta t$, the interaction (merging/breaking) between haloes can involve multiple haloes of different sizes. However, for an infinitesimal time interval $\Delta t \to 0$, the interaction is most likely between two haloes of very different sizes (haloes of a similar size and a single merger).

(ii) Asymmetry: mass transfer across halo groups is a two-way process with mass cascading both upward and downward in the
halo mass space. However, the net mass transfer is upward, i.e., the structure redistribution functions of a given halo size \( n_{p1} \) are asymmetric about \( n_{p1} \), with more mass inherited from halo groups smaller than \( n_{p1} \) (via halo merging) and less mass inherited from halo groups larger than \( n_{p1} \) (via halo breaking-up). Larger-size haloes accrete their mass from a relatively wider size range of haloes, while smaller-size haloes accrete their mass from a narrower size range of haloes.

(iii) Inverse: mass cascade through halo groups of different sizes is two-way and asymmetric. The net effect is that haloes are transferring mass from smaller to larger mass scales, with halo merging being dominant over the halo breaking up, i.e., an inverse mass cascade.

A2 Time scales in mass cascade

In this section, we try to develop a few elementary ideas about the time and mass scales for inverse mass cascades. We observe that there exists a broad spectrum of halo sizes. This can be a direct result of maximizing system entropy [30]. The smallest halo is often created by gravitational collapse at the smallest scale that merges with other haloes and passes their mass onto larger haloes. The larger haloes are themselves transitory and pass their mass to even larger haloes, and so on. At every instant \( t \), there is a continuous cascade of mass from the smallest to the largest mass scales that we assumed to be a characteristic mass \( m^*_h \).

Let the time scale \( \tau_h (m_h, a) \) be the average waiting time of a single merging event with a single merger for a halo group of mass \( m_h < m^*_h \) at scale factor \( a \). The rate at which mass is passed up from this group is \( \epsilon_m \sim -m_h/\tau_h \) (negative sign for inverse mass cascade). When the system is in a statistically steady state, this rate of mass transfer must match exactly the rate of mass injection into the halo sub-system at the smallest scales \( m_h \to 0 \) and the rate of mass dissipation at the largest mass scale \( m^*_h \). If this is not the case, there would be an accumulation of mass at some intermediate scale below \( m^*_h \). We exclude this possibility because we want the statistics of halo groups to be self-similar and scale-free for halo groups of mass less than \( m^*_h \) once a statistically steady state is established. This means a mass propagation range with halo mass \( m_h < m^*_h \),

\[
-\epsilon_m \sim m_h/\tau_h = m^*_h/\tau^*_h, \tag{A4}
\]

where the mass flux \( \epsilon_m \) is independent of the halo mass \( m_h \).

With the system in the statistically steady state, halo groups with mass below \( m^*_h \) \( (m_h < m^*_h) \) propagate the mass to larger scales without any net accumulation of mass in that group. The total mass in the group \( m_g = m_h n_h \) should be time-invariant, where \( n_h \) is the number of haloes in that group that should also be time-invariant. Mass cascade in this range does not contribute to growing the halo group mass \( m_g \). The average waiting time \( \bar{\tau}_h \) (halo lifespan) for a given halo in the group can be calculated,

\[
\bar{\tau}_h (m_h, a) = \sum_{k=1}^{\infty} \frac{k \tau_k}{n_h} \left( \frac{n_h - 1}{n_h} \right)^{k-1}
= \frac{\tau_h}{n_h} \left( \frac{n_h - 1}{n_h} \right)^{k-1} + \ldots + n_h \tau_h,
\]

where \( k \) is the number of time intervals \( \tau_k \) for that halo to merge with a single merger. All haloes in the same group are assumed to merge with a single merger with the same probability during the time interval of \( \tau_k \). A second time scale \( \tau_g (m_h, a) \) reads

\[
\tau_g (m_h, a) = \bar{\tau}_h (m_h, a) = n_h \tau_h = -m_g/\epsilon_m. \tag{A6}
\]

which is the average waiting time (lifespan) for a given halo in a halo group of mass scale \( m_h \), or equivalently the time required to cascade the entire mass \( m_g \) of that halo group. The time scale \( \tau_g \) should decrease with \( m_h \) due to faster mass accretion of larger haloes.

Let \( M_h (a) \) be the total mass in the halo sub-system at physical time \( t \) or scale factor \( a \). The third time scale \( \tau_M (a) \) is introduced as the time required to cascade the entire mass in the halo sub-system,

\[
\tau_M (a) = -M_h (a)/\epsilon_m (a) \sim t, \tag{A7}
\]

which is expected to be on the order of the current cosmic time \( \tau \).

We are now ready to determine the characteristic mass scale \( m^*_h \) for mass cascade. Let \( \tau_g (m_h, a) \) be the average waiting time for a halo of mass \( m_h \) to merge with a single merger of mass \( m_p \) at physical time \( t \). The fourth time scale \( \tau_f (m_h, a) \) that we will introduce is

\[
\tau_f (m_h, a) = \tau_g (m_h, a) n_p = \tau_g (m_h, a) m_h/m_p, \tag{A8}
\]

where \( n_p \) is the number of particles in that given halo. The time scale \( \tau_f \) approximately represents the time the entire halo of mass \( m_h \) was formed via a sequence of merging events \( (n_p \) times) with single mergers of mass \( m_p \). Let’s assume a typical halo of mass \( m^*_h \) that is constantly growing with the waiting time exactly to be \( \tau_g \) for every single merging event during its entire mass accretion history. The actual waiting time of haloes can be random and either less or greater than \( \tau_g \). The mass accretion of that typical halo should read

\[
\frac{dm^L_h}{dt} = \frac{m_p}{\tau_L} = \frac{n^L_p m_p}{n^L_p \tau_L} = \frac{m^L_h}{n^L_p \tau_L},
\]

where \( n^L (a) = \tau_g \left( m^L_h, a \right) \). We further have (from Eq. (A9)),

\[
\frac{d \ln m^L_h}{dt} = \frac{t}{n^L_p \tau_L} = \frac{t}{\tau_f \left( m^L_h, a \right)}, \tag{A10}
\]

where we should expect \( \tau_f \left( m^L_h, a \right) \sim t \) if haloes of mass \( m^L_h \) \( (t) \) \( \sim \) \( m^*_h \) is the largest halo at present epoch. Here \( n^L_p = m^L_h/m_p \) is the number of particles in that typical halo. It turns out that this is the case (Eq. (A51)). We expect that large haloes require more time to form, and the time scale \( \tau_f \) increases with the halo size \( m_p \). Obviously, the four time scales we introduced satisfy the inequality

\[
\tau_M (a) \geq \tau_f (m_h, a) \geq \tau_g (m_h, a) \geq \tau_h (m_h, a). \tag{A11}
\]

