

PAPER • OPEN ACCESS

Everything hot everywhere all at once: neutrinos and hot dark matter as a single effective species

To cite this article: Amol Upadhye *et al* JCAP01(2025)077

View the [article online](#) for updates and enhancements.

You may also like

- [On Space Charge and Spatial Distribution of Defects in Yttria-Stabilized Zirconia](#)
Lei Zhang and Anil V. Virkar
- [Development of a functional airway-on-a-chip by 3D cell printing](#)
Ju Young Park, Hyunryul Ryu, Byungjun Lee et al.
- [Predicting resting-state brain functional connectivity from the structural connectome using the heat diffusion model: a multiple-timescale fusion method](#)
Zhengyuan Lv, Jingming Li, Li Yao et al.

RECEIVED: October 21, 2024

REVISED: December 2, 2024

ACCEPTED: December 23, 2024

PUBLISHED: January 17, 2025

Everything hot everywhere all at once: neutrinos and hot dark matter as a single effective species

Amol Upadhye,^{a,b} Markus R. Mosbech^{Ⓜ,c,d} Giovanni Pierobon^b and Yvonne Y.Y. Wong^b

^aSouth-Western Institute for Astronomy Research, Yunnan University,
Kunming 650500, People's Republic of China

^bSydney Consortium for Particle Physics and Cosmology,
School of Physics, The University of New South Wales,
Sydney NSW 2052, Australia

^cInstitute for Theoretical Particle Physics and Cosmology (TTK), RWTH Aachen University,
D-52056 Aachen, Germany

^dInstitute for Theoretical Particle Physics (TTP), Karlsruhe Institute of Technology (KIT),
76128 Karlsruhe, Germany

E-mail: a.upadhye@ynu.edu.cn, mosbech@physik.rwth-aachen.de,
g.pierobon@unsw.edu.au, yvonne.y.wong@unsw.edu.au

ABSTRACT: Observational cosmology is rapidly closing in on a measurement of the sum M_ν of neutrino masses, at least in the simplest cosmologies, while opening the door to probes of non-standard hot dark matter (HDM) models. By extending the method of effective distributions, we show that any collection of HDM species, with arbitrary masses, temperatures, and distribution functions, including massive neutrinos, may be represented as a single effective HDM species. Implementing this method in the **FlowsForTheMasses** non-linear perturbation theory for free-streaming particles, we study non-standard HDM models that contain thermal QCD axions or generic bosons in addition to standard neutrinos, as well as non-standard neutrino models wherein either the distribution function of the neutrinos or their temperature is changed. Along the way, we substantially improve the accuracy of this perturbation theory at low masses, bringing it into agreement with the high-resolution TianNu neutrino N-body simulation to $\approx 2\%$ at $k = 0.1 h/\text{Mpc}$ and to $\leq 21\%$ over the range $k \leq 1 h/\text{Mpc}$. We accurately reproduce the results of simulations including axions and neutrinos of multiple masses. Studying the differences between the normal, inverted, and degenerate neutrino mass orderings on their non-linear power, we quantify the error in the common approximation of degenerate masses. We release our code publicly at <http://github.com/upadhye/FlowsForTheMassesII>.

KEYWORDS: cosmological neutrinos, dark matter theory, neutrino masses from cosmology, power spectrum

ARXIV EPRINT: [2410.05815](https://arxiv.org/abs/2410.05815)



Contents

1	Introduction	1
2	Background	3
2.1	Cosmic neutrinos	3
2.2	Multi-fluid perturbation theories	4
2.3	Gauss-Laguerre quadrature	8
2.4	Effective distribution functions	9
3	Hot dark matter as an effective particle	9
3.1	Effective HDM	10
3.2	Discrete momenta and Gauss-Laguerre quadrature	11
3.3	Clustering and free-streaming regimes	13
3.4	Convergence with N_{GLQ} and N_τ	15
3.5	Non-linear perturbation theory: <code>FlowsForTheMasses</code> – II	16
3.6	Numerical accuracy	18
4	Results I: non-linear enhancement of HDM clustering	21
4.1	Accuracy at low M_ν : solving a puzzle	21
4.2	Normal ordering and recovery of individual-species power spectra	24
4.3	Errors in the degenerate-ordering approximation	25
4.4	Extension to axionic models	28
5	Results II: evading cosmological neutrino bounds	30
5.1	Non-standard neutrinos and clustering suppression	30
5.2	Skewed neutrino models	32
5.3	Cool-neutrino models	33
6	Conclusions	37

1 Introduction

Cosmology is well on the way to measuring the sum M_ν of neutrino masses, a fundamental particle physics parameter, for the first time. The cosmological upper bounds $M_\nu \leq 120$ meV from joint constraints by the Planck survey [1] and $M_\nu \leq 72$ meV by the DESI survey [2] are rapidly converging upon the lower bound of 59 meV from terrestrial experiments [3–6]. However, these rely upon restrictive assumptions about the dark energy responsible for the cosmic acceleration, when allowing the dark energy equation of state and its derivative to vary weakens the M_ν bound by a factor of about 2–3 [7–9]. Persistent tensions and anomalies in the cosmological data, including the Hubble tension [10–13], the S_8 tension [14–17], and the CMB lensing anomaly [1, 18], hint at a more complicated dark-sector phenomenology. Moreover, some recent DESI+Planck analyses prefer M_ν below the terrestrial lower bound [19–22], a

tension possibly arising from a combination of the DESI preference for a lower matter density fraction $\Omega_{m,0}$ and the Planck preference for unexpectedly strong lensing of the CMB [23].

In this context, it is imperative that theoretical cosmologists study a broad range of phenomena associated with neutrinos and other hot dark matter (HDM) models. Within the neutrino sector, differences among the masses of the three species will be amplified by non-linear clustering at small scales, while non-standard models could modify the number, temperature, or distribution function of cosmological neutrinos [24–27]. Other HDM models include the axion, theorized as a solution to the strong CP problem in quantum chromodynamics [28, 29], whose thermal production in the early universe in the $m \lesssim \text{eV}$ regime was recently studied in refs. [30–39]. Each of these HDM models modify the gravitational clustering of HDM, leading to subtle differences in their clustering power as well as their impact upon scale-dependent halo bias [40, 41], differential HDM capture by halos between HDM-rich and HDM-poor regions [42], “wakes” caused by coherent HDM streaming past collapsed cold dark matter (CDM) halos [43], and other non-linear effects. As cosmological data improve, such higher-order effects could be used either to confirm the standard neutrino picture or to reject it in favor of a more complicated HDM sector, provided that we have the theoretical and numerical tools with which to quantify HDM clustering. Since the space of HDM models is large, tools such as non-linear perturbation theory allowing for a rapid exploration of the parameter space are desirable.

However, hot dark matter presents a host of numerical challenges due to its large velocity dispersion. A given HDM particle’s thermal velocity acts as an escape velocity allowing it to stream freely out of a sufficiently small and diffuse halo. Whereas the CDM and baryons are cold, in the sense that their velocities are determined entirely by their positions, flattening their six-dimensional phase space to a three-dimensional sheet, we must track all six dimensions for an HDM species. In an N-body particle simulation, sampling this six-dimensional phase space requires a large number of particles to avoid shot noise, while faster particles require a finer time stepping to track their gravitational deflection. Velocity dispersion also complicates a perturbative treatment of HDM, since non-linear perturbation theories typically begin with the continuity and Euler equations, which assume a well-defined velocity field. These challenges are compounded by the need to include standard neutrinos, at least two of which are massive, along with any additional HDM species. Each of these HDM species has its own mass and momentum distribution.

We make two contributions to this velocity dispersion problem. Firstly, we extend the effective distribution function method of ref. [44], originally developed for non-relativistic neutrinos, to represent any collection of HDM as a single *effective hot dark matter* (EHDM). Our extension applies to relativistic as well as non-relativistic species of different masses, temperatures, and distribution functions, provided that each species has decoupled from all non-gravitational interactions. Our key insight is that the clustering of such a particle depends only upon its four-velocity, rather than its mass and four-momentum separately, so that doubling both an HDM particle’s mass and its momentum simultaneously will not affect its clustering.

We implement the EHDM method in the `FlowsForTheMasses` perturbation theory of ref. [45], which provided the first non-linear perturbative power spectrum computation for

free-streaming HDM. `FlowsForTheMasses` functions by discretizing the HDM distribution function into “flows,” each characterized by its four-velocity, so that a single EHDM flow can represent many different HDM species. Furthermore, we show that the clustering power of an individual HDM component of the EHDM can be recovered from a linear combination of these same EHDM flows. This is crucial, for example, for a terrestrial detector that is sensitive only to the electron neutrino, or only to the axion, rather than to all HDM species making up the EHDM.

Secondly, we trade the uniform-density momentum binning of ref. [45] for a more efficient binning based upon Gauss-Laguerre quadrature, taking advantage of the fact that thermal distribution functions decay exponentially with particle momentum. Implementing our new procedure in the code `FlowsForTheMasses-II`, we demonstrate the effectiveness of this improved quadrature by reducing the low- M_ν , high- k error found in ref. [46] by more than a factor of two. We demonstrate the accuracy of `FlowsForTheMasses-II` for models containing axions and axion-like bosons in addition to neutrinos, and we apply it to quantifying the error in the commonly-used approximation treating all three neutrino masses as equal. Finally, we consider a pair of proposals to evade cosmological bounds on high-mass neutrinos, one by raising and the other by lowering their mean momentum [24, 25]. While the high-mean-momentum neutrinos are nearly indistinguishable from standard ones, the low-mean-momentum ones cluster more non-linearly, in a manner that will allow upcoming cosmological surveys to search for them. Thus we demonstrate the ability of non-linear perturbation theory to explore a large parameter space of non-standard models.

This article is organized as follows. After summarizing the numerical techniques used, in section 2, we derive and thoroughly study the EHDM technique in section 3, and describe its implementation along with Gauss-Laguerre quadrature in `FlowsForTheMasses-II`. Our results are split into two sections, beginning with section 4, which quantifies the non-linear enhancement of HDM clustering in a range of models. Section 5 studies the two proposals for evading cosmological M_ν bounds mentioned above, and section 6 concludes.

2 Background

2.1 Cosmic neutrinos

The Standard Model of particle physics predicts exactly three neutrinos, which are uncharged, weakly-interacting fermions whose small masses make them ultra-relativistic for much of the universe’s history up to CMB formation. Neutrinos decouple from photons, electrons, and positrons at a temperature of $T \sim 1$ MeV, shortly before electron-positron annihilation begins. In an idealized situation, the neutrino temperature at the end of e^+e^- annihilation is $(4/11)^{1/3}$ times that of the photons. Detailed calculations, however, have found that the neutrino-to-photon energy density is some 1.47% larger than that suggested by this simple temperature relation. This is equivalent to an increase in the effective number of neutrinos to 3.044 [47–51], which we approximate by raising the neutrino temperature by 0.365%, to $T_{\nu,0}a^{-1} = 1.9525a^{-1}$ K at scale factor a .

Much later, typically around or after electron-photon decoupling, the behavior of cosmic neutrinos is determined by their masses, at least two of which are required to be non-zero

by neutrino oscillation experiments. Such experiments can determine only mass-squared splittings $\Delta m_{21}^2 = 74.2_{-2.0}^{+2.1} \text{ meV}^2$ and $|\Delta m_{31}^2| = 2517_{-28}^{+26} \text{ meV}^2$ rather than the absolute neutrino mass scale M_ν [6]. Furthermore, the sign of the larger splitting may be either positive or negative, with the former implying a “normal order” (NO) of neutrino masses dominated by a single heavy neutrino and two light ones, and the latter implying an “inverted order” (IO) with two heavy neutrinos and a single light one. As M_ν rises, the fractional difference between the heaviest and lightest decreases, making a “degenerate order” (DO) of three equal neutrino masses a common approximation.

Even at late times, neutrinos’ Fermi-Dirac thermal velocity distribution profoundly affects their clustering. The subhorizon, non-relativistic clustering of a neutrino species of mass m_ν is characterized by a “free-streaming” length that is approximately the average distance travelled by neutrinos of that mass in a comoving Hubble time \mathcal{H}^{-1} . On much larger length scales, neutrinos’ free-streaming does not inhibit their clustering, and they cluster much the same as cold matter; we call this the “clustering regime.” On scales smaller than the free-streaming length, the “free-streaming regime,” neutrinos stream out of most overdense regions. We define the neutrino free-streaming wave number as [52]

$$k_{\text{FS}} := \sqrt{\frac{3\Omega_{\text{m}}(a)\mathcal{H}(a)^2}{2c_\nu(a)^2}}, \quad (2.1)$$

where the square of the neutrino sound speed is

$$c_\nu(a)^2 := \frac{3\zeta(3)T_{\nu,0}^2}{2\ln(2)m_\nu^2 a^2}, \quad (2.2)$$

and $\zeta(x)$ is the Riemann zeta function. Refs. [52, 53] demonstrated that the linear perturbation ratio $\delta\rho_\nu/\delta\rho_{\text{m}}/(\bar{\rho}_\nu/\bar{\rho}_{\text{m}})$ approaches unity at small k and k_{FS}^2/k^2 at large k , leading them to approximate

$$\frac{\delta\rho_\nu(k)}{\delta\rho_{\text{m}}(k)} \approx \frac{\bar{\rho}_\nu}{\bar{\rho}_{\text{m}}} \frac{1}{(1 + k/k_{\text{FS}})^2}, \quad (2.3)$$

by interpolating between the clustering and free-streaming limits.

2.2 Multi-fluid perturbation theories

The chief difficulty with applying standard cosmological perturbation theory to HDM species such as massive neutrinos is their significant velocity dispersion. Whereas a cold particle beginning at a given initial position with zero velocity can be tracked to a definite final position, HDM particles begin with a thermal distribution of initial velocities. Thus we must track their full six-dimensional phase space distribution, rather than the three-dimensional spatial distribution of cold particles.

In a series of articles, Dupuy and Bernardeau demonstrated that neutrinos could be treated perturbatively by splitting their population into multiple sub-populations, each characterized by a spatially-uniform zeroth-order velocity \vec{v} [54–56]. Since a particle with definite initial velocity can once again be tracked from a given initial position to a definite final position, standard perturbative techniques may be applied. Furthermore, the direction

of \vec{v} affects clustering only through its angle with the Fourier vector. Thus the multi-fluid method increases the dimensionality of the problem by two, v and $\hat{v} \cdot \hat{k}$, rather than three. References [54, 55] derived a fully relativistic linear theory for massive neutrinos in Newtonian gauge and a general gauge, respectively, while ref. [56] motivated a non-linear treatment; all of these results generalize to other HDM.

The EHDM formalism to be presented in section 3 is applicable to the relativistic perturbation theory of refs. [54, 55]. However, our focus here is clustering at late times, when each HDM species is non-relativistic, and particularly on small-scale non-linear HDM clustering. Thus we focus on multi-fluid perturbation theories in the subhorizon, non-relativistic case. In this limit, we may to excellent approximation treat fluids as obeying the continuity and Euler equations of classical fluid dynamics, with a gravitational potential determined from Poisson’s equation, in a universe whose uniform expansion is given by the Hubble rate. We restrict our consideration to spatially-flat cosmologies, though our results may be generalized to models with spatial curvature. Finally, we track only the scalar perturbations of each fluid, namely, the density contrast and velocity divergence. Although vector perturbations such as the velocity vorticity are important in small-scale, non-perturbative structures such as virialized halos, they are negligible at the linear and mildly non-linear scales accessible to perturbation theory [57–59].

References [54, 55] discretized the Fermi-Dirac distribution describing the initial neutrino momenta. Let P_i be one such lower-index three-momentum, and let τ_i be its value in the limit of a homogeneous universe. Then $P_i \rightarrow \tau_i$ at early times, and τ_i is itself time-independent, making it a useful quantity for a Lagrangian description of neutrinos in momentum space. Furthermore, physics depends upon the direction of $\vec{\tau}$ only through its angle $\mu = \cos^{-1}(\hat{\tau} \cdot \hat{k})$ with the Fourier vector \vec{k} , so we may approximate the neutrino population using N_τ values of the momentum magnitude $\tau = |\vec{\tau}| = \sqrt{\tau_1^2 + \tau_2^2 + \tau_3^2}$. We label these discrete momenta using Greek indices, as τ_α , with integer $\alpha \in [0, N_\tau - 1]$.

Reference [60] restricted the perturbation theory of refs. [54, 55] to the subhorizon, non-relativistic case, which is released as the code **MuFLR**.¹ The scalar perturbations of each “flow” α are the density contrast $\delta_\alpha(\vec{x}) := \rho_\alpha(\vec{x})/\bar{\rho}_\alpha - 1$ and the velocity divergence $\theta_\alpha(\vec{x}) := -\vec{\nabla} \cdot \vec{P}/(m_{\text{HDM}}a)$, where m_{HDM} is the HDM mass.² In Fourier space, δ_α and θ_α depend upon the magnitude k of the wave number as well as its angle with respect to the initial momentum $\vec{\tau}_\alpha$, through its cosine $\mu := \hat{k} \cdot \hat{\tau}$. The μ -dependence of these perturbations may be expanded in Legendre polynomials as

$$\delta_\alpha^{\vec{k}} := \sum_{\ell=0}^{\infty} (-i)^\ell \mathcal{P}_\ell(\mu) \delta_{\alpha\ell}^k, \quad \theta_\alpha^{\vec{k}} := \sum_{\ell=0}^{\infty} (-i)^\ell \mathcal{P}_\ell(\mu) \theta_{\alpha\ell}^k, \quad (2.4)$$

where \mathcal{P}_ℓ is the Legendre polynomial of order ℓ ; we use wave number superscripts to denote functional dependence, so $\delta_\alpha^{\vec{k}} = \delta_\alpha(k)$.

¹MuFLR is publicly available at <https://github.com/upadhye/MuFLR>.

²Strictly speaking, the continuity and Euler equations apply to a fluid with a well-defined momentum, rather than merely a momentum magnitude, at each point in space. However, all fluids with the same momentum magnitude obey the same set of equations of motion. We use the term “flow” to refer to all such fluids at once, and the same index α for all fluids with initial momentum magnitude τ_α .

Using $\eta := \log(a/a_{\text{in}})$ as our time variable,³ for initial scale factor a_{in} , and primes to denote derivatives with respect to η , the linear continuity and Euler equations for flow α are

$$\begin{aligned} (\delta_{\alpha\ell}^k)' &= \frac{kv_\alpha}{\mathcal{H}} \left(\frac{\ell}{2\ell-1} \delta_{\alpha,\ell-1}^k - \frac{\ell+1}{2\ell+3} \delta_{\alpha,\ell+1}^k \right) + \theta_{\alpha\ell}^k, \\ (\theta_{\alpha\ell}^k)' &= - \left(1 + \frac{\mathcal{H}'}{\mathcal{H}} \right) \theta_{\alpha\ell}^k - \delta_{\ell 0}^{(K)} \frac{k^2 \Phi^k}{\mathcal{H}^2} + \frac{kv_\alpha}{\mathcal{H}} \left(\frac{\ell}{2\ell-1} \theta_{\alpha,\ell-1}^k - \frac{\ell+1}{2\ell+3} \theta_{\alpha,\ell+1}^k \right), \end{aligned} \quad (2.5)$$

where $\delta^{(K)}$ is the Kronecker delta, $\mathcal{H} = aH$ the conformal Hubble rate, and $v_\alpha := \tau_\alpha/(m_{\text{HDM}}a)$ the flow speed. Since this perturbation theory is Lagrangian in momentum space, a particle cannot move from one flow to another. Thus the only interaction between different flows occurs through the gravitational potential Φ , given by Poisson's equation:

$$k^2 \Phi^k = -\frac{3}{2} \mathcal{H}^2 \left(\Omega_{\text{cb}}(\eta) \delta_{\text{cb}}^k + \sum_{\alpha=0}^{N_\tau-1} \Omega_\alpha(\eta) \delta_{\alpha 0}^k \right). \quad (2.6)$$

Here, δ_{cb} is the density contrast of the CDM and baryons, which we approximate as a single fluid labeled “cb” henceforth. It can be obtained either from a linear perturbative treatment of the cb fluid, resulting in a fully linear perturbation theory for the neutrinos, or from a non-linear calculation which then sources the linear eq. (2.5) through eq. (2.6), known as “linear response.” The time-dependent density fractions $\Omega_{\text{cb}}(\eta)$ and $\Omega_\alpha(\eta)$ are respectively given in terms of their values $\Omega_{\text{cb},0}$ and $\Omega_{\alpha,0}$ today by $\mathcal{H}^2 \Omega_{\text{cb}}(\eta) = \mathcal{H}_0^2 \Omega_{\text{cb},0}/a$ and $\mathcal{H}^2 \Omega_\alpha(\eta) = \mathcal{H}_0^2 \Omega_{\alpha,0}/(a\sqrt{1-v_\alpha^2})$, where \mathcal{H}_0 is the value of \mathcal{H} today.

Initial conditions at $a_{\text{in}} = 10^{-3}$ are given in ref. [60]. Specifically,

$$\delta_{\text{cb}}(a_{\text{in}}) = a_{\text{in}} + \frac{2}{3} a_{\text{eq}}, \quad \theta_{\text{cb}}(a_{\text{in}}) = a_{\text{in}}, \quad (2.7)$$

and

$$\delta_{\alpha,0}(a_{\text{in}}, k) = \frac{k_{\text{FS},\alpha}^2 (1 - f_\nu) \delta_{\text{cb}}(a_{\text{in}})}{(k + k_{\text{FS},\alpha})^2 - f_\nu k_{\text{FS},\alpha}^2}, \quad \theta_{\alpha,0} = \delta'_{\alpha,0}, \quad (2.8)$$

for cb and HDM, respectively, where

$$k_{\text{FS},\alpha}^2 = \frac{3\Omega_{\text{m}}(a)\mathcal{H}^2 a^2 m_{\text{HDM}}^2}{2\tau_\alpha^2} \quad (2.9)$$

is a generalization of eq. (2.2) to a flow of arbitrary velocity $v_\alpha = \tau_\alpha/(m_{\text{HDM}}a)$, shown in ref. [60] to obey $\delta_{\alpha,0}/\delta_{\text{m}} \rightarrow k_{\text{FS},\alpha}^2/k^2$ at large k . The HDM initial density monopole of eq. (2.8) is an interpolation between the clustering limit, $\delta_{\alpha,0} \approx \delta_{\text{cb}}$, and the free-streaming limit, $\delta_{\alpha,0} \approx (k/k_{\text{FS},\alpha})^2 (1 - f_\nu) \delta_{\text{cb}}$, similar to that of eq. (2.3), from refs. [52, 53, 61].

We may estimate the error in this initial condition approximation as follows. The interpolation of ref. [52] may be written $\delta_{\text{HDM}}/\delta_{\text{m}} \approx 1/(1 + \beta k/k_{\text{FS}} + k^2/k_{\text{FS}}^2)$, where their value $\beta = 2$ may be treated as a free parameter. Thus changing β by ~ 1 will allow us to estimate the magnitude of the error. Comparing our results for β of 2 and 4, we find that the power spectra agree up to a fraction $\approx a_{\text{in}}/a$, or a tenth of a percent at $z = 0$. Moreover, both of these approximations approach the linear solution in both the large- k and

³Throughout this article, we use \log to denote the natural logarithm.

small- k limits. Thus, this error peaks around the free-streaming wave number at a_{in} , which is $\lesssim 10^{-3} h/\text{Mpc}$ for cosmologically-preferred neutrino masses. For scales of interest to us, $0.01 h/\text{Mpc} \lesssim k \lesssim 1 h/\text{Mpc}$, we find the power spectrum error from eq. (2.8) to be $< 0.001\%$.

