Neutrino Masses in the Landscape and Global-Local Dualities in Eternal Inflation

By

Dan Mainemer Katz

A dissertation submitted in partial satisfaction of the
requirements for the degree of
Doctor of Philosophy
in
Physics
in the
Graduate Division
of the
University of California, Berkeley

Committee in charge:
Professor Raphael Bousso, Chair
Professor Petr Horava
Professor Nicolai Y Reshetikhin

Spring 2016
Neutrino Masses in the Landscape and Global-Local Dualities in Eternal Inflation

Copyright 2016
by
Dan Mainemer Katz
Abstract

Neutrino Masses in the Landscape and Global-Local Dualities in Eternal Inflation

by

Dan Mainemer Katz

Doctor of Philosophy in Physics

University of California, Berkeley

Professor Raphael Bousso, Chair

In this dissertation we study two topics in Theoretical Cosmology: one more formal, the other more phenomenological. We work in the context of eternally inflating cosmologies. These arise in any fundamental theory that contains at least one stable or metastable de Sitter vacuum. Each topic is presented in a different chapter:

Chapter 1 deals with the measure problem in eternal inflation. Global-local duality is the equivalence of seemingly different regulators in eternal inflation. For example, the light-cone time cutoff (a global measure, which regulates time) makes the same predictions as the causal patch (a local measure that cuts off space). We show that global-local duality is far more general. It rests on a redundancy inherent in any global cutoff: at late times, an attractor regime is reached, characterized by the unlimited exponential self-reproduction of a certain fundamental region of spacetime. An equivalent local cutoff can be obtained by restricting to this fundamental region. We derive local duals to several global cutoffs of interest. The New Scale Factor Cutoff is dual to the Short Fat Geodesic, a geodesic of fixed infinitesimal proper width. Vilenkin’s CAH Cutoff is equivalent to the Hubbletube, whose width is proportional to the local Hubble volume. The famous youngness problem of the Proper Time Cutoff can be readily understood by considering its local dual, the Incredible Shrinking Geodesic. The chapter closely follows our paper [1].

Chapter 2 deals with the question of whether neutrino masses could be anthropically explained. The sum of active neutrino masses is well constrained, $58 \text{ meV} \leq m_\nu \lesssim 0.23 \text{ eV}$, but the origin of this scale is not well understood. Here we investigate the possibility that it arises by environmental selection in a large landscape of vacua. Earlier work had noted the detrimental effects of neutrinos on large scale structure. However, using Boltzmann codes to compute the smoothed density contrast on Mpc scales, we find that dark matter halos form abundantly for $m_\nu \gtrsim 10 \text{ eV}$. This finding rules out an anthropic origin of $m_\nu$, unless a different catastrophic boundary can be identified. Here we argue that galaxy formation becomes inefficient for $m_\nu \gtrsim 10 \text{ eV}$. We show that in this regime, structure forms late and is dominated by cluster scales, as in a top-down scenario. This is catastrophic: baryonic gas will cool too slowly to form stars in an abundance comparable to our universe. With
this novel cooling boundary, we find that the anthropic prediction for $m_\nu$ agrees at better than $2\sigma$ with current observational bounds. A degenerate hierarchy is mildly preferred. The chapter closely follows our paper [2].
## Contents

1. Global-Local Dualities in Eternal Inflation ......................................................... 1
   1.1 Introduction ........................................................................................................... 1
   1.2 Causal Patch/Light-Cone Time Duality ................................................................. 4
      1.2.1 Causal Patch Measure ..................................................................................... 4
      1.2.2 Proof of Equivalence to the Light-Cone Time Measure ................................. 6
      1.2.3 Light-Cone Time Rate Equation and Attractor Solution ............................... 7
   1.3 Short Fat Geodesic/New Scale Factor Cutoff Duality ............................................ 11
      1.3.1 Short Fat Geodesic Measure ............................................................................ 11
      1.3.2 Proof of Equivalence to the New Scale Factor Cutoff ................................... 14
      1.3.3 New Scale Factor Cutoff Rate Equation and Attractor Solution ................... 17
   1.4 General Global-Local Dualities ............................................................................ 20
      1.4.1 \(X\)-fat Geodesic Measure ............................................................................... 20
      1.4.2 Proof of Equivalence to the \(T\)-cutoff Measure .............................................. 22
      1.4.3 \(T\)-cutoff Rate Equation and Attractor Solution ............................................ 23
      1.4.4 The Hubbletube and the CAH measure ......................................................... 25
      1.4.5 The Incredible Shrinking Geodesic and the Proper Time Cutoff ...................... 26

2. Anthropic Origin of the Neutrino Mass? ................................................................. 30
   2.1 Introduction .......................................................................................................... 30
   2.2 Predictions in a Large Landscape ........................................................................... 39
      2.2.1 Prior as a Multiverse Probability Force ......................................................... 40
      2.2.2 Anthropic Weighting ....................................................................................... 41
      2.2.3 Measure .......................................................................................................... 42
   2.3 Calculation of \(dP/d\log m_\nu\) .................................................................................. 44
      2.3.1 Fixed, Variable, and Time-dependent Parameters .......................................... 44
      2.3.2 Homogeneous Evolution .................................................................................. 46
      2.3.3 Halo Formation ............................................................................................... 47
      2.3.4 Galaxy Formation: Neutrino-Induced Cooling Catastrophe ............................. 49
A Appendix

A.1 Cosmological Constant and the Causal Patch

A.1.1 Weinberg’s Prediction: \( \Lambda \sim t_{\text{vir}}^{-2} \)  
A.1.2 Causal Patch Prediction: \( \Lambda \sim t_{\text{obs}}^{-2} \)

A.2 Structure Formation with Neutrinos

A.2.1 Neutrino Cosmology
A.2.2 Late-Time Extrapolation of Numerical Results

A.3 Cooling and Galaxy Formation
Chapter 1

Global-Local Dualities in Eternal Inflation

1.1 Introduction

Global-local duality is one of the most fascinating properties of eternal inflation. It lies at the heart of a profound debate: should we expect to understand cosmology by adopting a global, “bird’s eye” viewpoint that surveys many causally disconnected regions, as if we stood outside the universe as a kind of meta-observer? Or should fundamental theory describe only experiments that can be carried out locally, in accordance with the laws of physics and respectful of the limitations imposed by causality?

Global-local duality appears to reconcile these radically different perspectives. It was discovered as a byproduct of attempts to solve the measure problem of eternal inflation: the exponential expansion of space leads to infinite self-reproduction, and all events that can happen in principle will happen infinitely many times. Short reviews include Refs. [6, 7].

To compute relative probabilities, a regulator, or measure, is required.

Most measure proposals are based on geometric cutoffs: one constructs a finite subset of the eternally inflating spacetime according to some rule. Relative probabilities can then be defined as ratios of the expected number of times the corresponding outcomes that occur in these subsets. Geometric cutoffs can be divided into two classes. Very roughly speaking, global cutoffs act on time, across the whole universe; this is natural from the bird’s eye viewpoint. Local cutoffs act on space; this is more natural from the viewpoint of an observer within the spacetime.

1 The measure problem has nothing to do with how many vacua there are in the theory. It arises if there exists at least one stable or metastable de Sitter vacuum. The observed accelerated expansion of the universe [3,4] is consistent with a fixed positive cosmological constant and thus informs us that our vacuum is likely of this type [5].

2 As a consequence, probabilities behave as if the spacetime was extendible [8,9], a counterintuitive feature that underlies the phenomenological successes of some measures.
Global cutoffs define a parameter $T$ that can roughly be thought of as a time variable. Spacetime points with $T$ smaller than the cutoff form a finite set in which expected numbers of outcomes can be computed; then the limit $T \to \infty$ is taken to define probabilities:

$$\frac{P_I}{P_J} \equiv \lim_{T \to \infty} \frac{N_I(T)}{N_J(T)} .$$

(1.1)

Examples include the proper time cutoff \[10–14\], where $T$ is the proper time along geodesics in a congruence; the scale factor time cutoff \[11,15–20\], where $T$ measures the local expansion of geodesics; and the light-cone time cutoff \[21–23\], where $T$ is determined by the size of the future light-cone of an event.\(^3\)

Local cutoffs restrict to the neighborhood of a single timelike geodesic. The simplest local cutoff is the causal patch \[26,27\]: the causal past of the geodesic, which depends only on the endpoint of the geodesic. Another example is the fat geodesic \[19\], which restricts to an infinitesimal proper volume near the geodesic. Relative probabilities are defined by computing an ensemble average over different possible histories of the cutoff region:

$$\frac{P_I}{P_J} \equiv \frac{\langle N_I \rangle}{\langle N_J \rangle} .$$

(1.2)

Global-local duality is the statement that there exist pairs of cutoffs—one global, one local—that yield precisely the same predictions. Our goal will be to exhibit the generality of this property and the basic structure underlying it. This will allow us to identify new local duals to some global measures of particular interest.

Discussion  Global-local duality implies that the distinction between two seemingly disparate perspectives on cosmology is, at best, subtle. However, it is too early to conclude that the global and local viewpoints are as interchangeable as the position and momentum basis in quantum mechanics. Some important distinctions remain; and for now, each side, global and local, exhibits attractive features that the other lacks.

Advantages of the global viewpoint: A key difference between global and local cutoffs is that local measures are sensitive to initial conditions, whereas global measures exhibit an attractor regime that completely determines all probabilities. The attractor regime can only be affected by infinite fine-tuning, or by choosing initial conditions entirely in terminal vacua so that eternal inflation cannot proceed. Thus, global measures are relatively insensitive to initial conditions.

Therefore, a local cutoff can reproduce the predictions of its global dual only with a particular choice of initial conditions on the local side, given by the attractor solution of the

---

\(^3\)In the absence of a first-principles derivation, the choice between proposals must be guided by their phenomenology. Fortunately, different definitions of $T$ often lead to dramatically different predictions (see, e.g., \[24,25\]). In this paper, we do not consider phenomenology but focus on formal properties.
global cutoff. (The distribution over initial conditions is set by the field distribution on a slice of constant global cutoff parameter.) Ultimately, there appears to be no reason why initial conditions might not be dictated by aspects of a fundamental theory unrelated to the measure. In this case, global and local measures could be inequivalent. But for now, the global cutoff is more restrictive, and thus more predictive, than its local dual, some of whose predictions could be changed by a different choice of initial conditions.

Advantages of the local viewpoint: Through the study of black holes and the information paradox, we have learned that the global viewpoint must break down at the full quantum level. Otherwise, the black hole would xerox arbitrary quantum states into its Hawking radiation, in contradiction with the linearity of quantum mechanics [30]. By contrast, a description of any one causally connected region, or causal patch, will only contain one copy of the information. For example, an observer remaining outside the black hole will be able to access the Hawking radiation but not the original copy, which is accessible to an observer inside the black hole. Indeed, this was the key motivation for introducing the causal patch as a measure: if required in the context of black holes, surely the same restriction would apply to cosmology as well.