For small haloes with mass \( m_h < m^*_h \), we expect the time scale \( \tau_f (m_h, a) \ll t \) to have sufficient time to form these haloes. For haloes with mass \( m_h > m^*_h \), the time required to form that halo \( \tau_f (m_h, a) \gg t \) and these haloes are very rare to find at current time \( t \). The time required to form haloes of a characteristic mass \( m^*_h \) should be exactly on the order of the current physical time \( t \), i.e. \( \tau_f \left( m^*_h, a \right) \sim t \sim \tau_M (a) \), from which we can derive,

\[
\frac{m^*_h}{\epsilon_m} n^*_p \sim \frac{M_h (a)}{\epsilon_m} \sim \frac{1}{H^2} \tag{A12}
\]

and

\[
m^*_h \sim \frac{M_h (a)}{n^*_p n^*_p} \sim \frac{\epsilon_m (a)}{H n^*_p n^*_p}, \text{ or } \frac{M_h (a)}{m^*_h} \sim \frac{m^*_h}{m_p} \sim n^*_p. \tag{A13}
\]

Here \( n^*_p, n^*_h \), and \( m^*_h \) are the number of particles in haloes of the characteristic mass \( m^*_h \) number of haloes in halo group of mass \( m^*_h \), and the total mass of that group, respectively. There is not enough time to form haloes larger than the characteristic mass \( m^*_h \). This does not exclude the existence of these large haloes because of the
random nature of waiting time. In N-body simulations, the total number of particles in the system scales as \( N \sim M_h/m_p \sim n_p^\alpha n_h^{\alpha-2} \) from Eq. (A13). A dimensionless number \( z_h \) can be defined for each halo group to reflect the competition between the local rate of mass transfer \((1/\tau_f)\) and the Hubble constant \( H \),

\[
z_h = \frac{M_h(a)}{m_p n_p} - \frac{\varepsilon_m(a)}{m_p n_p} H \sim \frac{t}{\tau_f} \quad \text{and} \quad z^*_h = \frac{M_h(a)}{m^*_p n^*_p},
\]

(A14)

where \( z_h \) decreases with halo size and \( z^*_h \) for haloes with characteristic mass \( m^*_h \) should be on the order of one. The exact value of \( z^*_h \) can be determined with Eq. (A53) \( (z^*_h \approx 1/\beta_0) \).

In short, two distinct ranges can be identified for inverse mass cascade from time/mass scales: 1) mass propagation range with \( m_h < m^*_h \), where the system is in a statistically steady state with a scale-independent mass flux \( \varepsilon_m \) and a time-invariant group mass \( m_g = m_h n_h \) (Fig. A5); 2) mass deposition range with \( m_h > m^*_h \), where mass cascaded from small scales is actively consumed to grow haloes. Halo group mass \( m_g \) is increasing with time in this range.

### A3 Mass flux and mass transfer functions

To quantify the mass cascade across halo groups, we introduce the real-space mass flux function that quantifies the net transfer of mass from all haloes smaller than the size \( m_h \) to all haloes greater than \( m_h \). The mass flux function \( \Pi_m (m_h, a) \) can be defined as

\[
\Pi_m (m_h, a) = - \frac{\partial}{\partial t} \left[ M_h(a) \int_{m_h}^\infty m_f \left( m_h, m^*_h \right) dm \right].
\]

(A15)

Here \( M_h(a) \) is the total mass in the halo sub-system that increases with scale factor \( a \). The halo mass function \( m_f \left( m_h, m^*_h \right) \) is the probability distribution of total mass \( M_h(a) \) with respect to halo mass \( m_h \).

Since halo mass \( m_h \) and scale factor \( a \) are the only two independent variables, the mass flux function can be written as a function of \( m_h \) and \( a \), i.e. \( m_f \left( m_h, m^*_h \right) = m_f \left( m_h, m^*_h(a) \right) \). The characteristic mass scale \( m^*_h(a) \) varies with the scale factor \( a \) only, a monotonically increasing function reflecting the fact that larger haloes emerge at a later time. The mass flux function \( \Pi_m \) across halo groups should be independent of halo size \( m_h \) for halo groups smaller than \( m^*_h(a) \), where the mass flux function reduces to

\[
\varepsilon_m(a) = \Pi_m (m_h, a) \quad \text{for} \quad m_h < m^*_h.
\]

(A16)

The constant mass flux (or the mass dissipation rate \( \varepsilon_m \)) that is independent of mass scale \( m_h \) cascades mass from the smallest mass scale to the characteristic scale \((0 < m_h < m^*_h)\) in the mass propagation range. The total mass of the halo group of size \( m_h \) is

\[
m^*_g (m_h, a) = M_h(a) m_f \left( m_h, m^*_h \right) m_p.
\]

(A17)

A direct result of the scale-independent mass flux is that the group mass \( m_g (m_h, a) \) of a halo group of size \( m_h \) reaches a steady state (Not varying with time, see Eq. (A19)). The total mass injected at the smallest mass scale (mass continuously injected from the out-of-halo sub-system into the halo sub-system) passes through the propagation range. It is consumed to grow the mass of halo groups above the characteristic mass \( m^*_h \) (Fig. A5).

The real-space mass transfer function can be defined as the derivative of the mass flux function with respect to halo mass,

\[
T_m (m_h, a) = \frac{\partial \Pi_m (m_h, a)}{\partial m_h} = \frac{\partial}{\partial t} \left[ M_h(a) m_f \left( m_h, m^*_h \right) \right] = \frac{\partial m_g (m_h, a)}{m_p \partial t},
\]

(A18)

which quantifies the rate of change of group mass \( m_g (m_h, a) \). For the mass propagation range,

\[
T_m (m_h, a) = 0 \quad \text{and} \quad \frac{\partial m_g (m_h, a)}{\partial t} = 0 \quad \text{for} \quad m_h < m^*_h.
\]

(A19)

The mass transfer function \( T_m (m_h, a) \) describes the removal of mass from a small scale and the deposition of mass at a large scale \((T_m (m_h, a) > 0)\) for \( m_h > m^*_h \).