Reference [45] developed the first non-linear perturbative power spectrum calculation for free-streaming HDM, such as massive neutrinos, called **FlowsForTheMasses**.⁴ It did so by generalizing the Time-Renormalization Group (Time-RG) perturbation theory of refs. [62, 63] to the case of a fluid with zeroth-order bulk velocity \vec{v}_α . The result is a set of perturbative mode-coupling integrals that couple different multipoles ℓ as well as wave numbers k . Since non-linear corrections decorrelate δ_α and θ_α , ref. [45] introduced the decorrelation perturbation

$$\chi_{\alpha\ell}^k := 1 - P_{\alpha,01\ell}^k / \sqrt{P_{\alpha,00\ell}^k P_{\alpha,11\ell}^k}, \quad (2.10)$$

where $P_{\alpha,bc\ell}^k$ is the ℓ th Legendre moment of the power spectrum,

$$\begin{aligned} P_{\alpha,00}^{\vec{k}} &= \sum_{\ell} \mathcal{P}_{\ell}(\mu)^2 P_{\alpha,00\ell}^k, & P_{\alpha,11}^{\vec{k}} &= \sum_{\ell} \mathcal{P}_{\ell}(\mu)^2 P_{\alpha,11\ell}^k, \\ P_{\alpha,01}^{\vec{k}} &= \sum_{\ell} \mathcal{P}_{\ell}(\mu)^2 (1 - \chi_{\alpha\ell}^k) \sqrt{P_{\alpha,00\ell}^k P_{\alpha,11\ell}^k}, \end{aligned} \quad (2.11)$$

and its indices 0 and 1 refer respectively to δ and θ . The non-linear equations of motion of **FlowsForTheMasses**, which replace eq. (2.5), are then

$$\begin{aligned} (\delta_{\alpha\ell}^k)' &= \frac{kv_\alpha}{\mathcal{H}} \left(\frac{\ell}{2\ell-1} \delta_{\alpha,\ell-1}^k - \frac{\ell+1}{2\ell+3} \delta_{\alpha,\ell+1}^k \right) + \theta_{\alpha\ell}^k + \frac{2}{\delta_{\alpha\ell}^k} I_{\alpha,001,001,\ell}^k, \\ (\theta_{\alpha\ell}^k)' &= - \left(1 + \frac{\mathcal{H}'}{\mathcal{H}} \right) \theta_{\alpha\ell}^k - \delta_{\ell 0}^{(K)} \frac{k^2 \Phi^k}{\mathcal{H}^2} + \frac{kv_\alpha}{\mathcal{H}} \left(\frac{\ell}{2\ell-1} \theta_{\alpha,\ell-1}^k - \frac{\ell+1}{2\ell+3} \theta_{\alpha,\ell+1}^k \right) + \frac{1}{\theta_{\alpha\ell}^k} I_{\alpha,111,111,\ell}^k, \\ (\chi_{\alpha\ell}^k)' &= \frac{2(1-\chi_{\alpha\ell}^k)}{(\delta_{\alpha\ell}^k)^2} I_{\alpha,001,001,\ell}^k + \frac{1-\chi_{\alpha\ell}^k}{(\theta_{\alpha\ell}^k)^2} I_{\alpha,111,111,\ell}^k - \frac{2}{\delta_{\alpha\ell}^k \theta_{\alpha\ell}^k} I_{\alpha,001,101,\ell}^k - \frac{1}{\delta_{\alpha\ell}^k \theta_{\alpha\ell}^k} I_{\alpha,111,011,\ell}^k. \end{aligned} \quad (2.12)$$

Here, the bispectrum integrals $I_{\alpha,acd,bef,\ell}^k$ are defined by their equations of motion,

$$(I_{\alpha,acd,bef,\ell}^k)' = -\Xi_{\alpha,bg\ell}^k I_{\alpha,acd,gef,\ell}^k - \tilde{\Xi}_{\alpha,egl}^k I_{\alpha,acd,bgf,\ell}^k - \tilde{\Xi}_{\alpha,fgl}^k I_{\alpha,acd,beg,\ell}^k + 2A_{\alpha,acd,bef,\ell}^k, \quad (2.13)$$

with

$$\begin{aligned} \Xi_{\alpha,bcl}^k &= \begin{bmatrix} 0 & -1 \\ \frac{k^2 \Phi^i}{\mathcal{H}^2 \delta_{\alpha 0}^k} & 1 + \frac{\mathcal{H}'}{\mathcal{H}} \end{bmatrix} - \frac{kv_\alpha}{\mathcal{H}} \frac{\ell}{2\ell-1} \begin{bmatrix} \frac{\delta_{\alpha,\ell-1}^k}{\delta_{\alpha,\ell}^k} & 0 \\ 0 & \frac{\theta_{\alpha,\ell-1}^k}{\theta_{\alpha,\ell}^k} \end{bmatrix} + \frac{kv_\alpha}{\mathcal{H}} \frac{\ell+1}{2\ell+3} \begin{bmatrix} \frac{\delta_{\alpha,\ell+1}^k}{\delta_{\alpha,\ell}^k} & 0 \\ 0 & \frac{\theta_{\alpha,\ell+1}^k}{\theta_{\alpha,\ell}^k} \end{bmatrix}, \\ \tilde{\Xi}_{\alpha,bcl}^k &= \begin{bmatrix} 0 & -1 \\ 0 & 1 + \frac{\mathcal{H}'}{\mathcal{H}} \end{bmatrix}, \end{aligned} \quad (2.14)$$

where the indices b and c label the rows and columns respectively,

$$A_{\alpha,acd,bef}^{\vec{k}} = \int_{\vec{q}} \gamma_{acd}^{\vec{k}\vec{q}\vec{p}} \left[\gamma_{bgh}^{\vec{k}\vec{q}\vec{p}} P_{\alpha,ge}^{\vec{q}} P_{\alpha,hf}^{\vec{p}} + \gamma_{egh}^{\vec{q},-\vec{p},\vec{k}} P_{\alpha,gf}^{\vec{p}} P_{\alpha,hb}^{\vec{k}} + \gamma_{fgh}^{\vec{p},\vec{k},-\vec{q}} P_{\alpha,gb}^{\vec{k}} P_{\alpha,he}^{\vec{q}} \right] \quad (2.15)$$

$$=: \sum_{\ell} \mathcal{P}_{\ell}(\mu)^2 A_{\alpha,acd,bef,\ell}^k \quad (2.16)$$

⁴**FlowsForTheMasses** is publicly available at <https://github.com/upadhye/FlowsForTheMasses>.

is the mode-coupling integral, and

$$\gamma_{001}^{\vec{k}\vec{q}\vec{p}} = \gamma_{010}^{\vec{k}\vec{p}\vec{q}} = \frac{(\vec{q} + \vec{p}) \cdot \vec{p}}{2p^2}, \quad \gamma_{111}^{\vec{k}\vec{q}\vec{p}} = \frac{|\vec{q} + \vec{p}|^2 \vec{q} \cdot \vec{p}}{2q^2 p^2}, \quad (2.17)$$

while all other γ_{abc} vanish. Following ref. [45], we smoothly cut off the mode-coupling integrals below $0.01(1 + \ell^2) h/\text{Mpc}$ to suppress large-scale noise; we checked that doubling this cutoff scale affects the neutrino power at the sub-percent level. Initial conditions for the $I_{\alpha,acd,bef,\ell}^k$, set at $\eta = 0$ (i.e., $a = a_{\text{in}}$), are

$$I_{\alpha,acd,bef,\ell}^k = 2A_{\alpha,acd,bef,\ell}^k. \quad (2.18)$$

On the right hand sides of eqs. (2.13), (2.15), we assume summation over repeated indices. Computation of the mode-coupling integrals $A_{\alpha,acd,bef,\ell}^k$ of eq. (2.16) is the most expensive part of **FlowsForTheMasses**, and its acceleration using Fast Fourier Transform (FFT) techniques is described thoroughly in ref. [45].

Compared with the perturbation theory of ref. [64], our treatment of eqs. (2.12)–(2.18) includes non-linear corrections to the HDM free-streaming terms themselves. Reference [65] includes these non-linear free-streaming corrections through a Green’s function that is non-local in time, but is unable to compute the HDM power spectrum, as done in eqs. (2.12)–(2.18).

Thus far we have not discussed precisely how τ_α are to be sampled from the Fermi-Dirac distribution function F_{FD} , or another distribution function appropriate to other HDM species. The **MuFLR** and **FlowsForTheMasses** perturbation theories considered in refs. [45, 60] used equal-number-density bins. That is, the range $0 \leq \tau < \infty$ was divided into N_τ intervals such that the integrals of $4\pi\tau^2 F_{\text{FD}}(\tau)$ over any two intervals are equal. For each $\alpha \in [0, N_\tau - 1]$, τ_α was chosen to be the median of the corresponding interval. In section 2.3 and section 3.2 we will discuss a more efficient sampling method.

2.3 Gauss-Laguerre quadrature

Consider a function $g(x)$, defined on $[0, \infty)$. Gauss-Laguerre quadrature (GLQ) approximates the integral of $e^{-x}g(x)$ on the semi-infinite interval using N_{GLQ} points x_α and weights w_α as

$$\int_0^\infty dx e^{-x} g(x) \approx \sum_{\alpha=0}^{N_{\text{GLQ}}-1} w_\alpha g(x_\alpha). \quad (2.19)$$

Here, N_{GLQ} is a positive integer; the x_α are the N_{GLQ} roots of the N_{GLQ} th Laguerre polynomial $\mathcal{L}_{N_{\text{GLQ}}}(x)$; and the weights are given by

$$w_\alpha = \frac{x_\alpha}{(N_{\text{GLQ}} + 1)^2 \mathcal{L}_{N_{\text{GLQ}}+1}(x_\alpha)^2}. \quad (2.20)$$

If $g(x)$ is a polynomial of degree no more than $2N_{\text{GLQ}} - 1$, then eq. (2.19) is exact rather than approximate [66]. We will find it convenient to define $G(x) = e^{-x}g(x)$, for which eq. (2.19) implies $\int_0^\infty dx G(x) \approx \sum_\alpha w_\alpha e^{x_\alpha} G(x_\alpha)$.

The error in approximation eq. (2.19), given by eq. (25.4.45) of ref. [67] is

$$\epsilon_{\text{GLQ}} = \frac{(N_{\text{GLQ}}!)^2}{(2N_{\text{GLQ}})!} g^{(2N_{\text{GLQ}})}(x_\epsilon) \approx \sqrt{\pi N_{\text{GLQ}}} 2^{-2N_{\text{GLQ}}} g^{(2N_{\text{GLQ}})}(x_\epsilon) \quad (2.21)$$

for some $0 \leq x_\epsilon < \infty$, where $g^{(2N_{\text{GLQ}})}$ is the $(2N_{\text{GLQ}})$ th derivative of g . The approximation in eq. (2.21) uses Stirling's formula for large N_{GLQ} . We will see that $g(x)$ is typically $x^2 e^x$ times a thermal distribution function. For the Boltzmann distribution e^{-x} , $g(x)$ is precisely a polynomial, making eq. (2.19) exactly correct for $N_{\text{GLQ}} \geq 2$. Errors for the Bose-Einstein and Fermi-Dirac distributions are due to their difference from the Boltzmann distribution, differences whose derivatives are at most $\mathcal{O}(1)$, meaning that $N_{\text{GLQ}} \gtrsim 10$ should be highly accurate. However, at very small length scales, we will see that $g(x)$ is suppressed by an additional two powers of x , leading to larger errors.

We find points and weights for GLQ using the `scipy.special.roots_laguerre` python function. This limits us to $N_{\text{GLQ}} \leq 186$; above this bound, double precision numbers are inadequate for evaluating the higher-order Laguerre polynomials required for determining the weights. We will see that $N_{\text{GLQ}} = 186$ is far larger than necessary for our applications.

2.4 Effective distribution functions

The method of effective distribution functions was introduced in ref. [44] and applied to the case of multiple neutrinos with non-degenerate masses. We summarize their method here before generalizing it in the next section.

Consider a neutrino species s with mass $m_\nu^{(s)}$, temperature constant $T_{\nu,0}^{(s)}$, and lower-index homogeneous-universe three-momentum $\tau_i^{(s)}$, hence $(\tau^{(s)})^2 = (\tau_1^{(s)})^2 + (\tau_2^{(s)})^2 + (\tau_3^{(s)})^2$. Its distribution function is the Fermi-Dirac distribution in the relativistic limit, $F_{\text{FD}}(\tau^{(s)}) = (2\pi)^{-3} [\exp(\tau^{(s)}/T_{\nu,0}^{(s)}) + 1]^{-1}$. The mass in a phase space volume element $d^3x d^3\tau^{(s)}$ is then $g_\nu^{(s)} m_\nu^{(s)} F_{\text{FD}}(\tau^{(s)}) d^3x d^3\tau^{(s)}$, where $g_\nu^{(s)} = 2$ accounts for a neutrino and an antineutrino.

Defining $\tau_i = \tau_i^{(s)} m_{\text{EHDM}}/m_\nu^{(s)}$ for some quantity m_{EHDM} with dimensions of mass, we may change phase space variables from $\tau_i^{(s)}$ to τ_i . The mass of multiple species in a phase space element may now be written as $\sum_s g_\nu^{(s)} m_\nu^{(s)} F_{\text{FD}}(\tau m_\nu^{(s)}/m_{\text{EHDM}}) (m_\nu^{(s)}/m_{\text{EHDM}})^3 d^3x d^3\tau$. Thus if we define a new EHDM particle with mass m_{EHDM} , momentum τ , and distribution

$$F_{\text{EHDM}}(\tau) = \sum_s g_\nu^{(s)} \left(\frac{m_\nu^{(s)}}{m_{\text{EHDM}}} \right)^4 F_{\text{FD}}\left(\frac{\tau m_\nu^{(s)}}{m_{\text{EHDM}}} \right), \quad (2.22)$$

then its mass density $m_{\text{EHDM}} F_{\text{EHDM}}(\tau) d^3x d^3\tau$ equals that of all neutrino species combined. Rather than including three different-mass neutrino species into an expensive calculation such as an N-body simulation, we may include a single particle with mass m_{EHDM} and the above distribution function.

3 Hot dark matter as an effective particle

We begin this section by generalizing the method of effective distribution functions to an arbitrary collection of HDM species, each with its own mass, temperature, and distribution function, for all times from the ultra-relativistic to the non-relativistic regimes. Sections 3.2–3.4 apply Gauss-Laguerre quadrature to the resulting EHDM species, with section 3.2 deriving the density contrast of the EHDM and each of its constituent HDM species from those of the GLQ flows, section 3.3 providing error estimates in the clustering and free-streaming regimes, and section 3.4 demonstrating that GLQ for the EHDM converges to $\sim 0.1\%$. The cost of

this precision, an increased sampling of lower neutrino momenta, exacerbates the numerical instabilities of the **FlowsForTheMasses** perturbation theory, and section 3.5 develops a new technique for mitigating these instabilities. Finally, section 3.6 assesses the accuracy of perturbation theory with EHDM, GLQ, and this numerical stabilization.

3.1 Effective HDM

Before implementing the method of effective distribution functions, we discuss its applicability to distribution functions that vary in time and space, as well as its limitations. Working in conformal Newtonian gauge and using conformal time \mathcal{T} , the line element is

$$ds^2 = a^2[-(1 + 2\Phi)d\mathcal{T}^2 + (1 - 2\Psi)|\vec{dx}|^2]. \quad (3.1)$$

The collisionless Boltzmann equation for particles of positive mass m and four-velocity U_μ , hence four-momentum $P_\mu = mU_\mu$, along a geodesic with affine parameter λ is

$$0 = U^\mu \frac{\partial F}{\partial x^\mu} + \frac{\partial U^i}{\partial \lambda} \frac{\partial F}{\partial U^i} = U^0 \frac{\partial F}{\partial \mathcal{T}} + U^i \frac{\partial F}{\partial x^i} - \Gamma_{\mu\nu}^i U^\mu U^\nu \frac{\partial F}{\partial U^i}, \quad (3.2)$$

where we have used the geodesic equation, and $\Gamma_{\mu\nu}^i$ is the Christoffel symbol. In particular, the evolution of the distribution function is independent of the particle mass. For a given initial position and four-velocity, the fractional change in every such distribution function is identical. Thus an EHDM with a distribution function of the form of eq. (2.22), defined at a time after all HDM species have decoupled from any non-gravitational interactions, will continue to represent those species thenceforth. This conclusion applies to all orders in perturbations of the metric and the distribution function.

The stress-energy tensor for this species of mass m may also be expressed as an integral over four-velocities:

$$T_{\mu\nu} = \int \frac{d^3 P_i}{\sqrt{-g}} \frac{P_\mu P_\nu}{P^0} F(x, U) = \int \frac{d^3 U_i}{\sqrt{-g}} \frac{U_\mu U_\nu}{U^0} m^4 F(x, U). \quad (3.3)$$

This clarifies the m^4 scaling of each component of the effective distribution function of eq. (2.22). Since the $T_{\mu\nu}$ integral scales as the fourth power of the four-momentum and contributions from multiple species add linearly, it follows that replacing P^μ with mU^μ must lead to an effective distribution for all species s at any given position x and four-velocity U proportional to the sum of $g^{(s)}(m^{(s)})^4 F^{(s)}(x, U)$ over all s .

Also evident from the above is a limitation of the effective distribution function approach. The distribution function $F^{(s)}(x, U)$ for an individual HDM species s records the phase-space number density of that species, and the sum over s the total HDM phase-space number density. However, the corresponding effective distribution function will not in general match the individual or total HDM number densities across all of phase space. The m^4 mass scaling of eq. (2.22) results in the correct $T_{\mu\nu}$, which has mass dimension four, but we cannot rely on it for other quantities. Fortunately for our purposes, $T_{\mu\nu}$ is sufficient for studying our observables of interest as well as their evolution and their impact on the spacetime metric through Einstein's equation.

Since we are particularly interested in density perturbations, we next simplify the T_0^0 component of the stress-energy tensor. Dupuy and Bernardeau point out in ref. [54] that the

spatial components of the lower-index four-velocity in the limit of a homogeneous universe, $U_i^{(0)}$, are constant in time. Thus they treat the constant $\vec{u} := (U_1^{(0)}, U_2^{(0)}, U_3^{(0)})^T$ as a Lagrangian coordinate for the particle velocity. Note that $u_0 := U_0^{(0)} = -\sqrt{a^2 + |\vec{u}|^2}$, where $|\vec{u}|^2 = \delta_{ij}^{(K)} u_i u_j$. Our treatment in section 2.2 further simplifies the perturbation theory by working in the subhorizon, non-relativistic limit, in which the flow velocity is $\vec{v} \approx \vec{u}/a$.

Making the cosmologically valid approximation that the metric potentials Φ and Ψ are small compared with unity, even though the matter perturbations may be large, we have $U_0 = (1 + \Phi)U_0^{(0)}$ and $U_i = (1 - \Psi)U_i^{(0)}$. Since $\sqrt{-g} = a^4(1 + \Phi - 3\Psi)$, we may simplify $d^3U_i/(U^0\sqrt{-g}) \approx -d^3\vec{u}/(a^2u_0)$ to linear order in Φ and Ψ . Up to the same order of approximation, these also cancel from the products $U^0U_0 = -a^{-2}u_0^2$ and $U^iU_j = a^{-2}u_iu_j$.

Next, let $\bar{T}_{\mu\nu}$ be the spatially-averaged stress-energy tensor, representing the homogeneous component of the matter. Because $F(x, U)$ is the only position-dependent quantity in T_0^0 , its perturbation may be written

$$-\delta\rho = \delta T_0^0 = T_0^0 - \bar{T}_0^0 = -\frac{m^4}{a^4} \int d^3\vec{u} \sqrt{a^2 + |\vec{u}|^2} \bar{F}(|\vec{u}|) \left[\frac{F(x, U)}{\bar{F}(|\vec{u}|)} - 1 \right], \quad (3.4)$$

where the background, homogeneous distribution function \bar{F} is assumed to be a function of the magnitude of \vec{u} alone. The quantity in square brackets is the only factor that depends on the angular components $\hat{u} = \vec{u}/|\vec{u}|$, and angular integration projects out its monopole, i.e.,

$$\delta\rho = \frac{m^4}{a^4} \int d|\vec{u}| 4\pi |\vec{u}|^2 \sqrt{a^2 + |\vec{u}|^2} \bar{F}(|\vec{u}|) \delta_{\ell=0}(x, |\vec{u}|), \quad (3.5)$$

where $\delta_{\ell=0} := \int d^2\hat{u}/(4\pi) [F(x, U)/\bar{F}(|\vec{u}|) - 1]$ is the monopole. Since all non-interacting particles beginning at the same x and U will move in the same way, regardless of their masses, all decoupled HDM species have the same $\delta(x, \vec{u}) = F(x, U)/\bar{F}(|\vec{u}|) - 1$. This allows us to reconstruct the perturbations of the individual component species by simply reweighing the flow perturbations. That is,

$$\delta\rho^{(s)} = \frac{g^{(s)}(m^{(s)})^4}{a^4} \int d|\vec{u}| 4\pi |\vec{u}|^2 \sqrt{a^2 + |\vec{u}|^2} \bar{F}^{(s)}(|\vec{u}|) \delta_{\ell=0}(x, |\vec{u}|) \quad (3.6)$$

for the density perturbation of the species s .

In summary, we have generalized the method of effective distributions of ref. [44] to the case of multiple HDM species with arbitrary masses, temperatures, and distribution functions, at all times after the species have decoupled. We have shown its applicability to relativistic as well as non-relativistic HDM, and demonstrated how to recover the density perturbations of the component HDM species. A key task of perturbation theory for HDM, then, is to determine $\delta(x, \vec{u})$. References [45, 60], summarized in section 2.2, did this for a discrete set of $|\vec{u}|$. We next consider how to choose these velocities.

3.2 Discrete momenta and Gauss-Laguerre quadrature

Section 2.3 summarizes the GLQ method for approximating the integral of an exponentially-decaying function in some parameter x , a set that includes the Bose-Einstein, Fermi-Dirac, and Maxwell-Boltzmann distribution functions. The complication is that each $F_{\text{HDM}}^{(s)}$ has its

own exponential decay behaviour, $\exp(-m_{\text{HDM}}^{(s)}|\vec{u}|/T_{\text{HDM},0}^{(s)})$. We therefore need to choose an effective mass m_{EHDM} and temperature constant $T_{\text{EHDM},0}$ for our effective HDM particle, such that for $q := m_{\text{EHDM}}|\vec{u}|/T_{\text{EHDM},0}$, $F_{\text{EHDM}}(q)$ declines as $\exp(-q)$ in the range of q contributing the most to the density, so that we can effectively apply the GLQ method.

There is no general prescription for selecting m_{EHDM} and $T_{\text{EHDM},0}$. Motivated by the fact that both the mean energy density and number density scale as $(1+z)^3$ at late times $z \sim 0$ to an excellent approximation, we define for N_{HDM} species the effective HDM mass to be

$$m_{\text{EHDM}} = \frac{\sum_{s=0}^{N_{\text{HDM}}-1} \bar{\rho}_{\text{HDM},0}^{(s)}}{\sum_{s=0}^{N_{\text{HDM}}-1} \bar{n}_{\text{HDM},0}^{(s)}}, \quad (3.7)$$

to ensure that the contribution of each s to m_{EHDM} is weighted by its contribution to the total density. We further set the effective temperature according to

$$N_{\text{HDM}} m_{\text{EHDM}} T_{\text{EHDM},0} = \sum_s m_{\text{HDM}}^{(s)} T_{\text{HDM},0}^{(s)}, \quad (3.8)$$

i.e., $T_{\text{EHDM},0}$ is the mass-weighted average temperature. Note that the effective distribution function will not, in general, be any equilibrium distribution function, so $T_{\text{EHDM},0}$ does not imply any physical system in thermal equilibrium at that temperature.

Reference [44] instead set m_{EHDM} to the largest of the individual-species masses. While this is a reasonable choice for their case of interest, in which all $T_{\text{HDM},0}^{(s)}$ and $T_{\text{EHDM},0}$ are equal, it is no longer optimal if the heaviest species is also several times colder than the rest, causing its contribution to the total density to be subdominant. Another choice is to demand that $F_{\text{EHDM}}(q)$ be proportional to $\exp(-q)$ at large q , which immediately sets $m_{\text{EHDM}}/T_{\text{EHDM},0}$ equal to the smallest value of $m_{\text{HDM}}^{(s)}/T_{\text{HDM},0}^{(s)}$ amongst all species s . This choice is however also not optimal in general, as the lightest and hottest species generally contribute the least to small-scale clustering. Moreover, reducing the mass of this species will compress the distribution functions of the more dominant species into a smaller range of q , leading to larger errors.