The global spacetime is obtained by pretending that the state of the universe is measured, roughly once per Hubble time in every horizon volume. It is not clear what the underlying process of decoherence is, since no natural environment is available (by definition, since we are considering the entire universe). By contrast, the local description (at least, the causal patch) exhibits decoherence at the semiclassical level, since matter can cross the event horizon [31]. This suggests that the global picture may be merely a convenient way of combining the different semiclassical histories of the causal patch into a single spacetime.

Aside from the fundamental questions raised by global-local duality, we expect that our results will aid future studies of measure phenomenology. Computations are significantly simpler in the local dual, because it strips away an infinite redundancy. The local cutoff region can be considered an elementary unit of spacetime, which (from the global viewpoint) is merely reproduced over and over by the exponential expansion of the eternally inflating universe.

Outline In Sec. 1.2, we set the stage by showing that the light-cone time cutoff is equivalent to the causal patch, with initial conditions in the longest-lived de Sitter vacuum. This is a known result [21, 22]. The measures on both sides of the duality are particularly simple; as a consequence, the duality proof is especially transparent.

In Sec. 1.3, we define the Short Fat Geodesic measure, and we show that it is the local dual to the recently proposed New Scale Factor Cutoff [20]. This generalizes to arbitrary eternally inflating spacetimes the known duality between the fat geodesic and the scale factor

---

4In a theory with a vacuum landscape large enough to solve the cosmological constant problem [28], most predictions of low-energy properties are fairly insensitive to initial conditions even in local measures [29].
time cutoff \[19\], which originally applied only to everywhere-expanding multiverse regions. We also discuss important formal differences between the causal patch/light-cone time pair and all other global-local pairs we consider. Unlike other local measures, which require the specification of spatial boundary conditions, the causal patch is entirely self-contained and can be evaluated without referring to a global viewpoint.

In Sec. [1.4], we generalize the proof of the previous section to relate a large class of global-local pairs. On the local side, one can consider modulations of the fatness of the geodesic; on the global side, this corresponds to particular modifications of the definition of the cutoff parameter \( T \), which we identify explicitly. We illustrate this general result by deriving local duals to two global proposals, the CAH cutoff \[32\] and the proper time cutoff. The local dual, the *Hubbletube*, naturally extends the range of applicability of the CAH cutoff to include decelerating regions; but unfortunately, an additional prescription (such as CAH+ \[32\]) is still required to deal with nonexpanding regions. The local dual to the proper time cutoff, the *Incredible Shrinking Geodesic*, makes the phenomenological problems of this simplest of global cutoffs readily apparent.

### 1.2 Causal Patch/Light-Cone Time Duality

In this section we show that the causal patch measure (with particular initial conditions) is equivalent to the light-cone time measure, i.e., that both define the same relative probabilities. We follow Ref. \[22\], where more details can be found. The proof is rather simple if one is willing to use the (intuitively natural) results for the attractor behavior of eternal inflation as a function of the global time coordinate. For this reason we will first present a proof of duality, while assuming the attractor behavior, in Sec. [1.2.1]. Then we will derive the attractor behavior, in Sec. [1.2.3].

#### 1.2.1 Causal Patch Measure

The causal patch is defined as the causal past of the endpoint of a geodesic. Consider two outcomes \( I \) and \( J \) of a particular observation; for example, different values of the cosmological constant, or of the CMB temperature. The relative probabilities for these two outcomes, according to the causal patch measure, is given by

\[
\frac{\hat{P}_I}{\hat{P}_J} = \frac{\langle N_I \rangle_{\text{CP}}}{\langle N_J \rangle_{\text{CP}}},
\]

(1.3)

Here, \( \langle N_I \rangle \) is the expected number of times the outcome \( I \) occurs in the causal patch.

---

5See Ref. \[33, 34\] for a simplified model that exhibits most of the essential features of eternal inflation, including causal patch/light-cone time duality.
Computing $\langle N_I \rangle_{\text{CP}}$ involves two types of averaging: over initial conditions, $p_i^{(0)}$, and over different decoherent histories of the patch. We can represent the corresponding ensemble of causal patches as subsets of a single spacetime. Namely, we consider a large initial hypersurface (a moment of time), $\Sigma_0$, containing $Z \to \infty$ different event horizon regions, with a fraction $p_i^{(0)}$ of them in the vacuum $i$. Event horizons are globally defined, but we will be interested in cases where the initial conditions have support mainly in long-lived metastable de Sitter vacua. Then we make a negligible error by assuming that the event horizon on the slice $\Sigma_0$ contains a single de Sitter horizon volume, of radius $H_\alpha^{-1} = (3/\Lambda_\alpha)^{1/2}$, where $\Lambda_\alpha$ is the cosmological constant of vacuum $\alpha$. and we may take the spatial geometry to be approximately flat on $\Sigma_0$. More general initial conditions can be considered [22].

At the center of each initial horizon patch, consider the geodesic orthogonal to $\Sigma_0$, and construct the associated causal patch. We may define $\langle N_I \rangle_{\text{CP}}$ as the average over all $Z$ causal patches thus constructed, in the limit $Z \to \infty$. So far, each causal patch is causally disconnected from every other patch. It is convenient to further enlarge the ensemble by increasing the density of geodesics to $z$ geodesics per event horizon volume, and to take $z \to \infty$:

$$\langle N_I \rangle_{\text{CP}} = (zZ)^{-1} \sum_{\nu=1}^{zZ} N_{\nu,\text{CP}}^I,$$

where the sum runs over the $zZ$ causal patches, and $N_{\nu,\text{CP}}^I$ is the number of times $I$ occurs in the causal patch $\nu$. The causal patches will overlap, but this will not change the ensemble average.

At finite large $z$, a sufficiently early event $Q$ in the future of $\Sigma_0$ will thus be contained in a number of causal patches. The later $Q$ occurs, the fewer patches will contain it (Fig. 1.1). In the limit $z \to \infty$, every event will be contained in an infinite number of patches, but there is still a sense in which later events are overcounted less. This can be captured by defining the quantity $\pi(Q)$, as $z^{-1}$ times the number of causal patches containing a given event $Q$.

By causality, the causal patch of a geodesic contains $Q$ if and only if that geodesic enters the future of $Q$. Therefore, $\pi(Q)$ is the volume, measured in units of horizon volume, on $\Sigma_0$, of the starting points of those geodesics that eventually enter the future light-cone of $Q$. This allows us to reorganize the sum in Eq. (1.4). Instead of summing over causal patches, we may sum over all events $Q$ where outcome $\check{I}$ occurs, taking into account that each such instance will be “overcounted” by the ensemble of causal patches, by a factor proportional to $\pi(Q)$:

$$\langle N_I \rangle_{\text{CP}} = Z^{-1} \sum_{Q \in \check{I}} \pi(Q).$$

Light-cone time is defined precisely so that it is constant on hypersurfaces of constant

---

6We will use Greek indices to label de Sitter vacua ($\Lambda > 0$), indices $m,n,...$ to label terminal vacua ($\Lambda \leq 0$), and $i,j$ for arbitrary vacua.
Figure 1.1: Discrete ensemble of causal patches [22]. The event $Q$ is contained in those causal patches whose generating geodesics (blue) enter the causal future of $Q$, $I^+(Q)$ (shaded green/dark). In the continuous limit, $z \to \infty$, the causal patch measure weights $Q$ in proportion to the volume of its future light-cone on the future boundary. Thus, the weight of $Q$ depends only on light-cone time $t_{LC}$. This underlies the equivalence of the causal patch measure (with particular initial conditions) and the light-cone time cutoff. This is a conformal (or Penrose) diagram; the spacetime metric is rescaled but light-rays still travel at 45 degrees.

The exact definition is not essential but it is convenient to choose

$$t_{LC}(Q) \equiv -\frac{1}{3} \log \pi(Q).$$

(1.6)

This defines a time variable at every event $Q$ in the future of the initial hypersurface $\Sigma_0$. We may reorganize the sum once more, as an integral over light-cone time:

$$\langle N_I \rangle_{CP} = \int dt_{LC} \frac{dN_I}{dt_{LC}} e^{-3t_{LC}},$$

(1.7)

where $dN_I$ is the number of events of type $I$ that occur in the time interval $(t_{LC}, t_{LC} + dt_{LC})$, and the integral ranges over the future of $\Sigma_0$.

### 1.2.2 Proof of Equivalence to the Light-Cone Time Measure

So far, we have been dealing with a local measure, the causal patch. We have merely represented the causal patch ensemble in terms of a single global spacetime. Moreover, we have rewritten the ensemble average, as an integral over a time variable $t_{LC}$, adapted to the factor $\pi(Q)$ by which events in the global spacetime are weighted in the ensemble.

We will now show that with a particular, simple choice of initial conditions, the causal patch probabilities $\hat{P}_I$ (i.e., the ensemble averages $\langle N_I \rangle_{CP}$) agree with the probabilities computed from a global measure, the light-cone time cutoff. These probabilities are defined
by
\[
\frac{\dot{P}_I}{\dot{P}_J} = \lim_{t_{LC} \to \infty} \frac{N_I(t_{LC})}{N_J(t_{LC})},
\]
(1.8)
where \(N_I(t_{LC})\) is the number of events of type \(I\) prior to the light-cone time \(t_{LC}\).

As we shall review below, the cosmological dynamics, as a function of light-cone time, leads to an attractor regime:
\[
N_I(t_{LC}) = \tilde{N}_I e^{\gamma t_{LC}} + O(e^{\varphi t_{LC}}),
\]
(1.9)
where \(\varphi < \gamma < 3\). Therefore, the light-cone time probabilities are given by
\[
\frac{\dot{P}_I}{\dot{P}_J} = \frac{\tilde{N}_I}{\tilde{N}_J}.
\]
(1.10)

The causal patch probabilities can also be evaluated using Eq. (1.9), if we choose initial conditions in the attractor regime, i.e., if we take \(\Sigma_0\) to be a slice of constant, very late light-cone time. Substituting into Eq. (1.7), one finds
\[
\langle N_I \rangle_{CP} = \tilde{N}_I \int dt_{LC} \gamma e^{(\gamma-3)t_{LC}},
\]
(1.11)
Since \(\gamma < 3\), the integral converges to an \(I\)-independent constant, so relative probabilities in the causal patch measure are given by
\[
\frac{\dot{P}_I}{\dot{P}_J} = \frac{\tilde{N}_I}{\tilde{N}_J}.
\]
(1.12)
This agrees with the light-cone time probabilities, Eq. (1.10). Therefore, the two measures are equivalent.