Since the mass dissipation rate \( \varepsilon_m(a) \) is independent of halo size for \( m_h < m^*_h \), we may compute the mass flux function at the smallest scales using Eq. (A15) with \( m_h = 0 \),

\[
\varepsilon_m(a) = \Pi_m (m_h = 0, a) = -\frac{\partial M_h(a)}{\partial t} \quad \text{for} \quad m_h < m^*_h.
\]

(A20)

Let the time scale \( \tau_h (m_h, a) \) be the average time for a single merging event in a halo group of mass \( m_h \), or equivalently an event frequency \( f_h (m_h, a) \). The rate of mass transfer from the scale below \( m_h \) to the scale above \( m^*_h \) is,

\[
\varepsilon_m(a) = -a_0 f_h (m_h, a) \quad \text{for} \quad m_h < m^*_h.
\]

(A21)

where \( a_0 \) is a numerical factor on the order of unity. The event frequency \( f_h (m_h, a) \) should be proportional to the number of haloes in the group (term 1 in Eq. (A22)) and the surface area of the haloes (term 2). Because the halo interactions in mass space are local, we can assume the mass cascade involves merging events between a halo of similar size and a single merger (Figs. A2 and A3). Halo group with more haloes (term 1 in Eq. (A22)) and haloes with a larger surface area (proportional to \( m^*_h \) i.e., the term 2 in Eq. (A22)) have a greater probability of merging with a single merger,

\[
f_h (m_h, a) = f_0(a) M_h(a) m_f \left( m_h, m^*_h \right) m_p \left( \frac{m_h}{m_p} \right)^A \left( \frac{1}{2} \right),
\]

(A22)

where \( f_0(a) \) is a fundamental frequency for the merging between two single mergers at a given redshift \( z \) or scale factor \( a \) and may be used to determine dark matter particle mass \( m_p \) (Eq. (A32)).
The characteristic time of a single merging event can be written as $$\tau_n (m_h, a) = 1/(\alpha_0 f_0).$$

Without loss of generality, the exponent $$\lambda$$ is a halo geometry parameter that represents the effect of halo surface area on the merging frequency $$f_0 (m_h, a)$$. For two haloes of very different sizes (merging between a large halo and a single merger), it is estimated that $$\lambda = 2/3$$ with $$m_h \propto r_h^{3/2} \propto A_h^{3/2}$$, where $$A_h$$ is the halo surface area. For small haloes, merging is more likely between two haloes of comparable sizes where $$\lambda$$ can deviate from 2/3 and approach 1.

Substitution of Eq. (A22) for the event frequency into the Eq. (A21) leads to the mass flux,

$$\varepsilon_m (a) = -\alpha_0 H (a) M_h (a) f_M (m_h, m_h^*) (a) m_p \frac{m_h^*}{m_p} \lambda.$$  \hfill (A23)

For self-similar gravitational clustering in the mass propagation range ($$m_h < m_h^*$$), the halo mass $$m_h$$ and characteristic mass scale $$m_h^*$$ are the only two controlling variables. We can simply express the mass function as $$f_M (m_h, m_h^*) \sim (m_h)^x (m_h^*)^{-x-1}.$$ Now using dimensional analysis, the only possible form of the mass function $$f_M$$ that satisfies Eq. (A23) is ($$f_M$$ should have a unit of kg in SI units and $$\varepsilon_m$$ is a function of a only and is independent of $$m_h$$),

$$f_M (m_h, m_h^*) = \beta_0 m_h^{-d} (m_h^*)^{\lambda-1}$$ for $$m_h < m_h^*$$,  \hfill (A24)

where $$\beta_0 \propto O (1)$$ is a numerical constant. The mass flux and event frequency in the mass propagation range can be expressed as (after substituting Eq. (A24) into (A23)),

$$\varepsilon_m (a) = -\alpha_0 \beta_0 \lambda_0 \lambda_0^* N m_p f_0 (a)$$ for $$m_h < m_h^*$$,  \hfill (A25)

and

$$f_h (m_h, a) = \beta_0 \lambda_0^* N m_p / m_h$$ for $$m_h < m_h^*$$,  \hfill (A26)

where a dimensionless constant $$\lambda_0$$ is defined as

$$\lambda_0 = \frac{m_h}{N m_p} \left( \frac{m_h^*}{m_p} \right) \lambda,$$  \hfill (A27)

which is a time-invariant constant and dependent only on the mass resolution $$m_p$$. For $$m_p = 2.27 \times 10^{11} M_{\odot}/h$$ from Table 1 and $$m_h^* (z = 0) \approx 2 \times 10^{13} M_{\odot}/h$$, $$\lambda_0$$ is on the order of 0.13.

The halo group mass in the propagation range (Eq. (A24)) reads

$$m_G (m_h, a) \equiv m_G (m_h) = \lambda_0 \lambda_0^* N m_p (m_p / m_h)^{\lambda},$$  \hfill (A28)

Equivalently, we have

$$m_G (m_p) = \lambda_0 \lambda_0^* N m_p,$$  \hfill (A29)

for the group mass of single mergers with $$m_h = m_p$$. The group mass $$m_G$$ is proportional to $$m_h^{-\lambda}$$ (Fig. A5). The relation between the rate of change of mass $$m_h$$ and $$m_h^*$$ can be found from Eq. (A27),

$$\frac{\partial \ln m_h}{\partial \ln a} = 1 - \lambda \frac{\partial \ln m_h^*}{\partial \ln a}.$$  \hfill (A30)

With Eqs. (A20) and (A27), we may derive the mass dissipation rate as a function of $$m_h^*$$,

$$\varepsilon_m (a) = - (1 - \lambda) \frac{\partial \ln m_h^*}{\partial \ln a} H (a) M_h (a) = -\lambda_0 (1 - \lambda) \frac{\partial \ln m_h^*}{\partial \ln a} H (a) N m_p \left( \frac{m_h^*}{m_p} \right)^{1 - \lambda}.$$  \hfill (A31)