We emphasize that the effective distribution function technique is most effective when all HDM species have similar mass-to-temperature ratios. In this regime, all $m_{\text{EHDM}}/T_{\text{EHDM},0}$ choices above are likely to give similar results: we have tested this proposition by increasing $m_{\text{EHDM}}/T_{\text{EHDM},0}$ by a factor of two relative to eqs. (3.7)–(3.8), for a three-neutrino model with masses 42 meV, 43 meV, and 65 meV, and found a similar accuracy. Outside of this regime, the effective distribution function technique remains mathematically valid, but accuracy will require a large number of quadrature points. Henceforth we fix m_{EHDM} and $T_{\text{EHDM},0}$ as per eqs. (3.7)–(3.8).

Given these definitions, we can now change the variables in F_{EHDM} from $|\vec{u}|$ to $q = m_{\text{EHDM}}|\vec{u}|/T_{\text{EHDM},0}$, i.e.,

$$F_{\text{EHDM}}(q) = \frac{1}{m_{\text{EHDM}}^4} \sum_{s=0}^{N_{\text{HDM}}-1} g_{\text{HDM}}^{(s)} (m_{\text{HDM}}^{(s)})^4 F_{\text{HDM}}^{(s)} \left(q \frac{T_{\text{EHDM},0}}{T_{\text{HDM},0}^{(s)}} \frac{m_{\text{HDM}}^{(s)}}{m_{\text{EHDM}}} \right). \quad (3.9)$$

Section 2.3 then tells us that an integrand should be evaluated at values q_α , for $0 \leq \alpha \leq N_{\text{GLQ}} - 1$ equal to the N_{GLQ} roots of the N_{GLQ} th Laguerre polynomial, for a positive integer

N_{GLQ} . Taking the Fourier transform of \vec{x} and suppressing the time-dependence of $\delta_{\ell=0}$, we may now approximate the perturbed density of eq. (3.5) as

$$\begin{aligned}\delta\rho^{\vec{k}} &= \frac{4\pi m_{\text{EHDM}} T_{\text{EHDM},0}^3}{a^3} \int dq q^2 \sqrt{1 + \frac{T_{\text{EHDM},0}^2 q^2}{m_{\text{EHDM}}^2 a^2}} F_{\text{EHDM}}(q) \delta_{\ell=0}^{\vec{k}}(q) \\ &\approx \frac{4\pi m_{\text{EHDM}} T_{\text{EHDM},0}^3}{a^3} \sum_{\alpha=0}^{N_{\text{GLQ}}-1} w_{\alpha} e^{q_{\alpha}} q_{\alpha}^2 \sqrt{1 + \frac{T_{\text{EHDM},0}^2 q_{\alpha}^2}{m_{\text{EHDM}}^2 a^2}} F_{\text{EHDM}}(q_{\alpha}) \delta_{\alpha 0}^{\vec{k}} \\ &= \sum_{\alpha} \bar{\rho}_{\alpha} \delta_{\alpha 0}^{\vec{k}} \text{ where } \bar{\rho}_{\alpha} := \frac{4\pi m_{\text{EHDM}} T_{\text{EHDM},0}^3}{a^3} \sum_{\alpha} w_{\alpha} e^{q_{\alpha}} q_{\alpha}^2 \sqrt{1 + \frac{T_{\text{EHDM},0}^2 q_{\alpha}^2}{m_{\text{EHDM}}^2 a^2}} F_{\text{EHDM}}(q_{\alpha}),\end{aligned}\tag{3.10}$$

and $\delta_{\alpha 0}$ is evaluated at q_{α} , that is, $|\vec{u}| = q_{\alpha} T_{\text{EHDM},0}/m_{\text{EHDM}}$, corresponding to momentum $\tau_{\alpha} = q_{\alpha} T_{\text{EHDM},0}$. The weight w_{α} is given by eq. (2.20). Recall also our convention of section 2.2 that a wave number superscript denotes a functional dependence upon that wave number.

These GLQ density perturbations $\delta_{\alpha}^{\vec{k}}$ correspond precisely to the flow perturbations of refs. [45, 60], summarized in section 2.2. The effective HDM method has shown us how to use the same flow perturbations for an arbitrary collection of HDM particles, while GLQ has shown us how to choose the flow momenta τ_{α} efficiently. We may use these same $\delta_{\alpha}^{\vec{k}}$ to recover the density perturbations of individual species via

$$\delta\rho^{(s)} \approx \frac{4\pi g_{\text{HDM}}^{(s)} T_{\text{EHDM},0}^3 (m_{\text{HDM}}^{(s)})^4}{m_{\text{EHDM}}^3 a^3} \sum_{\alpha=0}^{N_{\text{GLQ}}-1} w_{\alpha} q_{\alpha}^2 e^{q_{\alpha}} \sqrt{1 + \frac{T_{\text{EHDM},0}^2 q_{\alpha}^2}{m_{\text{EHDM}}^2 a^2}} F_{\text{HDM}}^{(s)}\left(q_{\alpha} \frac{T_{\text{EHDM},0}}{T_{\text{HDM},0}^{(s)}} \frac{m_{\text{HDM}}^{(s)}}{m_{\text{EHDM}}}\right) \delta_{\alpha}^{\vec{k}},\tag{3.11}$$

which is the discretized version of eq. (3.6) within the GLQ scheme.

3.3 Clustering and free-streaming regimes

Our main goal is to quantify HDM clustering through its monopole density perturbation $\delta_{\alpha,\ell=0}^k$. We approach this goal by considering $\delta_{\alpha,0}^k$ in two different regimes, the clustering and free-streaming regimes. In the clustering regime, at length scales much larger than the free-streaming scale, each HDM flow clusters like CDM, so $\delta_{\alpha,0}^k = \delta_{\text{m}}^k$. In the free-streaming regime, at small scales, HDM clustering is strongly suppressed with respect to the total matter clustering. For linear HDM, $\delta_{\alpha,0}^k = (k_{\text{FS},\alpha}/k)^2 \delta_{\text{m}}^k$, where the free-streaming wave number of eq. (2.9) separating the clustering and free-streaming regimes can be written

$$k_{\text{FS},\alpha}(a)^2 = \frac{3\Omega_{\text{m}}(a)\mathcal{H}^2}{2v_{\alpha}^2} = \frac{3\Omega_{\text{m},0}\mathcal{H}_0^2}{2av_{\alpha}^2} = \frac{3\Omega_{\text{m},0}\mathcal{H}_0^2 m_{\text{EHDM}}^2 a}{2\tau_{\alpha}^2} = \frac{3\Omega_{\text{m},0}\mathcal{H}_0^2 m_{\text{EHDM}}^2 a}{2T_{\text{EHDM},0}^2 q_{\alpha}^2},\tag{3.12}$$

as shown in refs. [52, 53, 60, 68]. They also confirm the $\sim 10\%$ accuracy of the interpolation $\delta_{\alpha,0}^k \approx (1 + k/k_{\text{FS},\alpha})^{-2} \delta_{\text{m}}^k$ between the two regimes.

We are especially interested in late-time non-relativistic clustering, $qT_{\text{EHDM},0} \ll am_{\text{EHDM}}$, for which the momentum integral in eq. (3.10) simplifies to $\int dq q^2 F_{\text{EHDM}}(q) \delta^{\vec{k}}(q)$. The clustering limit is simple, as $\delta^{\vec{k}}(q) \approx \delta_{\text{m}}^k$ is momentum-independent and can be factored out of the integral. Thus $\delta\rho_{\text{HDM,clus}}^k \approx \bar{\rho}_{\text{HDM}} \delta_{\text{m}}^k$. The momentum integral is then the same one

that we need to compute the average HDM density,

$$\bar{\rho}_{\text{HDM}}(a) = \frac{4\pi m_{\text{EHDM}} T_{\text{EHDM},0}^3}{a^3} \int dq q^2 F_{\text{EHDM}}(q) \approx \frac{4\pi m_{\text{EHDM}} T_{\text{EHDM},0}^3}{a^3} \sum_{\alpha=0}^{N_{\text{GLQ}}-1} w_{\alpha} q_{\alpha}^2 F_{\text{EHDM}}(q_{\alpha}). \quad (3.13)$$

Thus, in the clustering limit, we must choose N_{GLQ} high enough for $\int dq q^2 F_{\text{EHDM}}(q) \approx \sum_{\alpha=0}^{N_{\text{GLQ}}-1} w_{\alpha} q_{\alpha}^2 F_{\text{EHDM}}(q_{\alpha})$ to our desired level of accuracy.

The free-streaming regime is more complicated. As a rough estimate using eq. (3.12), we may substitute $\delta_{\alpha,0}^k = (k_{\text{FS},\alpha}^2/k^2)\delta_m$ for δ_m in the integral over q in eq. (3.10). Since $k_{\text{FS},\alpha}^2 \propto q_{\alpha}^{-2}$, the result is proportional to $\int dq F_{\text{EHDM}}(q)$. This means that, unlike the clustering limit, the free-streaming regime is dominated by lower q , and the convergence of GLQ must be considered separately in this limit. If every single HDM species has a distribution function that is finite down to $q = 0$, then the integral $\int dq F_{\text{EHDM}}(q)$ converges. If however a single bosonic species is present, then $F_{\text{EHDM}}(q) \propto 1/q$ as $q \rightarrow 0$, leading to a logarithmic divergence of the integral as the lower integration limit approaches zero. Thus we cannot speak rigorously of a free-streaming limit in the general case. Furthermore, the dominance of low- q_{α} flows means that, at large but finite k , the total $\delta_{\alpha,0}^k$ will receive significant contributions from flows that are not yet in the free-streaming regime, that is, $(k_{\text{FS},\alpha}^2/k^2)\delta_m \ll \delta_{\alpha,0}^k \lesssim \delta_m$. Thus, to determine the convergence criterion for GLQ, we should instead focus on the intermediate regime between clustering and free-streaming.

To this end, we note that each individual flow has an approximate interpolated solution given by [52]:

$$\delta_{\alpha,0}^k \approx \frac{\delta_m^k}{(1 + k/k_{\text{FS},\alpha})^2} = \frac{\delta_m}{(1 + q_{\alpha}/q_{\text{cut}})^2}, \quad (3.14)$$

where at the second equality we have recast the $k/k_{\text{FS},\alpha}$ in terms of an infrared cutoff defined as $q_{\text{cut}}(a, k)^2 := (3\Omega_{\text{m},0}\mathcal{H}_0^2 m_{\text{EHDM}}^2 a)/(2k^2 T_{\text{EHDM},0}^2)$. The expression reproduces the correct behaviors at both $k \ll k_{\text{FS},\alpha}$ and $k \gg k_{\text{FS},\alpha}$ limits, making it suitable for the intermediate regime. Clearly, integrating over q in the monopole density perturbation now returns an expression proportional to $\int dq q^2 F_{\text{EHDM}}(q)/(1 + q/q_{\text{cut}})^2$, which converges at all finite k , even for distribution functions $\propto 1/q$ at low q , such as the Bose-Einstein distribution.

Thus we arrive at the convergence criteria for the application of the GLQ scheme. Suppose we are given an effective HDM whose clustering at $k \leq k_{\text{max}}$ and $a \geq a_{\text{min}}$ we would like to compute. We must choose a sufficiently high N_{GLQ} so that the conditions

$$\int dq q^2 F_{\text{EHDM}}(q) \approx \sum_{\alpha=0}^{N_{\text{GLQ}}-1} w_{\alpha} q_{\alpha}^2 e^{q_{\alpha}} F_{\text{EHDM}}(q_{\alpha}), \quad (3.15)$$

and

$$\int dq \frac{q^2 F_{\text{EHDM}}(q)}{(1 + q/q_{\text{cut}}(a_{\text{min}}, k_{\text{max}}))^2} \approx \sum_{\alpha=0}^{N_{\text{GLQ}}-1} \frac{w_{\alpha} q_{\alpha}^2 e^{q_{\alpha}} F_{\text{EHDM}}(q_{\alpha})}{(1 + q_{\alpha}/q_{\text{cut}}(a_{\text{min}}, k_{\text{max}}))^2} \quad (3.16)$$

are both satisfied to our desired accuracy. For practical purposes, we choose $k_{\text{max}}/\sqrt{a_{\text{min}}} = 10 \, h/\text{Mpc}$. We may truncate the GLQ series to $\alpha < N_{\tau}$ flows for $N_{\tau} < N_{\text{GLQ}}$ provided that this truncation keeps the sums on the right hand side within our error threshold.

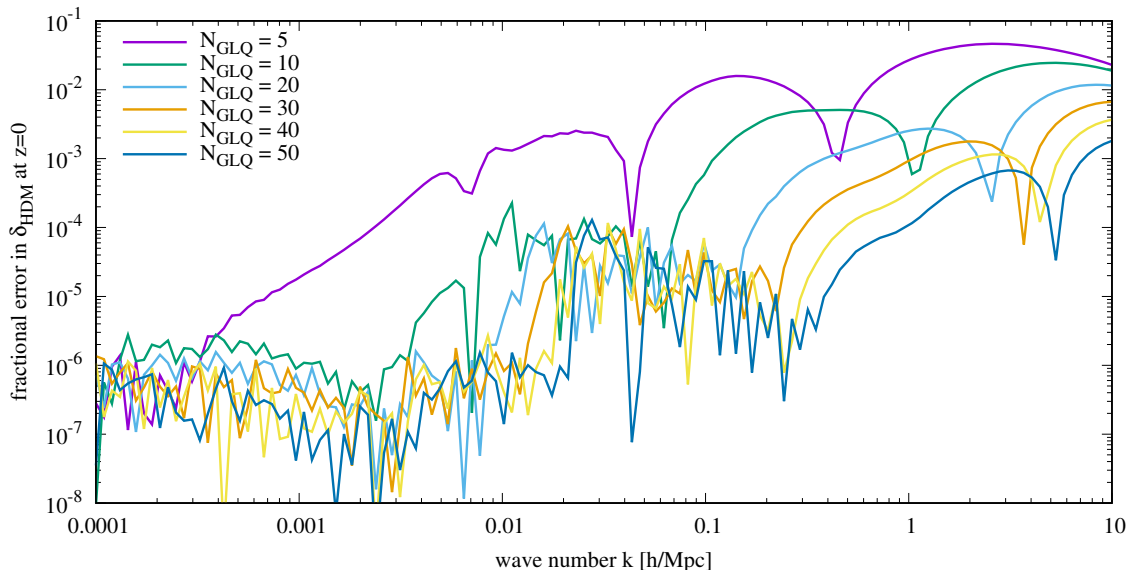


Figure 1. Convergence of Gauss-Laguerre quadrature for a model with three NO neutrinos of total mass $M_\nu = 150$ meV and other parameters given by eq. (3.17) for various choices of N_{GLQ} . For each N_{GLQ} considered, we plot the fractional error incurred in the estimate of the effective HDM density monopole δ_{HDM} , relative to an estimate of the same using $N_{\text{GLQ}} = 70$.

3.4 Convergence with N_{GLQ} and N_τ

Figure 1 quantifies the k -dependent errors in GLQ as N_{GLQ} is raised, by comparing estimates of the effective HDM density contrast using various choices of N_{GLQ} relative to a large- $N_{\text{GLQ}} (= 70)$ estimate. We have assumed a $\nu\Lambda\text{CDM}$ model with three NO neutrinos of mass $M_\nu = 150$ meV, and other parameters

$$\Omega_{\text{m},0}h^2 = 0.1518; \quad \Omega_{\text{b},0}h^2 = 0.02242; \quad A_{\text{s}} = 2.2 \times 10^{-9}; \quad n_{\text{s}} = 0.9665; \quad h = 0.6766. \quad (3.17)$$

The CDM+baryon (cb) fluid is evolved using Time-RG perturbation theory, as implemented in the `redTime` code of refs. [7, 69], to which the neutrinos respond linearly. For each choice of N_{GLQ} , we use as large an N_τ as possible while keeping $q_{N_\tau-1} < 100$, a convention that we adopt henceforth unless otherwise mentioned. That is, for our choices of N_{GLQ} of 5, 10, 20, 30, 40, 50, 70, and 100, the corresponding settings of N_τ are, respectively, 5, 10, 20, 29, 36, 41, 50, and 61. As expected, errors grow with k beyond the free-streaming wave number $k_{\text{FS}} = 0.04$ h/Mpc. Encouragingly, they remain under 0.2% for $N_{\text{GLQ}} = 50$ and under 1.2% for $N_{\text{GLQ}} = 20$, meaning that high precision can be attained with a modest number of flows.

We may also use eqs. (3.15)–(3.16) to estimate the error in GLQ for given N_{GLQ} by comparing a slow but accurate numerical quadrature of the left-hand side to GLQ on the right-hand side. Consider the high- k error in particular, with q_{cut} specified by $k_{\text{max}} = 10$ h/Mpc and $a_{\text{min}} = 1$. For N_{GLQ} of 5, 10, 20, 30, 40, and 50, eq. (3.16) estimates errors of 0.84%, 0.82%, 0.74%, 0.67%, 0.59%, and 0.52%, respectively, compared with actual errors of 2.3%, 1.9%, 1.1%, 0.67%, 0.37%, and 0.18% in figure 1. Thus the error estimate of eq. (3.16) is accurate at the order-of-magnitude level, providing a rough guide to the necessary N_{GLQ} for a given error tolerance. Meanwhile, at low k , the error estimates of eq. (3.15) for N_{GLQ} of 5

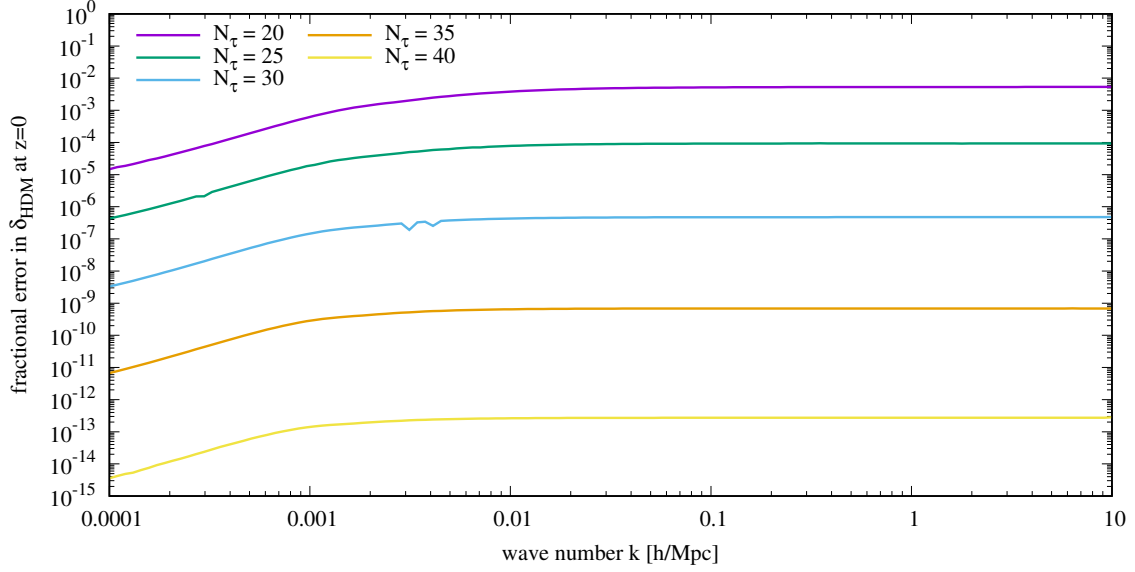


Figure 2. Convergence of GLQ with N_τ , for a model with three neutrinos of total mass $M_\nu = 150$ meV in the normal mass hierarchy, and other parameters given by eq. (3.17). $N_{\text{GLQ}} = 100$ is fixed, while the number of flows N_τ is varied, and the result is compared with $N_\tau = 60$.

and 10, respectively 5×10^{-4} and 4×10^{-6} , somewhat overestimate the errors of figure 1 at $k \sim 10^{-3} h/\text{Mpc}$. However, as N_{GLQ} is increased in the figure, the error seems to hit a floor around 10^{-7} . A larger floor $\sim 10^{-5}$ is also evident at intermediate scales $k \sim 0.1 h/\text{Mpc}$. As these are well within our error budget and subdominant to high- k errors, we do not investigate them further.

Figure 2 considers the convergence of GLQ as we vary N_τ in the estimate of the effective HDM monopole density contrast. With $N_{\text{GLQ}} = 100$ fixed, $N_\tau = 60$ suffices to reach $q_{N_\tau-1} = 95$, and we test smaller values of N_τ against the case of $N_\tau = 60$. Evidently, for every choice of N_τ shown, the corresponding estimate of δ_{HDM} has converged to better than 1%. Even a modest increase in N_τ rapidly decreases the error across the whole k -range; at $N_\tau \geq 45$, the errors fall below the numerical precision of $\sim 10^{-16}$ and are thus not shown in the figure. The corresponding low- k error estimates from comparing the two sides of eq. (3.15) are 5×10^{-3} for $N_\tau = 20$, 9×10^{-5} for $N_\tau = 25$, 5×10^{-7} for $N_\tau = 30$, 7×10^{-10} for $N_\tau = 35$, and 2×10^{-13} for $N_\tau = 40$, about an order-of-magnitude larger than the low- k errors in figure 2. At high k , the error estimates of eq. (3.16) are: 9×10^{-5} for $N_{\text{GLQ}} = 20$; 7×10^{-7} for $N_{\text{GLQ}} = 25$; 2×10^{-9} for $N_{\text{GLQ}} = 30$; 1.5×10^{-12} for $N_{\text{GLQ}} = 35$; and 3×10^{-16} for $N_{\text{GLQ}} = 40$. These are about two orders of magnitude smaller than the high- k errors of figure 2, but scale similarly with N_τ , demonstrating that eqs. (3.15)–(3.16) are reasonable approximate guides to GLQ convergence with N_τ .

3.5 Non-linear perturbation theory: FlowsForTheMasses – II

In principle, the implementation of GLQ and EHDM in `FlowsForTheMasses` is straightforward. EHDM simply means replacing the Fermi-Dirac distribution for neutrinos by the effective distribution function $F_{\text{EHDM}}(q)$ of eq. (3.9). GLQ requires $\tau_\alpha = q_\alpha T_{\text{EHDM},0}$ for each flow α ,

and its corresponding late-time density fraction is proportional to $w_\alpha q_\alpha^2 \exp(q_\alpha) F_{\text{EHDM}}(q_\alpha)$, as discussed in section 3.2. However, problems arise for the smallest q_α .

Reference [45] encountered numerical instabilities in the **FlowsForTheMasses** perturbation theory applied to massive neutrinos in the high- k and high- ℓ regime. A stability threshold k_{st} had to be introduced, above which evolution of the perturbations was no longer tracked, and k_{st} was reduced dynamically as instabilities caused a reduction of the integration step size to below $\Delta\eta = 10^{-6}$. For the fiducial model used, with $\Omega_{\nu,0}h^2 = 0.005$, ref. [45] found it possible to stabilize **FlowsForTheMasses** up to $k_{\text{st}} \gtrsim 3 \text{ h/Mpc}$ by truncating the Legendre moment expansion of the power spectrum input to the mode-coupling integrals, $\ell < N_{\mu,\text{NL}}$, meaning that $A_{\alpha,\text{acd},\text{bef},\ell}^k$ was nonzero only for $\ell < 2N_{\mu,\text{NL}} - 1$. A choice of $N_{\mu,\text{NL}}$ as high as 8 was found to be computationally tractable, but $N_{\mu,\text{NL}} = 6$ achieved a reasonable balance between precision and computational cost. Therefore, following [45], we adopt the choice $N_{\mu,\text{NL}} = 6$ henceforth, unless stated otherwise.