### 1.2.3 Light-Cone Time Rate Equation and Attractor Solution

We will now complete the proof by deriving the attractor regime, Eq. (1.9). (We will follow [22] and will make use of certain general properties of rate equations in eternal inflation [40].) It is convenient to do this in two steps. Treating each long-lived metastable de Sitter vacuum as pure, empty de Sitter space, one derives the number \(n_\alpha(t_{LC})\) of horizon patches of vacuum \(\alpha\). Because of the slow decays, most regions are indeed empty, and slices

\footnote{Interestingly, the light-cone time cutoff was not discovered as the global dual to the causal patch. It was proposed independently [21] as a covariant implementation [35, 36] of a suggestion by Garriga and Vilenkin [37] that an analogue of the UV/IR relation [38] of gauge/gravity duality [39] would yield a preferred global time variable in eternal inflation. An apparent relation to the causal patch was immediately noted [21], but the exact duality was recognized only later [22].}
of constant light-cone time are spatially flat on the horizon scale. Thus, a horizon patch at constant time $t_{\text{LC}}$ can be defined as a physical volume

$$v_\alpha = \frac{4\pi}{3} \tau_{\Lambda,\alpha}^3,$$  \hspace{1cm} (1.13)

where

$$\tau_{\Lambda,\alpha} \equiv \sqrt{\frac{3}{\Lambda_\alpha}},$$  \hspace{1cm} (1.14)

is the time and distance scale associated with the cosmological constant in vacuum $\alpha$. The number $n_\alpha$ of horizon patches at the time $t_{\text{LC}}$ is related to the physical volume $V_\alpha$ occupied by vacuum $\alpha$, as

$$n_\alpha(t_{\text{LC}}) = \frac{V_\alpha(t_{\text{LC}})}{v_\alpha}.$$  \hspace{1cm} (1.15)

In the second step, one focuses on the decay events in this distribution, i.e., the production of new bubbles. These bubbles can then be considered in detail. In general they will be not be empty, and they need not have positive cosmological constant.

The rate equation for the number of horizon patches of metastable de Sitter vacua is

$$\frac{dn_\alpha}{dt_{\text{LC}}} = (3 - \kappa_\alpha)n_\alpha + \sum_\beta \kappa_{\alpha\beta}n_\beta,$$  \hspace{1cm} (1.16)

where $\kappa_{i\beta} = v_i\tau_{\Lambda,\beta}\Gamma_{i\beta}$ is the dimensionless decay rate from $\beta$ to $i$. That is, $\Gamma_{i\beta}$ is the rate at which $i$-bubbles are produced inside the $\beta$-vacuum, per unit four-volume; and $\kappa_{i\beta}$ is the decay rate per unit horizon volume and unit de Sitter time scale. Also, $\kappa_\alpha \equiv \sum_i \kappa_{i\alpha}$ is the total dimensionless decay rate of vacuum $\alpha$. We will now explain the origin of each term on the right-hand side.

The first term, $3n_\alpha$, arises from the exponential volume growth of de Sitter space. In regions occupied by vacuum $\alpha$, the metric behaves locally as $ds^2 = -d\tau^2 + e^{2t/\tau_{\Lambda,\alpha}}dx^2$, where $t$ is proper time. The relation between proper time and light-cone time is

$$dt_{\text{LC}} = \frac{d\tau}{\tau_{\Lambda,\alpha}},$$  \hspace{1cm} (1.17)

in pure de Sitter space. In metastable de Sitter space this relation is modified, on average, by a relative correction not exceeding $\kappa_\alpha$, which can be neglected for the purposes of the rate equation.

The second term, $-\kappa_\alpha n_\alpha$ is an effective term that takes into account the decay of vacuum $\alpha$ into other vacua. Decays of this type proceed by the formation of a bubble of the new vacuum [41]. Typically, the spherical domain wall separating the vacua will be small initially, compared to the size of the event horizon of the parent vacuum. The domain wall will then expand at a fixed acceleration, asymptotically approaching the future light-cone of the
nucleation event. A detailed treatment of this dynamics would enormously complicate the rate equation, but fortunately an exquisite approximation is available. Even at late times, because of de Sitter event horizons, only a portion of the parent vacuum is destroyed by the bubble. This portion is the causal future of the nucleation point, and at late times it agrees with the comoving future of a single horizon volume centered on the nucleation point, at the nucleation time. Because the bubble reaches its asymptotic comoving size very quickly (exponentially in light-cone time), only a very small error, of order $\kappa$, is introduced if we remove this comoving future, rather than the causal future, from the parent vacuum. That is, for every decay event in vacuum $\alpha$, the number of horizon patches of type $\alpha$ is reduced by 1 in the rate equation. This is called the square bubble approximation. The expected number of such events is $-\kappa\dot{n}_\alpha dt_{\text{LC}}$.

The third term, $\sum_\beta \kappa_{\alpha\beta}n_\beta$, captures the production of bubbles of vacuum $\alpha$ by the decay of other vacua. The prefactor of this term is fixed by the continuity of light-cone time. This is the requirement that the future light-cone of an event $Q-\epsilon$ just prior to the nucleation event has the same asymptotic size $\pi(Q-\epsilon)$ as the future light-cone of an event $Q+\epsilon$ just after nucleation, $\pi(Q+\epsilon)$, as $\epsilon \to 0$. In the square bubble approximation, this implies that each nucleation event effectively contributes a comoving volume of new vacuum equivalent to one horizon patch at the time of nucleation. For the reasons described in the previous paragraph, one patch has the correct comoving size to eventually fill the future light-cone of $Q+\epsilon$\(^8\). Thus, for every decay event in which a bubble of vacuum $\alpha$ is produced, the number of horizon patches of type $\alpha$ is increased by 1 in the rate equation. The expected number of such events is $\sum_\beta \kappa_{\alpha\beta}n_\beta dt_{\text{LC}}$.

The rate equation (1.16) has the solution \[ n_\alpha(t_{\text{LC}}) = \tilde{n}_\alpha e^{\gamma t_{\text{LC}}} + O(e^{\varphi t_{\text{LC}}}) , \] where $\varphi < \gamma < 3$. (The case $\gamma = 3$ arises if and only if the landscape contains no terminal vacua, i.e., vacua with nonpositive cosmological constant, and will not be considered in this paper.) Here, $\gamma \equiv 3 - q$ is the largest eigenvalue of the matrix $M_{\alpha\beta}$ defined by rewriting Eq. (1.16) as \[ \frac{dn_\alpha}{dt_{\text{LC}}} = \sum_\beta M_{\alpha\beta}n_\beta; \] and $\tilde{n}_\alpha$ is the corresponding eigenvector. The terms of order $e^{\varphi t}$ are subleading and become negligible in the limit as $t_{\text{LC}} \to \infty$. To a very good approximation (better than $q \ll 1$), the eigenvector is dominated by the longest-lived metastable de Sitter vacuum in the theory, which will be denoted by $\ast$: \[ \tilde{n}_\alpha \approx \delta_{\alpha\ast} , \] and \[ q \approx \kappa_{\ast} \] \[ \text{(1.20)} \]

\(^8\)Note that this implies that the physical volume removed from vacuum $\beta$ is not equal to the physical volume added to vacuum $\alpha$ by the decay. This discontinuity is an artifact of the square bubble approximation and has no deeper significance. In the exact spacetime, the evolution of volumes is continuous.
is its total dimensionless decay rate.

Next, we compute number of events of type $I$ prior to the time $t_{LC}$. We assume that the events unfolding in a new bubble of vacuum $i$ depend only on $i$, but on the time of nucleation. This is true as long as the parent vacuum is long-lived, so that most decays occur in empty de Sitter space. For notational convenience, we will also assume that evolution inside a new bubble is independent of the parent vacuum; however, this could easily be included in the analysis. Then the number of events of type $I$ inside a bubble of type $i$, $dN_I/dN_i$ will depend only on the light-cone time since bubble nucleation, $u_{LC} \equiv t_{LC} - t_{nuc}^{\text{vac}}$. Therefore, we can write

$$N_I(t_{LC}) = \kappa_I n_*(t_{LC}) + \sum_{i \neq *} \int_0^{t_{LC}} \left( \frac{dN_I}{dN_i} \right)_{t_{LC} - t_{nuc}^{\text{vac}}} \frac{dN_i}{dt_{LC}} \, du_{LC}, \quad (1.21)$$

Because the dominant vacuum $*$ plays a role analogous to an equilibrium configuration, it is convenient to separate it out from the sum, and to define $\kappa_I$, as the dimensionless rate at which events of type $I$ are produced in $*$ regions. The rate at which vacua of type $i$ are produced is

$$\frac{dN_i}{dt_{LC}} = \sum_{\beta} \kappa_{i\beta} \tilde{n}_\beta. \quad (1.22)$$

By changing the integration variable to $u_{LC}$ in Eq. (1.21), and using Eq. (1.18), one finds that

$$N_I(t_{LC}) = \left( \kappa_I n_* + \sum_{i \neq *} \sum_{\beta} N_I \kappa_{i\beta} \tilde{n}_\beta \right) e^{\gamma_{t_{LC}}} + O(e^{\varphi_{t_{LC}}}), \quad (1.23)$$

where

$$N_I \equiv \int_0^\infty du_{LC} e^{-\gamma_{t_{LC}}} \left( \frac{dN_I}{dN_i} \right)_{u_{LC}}. \quad (1.24)$$

depends only on $I$ and $i$. The above integral runs over the interior of one $i$-bubble, excluding regions where $i$ has decayed into some other vacuum. Naively, the integral should range from 0 to $t_{LC}$. But the global measure requires us to take the limit $t_{LC} \to \infty$ in any case, and it can be done at this step separately without introducing divergences. Since $*$ does not appear in the sum in Eq. (1.23), and all other vacua decay faster than $*$, the interior of the $i$-bubble in Eq. (1.24) grows more slowly than $e^{\gamma_{t_{LC}}}$. Therefore, the integral converges, and we may write

$$N_I(t_{LC}) = \tilde{N}_I e^{\gamma_{t_{LC}}} + O(e^{\varphi_{t_{LC}}}), \quad (1.25)$$

where

$$\tilde{N}_I \equiv \kappa_I n_* + \sum_{i \neq *} \sum_{\beta} N_I \kappa_{i\beta} \tilde{n}_\beta. \quad (1.26)$$
1.3 Short Fat Geodesic/New Scale Factor Cutoff Duality

In this section we introduce the Short Fat Geodesic measure. We show that, with particular initial conditions, it is equivalent to the New Scale Factor Cutoff [20]. This generalizes to arbitrary eternally inflating spacetimes the duality between the (long) fat geodesic and (old) scale factor time cutoff discovered in Ref. [19], which applied only to everywhere-expanding multiverse regions.