With Eqs. (A25) and (A31), we find the fundamental frequency $$f_0 (a)$$ as a function of $$m_h^*$$,

$$f_0 (a) = \left( \frac{1 - \lambda}{\alpha_0 \beta_0} \right) \frac{\partial \ln m_h^*}{\partial \ln a} H (a) \left( \frac{m_h^*}{m_p} \right)^{1 - \lambda},$$  \hfill (A32)

that is also related to the Hubble constant $$H (a)$$ and mass resolution $$m_p$$. The characteristic time scale $$\tau_n (m_h^*, a)$$ associated with the characteristic mass $$m_h^*$$ is

$$\tau_n (a) = \frac{m_h^*}{H (a) m_p} \frac{1}{\partial \ln m_h^* / \partial \ln a} H (a) \left( \frac{m_h^*}{m_p} \right)^{1 - \lambda}.$$  \hfill (A33)

The fundamental frequency $$f_0 (a)$$ is the frequency for the elementary merging between two single mergers and is expected to decrease with time. Without loss of generality, let’s assume a power-law of $$f_0 (a) \propto a^{-\tau_0}$$ that leads to $$\varepsilon_m (a) \propto a^{-\tau_0}$$ (Eq. (A25)),

$$m_h^* \propto a^{3/2 - \tau_0} \quad (A31),$$

and $$M_h (a) \propto a^{3/2 - \tau_0} \quad (A27)$$. Once the statistically steady state is established for inverse mass cascade, the total mass of the halo sub-system increases as $$M_h (a) \propto a^{3/2 - \tau_0}$$ regardless of the value of $$\lambda$$. Obviously,

$$\frac{\partial \ln m_h}{\partial \ln a} = (1 - \lambda) \frac{\partial \ln m_h^*}{\partial \ln a} = \frac{3}{2} - \tau_0 > 0.$$  \hfill (A34)

With total mass in halo sub-system $$M_h (a)$$ increasing with the scale factor $$a$$, it requires $$0 < \tau_0 < 3/2$$. With $$\tau_0 > 0$$ and $$a \rightarrow \infty$$, the mass flux $$\varepsilon_m (a)$$ approaches zero with the entire system approaches the limiting thermodynamic equilibrium but can never reach. In this regard, the inverse mass cascade with a constant rate is a key feature of the statistically (intermediate) steady state when the system evolves toward the limiting equilibrium.

Number of haloes $$n_h (m_h)$$ in halo group with a given mass $$m_h$$ is

$$n_h (m_h) \equiv n_h (m_h) = \lambda_0 \beta_0 N m_p (m_p / m_h)^{\lambda},$$  \hfill (A35)

which does not vary with time once the statistically steady state is established. Substitution of Eq. (A34) into Eq. (A32), we can express the fundamental frequency $$f_0 (a)$$ as,

$$f_0 (a) = \frac{1}{\alpha_0 \beta_0} \left( \frac{3}{2} - \tau_0 \right) H (a) \left( \frac{m_h}{m_p} \right)^{1 - \lambda} = \frac{b_0}{\alpha_0 \beta_0} H_0 (a) a^{-\tau_0},$$  \hfill (A36)

where the mass resolution $$m_p$$ can be related to a fixed characteristic mass $$m_h^* (a)$$ at $$a = 1$$,

$$b_0 = \left( \frac{3}{2} - \tau_0 \right) \left( \frac{m_h^* (a = 1)}{m_p} \right)^{1 - \lambda},$$  \hfill (A37)

which is dependent on the mass resolution $$m_p$$ only (the mass of dark matter particle), with smaller particle mass $$m_p$$ giving rise to a greater fundamental frequency $$f_0 (a)$$ in $$N$$-body simulation. For mass resolution of $$m_p = 2.27 \times 10^{11} M_{\odot}/h$$ from Table 1 and $$m_h^* (a = 1) = 2 \times 10^{13} M_{\odot}/h$$, $$b_0 \approx 2.2$$ with $$\tau_0 = 1$$. In other words, the dark matter particle mass can be determined if the fundamental frequency $$f_0 (a)$$ can be precisely measured.

Now, we can introduce a numerical constant,

$$c_0 = b_0 \lambda_0 = \left( \frac{3}{2} - \tau_0 \right) \frac{M_h (a)}{N m_p} a^{3/2 - \tau_0},$$  \hfill (A38)

which is the mass fraction of halo sub-system $$M_h (a)$$ and not dependent on the mass resolution $$m_p$$ and scale factor $$a$$. We estimate $$c_0 \approx 0.29$$ with $$\tau_0 = 1$$, i.e. $$M_h \approx 0.58 N m_p$$ when $$a = 1$$. This information is used to study the density distributions for particles in haloes and out-of-haloes, respectively [55].

Finally we present the simplified expressions for time and mass scales and mass flux function that can be fully described as functions of halo mass $$m_h$$, scale factor $$a$$, and mass resolution $$m_p$$ with four numerical constants $$b_0$$ (pre-factor for halo mass function $$f_M$$), $$c_0$$
Mass and energy cascade in dark matter flow

\[ e_m(a) = -\left(\frac{3}{2} - \tau_0\right) H(a) M_h(a) = -c_0 H_0 N m_p a^{-\tau_0}, \quad (A39) \]

\[ \tau_h^* (a) = \frac{1}{(N \lambda_0 H(a))} \left( \frac{m_h^*}{m_p} \right)^{\frac{1}{3}} = m_h^* (a = 1) \left( \frac{\lambda_0}{H_0} a^{(3/2 - \tau_0)} \right)^{\frac{1}{3}}, \quad (A40) \]

\[ m_h^* (a) = \left( \frac{b_0}{3/2 - \tau_0} \right) \left( \frac{1}{a^{(3/2 - \tau_0)} m_p} \right) \quad (A41) \]

\[ M_h(a) = \frac{c_0}{3/2 - \tau_0} N m_p a^{(3/2 - \tau_0)}. \quad (A42) \]

Let us revisit the time scales we introduced in Section A2. The time scale \( \tau_M(a) \)

\[ \tau_M(a) = \frac{M_h(a)}{e_m(a)} = \frac{3/2}{3/2 - \tau_0} a^{3/2 \tau_0} = \frac{3/2}{3/2 - \tau_0} t, \quad (A43) \]

which is the time it takes to cascade all mass in the halo sub-system at a given scale factor \( a \) and \( \tau_M(a) \) is on the order of \( \tau_0 \). This relation might be used to determine the value of \( \tau_0 \) from N-body simulations.