However, applying **FlowsForTheMasses** with $N_{\mu,\text{NL}} = 6$ to a much broader range $\Omega_{\nu,0}h^2 \leq 0.01$, ref. [46] found that this truncation alone was inadequate for the smallest τ and largest neutrino masses, that is, for the smallest velocities. Numerical instabilities affected nearly 20% of their sample of 101 runs. They were able to stabilize **FlowsForTheMasses** for this larger mass range by introducing a further truncation of the ℓ range of the bispectrum integrals $I_{\alpha,\text{acd},\text{bef},\ell}^k$ and the mode-coupling integrals $A_{\alpha,\text{acd},\text{bef},\ell}^k$, restricting $\ell < N_{\mu,\text{AI}}$. They defined stability as the integration being able to reach $z = 0$ with $k_{\text{st}} \geq 1.2 \text{ h/Mpc}$, a definition which we adopt here. The choice of $N_{\mu,\text{AI}} = 4$ or 5 was found to stabilize the perturbation theory, at the cost of an additional error of 1% – 2%, which they quantified by comparison to higher $N_{\mu,\text{AI}}$. While the error associated with this truncation was larger for lower $\Omega_{\nu,0}h^2$, truncation was only necessary for $\Omega_{\nu,0}h^2 > 0.006$.

GLQ exacerbates this instability by requiring more low-velocity bins, necessary for accurately predicting the small-scale clustering of HDM. We find that even for $M_\nu = 59 \text{ meV}$, the minimum allowed value in the normal ordering, **FlowsForTheMasses** is unstable. Furthermore, reducing $N_{\mu,\text{AI}}$ to accommodate the lowest- α flows will introduce unacceptable errors into the larger α ; that is, the $N_{\mu,\text{AI}}$ truncation is too aggressive for our purposes. Thus we allow each flow α to have its own truncation, $\ell < N_{\mu,\text{AI}}^{(\alpha)}$, for its $I_{\alpha,\text{acd},\text{bef},\ell}^k$ and $A_{\alpha,\text{acd},\text{bef},\ell}^k$. A full exploration of the parameter space of all $N_{\mu,\text{AI}}^{(\alpha)}$ would be prohibitively expensive computationally, and we do not consider it here. However, for a modest GLQ order $N_{\text{GLQ}} = 20$, we find that reducing only $N_{\mu,\text{AI}}^{(0)}$ is sufficient to ensure the stability of models with densities $\Omega_{\nu,0}h^2 \lesssim 0.003$, while larger N_{GLQ} can be reached for smaller neutrino masses. Henceforth, unless otherwise mentioned, we only discuss non-linear perturbative results for which stability may be achieved by reducing $N_{\mu,\text{AI}}^{(0)}$ alone, leaving all others at their maximum value of $2N_{\mu,\text{NL}} - 1 = 11$.

Figure 3 compares the power spectra for the $M_\nu = 59 \text{ meV}$ normal and degenerate mass orderings as $N_{\mu,\text{AI}}^{(0)}$ is varied. Since our $N_{\mu,\text{AI}}^{(0)}$ truncation is less aggressive than the global $N_{\mu,\text{AI}}$ truncation of ref. [46], we find considerably smaller errors. Even $N_{\mu,\text{AI}}^{(0)} = 1$ reaches 2.2% accuracy all the way to $k = 1 \text{ h/Mpc}$ for the degenerate ordering, with smaller errors for the normal ordering. Furthermore, since these errors decrease with increasing neutrino mass, and figure 3 considers the smallest allowed M_ν , we may use 2.2% as an upper bound for all

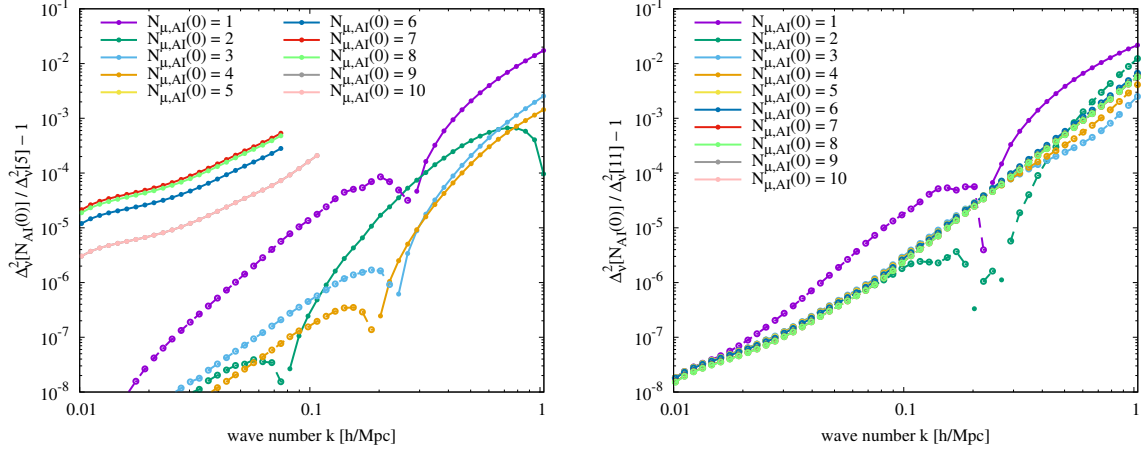


Figure 3. Fractional differences in the $z = 0$ neutrino power spectra computed using various choices of $N_{\mu,AI}^{(0)}$ relative to that computed using the largest stable $N_{\mu,AI}^{(0)}$ value, for a massive neutrino model with $M_\nu = 59$ meV. *Left:* normal mass ordering. The largest $N_{\mu,AI}^{(0)}$ achieving stability is 5. *Right:* degenerate mass ordering. All $N_{\mu,AI}^{(0)}$ are stable, so we have used $N_{\mu,AI}^{(0)} = 11$ as the reference for comparison.

larger masses in the case of $N_{\mu,AI}^{(0)} = 1$; 1.2% in the case of $N_{\mu,AI}^{(0)} = 2$; and 0.25% in the case of $N_{\mu,AI}^{(0)} = 3$, with higher $N_{\mu,AI}^{(0)}$ consistently under 1% for both mass orderings. Thus we regard this $N_{\mu,AI}^{(0)}$ truncation as sufficiently accurate for our purposes, though its extension to higher masses than those considered in this article will require further consideration. Henceforth, we refer to this new version of `FlowsForTheMasses`, implementing EHDM and GLQ, with both $N_{\mu,NL}$ and $N_{\mu,AI}^{(\alpha)}$ truncations, as `FlowsForTheMasses-II`.

3.6 Numerical accuracy

We conclude this section by assessing the numerical accuracy of GLQ and EHDM in perturbation theory. We consider $\nu\Lambda$ CDM models with the cosmological parameters of eq. (3.17) and $M_\nu = 150$ meV neutrinos in either the normal or degenerate mass ordering. In the case of linear perturbation theory, we compare our results to the `CLASS` code, in which we set the three neutrino masses individually. Our `CLASS` runs fix all non-cold dark matter (NCDM) tolerances to 10^{-9} and set `l_max_ncdm=500`.

Figure 4 considers the DO case in order to facilitate comparison with the degenerate-mass multi-fluid perturbation theory of ref. [60], which uses uniform-number-density neutrino bins. It demonstrates that GLQ is far more efficient than uniform bins. The number of flows, N_τ , is directly proportional to the computational cost, since the neutrinos are the most computationally-demanding part of the linear perturbative calculation. Evidently, GLQ with $N_\tau = 35$ has a $k \geq 1$ h/Mpc error two orders of magnitude below that of uniform binning with $N_\tau = 100$. GLQ even outperforms the thousand-bin calculation by more than an order of magnitude. Note that this $N_\tau = 35$ GLQ curve is most closely comparable to the $N_{GLQ} = 40$ curve in figure 1, which used $N_\tau = 36$. Both have errors $\lesssim 0.1\%$.

Next, we consider individual neutrino density monopoles in the NO case. Figure 5 demonstrates the accuracy of the combined GLQ and EHDM methods compared with `CLASS`,

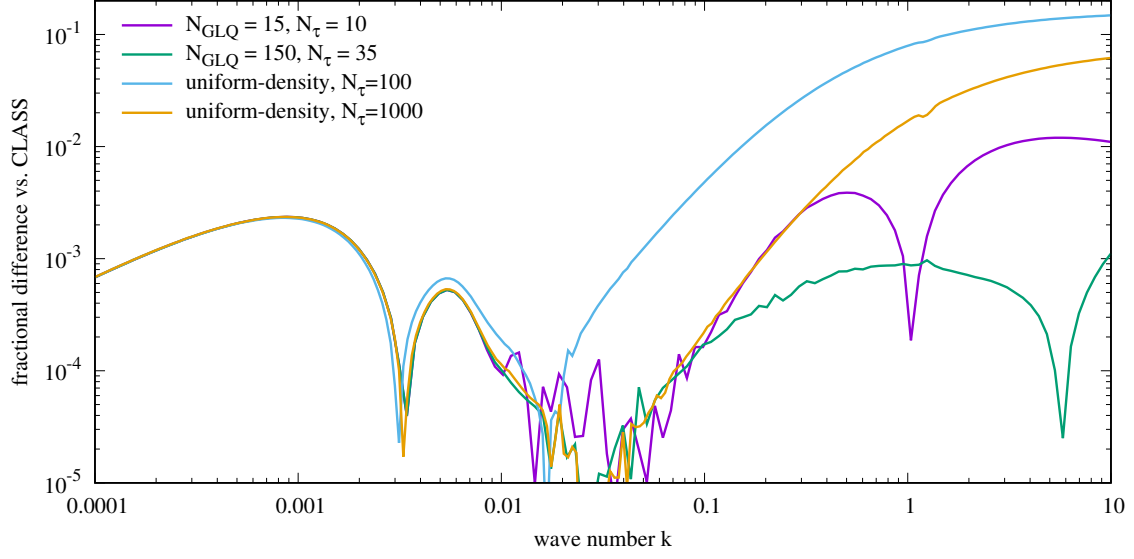


Figure 4. Comparison of HDM densities δ_{HDM} , computed using the linear perturbation theory of section 2.2 with GLQ or uniform-density binning, to CLASS. We consider three degenerate-mass neutrinos with $M_\nu = 150$ meV at $z = 0$.

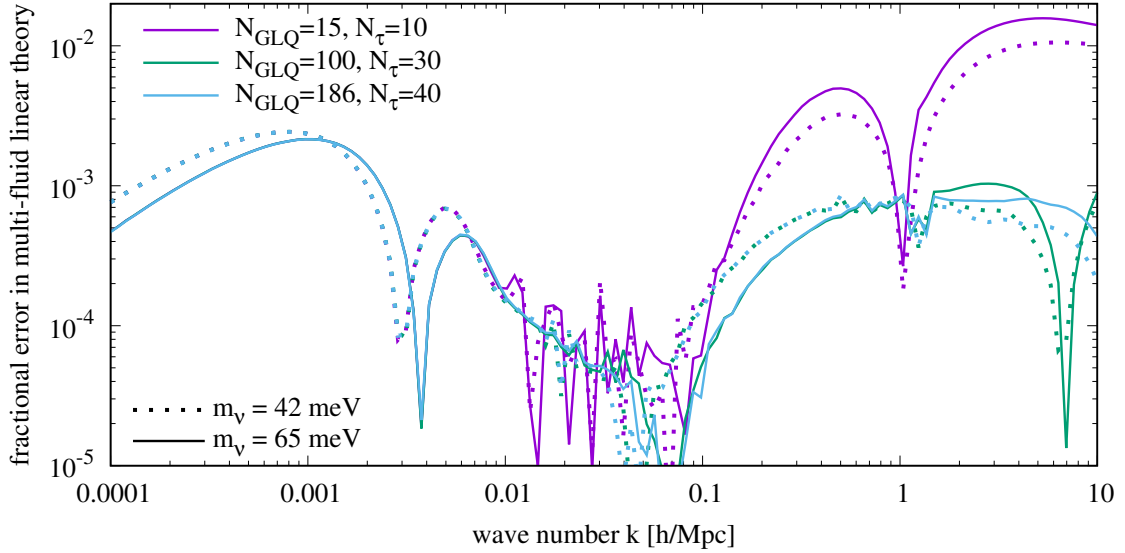


Figure 5. Comparison of HDM densities δ_{HDM} computed using the linear perturbation theory of section 2.2 implementing GLQ and EHDM, with the output of CLASS at $z = 0$. The model has three NO neutrinos with $M_\nu = 150$ meV. The 42 meV (65 meV) neutrino is shown using dotted (solid) lines; errors in the 42 meV and 43 meV neutrinos are nearly identical.

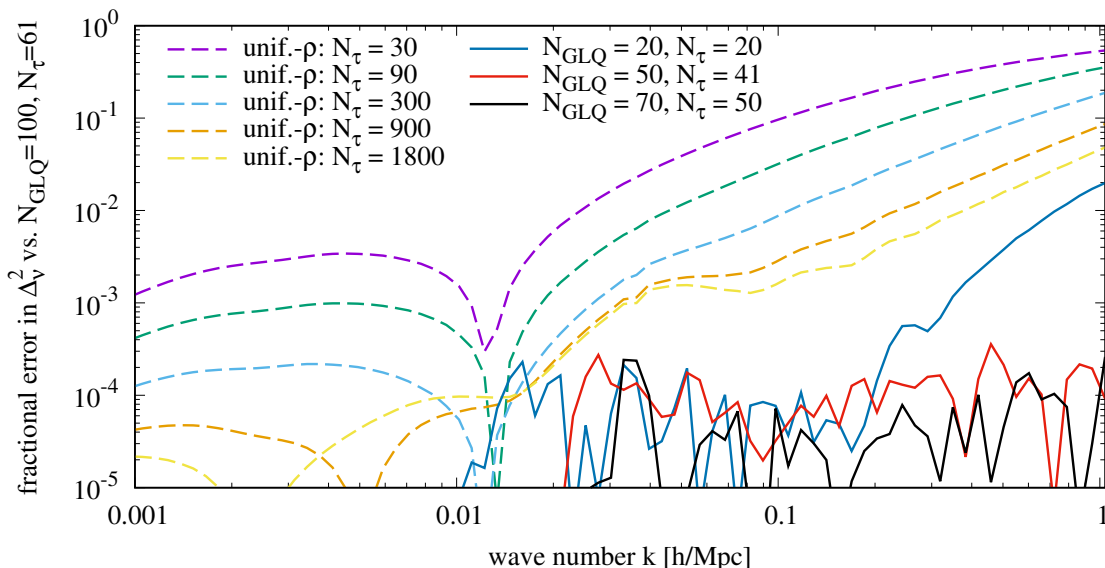


Figure 6. Fractional errors in $\Delta_\nu^2(k)$ at $z = 0$ computed using uniform-density momentum bins (dashed) and GLQ bins (solid) relative to a reference model with $N_{\text{GLQ}} = 100$ and $N_\tau = 60$. The neutrino model contains $M_\nu = 150$ meV NO masses.

in which the three neutrinos are considered separately. Even a modest number of GLQ points, $N_{\text{GLQ}} = 15$, agrees with **CLASS** to $< 2\%$ across the entire k range for the both the light and the heavy neutrinos separately, with somewhat larger errors for the more massive neutrino. Larger N_{GLQ} reduces these errors to $\leq 0.1\%$ for $k \geq 0.002$ h/Mpc . Increasing N_{GLQ} beyond ~ 100 yields no discernible improvements to the accuracy, showing that the EHDM momentum resolution is no longer a dominant source of error. We thus confirm that the EHDM method is fundamentally sound and that GLQ is accurate for $k \leq 10$ h/Mpc .

Finally, we compare uniform-density three-species binning to GLQ and EHDM in non-linear perturbation theory, for three NO neutrinos with $M_\nu = 150$ meV. The computational expense of **FlowsForTheMasses** rises in proportion to $N_{\mu,\text{NL}}^6$, so we restrict ourselves in this subsection to $N_{\mu,\text{NL}} = 3$, the minimum value recommended by ref. [45]. In the calculation with uniform-density bins, each of the three neutrinos is tracked using $N_\tau/3$ bins of equal density. Thus bins corresponding to different neutrino species have different densities. Since we showed GLQ to have converged by $N_{\text{GLQ}} = 100$, we use the GLQ power spectrum with $N_{\text{GLQ}} = 100$ and $N_\tau = 61$ as the reference model against which all others are compared.

Figure 6 compares several neutrino power spectrum computations using uniform-density as well as GLQ binning methods. Up to a numerical noise at the $\sim 0.01\%$ level, N_{GLQ} of 50 and 70 are nearly identical to the reference model for $k \leq 1$ h/Mpc . Even $N_{\text{GLQ}} = 20$ is consistent with that noise up to $k = 0.2$ h/Mpc and has $\leq 2\%$ errors up to $k = 1$ h/Mpc . At large scales, $k \leq 0.3$ h/Mpc , the highest-resolution uniform-density binning agrees with all of the GLQ calculations at the percent level. However, uniform-density binning converges slowly at high k , where a high resolution of the smallest momenta is essential to an accurate computation. At $k = 1$ h/Mpc , the lowest-resolution GLQ, with $N_{\text{GLQ}} = N_\tau = 20$, is more accurate than even the highest-resolution uniform-density binning using nearly a hundred times as many

bins, illustrating the advantages of Gauss-Laguerre quadrature. We also see that the two highest-resolution uniform-density calculations disagree by $\geq 1\%$ for $k \geq 0.4 \ h/\text{Mpc}$ and by $\approx 4\%$ at $k = 1 \ h/\text{Mpc}$, justifying our use of GLQ for the reference model.

We therefore conclude that **FlowsForTheMasses-II** with EHDM and GLQ has converged at about the percent level for N_{GLQ} of 15 or 20. Since a $N_{\text{GLQ}} = N_\tau = 20$, $N_\mu = 10$, $N_{\mu,\text{NL}} = 6$ calculation can be run in under 5 hours on a single 32-processor node of the Katana cluster at UNSW-Sydney, we treat this as our standard set of **FlowsForTheMasses-II** parameters. We proceed to apply it to neutrinos and other HDM models.

4 Results I: non-linear enhancement of HDM clustering

As **FlowsForTheMasses-II** has a wide range of applications, we divide our results into two sections. This section quantifies the enhancement of HDM clustering due to non-linear growth, for pure neutrino models as well as models including an axionic HDM component. We begin in section 4.1 by demonstrating that Gauss-Laguerre quadrature reduces the small-scale power spectrum error by more than a factor of two for light neutrinos, allowing our perturbative treatment to agree with high-quality simulations to $\approx 20\%$ up to $k = 1 \ h/\text{Mpc}$ when combined with the emulator of ref. [70]. Section 4.2 compares **FlowsForTheMasses-II** to an N-body simulation that included neutrino species of two different masses, demonstrating both that we can accurately recover the power spectrum of each species and that neglecting the mass difference leads to a significant error in the total non-linear neutrino power. This latter finding motivates us in section 4.3 to quantify the error in the commonly-made degenerate-mass-ordering approximation. Finally, section 4.4 confirms the accuracy of our EHDM treatment for models containing a single axion in addition to two massive neutrinos. Afterwards, section 5 explores recent attempts to evade cosmological neutrino bounds through a non-standard neutrino sector.

4.1 Accuracy at low M_ν : solving a puzzle

Reference [46] encountered a mysterious $\approx 50\%$ small-scale error in **FlowsForTheMasses** and its companion **Cosmic-Enu** emulator. This error at $k \approx 1 \ h/\text{Mpc}$ is evident in comparisons with a series of degenerate-mass $\nu\Lambda\text{CDM}$ simulations conducted by the Euclid code-comparison project in ref. [71] for M_ν ranging from 150 meV to 600 meV; the remaining parameters are $\Omega_{\text{m},0}h^2 = 0.1432$, $\Omega_{\text{b},0}h^2 = 0.022$, $A_s = 2.215 \times 10^{-9}$, $n_s = 0.9619$, and $h = 0.67$. This error is considerably larger than the $15\% - 20\%$ error expected from the N-body comparison of the original **FlowsForTheMasses** publication, ref. [45]. The simulation of that reference used a small volume, a box of edge length $128 \text{ Mpc}/h$, in order to reduce shot noise, at the cost of neglecting larger-scale power that could flow down to smaller scales due to non-linear clustering. This could explain some of the error, but not its M_ν -independence.

Although they were unable to find a conclusive explanation for this M_ν -independent error, ref. [46] suggested three possibilities:

1. Perturbation theory error. Non-linear perturbation theory is most accurate on quasi-linear scales, and smaller-scale accuracy would require higher-order perturbative corrections. However, lighter neutrinos cluster more linearly, so it is difficult to explain how this error remains $\approx 50\%$ while M_ν is varied by a factor of four.

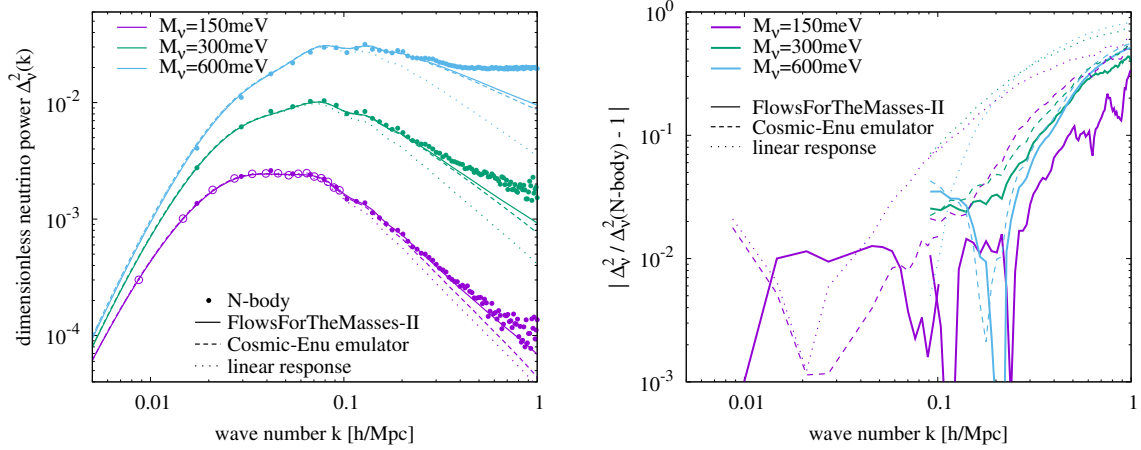


Figure 7. Predictions of the dimensionless neutrino power spectrum at $z = 0$ for the degenerate-ordering $M_\nu = 150 \text{ meV}$, 300 meV , and 600 meV $\nu\Lambda\text{CDM}$ models of ref. [71], using a variety of computation methods. *Left:* the absolute neutrino power spectrum computed using **FlowsForTheMasses-II** (solid), **Cosmic-Enu** (dashed), linear response (dotted), and N-body simulations (points). In the case of N-body predictions, filled points correspond to $(512 \text{ Mpc})^3$ simulation volumes and open points to a volume of $(1024 \text{ Mpc})^3$ for the $M_\nu = 150 \text{ meV}$ model. *Right:* fractional errors in $\Delta_\nu^2(k)$ relative to the N-body predictions.

2. Non-perturbative clustering. Perturbation theory cannot account for non-perturbative structures such as CDM halos, which are expected to capture some portion of the neutrino population. However, again, a fourfold change in M_ν will substantially change the number of neutrinos below the escape velocity of the typical halo, so such capture should be strongly M_ν -dependent.
3. N-body systematic biases. Reference [71] saw a 30% – 40% small-scale scatter among the different simulation methods. This scatter could be larger if errors due to imperfect convergence and incorrect initialization, as studied in ref. [72], are included.

Here, we are able to conclude definitively that this 50% error is actually the combination of two different errors. **FlowsForTheMasses** at $k \approx 1 \text{ h/Mpc}$ is indeed breaking down for higher M_ν . At low M_ν , an inadequate sampling of low neutrino momenta, responsible for most of the small-scale clustering, leads to large errors. This latter error can be substantially reduced, either by significantly increasing the number of momentum bins, or by switching to a more efficient quadrature method such as GLQ, as we proceed to show.