1.3.1 Short Fat Geodesic Measure

A fat geodesic is defined as an infinitesimal neighborhood of a geodesic. At each point on the geodesic, one can define an orthogonal cross-sectional volume \(dV\), which we imagine to be spherical. It is important to note that an orthogonal cross-section can be defined only infinitesimally—there is no covariant way of extending the cross-section to a finite volume. For example, the spacelike geodesics orthogonal to a point on the geodesic in question need not form a well-defined hypersurface.

Consider a family of geodesics orthogonal to an initial hypersurface \(\Sigma_0\). Along each geodesic, we may define the scale factor parameter

\[
\eta \equiv \int \frac{\theta(\tau)}{3} d\tau ,
\]

where

\[
\theta \equiv \frac{d}{d\tau} \log \frac{dV}{dV_0} \tag{1.28}
\]

is the expansion of the congruence. In Eq. (1.28), \(dV\) is the volume element at the proper time \(\tau\) along a geodesic spanned by infinitesimally neighboring geodesics in the congruence; \(dV_0\) is the volume element spanned by the same neighbors at \(\tau = 0\). In terms of the unit tangent vector field (the four-velocity) of the geodesic congruence, \(\xi = \partial_\tau\), the expansion can be computed as

\[
\theta = \nabla_a \xi^a .
\]

If geodesics are terminated at the first conjugate point\(^9\), this procedure assigns a unique scale factor parameter to every event in the future of \(\Sigma_0\).\(^{[20]}\)

A Short Fat Geodesic is a fat geodesic restricted to values of the scale factor parameter larger than that at \(\Sigma_0\), which we may choose to be zero. Thus, it consists of the portions of the fat geodesic along which neighboring geodesics are farther away than they are on \(\Sigma_0\). Typically, the congruence will expand locally for some time. Eventually, all but a set of measure zero of geodesics will enter a collapsing region, such as a structure forming region

\(^9\)also called focal point, or caustic; this is when infinitesimally neighboring geodesics intersect
such as ours, or a crunching $\Lambda < 0$ vacuum. In such regions, focal points will be approached or reached, where $\eta \to -\infty$. The Short Fat Geodesic is terminated earlier, when $\eta = 0$\textsuperscript{10}.

If the congruence is everywhere expanding, the Short Fat Geodesic reduces to the (long) fat geodesic defined in Ref. [19], as a special case. This is precisely the case in which the old scale factor time is well-defined and a duality between (long) fat geodesic and old scale factor time cutoff was derived. The duality derived below is more general and applies to arbitrary eternally inflating universes. If the expanding phase is sufficiently long, the terminal point where $\eta = 0$ can be less than one Planck time from the caustic [20]. This is expected to be generic if the initial conditions are dominated by a long-lived metastable de Sitter vacuum. In this approximation, the short fat geodesic could be defined equivalently as being terminated at the first caustic.

Let us pause to point out some important differences between the causal patch cutoff discussed in the previous section, and the Short Fat Geodesic.

- The causal patch depends only on the endpoint of the geodesic. It has (and needs) no preferred time foliation. That is, there is no preferred way to associate to every point along the generating geodesic a particular time slice of the causal patch containing that point. By contrast, a specific infinitesimal neighborhood is associated to every point on the Short Fat Geodesic, so the contents of the cutoff region depend on the entire geodesic. (The same will be true for the $X$-fat Geodesic considered in the following section.)

- As a consequence, the geodesic congruence could be eliminated entirely in the construction of the causal patch ensemble, in favor of a suitable ensemble of points on the future conformal boundary of the spacetime [23]. By contrast, the congruence is an inevitable element in the construction of all other measures considered in this paper.

- The causal patch can be considered on its own, whereas the Short Fat Geodesic is naturally part of a larger spacetime. In the construction of an ensemble of causal patches in Sec. I.2.1, the global viewpoint was optional. This is because the causal patch is self-contained: if the initial state is a long-lived de Sitter vacuum, no further boundary conditions are required in order to construct the decoherent histories of the patch. We chose a global representation (a large initial surface with many horizon patches) only with a view to proving global-local duality. By contrast, the Short Fat Geodesic is greater than the domain of dependence of its initial cross-section. It has timelike boundaries where boundary conditions must be specified. The simplest way

\textsuperscript{10}For simplicity, we will assume that the congruence does not bounce, i.e., first decrease to negative values of $\eta$ and then expand again without first reaching a caustic. This would be guaranteed by the strong energy condition, but this condition is not satisfied in regions with positive cosmological constant. However, it is expected to hold in practice, since the cosmological constant cannot counteract focussing on sufficiently short distance scales.
to obtain suitable boundary conditions is from a global representation in terms of geodesics orthogonal to some surface $\Sigma_0$.

We will consider a dense family (a congruence) from the start, because of the infinitesimal size of the fat geodesic. We index each geodesic by the point $x_0 \in \Sigma_0$ from which it originates. For the same reason, it will be convenient to work with a (formally continuous) distribution $D_I(x)$ of events of type $I$. The distribution is defined so that the number of events of type $I$ in a spacetime four-volume $V_4$ is

$$N_I(V_4) = \int_{V_4} d^4x \sqrt{g} D_I(x) , \quad (1.30)$$

where $g = | \det g_{ab} |$. (The special case of pointlike events can be recovered by writing $D_I$ as a sum of $\delta$-functions.)

The infinitesimal number of events of type $I$ in the Short Fat Geodesic emitted from the point $x_0 \in \Sigma_0$ is

$$dN_{\text{FG}}^I = dV \int_{\eta>0} d\tau D_I(x(\tau)) , \quad (1.31)$$

where $\tau$ is the proper time along the geodesic, and the integral is restricted to portions of the geodesic with positive scale factor parameter. $dV$ is a fixed infinitesimal volume, which we may choose to define on $\Sigma_0$:

$$dV \equiv dV_0 = d^3x_0 \sqrt{h_0} . \quad (1.32)$$

The total number of events in the ensemble of fat geodesics is obtained by integrating over all geodesics emanating from $\Sigma_0$:

$$N_{\text{FG}}^I = \int_{\Sigma_0} d^3x_0 \sqrt{h_0} \int_{\eta>0} d\tau D_I(x(\tau)) . \quad (1.33)$$

where $h_0$ is the root of the determinant of the three-metric on $\Sigma_0$. We may take Eq. (1.33) as the definition of the fat geodesic measure, with relative probabilities given by

$$\frac{P_{\text{FG}}^I}{P_{\text{FG}}^J} = \frac{N_{\text{FG}}^I}{N_{\text{FG}}^J} . \quad (1.34)$$

Note that Eq. (1.33) is not a standard integral over a four volume; it is an integral over geodesics. We may rewrite it as an integral over a four-volume because the definition of the Short Fat Geodesic ensures that the geodesics do not intersect.\textsuperscript{11} Then the coordinates

\textsuperscript{11}Strictly, the definition only ensures that infinitesimally neighboring geodesics do not intersect. We assume that $\Sigma_0$ is chosen so as to avoid nonlocal intersections (between geodesics with distinct starting points on $\Sigma_0$). We expect that this is generic due to the inflationary expansion, and in particular that it is satisfied for $\Sigma_0$ in the attractor regime of the New Scale Factor Cutoff. In structure forming regions, we expect that caustics occur before nonlocal intersections. If not, then some regions may be multiply counted \textsuperscript{20}; this would not affect the duality.
CHAPTER 1. GLOBAL-LOCAL DUALITIES IN ETERNAL INFLATION

The four-volume element is \( d^3x_0 d\tau \sqrt{h_0} \). Because the geodesics in the volume element \( d^3x_0 \sqrt{h_0} \) expand along with the congruence, the correct four-volume element is

\[
d^4x \sqrt{g} = d^3x_0 d\tau \sqrt{h_0} e^{3\eta}.
\]

This follows from the definition of expansion and scale factor parameter, Eqs. (1.27) and (1.28).

Returning to the event count, we can now write Eq. (1.33) as an integral over the spacetime region \( V_\eta(\eta > 0) \) traced out by the congruence of Short Fat Geodesics:

\[
N_{FG}^I = \int_{V_\eta(\eta > 0)} d^4y \sqrt{g(y)} e^{-3\eta} D_I(y) .
\]

The weighting factor \( e^{-3\eta} \) can be understood intuitively as the number of fat geodesics that overlap at each spacetime point; see Fig. 1.2.

In particular, we may choose the scale factor parameter \( \eta \) as a coordinate. However, the coordinate is one-to-one only if we restrict to the expanding or the collapsing portion of each geodesic. Thus, we may write

\[
N_{FG}^I = N_{FG,^+}^I + N_{FG,-}^I ,
\]

where

\[
N_{FG,^\pm}^I = \int_{0}^{\infty} d\eta' e^{-3\eta'} \int_{\Sigma_{\eta'}} d^3x \sqrt{g(\eta',x)} D_I(\eta',x) .
\]

Here, \( \Sigma_{\eta'}^\pm \) are hypersurfaces of constant scale factor parameter \( \eta' \) in the expanding (+) or contracting (−) portion of the congruence.

1.3.2 Proof of Equivalence to the New Scale Factor Cutoff

We now turn to the global side of the duality. Again, we consider the congruence of geodesics orthogonal to an initial surface \( \Sigma_0 \). The New Scale Factor Cutoff measure is defined as

\[
\frac{P_{SF}^I}{P_{SF}^J} = \lim_{\eta \to \infty} \frac{N_I(\eta)}{N_J(\eta)} ,
\]

where \( N_I(\eta) \) is the number of events of type \( I \) that have taken place in the spacetime regions with scale factor parameter less than \( \eta \)[20]. Using Eq. (1.30), we may write this as

\[
N_I(\eta) = \int_{M(\eta)} d^4x \sqrt{g} D_I(x) .
\]

The integral runs over the spacetime four-volume \( M(\eta) \), defined as the set of points that lie in the future \( \Sigma_0 \) (on which we set \( \eta = 0 \)) and whose scale factor time, Eq. (1.27), is less than
η. In order to make the assignment of a scale factor parameter to every spacetime point
unique, each geodesic is terminated immediately prior to caustic points, when neighboring
geodesics intersect.