The time scale \( \tau_h \) for a single merging in halo group of mass \( m_h \) is,

\[ \tau_h (m_h, a) = \frac{m_h}{e_m} \left( \frac{1}{a_0 f_0(a)} N m_p \right) = \frac{m_h}{c_0 H_0 N m_p} a^{\tau_0}. \quad (A44) \]

The time scale \( \tau_g \) that takes to cascade the group mass \( m_g \) for the halo group of mass \( m_h \) is,

\[ \tau_g (m_h, a) = \frac{m_h m_h}{e_m} \left( \frac{1}{a_0 f_0(a)} \right) \left( \frac{m_h}{m_p} \right)^{-\frac{1}{3}} = \frac{\beta_0}{(3/2 - \tau_0) H_0} \left( \frac{m_h}{m_h^* (a = 1)} \right)^{\frac{1}{3} - \frac{1}{3}} a^{\tau_0}. \quad (A45) \]

As expected, the mean waiting time (lifespan) \( \tau_g \) of a given halo decreases with halo mass as \( \tau_g \propto m_h^{-1} \) and increases with \( \alpha \). Larger haloes have a shorter lifespan. Extremely large haloes have very fast mass accretion and an infinitesimal lifespan.

The time scale \( \tau_f \) is introduced as the average time it takes to form the halo of mass \( m_h \) that reads

\[ \tau_f (m_h, a) = \frac{m_h m_h}{e_m} \left( \frac{1}{a_0 f_0(a)} \right) \left( \frac{m_h}{m_p} \right)^{-\frac{1}{3}} = \frac{\beta_0}{(3/2 - \tau_0) H_0} \left( \frac{m_h}{m_h^* (a = 1)} \right)^{\frac{1}{3} - \frac{1}{3}} a^{\tau_0}, \quad (A46) \]

which increases with halo mass as \( \tau_f \propto m_h^{-1} \) so that smaller haloes were formed earlier with smaller \( \tau_f \). The relation between time scales \( \tau_f (m_h^*, a) = \beta_0 \tau_M(a) \) can be easily obtained from Eqs. (A43) and (A46) that is consistent with our analysis in Section A2 (Eq. (A13)), i.e., \( \beta_0 M_h(a) = m_h n_p n_p m_p \) and \( \tau_f \beta_0 = 1 \) (from Eq. (A14)). Different time scales can be related to the mass flux as

\[ e_m = -\frac{M_h(a)}{\tau_M(a)} = -\frac{m_g (m_h)}{\tau_g (m_h, a)} = -\frac{m_e (m_h)}{\tau_f (m_h, a)} = -\frac{m_h}{\tau_h (m_h, a)}. \quad (A47) \]

Figure A6. The halo mass (normalized by \( 10^{12} M_\odot \)) accretion history for type II haloes, i.e. the dominant type for large haloes [see 56, Fig. 2], exhibits a power law scaling \( \propto a^{3/2} \).

The corresponding time scales at characteristic mass \( m_h^* \) are

\[ \tau_g (a) = \frac{\beta_0 m_p}{(3/2 - \tau_0) H_0 m_h^* (a = 1)} a^{-\frac{3}{2} - \frac{1}{2}} \tau_0, \quad (A48) \]

\[ \tau_f(a) = \frac{\beta_0}{(3/2 - \tau_0) H_0}. \quad (A49) \]

The two time scales \( \tau_g \propto a^{\tau_0} m_h^{-\frac{1}{2}} \) (Eq. (A45)) and \( \tau_f \propto a^{\tau_0} m_h^{-\frac{1}{2}} \) (Eq. (A46)), where larger haloes have a shorter lifespan but take more time to form. For \( \lambda = 2/3 \) and \( \tau_0 = 1 \), the lifespan \( \tau_g \) of characteristic haloes is independent of time that is consistent with Eq. (24).

Now we can track the growth of typical haloes by integrating Eq. (A9) with respect to the scale factor \( a \) and using expression of time scale \( \tau_g \) (Eq. (A45)) with initial condition \( m_h^* (a = 0) = 0 \) to obtain

\[ \frac{m_h^* (a)}{m_h (a = 1)} = \left( \frac{1 - \lambda}{\beta_0} \right)^{(1/4 - \lambda)} a^{3/2 - \tau_0}, \quad (A50) \]

and

\[ \frac{m_h^* (a)}{m_h (a)} = \left( \frac{1 - \lambda}{\beta_0} \right)^{(1/4 - \lambda)}, \quad (A51) \]

which follows the same scaling as characteristic mass scale \( m_h^* (a) \).

For large haloes with \( \tau_0 = 1 \) and \( \lambda = 2/3 \), \( m_h^* (a) \sim m^* (a) \sim a^{3/2} \). This scaling matches the mass growth of type II haloes, i.e., the dominant type for large haloes [see 56, Fig. 2], as shown in Fig. A6.

With the help of Eq. (A10), we confirm that time scale \( \tau_f \) (Eq. (A46)) to form the typical halo is on the same order of \( t \) (Eq. (A10)),

\[ \tau_f (m_h^*, a) = \frac{1 - \lambda}{1 - 2 \tau_0 / 3} t \sim t. \quad (A52) \]

Finally, it can be easily confirmed that (from Eq. (A31)),

\[ \Delta_m (a) = -\frac{d M_h (a)}{d t} = -\left( \frac{3}{2} - \tau_0 \right) H M_h (a). \quad (A53) \]

The relations between total mass of all haloes \( M_h \), typical halo mass \( m_h^* \) and mass scale \( m_h^* \) are (from Eqs. (A27), (A35) and (A50)),

\[ M_h (a) = \frac{1 - \lambda}{\beta_0} m_h^* (a) \frac{d f_0 (a)}{d t} = \frac{1 - \lambda}{\beta_0} m_h^* m_h^* n_p, \quad (A54) \]