Our results for a range of M_ν are demonstrated in figure 7. The **FlowsForTheMasses-II** curves use GLQ with $N_{\text{GLQ}} = 50$ and $N_\tau = 41$, while the **Cosmic-Enu** curves emulate **FlowsForTheMasses** using $N_\tau = 50$ uniform-density bins. In both cases, we approximate the clustering of the CDM and baryons using the Mira-Titan IV $\Delta_{\text{cb}}^2(k)$ Emulator (MT4) of ref. [70] for $z \leq 2$, and Time-RG for $z > 2$. Up to shot noise in the simulation, **FlowsForTheMasses-II** is accurate to $\approx 20\%$ for $M_\nu = 150 \text{ meV}$, $\approx 40\%$ for $M_\nu = 300 \text{ meV}$, and $\approx 50\%$ for $M_\nu = 600 \text{ meV}$. The switch to GLQ has only a minor impact at the highest mass shown, while the error is dominated by either higher-order or non-perturbative clustering. By increasing significantly with M_ν , this error behaves as expected of a breakdown in perturbation theory.

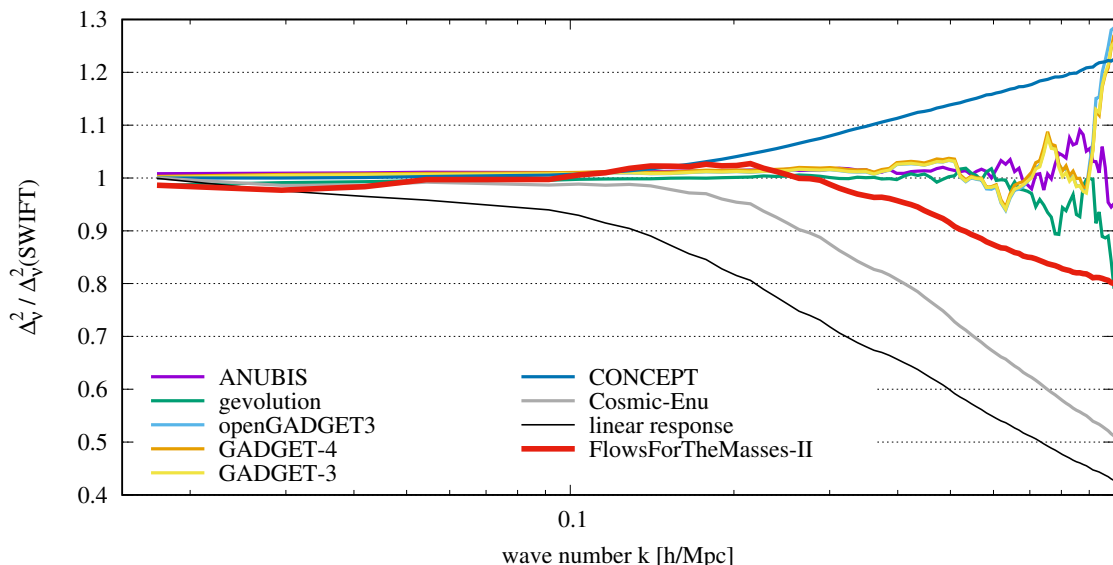


Figure 8. Comparison of `FlowsForTheMasses-II` in its prediction of the $z = 0$ dimensionless neutrino power spectrum $\Delta_\nu^2(k)$ for the degenerate-ordering $M_\nu = 150$ meV $\nu\Lambda$ CDM model of ref. [71], against a variety of neutrino simulation methods. Also shown are the linear response power spectrum of ref. [60] and the `Cosmic-Enu` emulator of ref. [46].

We find, however, that the numerical instabilities discussed in section 3.5 become severe for $M_\nu = 600$ meV. That section defined a flow-dependent truncation $\ell < N_{\mu, AI}^{(\alpha)}$ of the number of angular modes passed to the non-linear mode-coupling integrals. Thus far, with $M_\nu \leq 300$ meV, we have found that reducing $N_{\mu, AI}^{(0)}$ is sufficient to stabilize the perturbation theory to $k = 1.2$ h/Mpc. However, we find for $M_\nu = 600$ meV that we must extend this truncation to the first three flows, reducing each of $N_{\mu, AI}^{(0)}$, $N_{\mu, AI}^{(1)}$, and $N_{\mu, AI}^{(2)}$ to 3. Evidently, GLQ exacerbates the numerical instabilities of `FlowsForTheMasses-II` for large M_ν .

Figure 8 provides, for the $M_\nu = 150$ meV model, further detail about the improved low-momentum sampling in `FlowsForTheMasses-II`. Here, several different N-body methods [73–84] are compared with linear response [60, 85], the `Cosmic-Enu` emulator, and `FlowsForTheMasses-II`; the low-shot-noise SWIFT simulation of ref. [86], based upon the δf method of refs. [87, 88], is used as a reference. We can now see more clearly that GLQ reduces the `FlowsForTheMasses-II` error to $< 10\%$ up to $k = 0.5$ h/Mpc and $< 20\%$ up to $k = 1$ h/Mpc; its errors are now comparable to the scatter among the different N-body methods themselves. Thus GLQ has reduced the error of `FlowsForTheMasses-II` relative to `FlowsForTheMasses` and `Cosmic-Enu` by more than a factor of two. This represents a significant improvement over the uniform-density-binned codes for $k \gtrsim 0.2$ h/Mpc and over the linear response method for $k \gtrsim 0.05$ h/Mpc.

The highest-resolution massive neutrino N-body simulation conducted thus far is the TianNu simulation of refs. [42, 89, 90], which tracked 2.6 trillion neutrino particles in a cubic box of edge length 1200 Mpc/h. It approximated the normal mass ordering by simulating a single 50 meV neutrino, with the remaining two assumed to be massless. Figure 9 compares `FlowsForTheMasses-II`, with $N_{\text{GLQ}} = 20$ GLQ bins and $N_\tau = 20$ momentum bins, to the

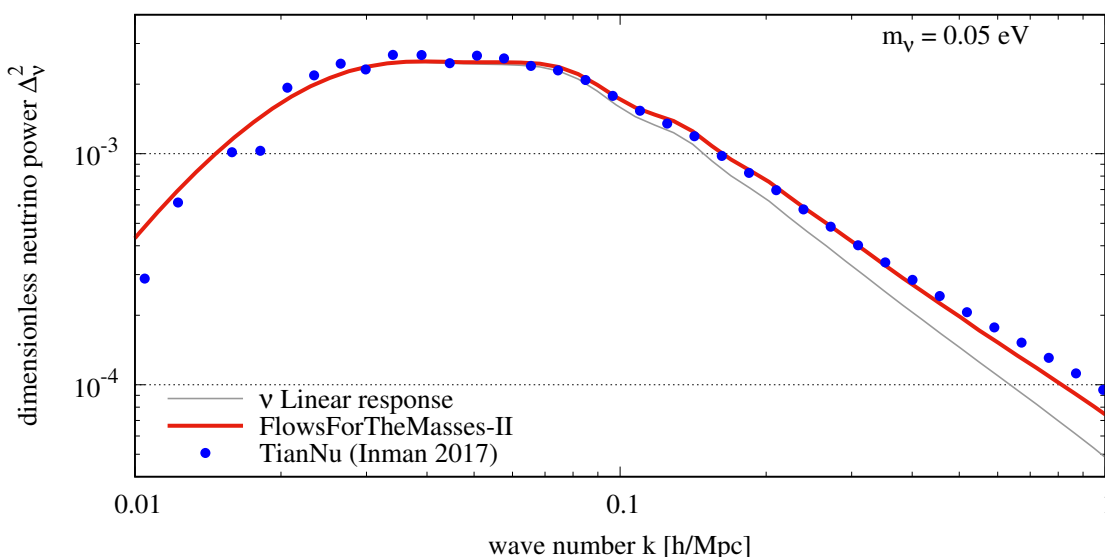


Figure 9. The $z = 0$ dimensionless neutrino power spectrum from the TianNu simulation of ref. [89], compared with **FlowsForTheMasses-II** (thick line) and linear response (thin line). The cosmological model contains one massive neutrino at 50 meV and two massless states.

TianNu power spectrum, finding an accuracy of 21% at $k = 1 \text{ h/Mpc}$. We have thus confirmed the accuracy of our GLQ-based **FlowsForTheMasses-II** perturbation theory to $\approx 20\%$ for $k \leq 1 \text{ h/Mpc}$ for neutrino masses up to 50 meV.

We have considered building a new emulator using GLQ momentum bins. However, the high- M_ν numerical instabilities noted above for $M_\nu = 600 \text{ meV}$ required the adjustment of three separate truncation parameters to resolve, and masses $M_\nu \approx 930 \text{ meV}$ at the upper end of the emulation range may require more. Individually adjusting this many parameters and demonstrating the insensitivity of the resulting power spectra to their precise values, for every single high- M_ν model in the training set, would be prohibitively computationally expensive. Alternatively, we may regard **Cosmic-Enu** as well-suited to $M_\nu \gtrsim 300 \text{ meV}$, where the benefits of GLQ are diminishing and where the degenerate mass ordering becomes accurate to $\approx 2\%$, as we shall see in section 4.3. Then we may use GLQ to construct two separate $M_\nu \leq 300 \text{ meV}$ emulators for the normal and inverted mass orderings. We leave this for future work.

4.2 Normal ordering and recovery of individual-species power spectra

Next, we apply the effective HDM method to the normal hierarchy, with the ultimate goal of verifying the individual-species power spectra implied by eq. (3.11). We compare our results to the **gevolution** N-body simulation of ref. [77], which uses an approximation to the normal mass ordering. Since multiple neutrino species substantially increase the computational costs of simulations, ref. [77] simulated a doubly-degenerate 60 meV neutrino and a singly-degenerate 80 meV neutrino, for a total mass $M_\nu = 200 \text{ meV}$. That is, they neglected the smaller mass splitting, Δm_{21}^2 , and approximated the larger one, Δm_{31}^2 , as 0.0028 eV^2 . Their simulation tracked 4096^3 CDM+baryon particles and 1.7×10^{11} neutrino particles in a $(2 \text{ Gpc}/h)^3$ box with a force resolution of $0.5 \text{ Mpc}/h$.

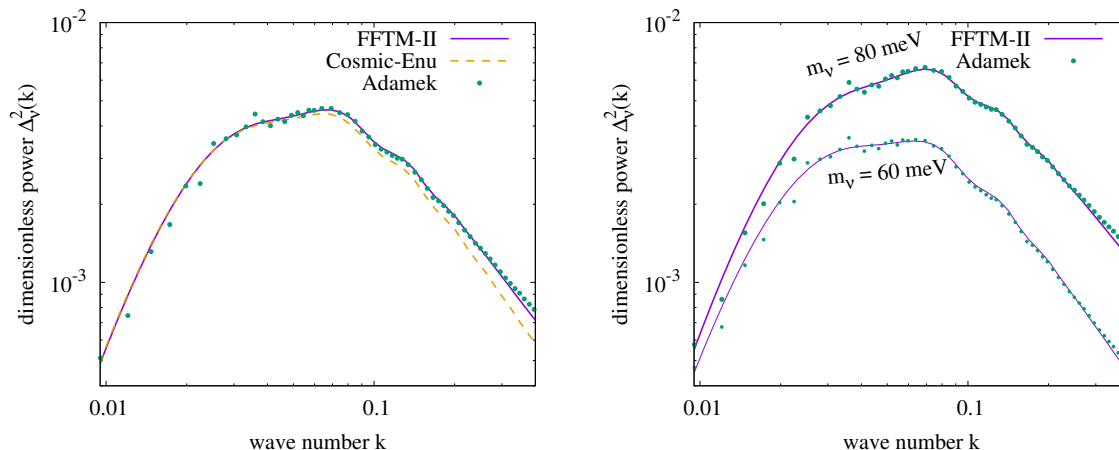


Figure 10. Comparison of the **FlowsForTheMasses-II** neutrino power spectrum at $z = 0$, using GLQ momentum binning, to the N-body neutrino simulation results of ref. [77], assuming $M_\nu = 200$ meV and an approximate normal mass ordering. *Left:* total neutrino power spectrum. For comparison, we plot also the emulated **Cosmic-Enu** power spectrum of ref. [46] for the same M_ν but assuming a degenerate ordering. *Right:* separate power spectra of the 80 meV and 60 meV species.

Figure 10 compares **FlowsForTheMasses-II** with $N_{\text{GLQ}} = 50$ and $N_\tau = 41$ to ref. [77] up to $k = 0.4$ h/Mpc , after which the N-body power spectra become dominated by shot noise. The other cosmological parameters of this $\nu\Lambda\text{CDM}$ model are $\Omega_{\text{m},0}h^2 = 0.142412$, $\Omega_{\text{b},0}h^2 = 0.022032$, $A_s = 2.215 \times 10^{-9}$, $n_s = 0.9619$, and $h = 0.67556$. Also shown in the left panel is the emulated power spectrum of **Cosmic-Enu**, which assumes a degenerate neutrino mass ordering. Relative to **Cosmic-Enu**, **FlowsForTheMasses-II** represents a nearly fourfold reduction in RMS fractional error over the range 0.1 $h/\text{Mpc} < k < 0.15$ h/Mpc , from 6.2% to 1.6%, and over a threefold reduction over 0.35 $h/\text{Mpc} < k < 0.4$ h/Mpc , from 24% to 7%.

Moreover, **FlowsForTheMasses-II** accurately recovers the power spectra of individual neutrino species, as seen in the right panel of figure 10. Its RMS fractional errors are comparable to those of the total neutrino power spectrum, though the heavier neutrino species has slightly smaller errors at low k and larger ones at high k . For example, in the 0.1 $h/\text{Mpc} < k < 0.15$ h/Mpc range, the **FlowsForTheMasses-II** error is 0.9% for the 80 meV species and 2.2% for the 60 meV species, while in the 0.35 $h/\text{Mpc} < k < 0.4$ h/Mpc range, these errors grow to 8.2% and 5.7%, respectively. This rise in the error of the heavier neutrino suggests a small-scale non-linear effect not captured by perturbation theory. Also evident from the smallest scales in the same plot is the fact that the 80 meV neutrino power spectrum is about three times that of the 60 meV neutrino, consistent with the $\Delta_\nu^2 \propto m_\nu^4$ scaling of refs. [52, 53].

4.3 Errors in the degenerate-ordering approximation

Bounds on M_ν commonly make the approximation of a degenerate mass ordering, which reduces the computational cost of their power spectrum computations. We next investigate the error in the total neutrino power spectrum arising from this approximation. This error necessarily vanishes in both the clustering limit, when all neutrino masses cluster the same,

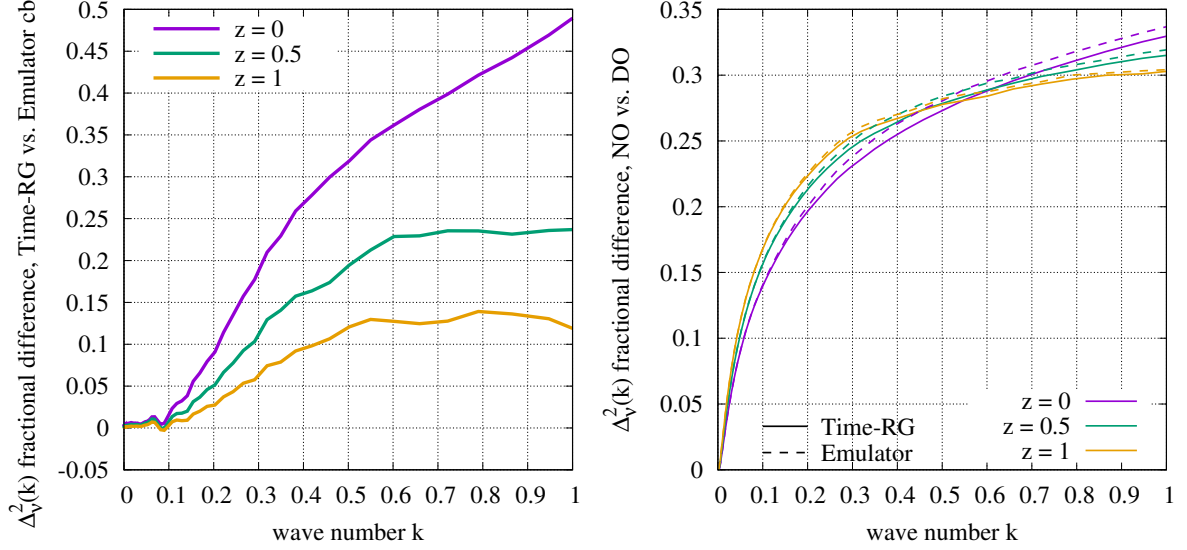


Figure 11. Dependence of the total neutrino power spectrum $\Delta_\nu^2(k)$, with $M_\nu = 150$ meV, upon the method used to compute the cb clustering. (Left) Time-RG perturbation theory is compared with the MT4 emulator of ref. [70] for DO masses. (Right) Sensitivity of $\Delta_\nu^2(k)$ to the difference between NO and DO neutrino masses. Time-RG and MT4 sensitivities agree to 3% at $z = 0$ and more closely at higher redshifts.

and the high- M_ν limit, in which fractional differences between the neutrino masses vanish. Thus we focus on $M_\nu \leq 300$ meV and $0.01 \, h/\text{Mpc} \leq k \leq 1 \, h/\text{Mpc}$. Since the free-streaming-limit power spectrum for a single neutrino of mass m_ν scales as m_ν^4 in the linear case [52], and as an even higher power of m_ν with non-linear corrections [46], we expect the degenerate mass ordering to underestimate the neutrino power in all cases. As a numerical example, we consider $\nu\Lambda\text{CDM}$ cosmologies with the parameters of eq. (3.17) fixed.

Beginning in this subsection, we switch to using Time-RG perturbation theory for the cb fluid. While ref. [60] showed Time-RG to overestimate the cb power by $\approx 30\%$ for $0.3 \, h/\text{Mpc} \lesssim k \lesssim 1 \, h/\text{Mpc}$ at $z = 0$, the authors found considerably greater accuracy when studying the sensitivity of the cb power to small parameter variations. Reference [91] similarly found that the Time-RG-based two-halo power spectrum accurately predicted the effects of a small change in M_ν .

Figure 11 considers the impact of the cb clustering overestimate upon the neutrino power spectrum, using as a reference the MT4 emulator, which assumes DO neutrinos. In the left plot, a direct comparison of $\Delta_\nu^2(k)$ using Time-RG and MT4 cb fluids shows that Time-RG also leads to an overestimate of the neutrino power, with the corresponding error exceeding 20% for $k \geq 0.3 \, h/\text{Mpc}$ at $z = 0$ and $k \geq 0.5 \, h/\text{Mpc}$ at $z = 0.5$. However, in figure 11 (Right), we find a remarkably close $\leq 3\%$ agreement in the enhancement of the NO over the DO $\Delta_\nu^2(k)$, when using Time-RG vs. MT4 for the cb power, despite MT4 assuming DO neutrinos to compute the cb power.

We may estimate the enhancement in the linear density contrast $\sum_s (f_{\nu,s}/f_\nu) \delta_{\nu,s}$ averaged over species s , in the free-streaming regime $k > k_{\text{FS},s}$, using $f_{\nu,s}/f_\nu = m_{\nu,s}/M_\nu$ and $\delta_{\nu,s} \approx \delta_m(k_{\text{FS},s}/k)^2$ with $k_{\text{FS},s} \propto m_{\nu,s}$. For $M_\nu = 150$ meV and NO masses 42 meV, 43 meV, and

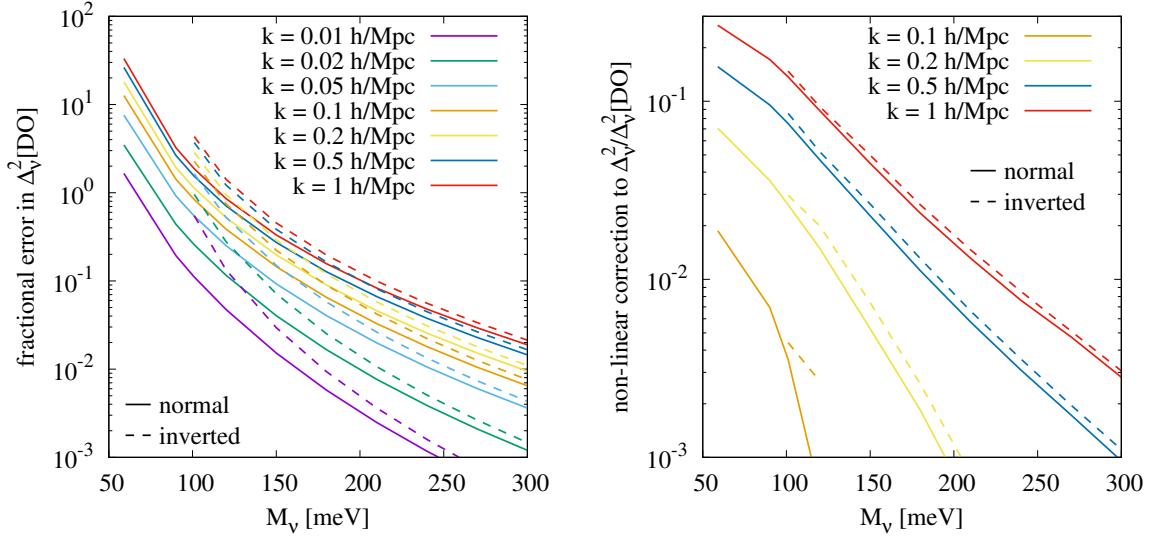


Figure 12. Errors in the degenerate-ordering approximation in the neutrino power spectrum. *Left:* total fractional error in the DO power spectrum $\Delta_\nu^2[\text{DO}]$ as an approximation to $\Delta_\nu^2[\text{NO}]$ (solid) and $\Delta_\nu^2[\text{IO}]$ (dashed). *Right:* fractional non-linear contribution to the error in the DO approximation. In all cases, we have used $N_{\text{GLQ}} = N_\tau = 20$ and $N_\mu = 10$.

65 meV, this implies a 14% linear enhancement in the density contrast and a 30% enhancement in the power, consistent with figure 11 (Right) at $z = 1$. Since neutrinos couple to the cb fluid only through Φ , to which their contribution is suppressed by their mass fraction $f_\nu \sim 1\%$ for $M_\nu = 150$ meV, we may expect this NO neutrino clustering enhancement to affect the cb power at the $\sim 0.3\%$ level. Repeating for IO masses implies an enhancement of $\sim 0.45\%$ in the cb power. Even allowing for non-linear clustering, we can trust the MT4 cb power spectrum to $\lesssim 1\%$ for NO and IO masses with $M_\nu \gtrsim 150$ meV. Figure 11 (Right), then, demonstrates that Time-RG predicts the hierarchy-dependence of $\Delta_\nu^2(k)$ to 3% – 4%.

On the other hand, the error in using the MT4 cb power for NO and IO neutrinos will depend significantly upon M_ν . Thus our compromise is to use the MT4 cb power when computing the neutrino power spectrum for any mass hierarchy, but to use the Time-RG cb power to compute (i) small fractional differences between neutrino models, and (ii) the total power spectrum and ratios for HDM beyond standard neutrinos. In practice, that means that sections 4.1 and 4.2 applied the MT4 emulator, while we will use Time-RG henceforth.

Figure 12 shows how the DO approximation fares in the prediction of the neutrino power spectrum, when it is used at several k to estimate the NO and IO power spectra over their respective mass ranges $M_\nu \geq 59$ meV and $M_\nu \geq 101$ meV. As expected, the relative excess of the actual Δ_ν^2 over $\Delta_\nu^2[\text{DO}]$ rises with decreasing M_ν and increasing k , with the degenerate ordering underestimating the power spectrum at $k = 1$ h/Mpc and $M_\nu = 59$ meV by a factor of more than thirty. Figure 12 (Right) further shows that non-linear corrections alone, computed here using Time-RG for the cb fluid and `FlowsForTheMasses-II` for the neutrinos, represent more than 10% of this increase for $k \geq 0.5$ h/Mpc. As constraints improve, we must be increasingly cautious about applying the degenerate-ordering approximation to studies of small-scale neutrino effects.