To compute the probabilities defined by the scale factor time cutoff, we note that the
cosmological dynamics of eternal inflation leads to an attractor regime [20]:
\[ N_I(\eta) = \bar{N}_I e^{\gamma \eta} + O(e^{\phi \eta}) , \]
where \( \phi < \gamma < 3 \). This will be reviewed in the next subsection; for now, we will simply use
this result. With Eq. (1.39), it implies that the scale factor time probabilities are given by
\[ \frac{P_{SF}^I}{P_{SF}^J} = \frac{\bar{N}_I}{\bar{N}_J} . \]

The Short Fat Geodesic measure can also be evaluated using Eq. (1.41), if initial condi-
tions on \( \Sigma_0 \) are chosen to lie in the attractor regime. A suitable \( \Sigma_0 \) can be constructed as
as a late-time hypersurface orthogonal to the congruence constructed from a much earlier,
arbitrary initial hypersurface, and resetting \( \eta \rightarrow 0 \) there. The proof will exploit the fact
that the Short Fat Geodesic probabilities, Eqs. (1.37) and (1.38), involve an integral over
a spacetime set closely related to \( M(\eta) \), reweighted relative to Eq. (1.40) by a factor that
depends only on \( \eta \) and thus does not change relative probabilities in the attractor regime.
The set $M(\eta)$ will contain one connected expanding region, $M^+(\eta)$, bounded from below by $\Sigma_0$, in which scale factor time is growing towards the future. In any model with collapsing regions (structure forming regions or crunches), $M(\eta)$ will also contain infinitely many mutually disconnected collapsing regions inside bubbles near the future conformal boundary of the spacetime. (The total contribution to the measure from such regions to the New Scale Factor Cutoff measure is finite at any finite value of $\eta$.) We denote the union of all collapsing regions by $M^-(\eta)$.

Let us split the integral in Eq. (1.40) into expanding and contracting portions:

$$N_I(\eta) = N_I^+(\eta) + N_I^-(\eta), \quad (1.43)$$

where

$$N_I^\pm(\eta) \equiv \int_{M^\pm(\eta)} d^4x \sqrt{|g|} D_I(x). \quad (1.44)$$

Since this division depends only on local properties, each portion has its own attractor solution:

$$N_I^\pm(\eta) = \bar{N}_I^\pm e^{\gamma \eta} + O(e^{\phi \eta}), \quad (1.45)$$

with $\bar{N}_I^+ + \bar{N}_I^- = \bar{N}_I$. This will be shown explicitly in the following subsection. In each portion, the scale factor parameter is monotonic along the geodesics in the congruence, and we may use it as an integration variable:

$$N_I^+(\eta) = \int_{0}^{\eta} d\eta' \int_{\Sigma_{\eta'}^+} d^3x \sqrt{g(\eta',x)} D_I(\eta',x); \quad (1.46)$$

$$N_I^-(\eta) = \int_{-\infty}^{\eta} d\eta' \int_{\Sigma_{\eta'}^-} d^3x \sqrt{g(\eta',x)} D_I(\eta',x), \quad (1.47)$$

where the hypersurface $\Sigma_{\eta'}^\pm$ consists of the points with fixed $\eta'$ in the expanding (+) or contracting (−) region. Therefore

$$\frac{dN_I^\pm}{d\eta} = \int_{\Sigma_{\eta}^\pm} d^3x \sqrt{g(\eta,x)} D_I(\eta,x). \quad (1.48)$$

so we may rewrite Eq. (1.38) as

$$N_I^{FG,\pm} = \int_{0}^{\infty} d\eta' e^{-3\eta'} \frac{dN_I^\pm}{d\eta'}. \quad (1.49)$$

We can now use the attractor solutions for the New Scale Factor Cutoff, Eq. (1.45), to
evaluate the Short Fat Geodesic cutoff:

\[
N_{I}^{FG} = \int_{0}^{\infty} d\eta' e^{-3\eta' \left(\frac{dN_{I}^{+}}{d\eta'} + \frac{dN_{I}^{-}}{d\eta'}\right)} \tag{1.50}
\]

\[
= \int_{0}^{\infty} d\eta' e^{\gamma(\gamma-3)\eta' \gamma} \tilde{N}_{I} \tag{1.51}
\]

\[
= \frac{\gamma}{3-\gamma} \tilde{N}_{I} . \tag{1.52}
\]

The prefactor is \( I \)-independent, so relative probabilities in the Short Fat Geodesic measure are given by

\[
P_{I}^{FG} = \tilde{N}_{I} , \tag{1.53}
\]

This agrees with the New Scale Factor Cutoff probabilities, Eq. (1.42). Therefore, the two measures are equivalent.

This result is somewhat counterintuitive. If the Short Fat Geodesic is defined for regions where \( \eta > 0 \), its global dual should be the New Scale Factor Cutoff not as defined above, but restricted to \( \eta > 0 \) both in the expanding and collapsing regions. In fact, it is. Both versions of the New Scale Factor Cutoff, with and without this additional restriction in the collapsing regions, are dual to the Short Fat Geodesic, because both have the same attractor regime. It is important to distinguish between \( \eta < 0 \) regions and decreasing-\( \eta \) regions in the global cutoff. The former quickly become unimportant in the attractor regime; the latter are always important. For the cumulative quantity \( N_{I}(\eta) \) to be exactly in the attractor regime, it would be necessary to restrict to \( \eta > 0 \) on both the expanding and collapsing side, and thus to exclude a few collapsing regions; otherwise, there may be a small transient of order \( e^{\phi_{\eta}} \) from those initial collapsing regions that have \( \eta < 0 \). However, the duality relies on evaluating the local measure by integrating up the global “derivative” \( dN_{I}/d\eta \) with weighting \( e^{-3\eta} \). Since the calculation makes reference only to the derivative in the region \( \eta > 0 \), the regions with \( \eta < 0 \) do not enter into the duality.

### 1.3.3 New Scale Factor Cutoff Rate Equation and Attractor Solution

In this subsection we derive the attractor solution, Eq. (1.41), starting from the rate equation for the New Scale Factor Cutoff. We will follow [20] and use the seminal results of [40]. As for the case of light-cone time, we will proceed in two steps. We first consider the rate equation for de Sitter vacua; then we include the detailed consequences of decays within this distribution, and explain how to treat collapsing regions.

Naively, the rate equation should follow from the result for light-cone time, Eq. (1.9), by an appropriate substitution. In empty de Sitter space, \( \theta/3 = \tau_{\Lambda,\alpha}^{-1} \), so Eqs. (1.27) and
\[ (1.17) \text{ imply that } dt_{\text{LC}} = d\eta \text{ in this regime. Setting } dt_c \rightarrow d\eta \text{ in Eq. (1.9), however, yields an incorrect equation:} \\
\frac{dn_\alpha}{d\eta} = (3 - \kappa_\alpha)n_\alpha + \sum_\beta \kappa_{\alpha\beta}n_\beta \tag{1.54} \]

The last term on the right hand side is incorrect. In the light-cone time rate equation, this term arose from the square bubble approximation. It is an effective term that anticipates the asymptotic size of bubbles of new vacua instead of treating their growth in detail. It subsumes, in particular, the cumulative effects of the early era within a new bubble (less than \( \tau_{\Lambda,\alpha} \) after nucleation). During this era the relation \( dt_{\text{LC}} = d\eta \) does not hold, so the substitution that led to Eq. (1.54) is unjustified. Another way of saying this is that the square bubble approximation is a different procedure for different time variables.

The correct rate equation for the New Scale Factor Cutoff contains an extra factor of \( v_\alpha/v_\beta \) in the final sum, where \( v_\alpha \) is the proper volume of a horizon patch of type \( \alpha \). It thus takes a particularly simple form,

\[ \frac{dV_\alpha}{d\eta} = (3 - \kappa_\alpha)V_\alpha + \sum_\beta \kappa_{\alpha\beta}V_\beta, \tag{1.55} \]

when expressed in terms of the proper volumes \( V_\alpha \) occupied by metastable de Sitter vacua at scale factor time \( \eta \), instead of the number of horizon patches \( n_\alpha = V_\alpha/v_\alpha \). More generally, one finds that the rate equation takes the above form, with \( V \rightarrow X \), \( \eta \rightarrow T \), if \( T \) measures the growth of the overall volume of space in units of \( X \). For example, scale factor parameter measures the growth of proper volume \( (T = \eta, X = 1) \) and light-cone time measures the growth of volume in units of horizon volume \( (T = t_{\text{LC}}, X = v_\alpha) \).

To derive Eq. (1.55), we note that the first two terms on the right hand side follow from the arguments given for the analogous terms in Sec. 1.3.3. They would also follow from Eq. (1.16) by substituting \( dt_{\text{LC}} \rightarrow d\eta \) and using \( V_\alpha = n_\alpha v_\alpha \); but the third term, as explained above, cannot be so obtained. It must be derived from a first principle argument identical to that given in Sec. 1.3.3; except that it is now the continuity of New Scale Factor parameter, not light-cone time, that must be ensured when a new bubble is formed. This means that instead of requiring that the number of horizon patches of \( \beta \)-vacuum lost must equal the number of horizon patches of \( \alpha \)-vacuum gained in \( \beta \rightarrow \alpha \) transitions, we now require that the proper volume of \( \beta \)-vacuum lost must equal the proper volume of \( \beta \)-vacuum gained. The amount lost in \( \beta \), per nucleation of \( \alpha \), is always one horizon volume of \( \beta \); this follows from causality. In the rate equation in terms of scale factor parameter, this must be converted into proper volume, and the same proper volume must be assigned to the new vacuum, \( \alpha \). This leads to the final term in Eq. (1.55). It also explains why the New Scale Factor rate equation looks simplest in terms of proper volume.

The solution of the rate equation for New Scale Factor time can be obtained from
Eq. (1.18) by substituting $t_{LC} \rightarrow \eta$ and $n \rightarrow V$:

$$V_\alpha(\eta) = \bar{V}_\alpha e^{\gamma \eta} + O(e^{\phi \eta}) .$$ (1.56)

That is, we must set $\bar{V}_\alpha$ (and not, as for light-cone time, $\bar{n}_\alpha$), equal to the dominant eigenvector of the transition matrix $M_{\alpha \beta}$. As before $\gamma = 3 - q$ is the largest eigenvalue. Note that this eigenvector and the dominant vacuum $\star$ are exactly the same as in the case of light-cone time. On a slice of constant light-cone time, the $\star$ vacuum dominates the number of horizon patches; on a slice of constant New Scale Factor Cutoff, it dominates the volume.