Vol. 000, 1–28 (2024)
The mass flux function $\varepsilon_m (a)$ can be interpreted as the rate of mass change of a typical halo $m_h^*$ or $m_p^*$ multiplied by the equivalent number of that halo in the system,

$$
\varepsilon_m (a) = -\frac{dm^L}{dt} n_h^L n_p^L = -\beta_0 \omega_0 N \frac{d}{dt} \left( \frac{m_h^* n_h^* n_p^*}{m_p^*} \right)^{1-\lambda} \quad (A54)
$$

or

$$
\varepsilon_m (a) = -\frac{1-\lambda}{\beta_0} \frac{dm^*_h}{dt} n_h^* n_p^* = -\beta_0 N \frac{1}{\beta_0} \frac{d}{dt} \left( \frac{m_h^* n_h^* n_p^*}{m_p^*} \right)^{1-\lambda} \quad (A55)
$$

In summary, the mathematical model for inverse mass cascade provides the complete dependence of time/mass scales and mass flux/transfer functions on the scale factor $a$, halo mass $m_h$, and mass resolution $m_p$. An interesting case is that $\tau_0 = 1$, where the fundamental frequency $f_0 (a) \propto a^{-1}$. Note that this is the same scaling as the photon frequency decaying due to the cosmological redshift. Table A1 lists the scaling exponents with respect to $a$ for different values of $\tau_0$ and $\lambda$. The scalings of $\varepsilon_m (a)$, $f_0 (a)$ and $M_h (a)$ are only dependent on $\tau_0$, while $m_h^* (a)$ and $\tau_h^* (a)$ depend on both $\tau_0$ and $\lambda$. Table A2 summarizes the dependence on the halo size $m_h$, where $\tau_g \sim m_h^{-1}$ and $\tau_f \sim m_h^{-1}$. Table A3 presents the dependence of relevant parameters on the mass resolution $m_p$.

A4 The probability distribution of waiting time

This section discusses the probability distribution of the waiting time for a given halo to merge with a single merger. The mean waiting time for a merging event in a halo group of mass $m_h$ is $\tau_h = \tau_g / n_h$ from Eq. (A6), where $\tau_g$ is the mean waiting time for the merging of a given halo. Let the actual time interval of a single merging for a given halo be a random variable $\tau^*_g$ with its mean given by $\tau^*_g = \langle \tau^*_g \rangle = n_h \tau_h \gg \tau_h$ (from Eq. (A5)). Typical haloes accreting a mass at a fixed waiting time have a direct delta distribution with a deterministic $\tau^*_g \equiv \gamma_g$ (Eq. (A9)). The probability distribution of time $\tau^*_g = k \tau_h$ can be described by a discrete distribution $P (k, n_h)$, where $k$ is the number of relevant intervals $\tau_h$ for a given halo to wait till the first merging with a single merger (See Eq. (A55)),

$$
P (k, n_h) = P_r (X = k) = \frac{1}{n_h} \left( 1 - \frac{1}{n_h} \right)^{k-1} \quad (A56)
$$

with $\sum_{k=1}^{\infty} P (k, n_h) = 1$.

The cumulative function and moments of the probability mass function $P (k, n_h)$ are given by,

$$
Q (k, n_h) = \sum_{m=1}^{k} P (m, n_h) = 1 - \left( 1 - \frac{1}{n_h} \right)^k \quad (A57)
$$

$$
\langle k \rangle = \sum_{k=1}^{\infty} \left[ P (k, n_h) k \right] = n_h, \quad (A58)
$$

$$
\langle k^2 \rangle = \sum_{k=1}^{\infty} \left[ P (k, n_h) k^2 \right] = n_h (2n_h - 1), \quad (A59)
$$

$$
\langle k^m \rangle = \sum_{k=1}^{\infty} \left[ P (k, n_h) k^m \right] = PolyLog (-m, 1 - 1/n_h) \frac{1}{n_h - 1} \quad (A60)
$$

The probability distribution of time interval $\tau^*_g$ of a given halo finally reads (from Eq. (A56)),

$$
P (\tau^*_g, n_h) = \frac{1}{n_h (1 - 1/n_h)} \left( \frac{n_h \tau^*_g - 1}{\tau^*_g} \right) = \frac{1}{n_h} \exp \left( -\frac{\tau^*_g}{\tau_g} \right) \quad (A61)
$$

with the mean $\tau_g \sim m_h^{-1}$. The exponential distribution of waiting time $\tau^*_g$ is obtained by taking the limit $n_h \to \infty$. The probability density function of the continuous random waiting time $\tau^*_g$ reads

$$
P (\tau^*_g) = \frac{1}{\tau_g} \exp \left( -\frac{\tau^*_g}{\tau_g} \right) \quad (A62)
$$

Clearly, the waiting time $\tau^*_g$ for a given halo follows an exponential distribution that is dependent on its mean value $\tau_g$ (Eq. (A45)). Therefore, the distribution of $\tau^*_g$ is position-dependent (i.e., dependent on the halo mass $m_h$) and scale factor $a$.

A5 Heterogeneous diffusion model and halo mass functions

The heterogeneous diffusion with spatially dependent diffusivity plays an important role in many physical problems. Examples are the mass transport in porous, inhomogeneous media and plasmas. The transport in these examples involves the waiting time that is explicitly dependent on the position. Particularly, the power-law dependence of the diffusivity is natural for many systems exhibiting self-similarity, for example, the disordered materials, the diffusion on fractals, and the mass cascade of self-gravitating collisionless dark matter flow (SG-CFD) in this work. Here, a heterogeneous diffusion model can be established for inverse mass cascade with waiting time explicitly dependent on the halo mass ($\tau_g \sim m_h^{-1}$). First, the group mass $m_g$ for halo groups reads

$$
m_g (m_h, a) = M_h (a) f_M (m_h, a) m_p, \quad (A63)
$$

where $M_h$ is the total mass in halo sub-system, $f_M (m_h, a)$ is the probability distribution with respect to the haloes mass $m_h$ (mass function), and $m_p$ is the mass resolution (particle mass). The time variation of $m_g$ has two contributions from Eq. (A63),

$$
\frac{\partial m_g}{\partial a} = M_h (a) m_p \frac{\partial f_M}{\partial a} + M_h (m_h, a) m_p \frac{\partial M_h}{\partial a}, \quad (A64)
$$

where term 1 is due to the time variation of $f_M (m_h, a)$ and term 2 is from the variation of total halo mass $M_h$. For position-dependent power-law diffusivity $D_{md} = D_{m0} m_h^{2 z} a$, the dynamics of $m_g$ can be described by the heterogeneous diffusion model and transforming the derivative from $t$ to $a$,

$$
\frac{\partial m_g}{\partial a} = \frac{\partial}{\partial m_h} \left\{ D_{m0} m_h^{2 z} \frac{\partial}{\partial m_h} \left( \sqrt{D_{m0} m_h^{2 z} m_g} \right) \right\} + \frac{\partial \ln M_h m_g}{\partial a} \cdot \frac{\partial m_g}{\partial a}, \quad (A65)
$$

where term 1 describes the heterogeneous diffusion of $m_g$ in mass space and term 2 describes the source term due to the increasing total mass $M_h$ in all haloes. The boundary conditions are:

$$
\frac{\partial m_g}{\partial a} \bigg|_{m_h=0} = \frac{m_p}{\tau_g} T_{m_h} \bigg|_{m_h=0} = 0, \quad (A66)
$$

Vol. 000, 1–28 (2024)
diffusivity in halo mass space can be expressed in terms of \( \eta \).