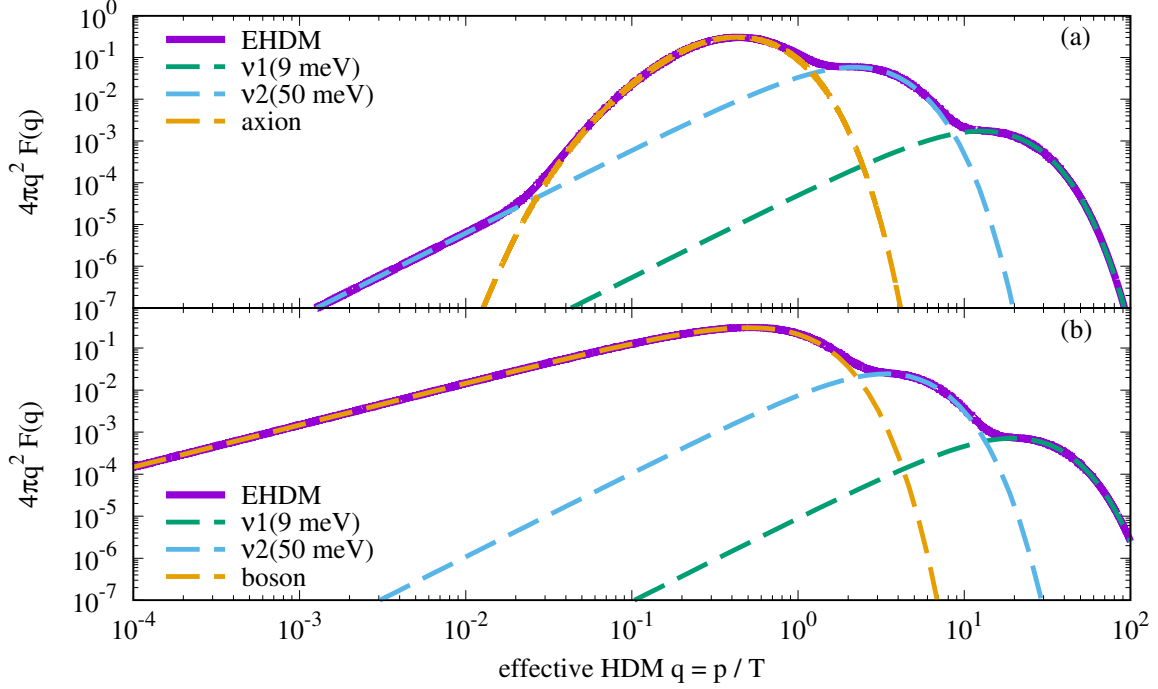


Figure 13. Effective distribution functions for two HDM models containing $M_\nu = 59$ meV NO neutrinos, in addition to (a) an axion whose distribution function was computed in ref. [35], or (b) a generic thermal boson following a Bose-Einstein distribution.

4.4 Extension to axionic models

We demonstrated in section 3.1 that the EHDM formalism is not limited to neutrinos, but applies to any set of HDM species. Here, we test its accuracy for models containing either axions or axion-like bosons, along with massive neutrinos. We assume a cosmological constant as the dark energy, and we fix the following cosmological parameters:

$$\Omega_{m,0}h^2 = 0.1424; \quad \Omega_{b,0}h^2 = 0.02242; \quad A_s = 2.1 \times 10^{-9}; \quad n_s = 0.966; \quad h = 0.6766. \quad (4.1)$$

We assume minimal-mass NO neutrinos, $M_\nu = 59$ meV, along with another particle with mass 228 meV, temperature $T_{\text{HDM},0}^{(s)} = 1.86$ K, and one of the following distribution functions:

- (a) axionic, as computed in ref. [35] for QCD axion production rates based upon pion-pion scattering data, and provided to us by the authors;
- (b) bosonic, that is, the Bose-Einstein distribution function $F_{\text{BE}}(q) = (2\pi)^{-3}(e^q - 1)^{-1}$.

The mass 228 meV is chosen so that in the axionic case, the thermal axion population's contribution to N_{eff} of $\Delta N_{\text{eff}} = 0.19$ remains slightly below observational bounds. The bosonic case exceeds these bounds and is included for illustrative purposes. Figure 13 shows the distribution functions of these two models.

As a high-accuracy reference calculation, we use a set of hybrid N-body simulations based upon the code of ref. [92], extended to EHDM models and using GLQ flows, as implemented in our companion paper, ref. [61]. We conducted one hybrid simulation for each of the

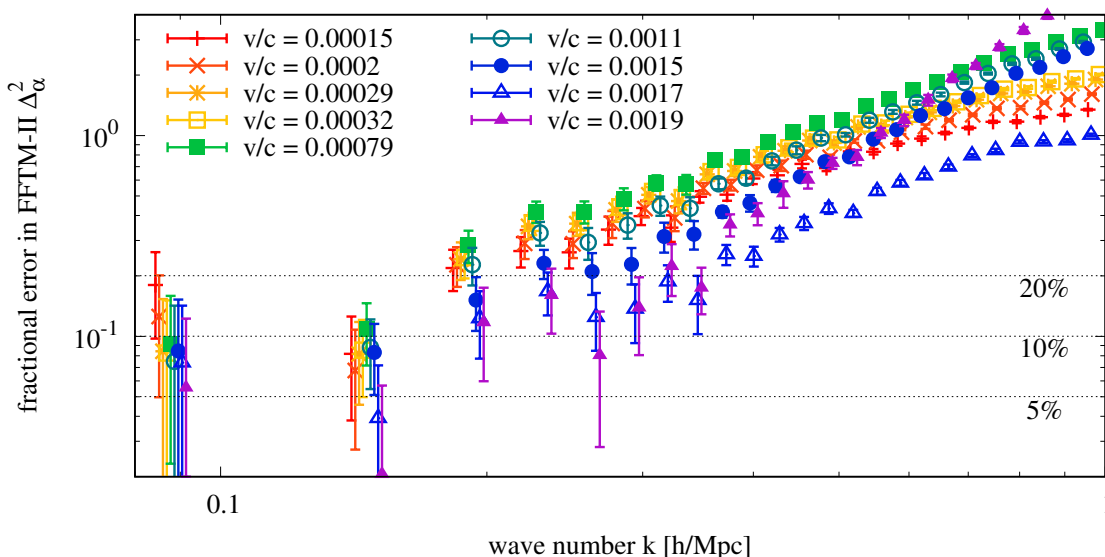


Figure 14. Errors the in **FlowsForTheMasses-II** power spectra of the individual EHDM flows for a range of flow velocities relative to hybrid N-body simulation. The simulation results have been binned, with each data point representing the average over 20 points per bin and the error bars the standard deviation. The full set of data points are collated from four HDM models described in section 4.4, with other cosmological parameters fixed to eq. (4.1). For visual clarity, we have applied a horizontal offset of up to a few percent to the data points.

axionic and bosonic models above, and two more for models that include only standard neutrinos as the HDM, with NO masses totally $M_\nu = 161$ meV and 315 meV, respectively. All four simulation runs and their corresponding **FlowsForTheMasses-II** runs use cosmological constant models with parameters given in eq. (4.1), as well as $N_{\text{GLQ}} = N_\tau = 15$ and $N_\mu = 10$.

Reference [92] showed for standard neutrino models that **FlowsForTheMasses** is accurate for flow velocities $v_\alpha/c \geq 0.0017 - 0.002$, that is, about 500 km/sec–600 km/sec. The four models considered here together have a total of nine flows with $v_\alpha/c \leq 0.002$: 0.00032 and 0.0017 for the axionic model; 0.0002 and 0.0011 for the bosonic model; 0.00029 and 0.0015 for the lighter neutrino model; and 0.00015, 0.00079, and 0.0019 for the heavier neutrino model. Figure 14 shows the fractional error in **FlowsForTheMasses-II** for each of these flows. Evidently, low- k error tends to decrease with rising flow velocity. At larger wave numbers, $0.1 \text{ h/Mpc} < k < 0.4 \text{ h/Mpc}$, flows with $v/c \leq 0.001$ have similar errors, while errors in the higher-velocity flows decrease with increasing v . Further, the two fastest flows shown, identified by filled and open triangles in the figure, have errors consistent with $\lesssim 10\%$ for $k \leq 0.2 \text{ h/Mpc}$ and $\lesssim 20\%$ up to $k \approx 0.35 \text{ h/Mpc}$. One of these two comes from the axion+neutrino model and the other one from a neutrino-only model. Thus we see that the guideline of ref. [92], that perturbation theory is adequate for flows with v/c larger than about 0.0017–0.002, holds even for very different HDM species.

Figure 15 uses $N_{\text{GLQ}} = 15$ flows to compare **FlowsForTheMasses-II** and hybrid N-body power spectra for the axionic and bosonic models above. In each case, the hybrid simulation converts into particles every flow with a velocity $v/c \leq 0.002$, i.e., the slowest two flows in each

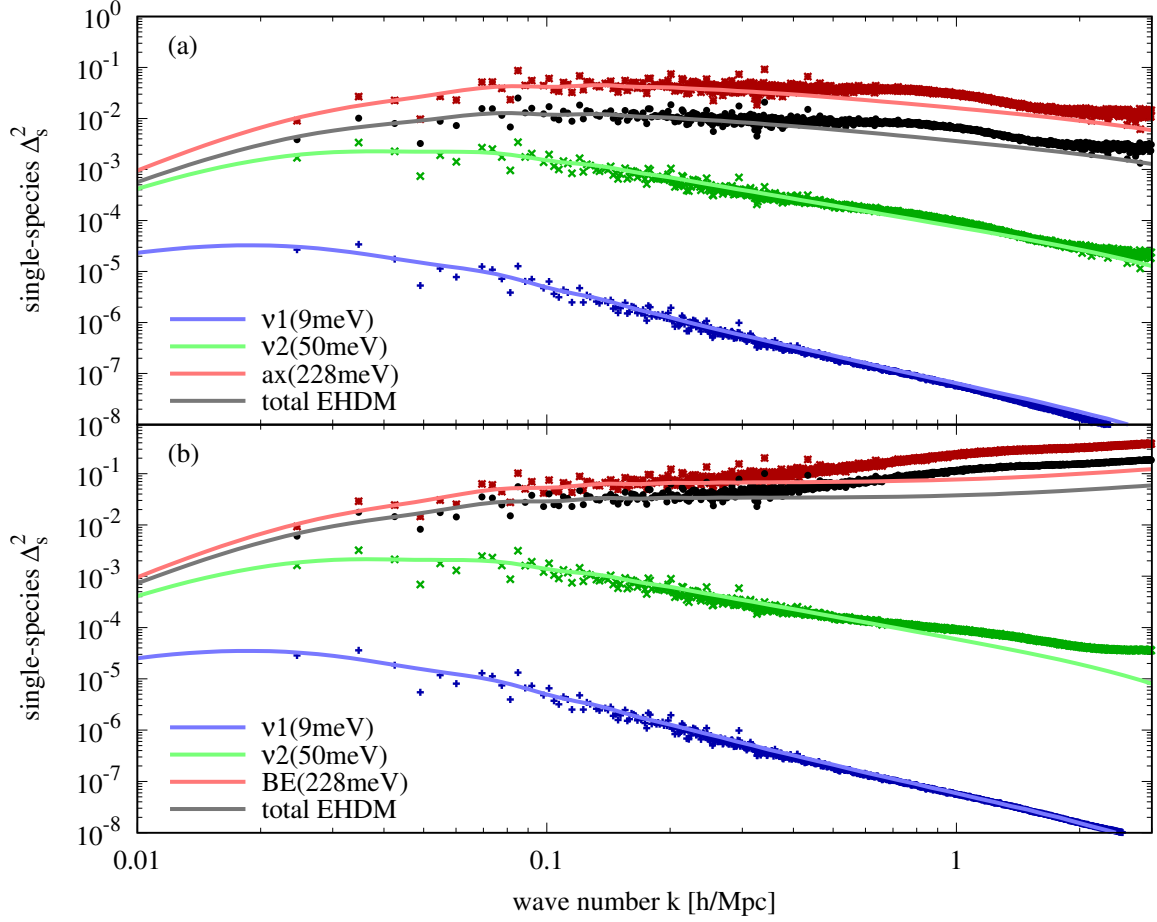


Figure 15. Power spectra at $z = 0$ for individual HDM species, determined from the GLQ flows as per eq. (3.11), for two HDM models. Each model assumes a cosmological constant and the parameters of eq. (4.1), with NO neutrino masses totally $M_\nu = 59$ meV plus a non-standard HDM: (a) axion+ ν model, where the axion distribution was computed in ref. [35], and (b) boson+ ν model, where the boson follows the relativistic Bose-Einstein distribution.

of these models, while the remaining flows are tracked using multi-fluid linear perturbation theory. **FlowsForTheMasses-II** agrees closely with N-body power spectra in all cases up to $k \lesssim 0.4 h/\text{Mpc}$, and for the 9 meV neutrinos to $k \approx 1 h/\text{Mpc}$. We have thus demonstrated that 15 flows are sufficient for predicting the clustering of the total EHDM as well as the component HDM species, even for species with very different distribution functions and an axion-to-neutrino mass ratio of 25.

5 Results II: evading cosmological neutrino bounds

5.1 Non-standard neutrinos and clustering suppression

Cosmological upper bounds on M_ν are considerably stronger than those from laboratory experiments, such as $M_\nu < 2400$ meV from KATRIN [93], but also more dependent upon our assumptions that the neutrinos have only Standard Model interactions and an approximately Fermi-Dirac distribution function. Moreover, the most stringent cosmological constraints

rely upon the assumption of Λ CDM cosmology. Recently, refs. [24, 25, 94] and others have proposed models that allow massive neutrinos to evade cosmological bounds but whose mass remains measurable in ongoing and near-future laboratory β -decay end-point experiments including KATRIN. Following ref. [25], we divide these models into two classes: “skewed- ν ” models, in which neutrinos’ distribution function is skewed away from the relativistic Fermi-Dirac distribution by a momentum-dependent factor so as to increase their mean momentum; and “cool- ν ” models, which lower the neutrinos’ temperature, hence their number density, in the early universe.

Reference [95] argues that cosmology chiefly constrains two properties of the relic neutrino background, its energy densities in the early and late universe, parameterized respectively as N_{eff} and $\Omega_{\nu,0}$. Thus, the goal of these alternative neutrino models is to preserve these two parameters while increasing M_ν into the range $600 \text{ meV} \lesssim M_\nu \lesssim 2400 \text{ meV}$ between the current laboratory bounds and the design sensitivity of ongoing experiments such as KATRIN. The argument that late-time cosmology only constrains $\Omega_{\nu,0}$ follows from the standard result that neutrinos making up a matter fraction $f_\nu = \Omega_{\nu,0}/\Omega_{\text{m},0}$ cause an $\sim 8f_\nu$ fractional suppression of the matter power spectrum on scales much smaller than their free-streaming scale. This is equivalent to a fractional suppression of δ_{m} by $4f_\nu$ and of δ_{cb} by $3f_\nu$. We begin this section by deriving this result in order to show its breadth as well as its limitations.

Applying eq. (2.5) to the CDM+baryon fluid (which has zero velocity) implies $-ak^2\Phi = \mathcal{H}[a\mathcal{H}\delta'_{\text{cb}}]'$. Working in the Einstein-de Sitter model and considering k sufficiently large that neutrino clustering can be neglected from eq. (2.6), we find $\delta_{\text{cb}} \propto a^{1-3f_\nu/5}$ for small f_ν , as compared with $\delta_{\text{cb}} \propto a$ in the massless-neutrino case. However, these are only valid while neutrinos are non-relativistic, $a \gtrsim a_{\text{nr}} = \bar{p}_{\nu,0}/m_\nu$, where the mean momentum $\bar{p}_{\nu,0} \approx 3.15T_{\nu,0}$ for the Fermi-Dirac distribution; before this time, neutrino masses have a negligible impact upon δ_{cb} . Thus the late-time suppression factor of δ_{cb} is $\approx (a_{\text{nr}}/a)^{3f_\nu/5}$. This corresponds to a fractional suppression $1 - (a_{\text{nr}}/a)^{3f_\nu/5} \approx 3f_\nu$ for $a = 1$ and $\Omega_{\nu,0}h^2 \approx 0.003$, hence a suppression of $8f_\nu$ for the matter power spectrum. The fractional suppression is weakly dependent upon $\Omega_{\nu,0}h^2$ and the scale factor; increasing $\Omega_{\nu,0}h^2$ to 0.01 at $a = 1$ leads to a suppression of $3.3f_\nu$ in δ_{cb} and hence $8.6f_\nu$ in the matter power spectrum, while $\Omega_{\nu,0}h^2 = 0.003$ and $a = 0.5$ gives a suppression of $2.6f_\nu$ and $7.2f_\nu$ in δ_{cb} and the matter power spectrum, respectively.

They key point here is that, because a_{nr} depends on the ratio $p_{\nu,0}/m_\nu$, increasing the neutrino momenta and masses by the same amount preserves this suppression due to neutrino free-streaming. Thus skewed- ν models should have approximately the same small-scale power suppression as the corresponding standard- ν models characterized by $T_{\nu,0} = 1.9525 \text{ K}$ and Fermi-Dirac distribution functions. However, cool- ν models, which seek to lower the neutrinos’ temperature in order to increase their masses, will lead to different small-scale matter power spectra, requiring a modification to the arguments of ref. [95].

We consider in the following the skewed- ν and cool- ν models in turn. Our benchmark observable is the CMB lensing potential power spectrum $C_L^{\phi\phi}$, which can probe matter clustering in the quasi-linear regime (corresponding to multipoles $1000 \lesssim L \lesssim 2000$) while remaining relatively free of systematic biases; see ref. [96] for a review of CMB lensing and ref. [97] regarding biases from baryonic effects. As a criterion for discerning between the two models, we compare the differences between their $C_L^{\phi\phi}$ to the sensitivity forecast for a

sample CMB Stage-4 survey [98]. In order to predict $C_L^{\phi\phi}$, we combine the `hyphi` code of ref. [99] with the `FlowsForTheMasses-II` neutrino treatment of section 3.

5.2 Skewed neutrino models

Skewed- ν models were considered in, e.g., refs. [24, 100]. Evidently from our discussion in section 3.1, the clustering of any collection of HDM species is determined by its velocity distribution. Thus, increasing both their masses and their momenta so as to preserve the velocity distribution will have no impact upon the HDM clustering. In other words, neutrinos could exceed cosmological bounds if their distribution function were skewed away from the relativistic Fermi-Dirac distribution and towards higher momenta. The drawback of skewed- ν models is that no obvious mechanism for such a skew is known.

We parameterize the skewed neutrino distribution function as $F_{\text{sk}}(q) = N_\sigma \sigma(q) F_{\text{FD}}(q)$. By demanding that the skewed- ν model reproduces the correct N_{eff} , that is, the skewed- ν and standard- ν models must have the same energy densities at early (pre-Big Bang Nucleosynthesis) times:

$$N_\sigma \int_0^\infty \frac{dq q^3 \sigma(q)}{e^q + 1} = \int_0^\infty \frac{dq q^3}{e^q + 1}, \quad (5.1)$$

we find a normalization factor

$$N_\sigma = \frac{I_3^{(\text{FD})}}{I_3^{(\text{sk})}}, \quad (5.2)$$

where

$$I_n^{(\text{FD})} = \int_0^\infty \frac{dq q^n}{e^q + 1}, \quad \text{and} \quad I_n^{(\text{sk})} = \int_0^\infty \frac{dq q^n \sigma(q)}{e^q + 1}. \quad (5.3)$$

Here, $I_n^{(\text{FD})}$ is the standard Fermi-Dirac integral. Next, we equate the models' late-time densities $\bar{\rho}_{\nu,0} \propto \Omega_{\nu,0}$ in the non-relativistic approximation,

$$M_\nu^{(\text{skew})} N_\sigma \int_0^\infty dq \frac{q^2 \sigma(q)}{e^q + 1} = M_\nu \int_0^\infty dq \frac{q^2}{e^q + 1}, \quad (5.4)$$

which leads to

$$\frac{M_\nu}{M_\nu^{(\text{skew})}} = \frac{I_2^{(\text{sk})}}{I_2^{(\text{FD})}} \frac{I_3^{(\text{FD})}}{I_3^{(\text{sk})}}. \quad (5.5)$$

On the other hand, the mean momentum is readily seen to be $a^{-1} T_{\nu,0} I_3^{(\text{FD})} / I_2^{(\text{FD})} \approx 3.15 T_{\nu,0} / a$ for standard neutrinos and $a^{-1} T_{\nu,0} I_3^{(\text{sk})} / I_2^{(\text{sk})}$ for skewed neutrinos. Then, by comparison with eq. (5.5), we see immediately that the ratios of mean momentum to neutrino mass are the same in both models. Thus, skewed- ν models must also cause a small-scale fractional suppression of $\sim 8f_\nu$ in the late-time matter power spectrum.

As a phenomenological example, we consider a neutrino distribution function $F_{\text{sk}}(q) = N_{n_{\text{sk}}} q^{n_{\text{sk}}} F_{\text{FD}}(q)$, where $N_{n_{\text{sk}}}$ is the normalization constant, and a larger $n_{\text{sk}} > 0$ implies a larger mean momentum. We assume that the normalization is fixed in the early universe so as to preserve $N_{\text{eff}} = 3.044$. This means that increasing n_{sk} reduces the neutrinos'

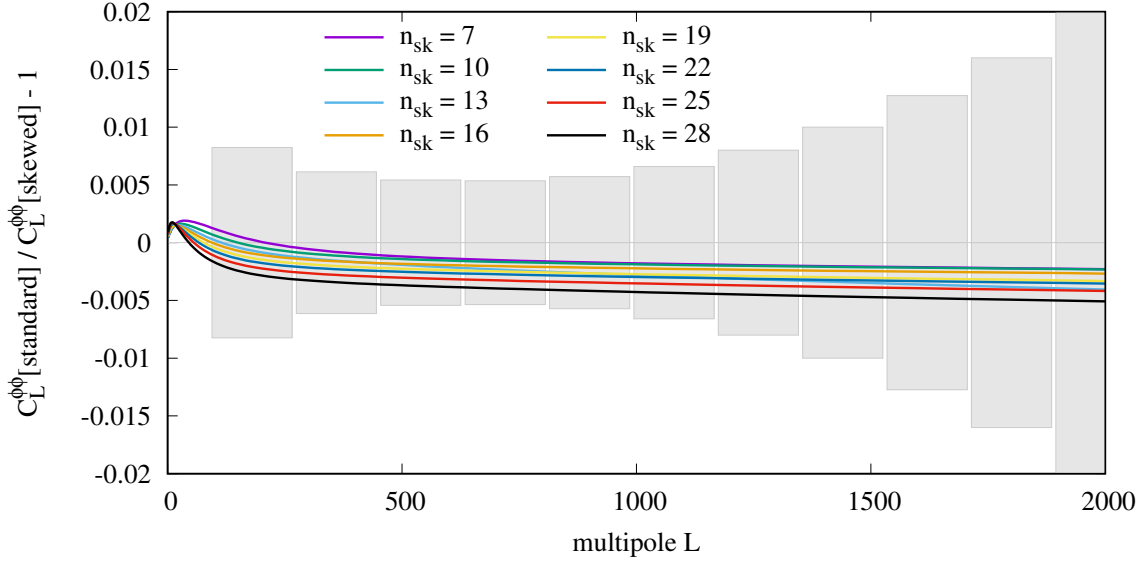


Figure 16. Fractional difference in the CMB lensing potential power spectra between skewed- ν models with distribution functions $\propto p^{n_{\text{sk}}} F_{\text{FD}}(p)$ and $M_\nu^{(\text{skew})} = 600$ meV, and standard- ν models described by a relativistic Fermi-Dirac distribution $F_{\text{FD}}(p)$. The skewed- ν distribution functions have been normalized to ensure $N_{\text{eff}} = 3.044$. For each n_{sk} , the corresponding standard- ν mass M_ν has been chosen such that it gives an energy density $\Omega_{\nu,0}$ matching its skewed- ν counterpart. That is, $\Omega_{\nu,0}$ decreases with rising n_{sk} . The shaded error bands correspond to the forecasted CMB Stage-4 sensitivities taken from ref. [98].

number density, such that fixing $\Omega_{\nu,0}$ would require a higher M_ν . Alternatively, we could fix $M_\nu^{(\text{skew})} = 600$ meV while increasing n_{sk} : doing so would allow the skewed- ν model to mimic a standard neutrino model with smaller M_ν ; indeed, we find that a skewed model with $n_{\text{sk}} = 13$ ($n_{\text{sk}} = 28$) has the same $\Omega_{\nu,0}$ as standard neutrinos with $M_\nu = 118$ meV (61 meV). Although this example does not exactly preserve the velocity distribution function, we will see that it is indistinguishably close to the standard neutrino case for near-future cosmological surveys.