It is convenient to define

$$\bar{n}_\alpha \equiv \frac{\bar{V}_\alpha}{v_\alpha} .$$ (1.57)

The number of horizon patches of type $\alpha$ at New Scale Factor parameter $\eta$ obeys

$$n_\alpha(\eta) = \bar{n}_\alpha e^{\gamma \eta} + O(e^{\phi \eta}) .$$ (1.58)

We now derive $N_I(\eta)$. The procedure will be slightly different from the one in Ref. [20] in that we will keep expanding and collapsing regions explicitly separated in all expressions. $N_I(\eta)$ receives a contribution from the expanding ($+$) regions, and one from the contracting ($-$) regions, $N_I(\eta) = N_I^+(\eta) + N_I^-(\eta)$. Analogous to Eq. (1.21) for the Lightcone Time Cutoff, one finds for the New Scale Factor Cutoff:

$$N_I^+(\eta) = \sum_{i \neq \star} \int_0^\eta \left( \frac{dN_i^+}{d\eta_i} \right)_{\eta - \eta_{\text{huc}}} \left( \frac{dN_i}{d\eta} \right)_{\eta_{\text{huc}}} d\eta_{\text{huc}} ,$$ (1.59)

and

$$N_I^-(\eta) = \sum_{i \neq \star} \int_{-\infty}^\eta \left( \frac{dN_i^-}{d\eta_i} \right)_{\eta - \eta_{\text{huc}}} \left( \frac{dN_i}{d\eta} \right)_{\eta_{\text{huc}}} d\eta_{\text{huc}} .$$ (1.60)

As in the previous section, we can now change the integration variable to $\zeta = \eta - \eta_{\text{huc}}$ in Eqs. (1.59) and (1.60), and use (1.58) to get

$$N_I^\pm(\eta) = \bar{N}_I^\pm e^{\gamma \eta} + O(e^{\phi \eta}) ,$$ (1.61)

where

$$\bar{N}_I^+ \equiv \sum_{i \neq \star} \sum_{\beta} N_{i\beta}^+ \bar{n}_\beta ,$$ (1.62)

and

$$\bar{N}_I^- \equiv \sum_{i \neq \star} \sum_{\beta} N_{i\beta}^- \bar{n}_\beta ,$$ (1.63)

with

$$N_{i\beta}^+ \equiv \int_0^\infty d\zeta e^{-\gamma \zeta} \left( \frac{dN_i^+}{d\eta_i} \right)_{\zeta} ,$$ (1.64)
and
\[ N_i^- \equiv \int_{-\infty}^{\infty} d\zeta e^{-\gamma \zeta} \left( \frac{dN_I^-}{dN_i^-} \right)_\zeta. \] (1.65)

Again, (1.64) and (1.65) converge because all vacua decay faster than the dominant vacuum. This yields Eq. (1.41), with \( \bar{N}_I = \bar{N}_I^- + \bar{N}_I^+ \).

1.4 General Global-Local Dualities

It is easy to generalize the duality studied in the previous section, between the Short Fat Geodesic cutoff and the New Scale Factor Cutoff. On the local side, the fatness of the geodesic can be allowed to vary along the geodesic. On the global side, this corresponds to a different choice of time variable. In this section, we mostly consider a generalization that preserves the key feature that the fatness of the geodesic does not depend explicitly on the time along the geodesic, but only on local features. This restriction defines a family of measures that include the scale factor cutoff, the light-cone time cutoff, and the CAH cutoff as special cases. In the final subsection, we consider a further generalization that we exemplify by deriving a local dual to the proper time cutoff.

1.4.1 \(X\)-fat Geodesic Measure

Consider a family of geodesics orthogonal to an initial hypersurface \( \Sigma_0 \). We assign each of these geodesics a cross sectional volume, \( X \). Intuitively, we may picture \( X \) as modulating the infinitesimal fatness of the geodesic. Equivalently, \( X \) can be thought of as a weighting factor that allows events of type \( I \) to contribute differently to the probability for \( I \), depending on where they are encountered. We require that \( X \) be everywhere nonnegative to ensure that probabilities are nonnegative. We will assume, for now, that \( X \) depends only on local properties of the congruence, such as the expansion, the shear, and their derivatives:
\[ X = X(\theta, \sigma^{ab}\sigma_{ab}, \frac{d\theta}{d\tau}, \ldots). \] (1.66)

A simple example, which we will consider explicitly in Sec. 1.4.4, is the Hubbletube. It is obtained by setting \( X \) to the local Hubble volume. With \( X \equiv 1 \), the \( X \)-fat geodesic reduces to the ordinary fat geodesic. In Sec. 1.4.5 we will consider further generalizations, in which \( X \) is not restricted to a local function of congruence parameters.

Along each geodesic, we may define the \( T \) parameter:
\[ T \equiv \eta - \frac{1}{3} \log X \] (1.67)
CHAPTER 1. GLOBAL-LOCAL DUALITIES IN ETERNAL INFLATION

Geometrically, $e^{3T}$ is the factor by which a volume element has expanded along the congruence, in units of the volume $X$. Every $X$-Fat Geodesic will be restricted to values of the $T$ parameter larger than that at $\Sigma_0$, which we may choose to be zero.\footnote{Strictly, this should be called the short $X$-fat geodesic: as in Sec. 1.3.1 we will be restricting the congruence to regions where its density (in units of the local fatness, $X$) is below its initial value. This ensures the broadest possible applicability of the duality we derive. By including the entire future-directed geodesic irrespective of this conditions, one could consider a “long” $X$-fat geodesic. This local measure would not generally have a natural global dual.}

We would like to compute probabilities using our new local cut-off, the $X$-fat geodesic, by modifying the definition of the fat geodesic, Eq. (1.33):

$$N_{XG}^I = \int_{\Sigma_0} d^3x_0 \sqrt{h_0} \int_{T > 0} d\tau X(x(\tau)) D_I(x(\tau)).$$

Relative probabilities given by

$$P_{XG}^I / P_{XG}^J = N_{XG}^I / N_{XG}^J.$$ 

We now follow the steps leading to Eq. (1.36) in Sec. 1.3.1. Assuming that the geodesics do not intersect, Eq. (1.68) can be rewritten as an standard integral over the four-volume encountered by the congruence:

$$N_{XG}^I = \int_{V_4(T > 0)} d^4y \sqrt{g(y)} e^{-3\eta Y(y)} D_I(y).$$

In particular, we can pick $T$ as a coordinate. Like the scale factor time, $T$ in Eq. (1.67) is defined for every point on the nonintersecting congruence. Multiple points along the same geodesic may have the same $T$; this will not be a problem. However, the coordinate is one-to-one only if we restrict to the “expanding” or the “contracting” portion of each geodesic.\footnote{In this section, “expanding” and “contracting” regions are defined with respect to the $X$ volume. An expanding/contracting region will be one where $T$ increases/decreases.}

Thus, in terms of $T$, Eq. (1.70) becomes

$$N_{XG}^I = N_{XG,+}^I + N_{XG,-}^I,$$

where

$$N_{XG,\pm}^I = \int_0^\infty dT e^{-3T} \int_{\Sigma_{T'}^{\pm}} d^3x \sqrt{g(T', x)} D_I(T', x).$$

Here, $\Sigma_{T'}^{\pm}$ are hypersurfaces of constant $T$ parameter $T'$ in the expanding ($+$) or contracting ($-$) portion of the congruence.
1.4.2 Proof of Equivalence to the $T$-cutoff Measure

Let us consider the time variable $T$ defined in Eq. (1.67) as a global cutoff. Probabilities are defined by

$$P_T^I = \lim_{T \to \infty} \frac{N_I(T)}{N_J(T)},$$

where $N_I(T)$ is the number of events of type $I$ that take place in spacetime regions with time less than $T$. Because $T$ need not be monotonic along every geodesic, such regions may not be connected. As shown in Sec. 1.3.3, this does not affect the proof of equivalence, which proceeds as in Sec. 1.3. Again, $N_I(T)$ receives a contribution from expanding (+) and contracting (-) regions, $N_I(T) = N_I^+(T) + N_I^-(T)$.

In terms of the distribution $D$, we have

$$N_I^+(T) = \int_0^T dT' \int_{\Sigma_{T'}} d^3x \sqrt{g(T', x)} D_I(T', x);$$  (1.74)

$$N_I^-(\eta) = \int_{-\infty}^T dT' \int_{\Sigma_{T'}} d^3x \sqrt{g(T', x)} D_I(T', x),$$  (1.75)

and therefore

$$\frac{dN_I^\pm}{dT} = \int_{\Sigma_T^\pm} d^3x \sqrt{g(T, x)} D_I(T, x).$$  (1.76)

We make use of the attractor solution

$$N_I(T) = \tilde{N}_I e^{\gamma T} + O(e^{\phi T}),$$  (1.77)

where $0 < \phi < \gamma < 3$ (see the following subsection). With Eq. (1.73), it implies that the $T$-cutoff probabilities are given by

$$\frac{P_I^T}{P_J^T} = \frac{\tilde{N}_I}{N_J}. $$  (1.78)

The $X$-fat geodesic probabilities are also determined by Eq. (1.77), if initial conditions on $\Sigma_0$ are chosen to lie in the attractor regime. Plugging Eq. (1.76) into Eq. (1.72), and then using Eq. (1.71), we get

$$N_I^{XG} = \int_0^\infty dT' e^{-3T'} \frac{dN_I}{dT'},$$  (1.79)

In the attractor regime, by Eq. (1.77), one has

$$\frac{dN_I}{dT'} = \gamma \tilde{N}_I e^{\gamma T'}.$$  (1.80)
Substituting into Eq. (1.79) and using $\gamma < 3$, the integral converges to an $I$-independent constant. Thus, relative probabilities in the $X$-fat geodesic measure are given by

$$P^X_I = \tilde{N}_I \tilde{N}_J .$$

(1.81)

This agrees with the $T$-cutoff probabilities, Eq. (1.78). Therefore, the two measures are equivalent.

### 1.4.3 $T$-cutoff Rate Equation and Attractor Solution

In this subsection, we derive the rate equation for the number of horizon patches of de Sitter vacua $\alpha$ as a function of $T$, and the attractor solution, Eq. (1.77). For the rate equation, we treat all de Sitter vacua as empty at all times. We use the square bubble approximation which treats each bubble as comoving in the congruence at its asymptotic size.

Let $x_\alpha$ be the asymptotic value of $X$ in the vacuum $\alpha$. $X$ will converge rapidly to $x_\alpha$ in empty de Sitter regions because, by assumption, $X$ depends only on local properties of the congruence. By Eq. (1.67) this implies that $dT = d\eta$ in such regions. Thus, the rate equation is

$$\frac{dn_\alpha}{dT} = (3 - \kappa_\alpha)n_\alpha + \sum_\beta \kappa_{\alpha\beta} n_\beta \frac{v_\beta x_\alpha}{v_\alpha x_\beta} .$$

(1.82)

The term $3n_\alpha$ captures the exponential growth of the number of horizon patches, which goes as $e^{3n}$. The term $-\kappa_\alpha n_\alpha$ captures the decay of vacuum $\alpha$, per unit horizon patch and unit scale factor time in empty de Sitter space.