Since \( \beta \), \( \lambda \).

For a constant diffusivity \( D_{m0} \) with respect to \( a \), we also would expect \( \tau_0 = 1 \), i.e. the fundamental frequency \( f_0 (a) \sim a^{-1} \).

The 4th moments of the mass function \( f_M \) can be obtained as,

\[
\int_0^\infty f_M (m_h, a) (m_h)^4 \, dm_h = \frac{1}{\sqrt{\pi}} \left( \frac{4 \eta_0}{2 \lambda + 2 a} \right)^{3 \lambda / 2} \Gamma \left( \frac{k - \lambda + 1}{2} \right) (m_h^*)^4,
\]

where, in particular, the mean halo mass is proportional to the characteristic mass scale \( m_h^* \).

\[
\langle m_h \rangle = \frac{1}{\sqrt{\pi}} \left( \frac{4 \eta_0}{2 \lambda + 2 a} \right)^{3 \lambda / 2} \Gamma \left( \frac{k - \lambda + 1}{2} \right) m_h^*.
\]

Finally, solution of the group mass \( m_g (m_h, a) \) is (from Eq. (A63)),

\[
m_g = M_h (a) \left( \frac{1 - \lambda}{2 \eta_0 \sqrt{\pi} \eta_0} \right) \left( \frac{m_h^*}{m_h^*} \right)^{\lambda \lambda / 2} \exp \left( -\frac{1}{4 \eta_0} \left( \frac{m_h}{m_h^*} \right)^{2 - 2 \lambda} \right).
\]

The mass flux and transfer functions can be obtained from definitions (Eqs. (A15), (A18)) in Section A3. The mass transfer function reads

\[
T_m = \frac{1 - \lambda}{2 \eta_0 \sqrt{\pi} \eta_0} \left( \frac{m_h^*}{m_h^*} \right)^{1 - \lambda} \exp \left( -\frac{1}{4 \eta_0} \left( \frac{m_h}{m_h^*} \right)^{2 - 2 \lambda} \right). \tag{A78}
\]

Here \( T_m (m_h, a) \sim m_g (m_h / m_p)^{2 - 2 \lambda} \) is a typical feature of heterogeneous diffusion. The mass flux function is finally given by,

\[
\Pi_m (m_h, a) = m_h \left( \frac{m_h}{m_h^*} \right)^{1 - \lambda} \left[ \text{erf} \left( \frac{1}{2 \sqrt{\eta_0}} \left( \frac{m_h}{m_h^*} \right)^{1 - \lambda} \right) \right]^{-1}
\]

with term I (complementary error function) dominating for small \( m_h \) and term II (exponential function) dominating for large \( m_h \).

Additionally, \( \Pi_m (m_h, a) = \varepsilon_m \) with \( m_h \to 0 \) satisfies the boundary conditions (Eq. (A67)).

Figure A7 plots the variation of the halo mass function \( f_M \) (Eq. (A71)), mass transfer function \(-\Pi_m\) (Eq. (A79)) and mass flux function \( T_m \) (Eq. (A78)) with halo size \( n_p \) for a given set of parameters at the present time \( t_0 \). Two distinct ranges can be clearly identified from Fig. A7: the mass propagation range with a constant mass flux \( \varepsilon_m = \Pi_m \) for \( m_h < m_h^\ast \) and mass deposition range for \( m_h > m_h^\ast \).
is the physical background density at the current epoch smoothed with a tophat filter of size $\sigma$ from a spherical collapse model or a two-body collapse model [58].

for haloes of mass $\delta n = m_h$ where $\delta n$ of $28$ Z. Xu

to simplify the halo mass function, $\Pi$ size $\delta n$ with a constant mass flux $\Pi_\delta$ that can be used to predict the shape and evolution of the halo modeling of structure formation and evolution. The Press-Schechter the most fundamental quantities for analytically or semi-analytically

$\Pi_\delta (m_h) = \frac{1}{\sqrt{2\pi} \rho_0} \exp \left[ -\frac{1}{2} \frac{m_h}{\rho_0} \right] \left( \frac{\delta}{\sigma} \right)^2 \left( \frac{m_h}{m_\ast} \right)^{3 \nu/2}$, (A80)

where $n_{PS}$ is the effective index of the power spectrum of density fluctuation. A normalized dimensionless variable $v$ can be introduced to simplify the halo mass function,

where $\delta_c (a) \sim a^{-1}$ is the critical density that has to be determined from a spherical collapse model or a two-body collapse model [58]. Here $\sigma^2 (m_h)$ is the variance of the initial density fluctuation when smoothed with a tophat filter of size $R = (3 m_h / 4 \pi \rho_0)^{1/3}$. Here $\rho_0$ is the physical background density at the current epoch $a = 1$. The term $\sigma^2 (m_h)$ is the halo virial velocity dispersion. The second equality in Eq. (A81) comes from the linear theory prediction of $\sigma^2 (m_h) \sim m_h^{3 \nu/2}$ for a power-law power spectrum with an effective index of $n_{PS}$. The third equality in Eq. (A81) comes from the virial theorem for haloes of mass $m_h$. Here $\sigma^2 (m_h) \sim Gm_h / r_h^{\nu} \sim m_h^{1 + \nu/3}$. The parameter $n$ is the exponent of gravitational potential $V_p (r) \sim r^n$. Since $\delta_c (a) \sim a^{-1}$ from the spherical collapse model, linear theory predicts that $\sigma^2 (m_h, a) \sim a^{-1} m_h^{1 + \nu/3}$, $\sigma^2 (m_h) \sim a \cdot m_h^{1 + \nu/3}$ and $v \sim a^{-2} m_h^{1 + \nu/3}$.