Figure 16 shows our results with fixed $M_\nu^{(\text{skew})} = 600$ meV. Each skewed- ν model is compared with its standard- ν counterpart, with matching N_{eff} and $\Omega_{\nu,0}$. The fractional difference between the CMB lensing potential power spectra in each pair fits comfortably within the binned error bands forecast for the proposed CMB Stage-4 survey in the CMB-S4 Science Book [98]. Thus, as expected, raising the neutrino masses and momenta in tandem does not appreciably affect neutrinos' clustering properties. If a theoretically-suitable, experimentally-falsifiable mechanism could be found for introducing such a distortion to the neutrinos' distribution function, then skewed neutrino models would be viable ways to evade cosmological neutrino mass bounds.

5.3 Cool-neutrino models

Contrary to skewed- ν models, cool- ν models increase the neutrinos' masses while decreasing their mean momenta, both of which have the effect of reducing the neutrinos' velocities, thereby amplifying both their small-scale power suppression and their non-linear clustering

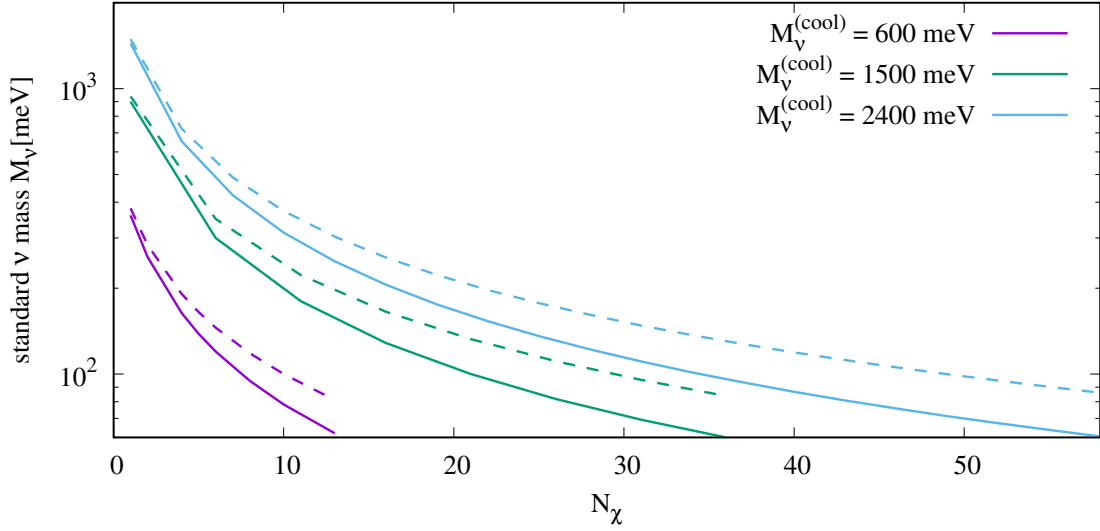


Figure 17. Neutrino mass sums M_ν in standard- ν models required to match the late-time energy density $\Omega_{\nu,0}$ (solid) and the small-scale matter power suppression (dashed) of cool- ν models with $M_\nu^{(\text{cool})} = 600$ meV (purple), 1500 meV (green), and 2400 meV (blue). The maximum N_χ shown is consistent with the equal- $\Omega_{\nu,0}$ mass sum M_ν being at least 60 meV.

while shifting their free-streaming wave numbers k_{FS} to larger values. We consider here the cool- ν model of ref. [25], where neutrinos thermalize with N_χ massless sterile particles χ via a massive mediator at a time after weak decoupling. The resulting neutrinos have a relativistic Fermi-Dirac distribution function. However, equipartition amongst the thermalized species leads to a neutrino temperature that is lower by a factor of $(1 + 2N_\chi/3)^{-1/3}$ compared with Standard Model neutrinos; the remainder of N_{eff} is made up by the massless particles. These cooler neutrinos have a correspondingly reduced number density, meaning that their masses must be larger than those of standard neutrinos for the same $\Omega_{\nu,0}$ value.

For example, a factor-of-two cooling in the neutrinos, corresponding to $N_\chi \approx 11$, implies an eightfold reduction in their number density, allowing for a corresponding eightfold increase in the sum of their masses, $M_\nu^{(\text{cool})}$. Since the neutrino free-streaming wave number $k_{\text{FS}} \propto m/T$, both the reduced temperature and the increased mass raise k_{FS} , pushing the neutrino suppression to smaller scales. In this particular example of $N_\chi \approx 11$, k_{FS} increases by a factor of sixteen.

We estimated in section 5.1 that the late-time fractional suppression in δ_{cb} by massive neutrinos is $\approx (a_{\text{nr}}/a)^{3f_\nu/5}$. Preserving f_ν while reducing $T_{\nu,0}$ and increasing m_ν will therefore reduce a_{nr} , increasing the suppression. Thus a cool- ν model with the same N_{eff} and $\Omega_{\nu,0}$ as a standard-temperature model will nevertheless have a smaller free-streaming scale and a greater suppression of the linear δ_{cb} and hence the linear matter power spectrum. Continuing with our $N_\chi \approx 11$ example of a halving of $T_{\nu,0}$ and an eightfold increase in $M_\nu^{(\text{cool})}$ relative to standard-temperature models, we find for $\Omega_{\nu,0}h^2 \approx 0.003$ and $a = 1$ that δ_{cb} and the matter power spectrum are suppressed at small scales by $\approx 4.5f_\nu$ and $\approx 11f_\nu$, respectively.

Since no single $M_\nu^{(\text{cool})}$ value in a cool- ν model can simultaneously match the late-time density parameter $\Omega_{\nu,0}$ and the small-scale matter power spectrum of a standard- ν model, we next ask which of these two choices of late-time phenomenology is the better approximation to cosmological constraints. We compare cool- ν models with fixed $M_\nu^{(\text{cool})}$ against two different standard- ν analogs: one with an equal $\Omega_{\nu,0}$, and another with an equal small-scale suppression. All models have the same N_{eff} by design. In the equal- $\Omega_{\nu,0}$ case, assuming N_χ massless sterile particles, the standard- ν model has $M_\nu = M_\nu^{(\text{cool})}/(1 + 2N_\chi/3)$.

In the equal-suppression case, since our suppression formula above is approximate, we match the $z = 0$ linear matter power suppression computed using CLASS at $k = 10 \text{ h/Mpc}$ to a fractional precision $< 10^{-5}$ by adjusting M_ν . We assume $\nu\Lambda\text{CDM}$ models with parameters

$$\Omega_{\text{m},0}h^2 = 0.14; \quad \Omega_{\text{b},0}h^2 = 0.022; \quad A_s = 2.2 \times 10^{-9}; \quad n_s = 0.96; \quad h = 0.67. \quad (5.6)$$

We further assume NO masses in the standard- ν case. Since we are interested in $M_\nu^{(\text{cool})} \geq 600 \text{ meV}$, we make the DO mass approximation for cool- ν models. Figure 17 shows equal-density and equal-suppression masses as functions of N_χ for $600 \text{ meV} \leq M_\nu^{(\text{cool})} \leq 2400 \text{ meV}$.

Implementing cool- ν models in **FlowsForTheMasses-II** presents new challenges. We noted in section 4.1 that reducing $N_{\mu,AI}^{(0)}$ was insufficient to stabilize the perturbation theory for $M_\nu \geq 600 \text{ meV}$. Lowering the neutrino temperature, hence the momentum of each flow, only exacerbates the problem. However, our focus in this section on the CMB lensing potential power spectrum mitigates this difficulty in a number of ways.

- (a) Since we consider the range $L \leq 2000$, we only need an accurate computation of the power in the $k \lesssim 0.5 \text{ h/Mpc}$ range, so we lower our minimum stability threshold from 1.2 h/Mpc to 1 h/Mpc .
- (b) The lensing kernel peaks around $z = 2$, where both **FlowsForTheMasses-II** for the neutrinos and Time-RG for the cb fluid are more accurate, as show in figure 11 (Left) and ref. [60].
- (c) The lensing potential is sourced by the total matter density contrast. In the cosmologically-viable models considered here, where neutrinos make up at most a few percent of the matter density, even a 10% – 20% error in the neutrinos is consistent with percent-level accuracy of $C_L^{\phi\phi}$.
- (d) As evident from figure 11 (Right) for **FlowsForTheMasses-II** as well as ref. [60] for Time-RG, our treatment is much more accurate at predicting small fractional corrections to the power spectrum than the power spectrum itself. Thus we focus here on fractional corrections to the matter and CMB lensing potential power spectra.

These accuracy requirements are reasonably forgiving. For cool- ν models with $M_\nu = 600 \text{ meV}$, we are able to reach the lower stability threshold of 1 h/Mpc for $N_{\text{GLQ}} = N_\tau = 20$ by lowering $N_{\mu,AI}^{(0)}$ and $N_{\mu,AI}^{(1)}$ to one, leaving all others at their maximal values. For $M_\nu \geq 1500 \text{ meV}$, following figure 1, we reduce N_{GLQ} and N_τ to 5 with the aim of 10% accuracy in the neutrino power spectrum. Further reducing $N_{\mu,AI}^{(0)}$, $N_{\mu,AI}^{(1)}$, and $N_{\mu,AI}^{(2)}$ to one is sufficient for stabilizing the perturbation theory up to $M_\nu = 2400 \text{ meV}$. In the case of $M_\nu = 1500 \text{ meV}$

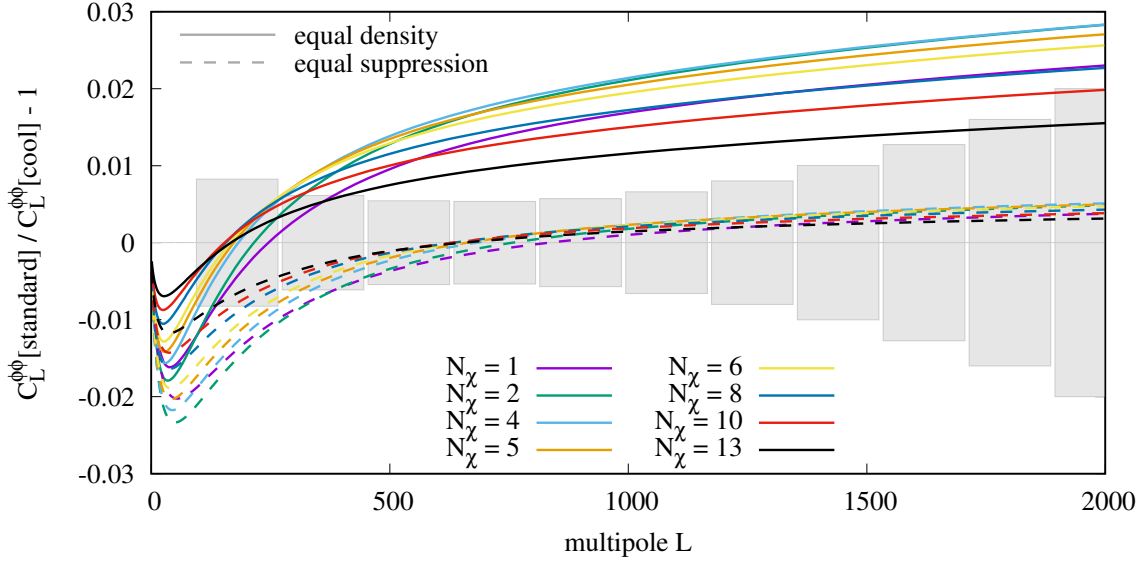


Figure 18. Fractional differences in the CMB lensing potential power spectrum $C_L^{\phi\phi}$ between standard- ν and cool- ν models with the cosmological parameters of eq. (5.6). We fix $M_\nu^{(\text{cool})} = 600$ meV, while varying the number N_χ of sterile species. Solid and dashed lines represent, respectively, standard- ν models matched to the same $\Omega_{\nu,0}$ and the same small-scale suppression as the corresponding cool- ν model. The shaded error bands denote the forecasted CMB Stage-4 sensitivities taken from ref. [98].

and $N_\chi = 36$, where $(N_{\mu, AI}^{(0)}, N_{\mu, AI}^{(1)}, N_{\mu, AI}^{(2)})$ of $(1, 1, 1)$, $(1, 1, 11)$, and $(1, 11, 11)$ are all stable for $k \leq 0.5$ h/Mpc , we find $< 2\%$ agreement between the neutrino power spectra over that range, and $< 0.06\%$ in $C_L^{\phi\phi}$ up to $L = 2000$. Thus, following figure 11 (Right), we are confident in our predictions of the neutrino-dependence of $C_L^{\phi\phi}$ at the few percent level. For the $z = 0$ matter power spectrum, ref. [60] implies a 15% accurate computation of the sensitivity to small model changes up to $k = 0.5$ h/Mpc , and 26% to $k = 1$ h/Mpc .

Figure 18 compares the CMB lensing potential power spectra of equal-density and equal-suppression standard- ν models, to the corresponding cool- ν model of a fixed $M_\nu^{(\text{cool})} = 600$ meV and various choices of $N_\chi \leq 13$. Increasing N_χ above 13 will imply equal-density M_ν below the lower bound from neutrino oscillation experiments, so we exclude these from consideration. For $N_\chi = 6$ ($N_\chi = 13$), the equal-density standard- ν model has $M_\nu = 120$ meV ($M_\nu = 62$ meV), while the corresponding equal-suppression masses are $\approx 21\%$ ($\approx 31\%$) larger. Evidently, equal-density standard- ν models will be distinguishable from the corresponding cool- ν models to a high significance using CMB Stage-4 data, even for the smallest masses $M_\nu^{(\text{cool})} = 600$ meV accessible to ongoing terrestrial experiments. Thus, from a phenomenological viewpoint, the statement that late-time cosmological observable depends only on the HDM density $\Omega_{\nu,0}$ is not correct, as demonstrated here by the cool- ν models.

On the other hand, as shown in figure 18, equal-suppression standard- ν and cool- ν models exhibit much the same CMB lensing potential power spectrum over a large range of multipoles L . In the case of $N_\chi \geq 10$, the equal-suppression standard- ν model lies within the error bars over the entire L range considered. Thus the HDM property actually constrained by the late-time cosmological data is closer to the small-scale matter power spectrum suppression than

the background HDM density. While this argument leaves intact the most important point made by ref. [25], namely, the ability of cool- ν models to evade neutrino mass constraints from observational cosmology, it also suggests new approaches by which cool- ν models may be excluded in the future. The key point is that the precise M_ν characterizing an equal-suppression model depends upon the observable as well as the scale factor. Specifically:

1. The linear suppression factor is weakly dependent upon the scale factor, as noted in section 5.1, with a threefold change in the scale factor leading to a $\approx 10\%$ change in the linear suppression fraction.
2. Cool- ν models, by virtue of having larger masses and lower temperatures than equal-suppression standard- ν models, have larger non-linear HDM clustering, itself depending rapidly upon the scale factor.
3. The cool- ν and equal-suppression standard- ν models have different f_ν , so the ratios of small-scale linear matter-to-cb suppression factors $\lim_{k \rightarrow \infty} P_m(k)/P_{cb}(k) = (1 - f_\nu)^2$ are also different.

CMB lensing and tomographic shear surveys probe the matter power spectrum at different redshifts, while galaxy clustering surveys trace the cb power spectrum. Thus a combination of all three may be able to eliminate or severely constrain cool- ν models designed to evade cosmological bounds. Furthermore, a large-volume survey may be able to constrain the difference between the free-streaming scales of cool- ν and standard- ν models. We leave forecasts of such joint constraints to future work.

Lastly, we elaborate upon item 2 in the list above, the non-linear contribution to HDM clustering. Figure 19 quantifies the fractional contribution of **FlowsForTheMasses-II** non-linear neutrino corrections to the matter power spectrum and the CMB lensing potential power spectrum. We consider cool- ν models with $M_\nu^{(\text{cool})}$ of 600 meV, 1500 meV, and 2400 meV, spanning the range of interest between the current terrestrial constraints of KATRIN [93] and its ultimate design sensitivity. Adopting a conservative cosmological upper bound on M_ν of about 380 meV [1],⁵ we see that in order for $M_\nu^{(\text{cool})} = 2400$ meV to stay within observational constraints, a N_χ of at least 10 would be required. In this case, the bottom panels of figure 19 show that the non-linear HDM correction to $C_L^{\phi\phi}$ is 0.16% at $L = 1000$ and 0.24% at $L = 2000$, while the small-scale $z = 0$ matter power correction is 0.52%. While even the $C_L^{\phi\phi}$ correction is a sizable fraction of the error bars and cannot be neglected, the rapid late-time increase of this correction could prove useful constraining it.

6 Conclusions

We have developed and derived a procedure for representing an arbitrary collection of HDM species, of any masses, temperatures, and distribution functions, from the relativistic to the non-relativistic regime, using a single EHDM species with an appropriately-chosen distribution

⁵Such a conservative bound could come from the simultaneous variation of the dark energy equation of state, its derivative, and the galaxy bias or Halo Occupation Distribution parameters, in the data analysis. See, e.g., ref. [7] for details.

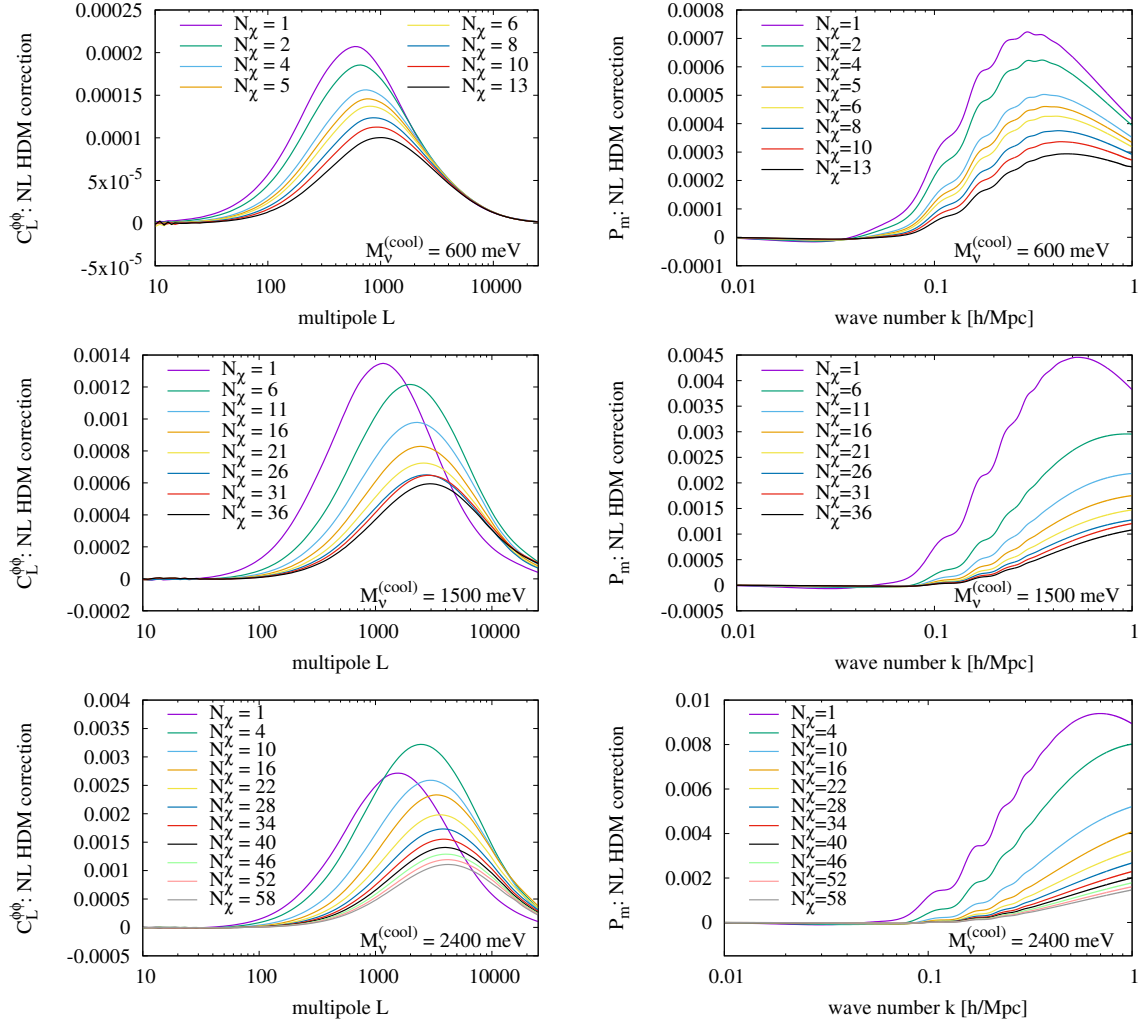


Figure 19. Fractional corrections in cosmological observables from neutrino non-linearity in cool- ν models. *Left:* corrections to the CMB lensing potential power spectrum $C_L^{\phi\phi}$ for $M_\nu^{(\text{cool})} = 600$ meV (top), 1500 meV (middle), and 2400 meV (bottom) for various choices of N_χ . *Right:* corrections to the $z = 0$ matter power spectrum $P_m(k)$ for the same set of $M_\nu^{(\text{cool})}$ and N_χ . The corresponding masses M_ν in standard- ν models, matching either the late-time density $\Omega_{\nu,0}$ or small-scale suppression of the cool- ν models, are shown in figure 17.

function. As this method follows directly from the full collisionless Boltzmann equation, it makes no assumption about the nature of the inhomogeneities in any distribution function and is equally applicable to linear and non-linear perturbation theory, as well as to N-body simulations. In this work, we have implemented this EHDM method in both linear (**MuFLR**) and non-linear (**FlowsForTheMasses-II**) multi-flow perturbation theories, which discretize the EHDM distribution into ~ 10 uniform-momentum flows. Furthermore, since terrestrial experiments are typically sensitive to only a single HDM species such as the electron neutrino, we have shown how an appropriate linear combination of these flows allows us to recover the power spectrum of each component HDM species that makes up the EHDM.

As cosmology enters the HDM era, perturbation theory is emerging as an indispensable tool, both for testing approximations within the standard neutrino picture and for exploring models well beyond it. Within the standard picture, we have shown that a more efficient choice of flow momenta improves the small-scale accuracy of `FlowsForTheMasses-II` at low M_ν by more than a factor of two relative to its predecessor, and we have quantified the differences among the normal, inverted, and degenerate neutrino mass orderings for a range of M_ν . Beyond standard neutrinos, we have considered mixed-HDM models that incorporate either a thermal QCD axion or a generic thermal boson in addition to a minimally-massive neutrino sector. We have also studied two attempts at evading cosmological neutrino bounds, either by skewing the neutrinos' distribution function or by reducing their temperature, and we have shown that the latter of these modifies the linear and non-linear clustering of neutrinos in a manner that could allow it to be distinguished from standard-temperature neutrinos using upcoming data from CMB Stage-4 experiments. In doing so, we have demonstrated `FlowsForTheMasses-II` to be an invaluable tool for studying light neutrinos as well as rapidly exploring the non-standard HDM parameter space.

Acknowledgments

We are grateful to J. Kwan, Y. Li, I. G. McCarthy, A. Notari, and T. Tan for insightful conversations. This research was undertaken using the HPC systems *Gadi* from the National Computational Infrastructure (NCI) supported by the Australian Government, and *Katana* at UNSW Sydney. AU acknowledges the support of the NSFC under grant No. 12473005. MM acknowledges support from the DFG grant LE 3742/8-1. GP and Y³W are supported by the Australian Government through the Australian Research Council's Future Fellowship (FT180100031) and Discovery Project (DP240103130) schemes.