The last term captures the creation of new regions of vacuum $\alpha$ by the decay of other vacua. In the square bubble approximation, one horizon patch of $\beta$ is lost when an $\alpha$-bubble forms in $\beta$ (see Sec. 1.2.3). Thus, $v_\beta/x_\beta$ $X$-patches of $\beta$-vacuum are lost, where $v_\beta$ is the volume of one horizon patch of $\beta$. Continuity of the time variable $T$ requires that the number of patches of size $X$ be continuous, so $v_\beta/x_\beta$ $X$-patches of $\alpha$ vacuum must be added. One $X$-patch of $\alpha$ vacuum equals $x_\alpha/v_\alpha$ horizon patches of $\alpha$ vacuum. Thus, the total number of horizon patches of $\alpha$-vacuum that are created per $\beta$-decay in the square bubble approximation is $v_\beta x_\alpha/v_\alpha x_\beta$. The number of such decays in the time interval $dT$ is $\kappa_{\alpha\beta} n_\beta dT$. This completes our derivation of the last term.

When expressed in terms of the number of $X$-patches,

$$n^X_\alpha = n_\alpha v_\alpha = \frac{V_\alpha}{x_\alpha} ,$$

(1.83)

the rate equation takes a very simple form:

$$\frac{dn^X_\alpha}{dT} = (3 - \kappa_\alpha)n^X_\alpha + \sum_\beta \kappa_{\alpha\beta} n^X_\beta .$$

(1.84)
This form is identical to that of Eqs. (1.16) and (1.55), and the general results of Ref. [40] apply. The late-time solution is again determined by the dominant eigenvalue, $\gamma = 3 - q$, of the transition matrix $M_{\alpha\beta}$, and by the associated eigenvector, which we now label $\tilde{n}^X_\alpha$:

$$n^X_\alpha(T) = \tilde{n}^X_\alpha e^{\gamma T} + O(e^{\phi T}) .$$  \hfill (1.85)

Next, we compute number of events of type $I$ prior to the time $T$. With

$$\tilde{n}_\alpha \equiv \frac{\tilde{n}^X_\alpha x_\alpha}{v_\alpha} ,$$  \hfill (1.86)

the number of horizon patches of type $\alpha$ at time $T$ obeys

$$n_\alpha(T) = \tilde{n}_\alpha e^{\gamma T} + O(e^{\phi T}) .$$  \hfill (1.87)

The remainder of the analysis is completely analogous to Sec. 1.3.3. When we include collapsing (i.e. decreasing $T$) regions at the future of $\Sigma_0$, we still obtain an attractor regime. Like in the New Scale Factor case [20], the corresponding $\Delta T_- = T_{\text{max}} - T_{\text{nuc}}$ and $\Delta T_+ = T_{\text{max}} - T_{\text{nuc}}$ will only depend on local physics in each bubble universe but not on $T_{\text{nuc}}$. This holds because we are assuming that $X$ only depends on local properties of the congruence. Therefore, $\Delta T_-$ and $\Delta T_+$ will increase by the same finite amounts during expansion and collapse phases in a particular pocket universe, no matter when the bubble universe is nucleated. There will be infinitely many collapsing regions at the future of $\Sigma_0$, but a finite number of bubbles contribute, namely the ones that formed before the time $T + \Delta T_{\text{sup}}$, where $\Delta T_{\text{sup}} \equiv \min\{0, \sup_0 (\Delta T_- - \Delta T_+)\}$

$$N_I(T) = \kappa_I n_s + \sum_{i \neq *} \int_0^{T + \Delta T_{\text{sup}}} \left( \frac{dN_I}{dN_i} \right)_{T_{\text{nuc}}} \left( \frac{dN_i}{dT} \right)_{T_{\text{nuc}}} dT_{\text{nuc}} .$$  \hfill (1.88)

Again, we conclude

$$N_I(T) = \tilde{N}_I e^{\gamma T} + O(e^{\phi T}) ,$$  \hfill (1.89)

where

$$\tilde{N}_I \equiv \kappa_I n_s + \sum_{i \neq *} \sum_\beta N_{i\beta} \tilde{n}_\beta ,$$  \hfill (1.90)

and

$$N_{i*} \equiv \int_{-\Delta T_{\text{sup}}}^{\infty} d\zeta e^{-\gamma \zeta} \left( \frac{dN_{i*}}{dN_i} \right) \zeta .$$  \hfill (1.91)

As in [20], this integral converges because all vacua decay faster than the dominant vacuum, and one obtains the same attractor behavior.

\footnote{Here, $T_{\text{f}}$ corresponds to the $T$ value at the regulated endpoint of the geodesic. Geodesics are terminated at some cutoff, for example one Planck time before they reach a point where $T \to -\infty$. The choice of cutoff depends on the definition of $T_i$; see the discussion in the next subsection. On the other hand, $T_{\text{max}}$ is the maximum $T$-value reached by the geodesic.}
1.4.4 The Hubbletube and the CAH measure

An example of particular interest is the Hubbletube: the $X$-fat geodesic whose fatness is proportional to the local Hubble volume $v_H$, as measured by the expansion of the congruence:

$$X \propto v_H = \frac{4\pi}{3} \left( \frac{3}{\theta} \right)^3. \quad (1.92)$$

Since constant numerical factors drop out of all relative probabilities, we simply set

$$X \equiv \theta^{-3}. \quad (1.93)$$

This measure is dual to a global cutoff at constant $T$, where

$$T \equiv \eta + \log \theta. \quad (1.94)$$

Equivalently, the global cutoff surfaces can be specified in terms of any monotonic function of $T$, e.g. $\exp(T)$. Note that

$$e^T = \theta a = \frac{da}{d\tau}, \quad (1.95)$$

where $a \equiv e^\eta$ is the scale factor and $\tau$ is proper time along the congruence. We thus recognize the global dual of the Hubbletube as Vilenkin's CAH-cutoff [32].

Naively, the CAH-cutoff is well-defined only in regions with accelerating expansion: $\ddot{a} > 0$, where the time variable $T$ increases monotonically along the geodesics. In this regime, the duality with the Hubbletube is obvious. But this regime is also extremely restrictive: it excludes not only gravitationally bound regions such as our galaxy, but also all regions in which the expansion is locally decelerating, including the homogeneous radiation and matter-dominated eras after the end of inflation in our vacuum.

However, if geodesics are terminated before caustics, the CAH cutoff can instead be defined as a restriction to a set of spacetime points with $T$ less than the cutoff value. This is similar to the transition from the old to the new scale factor measure: in the spirit of Ref. [20], one abandons the notion of $T$ as a time variable. In the case of the CAH parameter $T$, an infinite number of decelerating regions will be included under the cutoff for any finite $T$.

This possibility of increasing the regime of applicability of the CAH cutoff is particularly obvious from the local viewpoint. The local measure requires only $\theta > 0$ for positive fatness; this is strictly weaker than $\ddot{a} > 0$. It still excludes collapsing regions, but not regions undergoing decelerating expansion.

On either side of the duality, geodesics must be terminated at some arbitrarily small but finite proper time before they reach turnaround ($\theta = 0$), where $T \to -\infty$. Otherwise, events at the turnaround time receive infinite weight. This is needed only for finiteness; it eliminates an arbitrarily small region near the turnaround from consideration but does not affect other relative probabilities. However, this marks an important difference to the Short
Fat Geodesic and the New Scale Factor measure, where no additional cutoff near $\eta \to -\infty$ was needed.

In any case, the restriction to regions with $\theta > 0$ is necessary to make the Hubbletube well-defined. Unfortunately, this restriction is too strong to yield a useful measure since it excludes gravitationally bound regions like our own. Unlike in the case of the New Scale Factor Cutoff or the Causal Patch, there are thus large classes of regions to which the CAH cutoff cannot be applied. Additional rules must be specified, such as the CAH+ measure of Ref. \[32\].

### 1.4.5 The Incredible Shrinking Geodesic and the Proper Time Cutoff

The global proper time cutoff is defined as a set of points that lie on a geodesic from $\Sigma_0$ with proper length (time duration) less than $\tau$ along the geodesic \[10-14\]. (To make this well-defined, we terminate geodesics at the first caustic as usual, so that every point lies on only one geodesic.) Relative probabilities are then defined as usual, in the limit as the cutoff is taken to infinity.

The rate equation for the number of de Sitter horizon patches, in terms of proper time, is

$$
\frac{dn_\alpha}{d\tau} = (3 - \kappa_\alpha)H_\alpha n_\alpha + \sum_\beta \kappa_\alpha\kappa_\beta H_\beta n_\beta = M_{\alpha\beta} n_\beta ,
$$

(1.96)

where the transition matrix is given by

$$
M_{\alpha\beta} = (3 - \kappa_\alpha)H_\alpha \delta_{\alpha\beta} + \kappa_\alpha H_\beta .
$$

(1.97)

This differs from the transition matrix in all previous examples by the appearance of the Hubble constants of the de Sitter vacua, and so it will not have the same eigenvector and eigenvalues; it will have a completely different attractor regime. Instead of Planck units, it will be convenient to work in units of the largest Hubble constant in the landscape, $H_\alpha \to H_\alpha/H_{\text{max}}$ and $\tau \to H_{\text{max}}\tau$. We note that $H_{\text{max}}^{-1}$ is necessarily a microscopic timescale in any model where our vacuum contains a parent vacuum whose decay is sufficient for a reheat temperature consistent with nucleosynthesis. In the string landscape, one expects $H_{\text{max}}$ to be of order the Planck scale.

Due to the smallness of decay rates and the large differences in the Hubble rate between Planck-scale vacua ($H \sim 1$) and anthropic vacua ($H \ll 1$), we expect that again the largest eigenvalue is very close to the largest diagonal entry in the transition matrix, and that the associated eigenvector is dominated by the corresponding vacuum. In all previous examples, the dominant vacuum, $\ast$, was the longest-lived de Sitter vacuum. The associated eigenvalue was $\gamma \equiv 3 - \kappa_\ast$, where $\kappa_\ast$ is the decay rate of the $\ast$ vacuum. Now, however, the Hubble constant of each de Sitter vacuum, $H_\alpha$, is the more important factor. The dominant vacuum,
*, will be the fastest-expanding vacuum, i.e., the vacuum with the largest Hubble constant, which in our unit conventions is \( H_{\text{max}} = 1^{15} \). In the same units, the associated eigenvalue is again \( \gamma = 3 - \kappa_\alpha \).