With the dimensionless variable $v$ introduced in Eq. (A81), the equivalent dimensionless PS mass function in Eq. (A80) is

$$f_{PS} (v) = \frac{1}{\sqrt{2 \pi}} v^{-1/2} \exp \left( -\frac{v}{2} \right).$$ (A82)

Further improvement was achieved by extending the PS formalism with the elliptical collapse model [35, 59]. The modified PS model (ST model, hereafter ST) reads:

$$f_{ST} (v) = \frac{(1 + 1/(qv)^p) \sqrt{2q}}{\Gamma(1/2) + 2 \rho (1/2 - p)} 2 \sqrt{q} e^{-qv/2}.$$ (A83)

The best-fitted parameters from large-scale $N$-body simulations are $q = 0.75$ and $p = 0.3$ [60]. Many other forms of empirical mass functions were also proposed by fitting to the high-resolution simulation data. For example, a universal JK mass function was proposed to cover a wide range of simulation data with different cosmologies and redshifts [61].

$$f_{JK} (v) = \frac{0.315}{2 \pi} \exp \left[ -\frac{1}{2} \ln \left( \frac{v}{\delta_c} \right) + 0.61 \right]^{3.8}.$$ (A84)

With $\delta_c = 1.6865$ at $z = 0$. It should be noted that the empirical mass function does not satisfy the normalization constraint (Integral of mass function in mass space should give unity) and cannot extrapolate beyond the range of fitting data.

Now, let us return to our halo mass function from inverse mass cascade (Eq. (A71)), which does not rely on any particular collapse model (spherical or elliptical). If we introduce $v = \left( \frac{m_h}{m_\ast} \right)^{-2 - \Lambda}$, the halo mass function Eq. (A71) can be simplified to

$$f_{v} (v) = \frac{1}{2 \sqrt{\pi \eta_0}} v^{-1/2} \exp \left[ -\frac{v}{4 \eta_0} \right].$$ (A85)

Clearly, Eq. (A71) reduces to the Press-Schechter (PS) mass function if $\Lambda = (3 - n_{PS}) / 6$ and $\eta_0 = 1/2$ (See Eqs. (A81)). However, it should be noted that Eq. (A71) is more general and the parameter $\eta_0$ is related to the parameter $\beta_0$ (Eq. (A73)), where $\beta_0$ is the prefactor for power-law scaling in Eq. (A24). The halo geometry exponent $\Lambda$ has a fundamental connection to the effective index of power spectrum $n_{PS}$. In principle, the halo geometry exponent $\Lambda$ should be smaller than one such that $n_{PS} > -3$.

A universal halo mass function like Eq. (A85) is clearly a manifestation of a statistically steady state involving mass and energy cascade with scale-independent rates. All these results provide insights into a fundamental question: how does the non-equilibrium system maximize its entropy and approach the limiting equilibrium via a cascade process? Two typical examples are, of course, the hydrodynamic turbulence and SG-CFD (self-gravitating collisionless dark matter flow). Both examples exhibit an intermediate and statistically steady state where a direct energy (or inverse mass) cascade process is well established.

Figure A7. The variation of halo mass function $f_M$ (Eq. (A71)), mass transfer function $\Pi_\delta$ (Eq. (A79)) and mass flux function $T_M$ (Eq. (A78)) with halo size $n_M = m_h / m_p$ for a given set of parameters at the current physical time $t_0$. Two distinct ranges can be clearly identified: a mass propagation range with a constant mass flux $\epsilon M = \Pi_\delta$ for $m_h < m_\ast$ and a mass deposition range for $m_h > m_\ast$. For comparison, the mass transfer function $\Pi_\delta$ obtained using $N$-body simulation data at two different redshifts, $z = 0$ and $z = 0.1$, is also presented in the same plot as the dashed line.

A6 Existing halo mass functions

The abundance of haloes, i.e., halo mass function $f_M$, is one of the most fundamental quantities for analytically or semi-analytically modeling of structure formation and evolution. The Press-Schechter (PS) formalism is one of the first landmarks on the halo mass function [5, 57] that can be used to predict the shape and evolution of the halo mass distribution,

$$f_{PS} (m_h) = \frac{1}{\sqrt{2\pi} \rho_0} \left( 1 + n_{PS} \right) \frac{1}{m_h} \left[ \frac{m_h}{m_\ast} \right]^{3 \nu/2} \exp \left[ -\frac{1}{2} \frac{m_h}{m_\ast} \right],$$ (A80)

where $n_{PS}$ is the effective index of the power spectrum of density fluctuation. A normalized dimensionless variable $v$ can be introduced to simplify the halo mass function,

$$v = \frac{\delta_c (a)}{\sigma^2 (m_h)} = \left( \frac{m_h}{m_\ast (a)} \right)^{1 + 2 \nu} \left( \frac{m_\ast (a)}{m_\ast (a)} \right)^{3 \nu/2 + 3m_\ast} \left( \frac{m_h}{m_\ast (a)} \right)^{3 \nu/2 - 3m_\ast},$$ (A81)

where $\delta_c (a) \sim a^{-1}$ is the critical density that has to be determined from a spherical collapse model or a two-body collapse model [58]. Here $\sigma^2 (m_h)$ is the variance of the initial density fluctuation when smoothed with a tophat filter of size $R = (3 m_h / 4 \pi \rho_0)^{1/3}$. Here $\rho_0$ is the physical background density at the current epoch $a = 1$. The term $\sigma^2 (m_h)$ is the halo virial velocity dispersion. The second equality in Eq. (A81) comes from the linear theory prediction of $\sigma^2 (m_h) \sim m_h^{3 \nu/2}$ for a power-law power spectrum with an effective index of $n_{PS}$. The third equality in Eq. (A81) comes from the virial theorem for haloes of mass $m_h$. Here $\sigma^2 (m_h) \sim Gm_h / r_h^{\nu} \sim m_h^{1 + \nu/3}$. The parameter $n$ is the exponent of gravitational potential $V_p (r) \sim r^n$. Since $\delta_c (a) \sim a^{-1}$ from the spherical collapse model, linear theory