References

- [1] PLANCK collaboration, *Planck 2018 results. VI. Cosmological parameters*, *Astron. Astrophys.* **641** (2020) A6 [Erratum *ibid.* **652** (2021) C4] [[arXiv:1807.06209](#)] [[INSPIRE](#)].
- [2] DESI collaboration, *DESI 2024 VI: cosmological constraints from the measurements of baryon acoustic oscillations*, [arXiv:2404.03002](#) [[INSPIRE](#)].
- [3] F. Capozzi et al., *Global constraints on absolute neutrino masses and their ordering*, *Phys. Rev. D* **95** (2017) 096014 [Addendum *ibid.* **101** (2020) 116013] [[arXiv:2003.08511](#)] [[INSPIRE](#)].
- [4] S. Gariazzo et al., *Neutrino masses and their ordering: global data, priors and models*, *JCAP* **03** (2018) 011 [[arXiv:1801.04946](#)] [[INSPIRE](#)].
- [5] P.F. De Salas et al., *Neutrino mass ordering from oscillations and beyond: 2018 status and future prospects*, *Front. Astron. Space Sci.* **5** (2018) 36 [[arXiv:1806.11051](#)] [[INSPIRE](#)].
- [6] I. Esteban et al., *The fate of hints: updated global analysis of three-flavor neutrino oscillations*, *JHEP* **09** (2020) 178 [[arXiv:2007.14792](#)] [[INSPIRE](#)].
- [7] A. Upadhye, *Neutrino mass and dark energy constraints from redshift-space distortions*, *JCAP* **05** (2019) 041 [[arXiv:1707.09354](#)] [[INSPIRE](#)].
- [8] H. Shao et al., *Cosmological limits on the neutrino mass sum for beyond- Λ CDM models*, [arXiv:2409.02295](#) [[INSPIRE](#)].

- [9] S. Roy Choudhury and T. Okumura, *Updated cosmological constraints in extended parameter space with Planck PR4, DESI baryon acoustic oscillations, and supernovae: dynamical dark energy, neutrino masses, lensing anomaly, and the Hubble tension*, *Astrophys. J. Lett.* **976** (2024) L11 [[arXiv:2409.13022](#)] [[INSPIRE](#)].
- [10] W.L. Freedman et al., *Status report on the Chicago-Carnegie Hubble Program (CCHP): three independent astrophysical determinations of the Hubble constant using the James Webb space telescope*, [arXiv:2408.06153](#) [[INSPIRE](#)].
- [11] D. Pedrotti et al., *Multidimensionality of the Hubble tension: the roles of Ω_m and ω_c* , *Phys. Rev. D* **111** (2025) 023506 [[arXiv:2408.04530](#)] [[INSPIRE](#)].
- [12] A.G. Riess et al., *JWST observations reject unrecognized crowding of Cepheid photometry as an explanation for the Hubble tension at 8σ confidence*, *Astrophys. J. Lett.* **962** (2024) L17 [[arXiv:2401.04773](#)] [[INSPIRE](#)].
- [13] A.R. Khalife et al., *Review of Hubble tension solutions with new SH0ES and SPT-3G data*, *JCAP* **04** (2024) 059 [[arXiv:2312.09814](#)] [[INSPIRE](#)].
- [14] A. Leauthaud et al., *Lensing is low: cosmology, galaxy formation, or new physics?*, *Mon. Not. Roy. Astron. Soc.* **467** (2017) 3024 [[arXiv:1611.08606](#)] [[INSPIRE](#)].
- [15] V. Poulin, J.L. Bernal, E.D. Kovetz and M. Kamionkowski, *Sigma-8 tension is a drag*, *Phys. Rev. D* **107** (2023) 123538 [[arXiv:2209.06217](#)] [[INSPIRE](#)].
- [16] A. Amon and G. Efstathiou, *A non-linear solution to the S_8 tension?*, *Mon. Not. Roy. Astron. Soc.* **516** (2022) 5355 [[arXiv:2206.11794](#)] [[INSPIRE](#)].
- [17] I.G. McCarthy et al., *The FLAMINGO project: revisiting the S_8 tension and the role of baryonic physics*, *Mon. Not. Roy. Astron. Soc.* **526** (2023) 5494 [[arXiv:2309.07959](#)] [[INSPIRE](#)].
- [18] PLANCK collaboration, *Planck 2018 results. VIII. Gravitational lensing*, *Astron. Astrophys.* **641** (2020) A8 [[arXiv:1807.06210](#)] [[INSPIRE](#)].
- [19] N. Craig, D. Green, J. Meyers and S. Rajendran, *No νs is good news*, *JHEP* **09** (2024) 097 [[arXiv:2405.00836](#)] [[INSPIRE](#)].
- [20] D. Green and J. Meyers, *The cosmological preference for negative neutrino mass*, [arXiv:2407.07878](#) [[INSPIRE](#)].
- [21] W. Elbers et al., *Negative neutrino masses as a mirage of dark energy*, [arXiv:2407.10965](#) [[INSPIRE](#)].
- [22] J.-Q. Jiang et al., *Neutrino cosmology after DESI: tightest mass upper limits, preference for the normal ordering, and tension with terrestrial observations*, [arXiv:2407.18047](#) [[INSPIRE](#)].
- [23] M. Loverde and Z.J. Weiner, *Massive neutrinos and cosmic composition*, *JCAP* **12** (2024) 048 [[arXiv:2410.00090](#)] [[INSPIRE](#)].
- [24] I.M. Oldengott et al., *How to relax the cosmological neutrino mass bound*, *JCAP* **04** (2019) 049 [[arXiv:1901.04352](#)] [[INSPIRE](#)].
- [25] M. Escudero, T. Schwetz and J. Terol-Calvo, *A seesaw model for large neutrino masses in concordance with cosmology*, *JHEP* **02** (2023) 142 [Addendum *ibid.* **06** (2024) 119] [[arXiv:2211.01729](#)] [[INSPIRE](#)].
- [26] MICROBOONE collaboration, *First constraints on light sterile neutrino oscillations from combined appearance and disappearance searches with the MicroBooNE detector*, *Phys. Rev. Lett.* **130** (2023) 011801 [[arXiv:2210.10216](#)] [[INSPIRE](#)].

- [27] A.B. Balantekin, G.M. Fuller, A. Ray and A.M. Suliga, *Probing self-interacting sterile neutrino dark matter with the diffuse supernova neutrino background*, *Phys. Rev. D* **108** (2023) 123011 [[arXiv:2310.07145](#)] [[INSPIRE](#)].
- [28] R.D. Peccei and H.R. Quinn, *CP conservation in the presence of instantons*, *Phys. Rev. Lett.* **38** (1977) 1440 [[INSPIRE](#)].
- [29] R.D. Peccei and H.R. Quinn, *Constraints imposed by CP conservation in the presence of instantons*, *Phys. Rev. D* **16** (1977) 1791 [[INSPIRE](#)].
- [30] S. Hannestad, A. Mirizzi, G.G. Raffelt and Y.Y.Y. Wong, *Neutrino and axion hot dark matter bounds after WMAP-7*, *JCAP* **08** (2010) 001 [[arXiv:1004.0695](#)] [[INSPIRE](#)].
- [31] M. Archidiacono et al., *Axion hot dark matter bounds after Planck*, *JCAP* **10** (2013) 020 [[arXiv:1307.0615](#)] [[INSPIRE](#)].
- [32] R.Z. Ferreira and A. Notari, *Observable windows for the QCD axion through the number of relativistic species*, *Phys. Rev. Lett.* **120** (2018) 191301 [[arXiv:1801.06090](#)] [[INSPIRE](#)].
- [33] F. D’Eramo, F. Hajkarim and S. Yun, *Thermal axion production at low temperatures: a smooth treatment of the QCD phase transition*, *Phys. Rev. Lett.* **128** (2022) 152001 [[arXiv:2108.04259](#)] [[INSPIRE](#)].
- [34] F. D’Eramo, F. Hajkarim and S. Yun, *Thermal QCD axions across thresholds*, *JHEP* **10** (2021) 224 [[arXiv:2108.05371](#)] [[INSPIRE](#)].
- [35] A. Notari, F. Rompineve and G. Villadoro, *Improved hot dark matter bound on the QCD axion*, *Phys. Rev. Lett.* **131** (2023) 011004 [[arXiv:2211.03799](#)] [[INSPIRE](#)].
- [36] F. D’Eramo et al., *Cosmological bound on the QCD axion mass, redux*, *JCAP* **09** (2022) 022 [[arXiv:2205.07849](#)] [[INSPIRE](#)].
- [37] L. Caloni, M. Gerbino, M. Lattanzi and L. Visinelli, *Novel cosmological bounds on thermally-produced axion-like particles*, *JCAP* **09** (2022) 021 [[arXiv:2205.01637](#)] [[INSPIRE](#)].
- [38] F. Bianchini, G.G. di Cortona and M. Valli, *QCD axion: some like it hot*, *Phys. Rev. D* **110** (2024) 123527 [[arXiv:2310.08169](#)] [[INSPIRE](#)].
- [39] K. Bouzoud and J. Ghiglieri, *Thermal axion production at hard and soft momenta*, [arXiv:2404.06113](#) [[INSPIRE](#)].
- [40] M. LoVerde, *Halo bias in mixed dark matter cosmologies*, *Phys. Rev. D* **90** (2014) 083530 [[arXiv:1405.4855](#)] [[INSPIRE](#)].
- [41] C.-T. Chiang, M. LoVerde and F. Villaescusa-Navarro, *First detection of scale-dependent linear halo bias in N -body simulations with massive neutrinos*, *Phys. Rev. Lett.* **122** (2019) 041302 [[arXiv:1811.12412](#)] [[INSPIRE](#)].
- [42] H.-R. Yu et al., *Differential neutrino condensation onto cosmic structure*, *Nature Astron.* **1** (2017) 0143 [[arXiv:1609.08968](#)] [[INSPIRE](#)].
- [43] H.-M. Zhu, U.-L. Pen, X. Chen and D. Inman, *Probing neutrino hierarchy and chirality via wakes*, *Phys. Rev. Lett.* **116** (2016) 141301 [[arXiv:1412.1660](#)] [[INSPIRE](#)].
- [44] A.E. Bayer, A. Banerjee and Y. Feng, *A fast particle-mesh simulation of non-linear cosmological structure formation with massive neutrinos*, *JCAP* **01** (2021) 016 [[arXiv:2007.13394](#)] [[INSPIRE](#)].
- [45] J.Z. Chen, A. Upadhye and Y.Y.Y. Wong, *Flows for the masses: a multi-fluid non-linear perturbation theory for massive neutrinos*, *JCAP* **05** (2023) 046 [[arXiv:2210.16020](#)] [[INSPIRE](#)].

- [46] A. Upadhye et al., *Cosmic-Ev: an emulator for the non-linear neutrino power spectrum*, *Mon. Not. Roy. Astron. Soc.* **530** (2024) 743 [[arXiv:2311.11240](#)] [[INSPIRE](#)].
- [47] J.J. Bennett, G. Buldgen, M. Drewes and Y.Y.Y. Wong, *Towards a precision calculation of the effective number of neutrinos N_{eff} in the Standard Model. Part I. The QED equation of state*, *JCAP* **03** (2020) 003 [Addendum *ibid.* **03** (2021) A01] [[arXiv:1911.04504](#)] [[INSPIRE](#)].
- [48] K. Akita and M. Yamaguchi, *A precision calculation of relic neutrino decoupling*, *JCAP* **08** (2020) 012 [[arXiv:2005.07047](#)] [[INSPIRE](#)].
- [49] J. Froustey, C. Pitrou and M.C. Volpe, *Neutrino decoupling including flavour oscillations and primordial nucleosynthesis*, *JCAP* **12** (2020) 015 [[arXiv:2008.01074](#)] [[INSPIRE](#)].
- [50] J.J. Bennett et al., *Towards a precision calculation of N_{eff} in the Standard Model. Part II. Neutrino decoupling in the presence of flavour oscillations and finite-temperature QED*, *JCAP* **04** (2021) 073 [[arXiv:2012.02726](#)] [[INSPIRE](#)].
- [51] M. Drewes et al., *Towards a precision calculation of N_{eff} in the Standard Model. Part III. Improved estimate of NLO contributions to the collision integral*, *JCAP* **06** (2024) 032 [[arXiv:2402.18481](#)] [[INSPIRE](#)].
- [52] A. Ringwald and Y.Y.Y. Wong, *Gravitational clustering of relic neutrinos and implications for their detection*, *JCAP* **12** (2004) 005 [[hep-ph/0408241](#)] [[INSPIRE](#)].
- [53] Y.Y.Y. Wong, *Higher order corrections to the large scale matter power spectrum in the presence of massive neutrinos*, *JCAP* **10** (2008) 035 [[arXiv:0809.0693](#)] [[INSPIRE](#)].
- [54] H. Dupuy and F. Bernardeau, *Describing massive neutrinos in cosmology as a collection of independent flows*, *JCAP* **01** (2014) 030 [[arXiv:1311.5487](#)] [[INSPIRE](#)].
- [55] H. Dupuy and F. Bernardeau, *Cosmological perturbation theory for streams of relativistic particles*, *JCAP* **03** (2015) 030 [[arXiv:1411.0428](#)] [[INSPIRE](#)].
- [56] H. Dupuy and F. Bernardeau, *On the importance of nonlinear couplings in large-scale neutrino streams*, *JCAP* **08** (2015) 053 [[arXiv:1503.05707](#)] [[INSPIRE](#)].
- [57] S. Pueblas and R. Scoccimarro, *Generation of vorticity and velocity dispersion by orbit crossing*, *Phys. Rev. D* **80** (2009) 043504 [[arXiv:0809.4606](#)] [[INSPIRE](#)].
- [58] G. Jelic-Cizmek, F. Lepori, J. Adamek and R. Durrer, *The generation of vorticity in cosmological N-body simulations*, *JCAP* **09** (2018) 006 [[arXiv:1806.05146](#)] [[INSPIRE](#)].
- [59] O. Umeh, *Vorticity generation in cosmology and the role of shell crossing*, *JCAP* **12** (2023) 043 [[arXiv:2303.08782](#)] [[INSPIRE](#)].
- [60] J.Z. Chen, A. Upadhye and Y.Y.Y. Wong, *The cosmic neutrino background as a collection of fluids in large-scale structure simulations*, *JCAP* **03** (2021) 065 [[arXiv:2011.12503](#)] [[INSPIRE](#)].
- [61] G. Pierobon, M.R. Mosbech, A. Upadhye and Y.Y.Y. Wong, *One trick to treat them all: SuperEasy linear response for any hot dark matter in N-body simulations*, *JCAP* **12** (2024) 032 [[arXiv:2410.05816](#)] [[INSPIRE](#)].
- [62] M. Pietroni, *Flowing with time: a new approach to nonlinear cosmological perturbations*, *JCAP* **10** (2008) 036 [[arXiv:0806.0971](#)] [[INSPIRE](#)].
- [63] J. Lesgourgues, S. Matarrese, M. Pietroni and A. Riotto, *Non-linear power spectrum including massive neutrinos: the time-RG flow approach*, *JCAP* **06** (2009) 017 [[arXiv:0901.4550](#)] [[INSPIRE](#)].

- [64] M. Garny and P. Taule, *Loop corrections to the power spectrum for massive neutrino cosmologies with full time- and scale-dependence*, *JCAP* **01** (2021) 020 [[arXiv:2008.00013](#)] [[INSPIRE](#)].
- [65] F. F hrer and Y.Y.Y. Wong, *Higher-order massive neutrino perturbations in large-scale structure*, *JCAP* **03** (2015) 046 [[arXiv:1412.2764](#)] [[INSPIRE](#)].
- [66] S. Chandrasekhar, *Radiative transfer*, Dover, New York, NY, U.S.A. (1960).
- [67] M. Abramowitz and I.A. Stegun, *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, National Bureau of Standards, Washington, DC, U.S.A. (1972).
- [68] J.Z. Chen, A. Upadhye and Y.Y.Y. Wong, *One line to run them all: SuperEasy massive neutrino linear response in N-body simulations*, *JCAP* **04** (2021) 078 [[arXiv:2011.12504](#)] [[INSPIRE](#)].
- [69] A. Upadhye et al., *Redshift-space distortions in massive neutrino and evolving dark energy cosmologies*, *Phys. Rev. D* **93** (2016) 063515 [[arXiv:1506.07526](#)] [[INSPIRE](#)].
- [70] K.R. Moran et al., *The Mira-Titan Universe — IV. High-precision power spectrum emulation*, *Mon. Not. Roy. Astron. Soc.* **520** (2023) 3443 [[arXiv:2207.12345](#)] [[INSPIRE](#)].
- [71] EUCLID collaboration, *Euclid: modelling massive neutrinos in cosmology — a code comparison*, *JCAP* **06** (2023) 035 [[arXiv:2211.12457](#)] [[INSPIRE](#)].
- [72] J.M. Sullivan, J.D. Emberson, S. Habib and N. Frontiere, *Improving initialization and evolution accuracy of cosmological neutrino simulations*, *JCAP* **06** (2023) 003 [[arXiv:2302.09134](#)] [[INSPIRE](#)].
- [73] R. Teyssier, *Cosmological hydrodynamics with adaptive mesh refinement: a new high resolution code called RAMSES*, *Astron. Astrophys.* **385** (2002) 337 [[astro-ph/0111367](#)] [[INSPIRE](#)].
- [74] R. Mauland, O. Elgar y, D.F. Mota and H.A. Winther, *The void-galaxy cross-correlation function with massive neutrinos and modified gravity*, *Astron. Astrophys.* **674** (2023) A185 [[arXiv:2303.05820](#)] [[INSPIRE](#)].
- [75] J. Adamek, D. Daverio, R. Durrer and M. Kunz, *General relativity and cosmic structure formation*, *Nature Phys.* **12** (2016) 346 [[arXiv:1509.01699](#)] [[INSPIRE](#)].
- [76] J. Adamek, D. Daverio, R. Durrer and M. Kunz, *gevolution: a cosmological N-body code based on general relativity*, *JCAP* **07** (2016) 053 [[arXiv:1604.06065](#)] [[INSPIRE](#)].
- [77] J. Adamek, R. Durrer and M. Kunz, *Relativistic N-body simulations with massive neutrinos*, *JCAP* **11** (2017) 004 [[arXiv:1707.06938](#)] [[INSPIRE](#)].
- [78] A.M. Beck et al., *An improved SPH scheme for cosmological simulations*, *Mon. Not. Roy. Astron. Soc.* **455** (2016) 2110 [[arXiv:1502.07358](#)] [[INSPIRE](#)].
- [79] T. Marin-Gilabert, M. Valentini, U.P. Steinwandel and K. Dolag, *The role of physical and numerical viscosity in hydrodynamical instabilities*, *Mon. Not. Roy. Astron. Soc.* **517** (2022) 5971 [[arXiv:2205.09135](#)] [[INSPIRE](#)].
- [80] V. Springel, *The cosmological simulation code GADGET-2*, *Mon. Not. Roy. Astron. Soc.* **364** (2005) 1105 [[astro-ph/0505010](#)] [[INSPIRE](#)].
- [81] V. Springel et al., *The Aquarius project: the subhalos of galactic halos*, *Mon. Not. Roy. Astron. Soc.* **391** (2008) 1685 [[arXiv:0809.0898](#)] [[INSPIRE](#)].
- [82] V. Springel, R. Pakmor, O. Zier and M. Reinecke, *Simulating cosmic structure formation with the gadget-4 code*, *Mon. Not. Roy. Astron. Soc.* **506** (2021) 2871 [[arXiv:2010.03567](#)] [[INSPIRE](#)].

- [83] J. Dakin et al., *ν CONCEPT: cosmological neutrino simulations from the non-linear Boltzmann hierarchy*, *JCAP* **02** (2019) 052 [[arXiv:1712.03944](#)] [[INSPIRE](#)].
- [84] J. Dakin, S. Hannestad and T. Tram, *The cosmological simulation code CONCEPT 1.0*, *Mon. Not. Roy. Astron. Soc.* **513** (2022) 991 [[arXiv:2112.01508](#)] [[INSPIRE](#)].
- [85] Y. Ali-Haïmoud and S. Bird, *An efficient implementation of massive neutrinos in non-linear structure formation simulations*, *Mon. Not. Roy. Astron. Soc.* **428** (2012) 3375 [[arXiv:1209.0461](#)] [[INSPIRE](#)].
- [86] SWIFT collaboration, *SWIFT: a modern highly-parallel gravity and smoothed particle hydrodynamics solver for astrophysical and cosmological applications*, *Mon. Not. Roy. Astron. Soc.* **530** (2024) 2378 [[arXiv:2305.13380](#)] [[INSPIRE](#)].
- [87] W. Elbers et al., *An optimal non-linear method for simulating relic neutrinos*, *Mon. Not. Roy. Astron. Soc.* **507** (2021) 2614 [[arXiv:2010.07321](#)] [[INSPIRE](#)].
- [88] W. Elbers et al., *Higher order initial conditions with massive neutrinos*, *Mon. Not. Roy. Astron. Soc.* **516** (2022) 3821 [[arXiv:2202.00670](#)] [[INSPIRE](#)].
- [89] D. Inman et al., *Simulating the cold dark matter-neutrino dipole with TianNu*, *Phys. Rev. D* **95** (2017) 083518 [[arXiv:1610.09354](#)] [[INSPIRE](#)].
- [90] J.D. Emberson et al., *Cosmological neutrino simulations at extreme scale*, *Res. Astron. Astrophys.* **17** (2017) 085 [[arXiv:1611.01545](#)] [[INSPIRE](#)].
- [91] S. Hannestad, A. Upadhye and Y.Y.Y. Wong, *Spoon or slide? The non-linear matter power spectrum in the presence of massive neutrinos*, *JCAP* **11** (2020) 062 [[arXiv:2006.04995](#)] [[INSPIRE](#)].
- [92] J.Z. Chen, M.R. Mosbech, A. Upadhye and Y.Y.Y. Wong, *Hybrid multi-fluid-particle simulations of the cosmic neutrino background*, *JCAP* **03** (2023) 012 [[arXiv:2210.16012](#)] [[INSPIRE](#)].
- [93] KATRIN collaboration, *Direct neutrino-mass measurement with sub-electronvolt sensitivity*, *Nature Phys.* **18** (2022) 160 [[arXiv:2105.08533](#)] [[INSPIRE](#)].
- [94] C. Benso, T. Schwetz and D. Vatsyayan, *Large neutrino mass in cosmology and keV sterile neutrino dark matter from a dark sector*, [arXiv:2410.23926](#) [[INSPIRE](#)].
- [95] J. Alvey, M. Escudero and N. Sabti, *What can CMB observations tell us about the neutrino distribution function?*, *JCAP* **02** (2022) 037 [[arXiv:2111.12726](#)] [[INSPIRE](#)].
- [96] A. Lewis and A. Challinor, *Weak gravitational lensing of the CMB*, *Phys. Rept.* **429** (2006) 1 [[astro-ph/0601594](#)] [[INSPIRE](#)].
- [97] F. McCarthy, S. Foreman and A. van Engelen, *Avoiding baryonic feedback effects on neutrino mass measurements from CMB lensing*, *Phys. Rev. D* **103** (2021) 103538 [[arXiv:2011.06582](#)] [[INSPIRE](#)].
- [98] CMB-S4 collaboration, *CMB-S4 Science Book, First Edition*, [arXiv:1610.02743](#) [[INSPIRE](#)].
- [99] A. Upadhye et al., *Non-linear CMB lensing with neutrinos and baryons: FLAMINGO simulations versus fast approximations*, *Mon. Not. Roy. Astron. Soc.* **529** (2024) 1862 [[arXiv:2308.09755](#)] [[INSPIRE](#)].
- [100] A. Cuoco, J. Lesgourgues, G. Mangano and S. Pastor, *Do observations prove that cosmological neutrinos are thermally distributed?*, *Phys. Rev. D* **71** (2005) 123501 [[astro-ph/0502465](#)] [[INSPIRE](#)].