By decay chains, the dominant expansion rate \( \gamma \) drives both the growth rate of all other vacuum bubbles and all types of events, \( I \), at late times:

\[
N_I = \tilde{N}_I e^{\gamma \tau} + O(e^{\tau})
\]

where \( \varphi < \gamma < 3 \). Relative probabilities are given as usual by

\[
P_I \quad P_J = \frac{\tilde{N}_I}{\tilde{N}_J}.
\]

The proper time measure famously suffers from the youngness problem \([24, 43-48]\), or “Boltzmann babies” \([49]\). Typical observers are predicted to be thermal fluctuations in the early universe, and our own observations have probability of order \( \exp(-10^{60}) \). This holds in any underlying landscape model as long as it contains our vacuum. Thus the proper time measure is ruled out by observation at very high confidence level.

Explaining the origin of the youngness problem is somewhat convoluted in the global picture. Consider an event that occurs at 13.7 Gyr after the formation of the bubble universe it is contained in and that is included under the cutoff. For every such event, there will be a double-exponentially large number \( \exp(3H_{\text{max}} \Delta \tau) \) of events in the same kind of bubble universe that occur at 13.7 Gyr \(-\Delta \tau\) after the formation of the bubble. This is because new bubbles of this type are produced at an exponential growth rate with characteristic time scale \( H_{\text{max}} \). We will now show that the proper time cutoff has a local dual, the Incredible Shrinking Geodesic, in which the youngness problem is immediately apparent.

We now seek a local dual, i.e., a geodesic with fatness (or local weight) \( X(\tau) \), which will reproduce the same relative probabilities if initial conditions are chosen in the dominant (i.e., fastest-expanding) vacuum. To find the correct fatness, we invert Eq. (1.67):

\[
X(\tau) = e^{-3(\tau - \eta)} = \exp \int_0^\tau d\tau' [\theta(\tau') - 3] .
\]

Note that this result does not satisfy the constraint we imposed in all previous subsections, that the geodesic has constant fatness in asymptotic de Sitter regions.

Obtaining a local dual in this manner is somewhat brute-force. Recall that the duality relies on fact that the overcounting of events by overlapping fat geodesics depends only on the global time. Here this is accomplished in two steps. The factor \( e^{3\eta} \) undoes the dilution of geodesics: it fattens the geodesics by their inverse density, thus making the overcounting factor everywhere equal to one. The factor \( e^{-3\tau} \) is a regulator that depends only on the global time and renders the integral in Eq. (1.68) finite.

\[15\text{More precisely, the dominant vacuum will be the vacuum with largest } (3 - \kappa_\alpha)H_\alpha.\]
However, the result for $X(\tau)$ immediately makes the youngness problem apparent: note that $X$ is constant as long as the geodesic remains in the fastest expanding de Sitter vacuum, where $\theta = 3H_{\text{max}} = 3$ (see Fig. 1.3). However, in all other regions, $H < 1$, so $\theta - 3 < 0$ and the weight of events is suppressed exponentially as a function of the time after the decay of the dominant vacuum. In particular, in anthropically allowed regions, such as ours, the Hubble rate is very small compared to the microscopic rate $H_{\text{max}} = 1$. Thus, events are approximately suppressed as $e^{-3\tau}$, that is, exponentially with a microscopic characteristic timescale. For example, with $H_{\text{max}}$ of order the Planck scale, we thus find that events today are less likely than events yesterday by a factor of $\exp(-10^{48})$, and less likely than a billion years ago by a factor of $\exp(-10^{60})$. As a consequence, this measure assigns higher probability to (conventionally unlikely) observers arising from large quantum fluctuations in the early universe (and their bizarre observations) than to our observations \cite{24,43,49}. 
Figure 1.3: The Incredible Shrinking Geodesic. This is not a conformal diagram; the true proper fatness of the geodesic is shown as a function of proper time, $\tau$. As long as the geodesic remains in the dominant vacuum, its fatness is constant, i.e., it assigns the same weight to all events it encounters. In any other region, its fatness decreases exponentially with microscopic characteristic timescale of order the expansion rate of the dominant vacuum, $H_{\text{max}}$. Therefore, events occurring later than a few units of $H_{\text{max}}$ after the decay of the dominant vacuum have negligible probability. This includes our own observations, so the measure is ruled out.
Chapter 2

Anthropic Origin of the Neutrino Mass?

2.1 Introduction

In a theory with a large multidimensional potential landscape \[28\], the smallness of the cosmological constant can be anthropically explained \[30\]. The lack of a viable alternative explanation for a small or vanishing cosmological constant, the increasing evidence for a fine-tuned weak scale, and several other complexity-favoring coincidences and tunings in cosmology and the Standard Model, all motivate us to consider landscape models seriously, and to extract further pre- or post-dictions from them.

A large landscape can also explain an aspect of the Standard Model that has long remained mysterious: the origin of the masses and mixing angles of the quarks and leptons. Plausible landscape models allow for some of the first generation quark and lepton masses to be anthropically determined, while the remaining parameters are set purely by the statistical distribution of the Yukawa matrices. Results are consistent with the observed hierarchical, generation, and pairing structures \[75\]\[82\]. In such analyses, the overall mass scale of neutrinos may be held fixed and ascribed, e.g., to a seesaw mechanism. But ultimately, one expects that the mass scale will vary, no matter what the dominant origin of neutrino masses is in the landscape. For Dirac neutrinos, Yukawa couplings can vary; in the seesaw, a coupling or the right-handed neutrino mass scale can vary.

Thus we may ask whether anthropic constraints play a role in determining the overall

\[1\] It cannot be explained in a one-dimensional landscape, no matter how large \[51\]\[53\], because an empty universe is produced. The string theory landscape \[28\]\[54\]\[55\] is an example of a multidimensional landscape in which the cosmological constant scans densely and our vacuum can be produced with sufficient free energy. Related early work includes \[10\]\[56\]\[68\]. Reviews with varying ranges of detail and technicality are available, for example \[5\]\[69\]\[74\].
scale, or sum, of the standard model neutrino masses,

$$ m_\nu \equiv m(\nu_e) + m(\nu_\mu) + m(\nu_\tau) . $$

(2.1)

Current observational bounds imply

$$ 58 \text{ meV} \leq m_\nu \lesssim 0.23 \text{ eV} . $$

(2.2)

The lower bound comes from the mass splittings observed via solar and atmospheric neutrino oscillations [83]. The upper bound comes from cosmological observations that have excluded the effects that more massive neutrinos would have had on the cosmic microwave background and on large scale structure [84, 85]. The proximity of the lower to the upper bound gives us confidence that cosmological experiments in the coming decade will detect $m_\nu$, and that they may determine its value with a precision approaching the $10^{-2} \text{ eV}$ level [86].

An anthropic origin of the neutrino mass scale is suggested by the remarkable coincidence that neutrinos have affected cosmology just enough for their effects to be noticeable, but not enough to significantly diminish the abundance of galaxies. A priori, $m_\nu$ could range over dozens of orders of magnitude. If $m_\nu$ was only two orders of magnitude smaller than the observed value, its effects on cosmology would be hard to discern at all. If $m_\nu$ was slightly larger, fewer galaxies would form, and hence fewer observers like us. The goal of this paper is to assess this question quantitatively.

The basic framework for computing probabilities in a large landscape of vacua is reviewed in Sec. 2.2, and the probability distribution $dP/d\log m_\nu$ is computed in Sec. 2.3. In the remainder of this introduction, we will describe the key physical effects that enter into the analysis, and we will present our main results.

Summary: There are two competing effects that determine the neutrino mass sum. We assume that the statistical distribution of neutrino masses among the vacua of the landscape favors a large neutrino mass sum, with a force of order unity or less (see Sec. 2.2.1 for the definition of the multiverse force).

If the anthropic approach is successful, we must demonstrate a compensating effect: that neutrino masses much greater than the observed value are not frequently observed. That is, we must multiply the prior probability for some value of $m_\nu$ by the number of observers that will be produced in regions where $m_\nu$ takes this value. Observers are usually represented by some proxy such as galaxies. We consider two models for observations: at any given time, their rate is proportional to the number of Milky Way-like galaxies, or proportional to the growth rate of this galaxy population (see Sec. 2.2.2). We sum this rate over a spacetime region called the causal patch [26] (a standard regulator for the divergent spacetime that results from a positive cosmological constant; see Sec. 2.2.3). The product of prior distribution and the abundance of galaxies yields a predicted probability distribution. As usual, if the observed value lies some number of standard deviations from the mean of
Figure 2.1: (a) The dimensionless power spectrum at $z = 0$ for a range of neutrino masses with a normal hierarchy, computed using CAMB. From top to bottom at high wavenumber: $m_\nu = 0$ (red), $m_\nu = 5$ eV (purple), $m_\nu = 10$ eV (blue), and $m_\nu = 15$ eV (black). Free-streaming of massive neutrinos causes a suppression of power at high wavenumbers. Above a critical neutrino mass $m_\nu \sim 8$ eV, this effect is large enough so that the dimensionless power spectrum develops a peak near the free-streaming scale $k_\text{nr} < k_\text{gal}$. This implies that the first structure consists of cluster-size halos. (b) We obtain the smoothed density contrast $\sigma_R$ at $z = 0$ numerically, essentially by integrating $k^3P(k)$ up to the wavenumber $1/R$; see Eq. (2.3) and surrounding discussion. We take $R$ to be the comoving scale of the Milky Way ($10^{12} M_\odot$). The orange (upper) curve corresponds to a normal hierarchy; the green (lower) to a degenerate hierarchy. We see that neutrinos suppress halo formation only in the regime $m_\nu \lesssim 10$ eV where the dimensionless power spectrum has no maximum and the integral is dominated by the large-$k$ cutoff. For larger neutrino masses, the formation of galactic and larger halos is actually enhanced, because the dimensionless power spectrum develops a peak that dominates the integral. At higher neutrino masses, the peak power increases, due to a decrease of the free-streaming scale and a lengthening of the matter era. (This is more pronounced for a normal hierarchy.) Hence $\sigma_R$ increases. If observers formed in proportion to the mass fraction in large dark matter halos, this would rule out an anthropic origin of $m_\nu$; see Fig. 2.2.

The neutrino mass spectrum—the individual distribution of masses among the three active neutrinos—has a noticeable effect on structure formation. We consider two extreme cases. In the normal hierarchy, one neutrino contributes dominantly to the mass sum $m_\nu$; here we approximate the remaining two as massless. In this case the observed mass splittings require $m_\nu \geq 58$ meV. In the degenerate hierarchy, each mass is of order $m_\nu / 3$ (and here we approximate them as exactly equal). This case will soon be tested by cosmological
observations, since the observed mass splittings would imply $m_\nu \gtrsim 150$ meV, near the present upper limit. We do not explicitly consider the intermediate case of an inverted hierarchy, with two nearly degenerate massive neutrinos and one light or massless neutrino.

The main challenge lies in estimating the galaxy abundance as a function of $m_\nu$. The effects of one or more massive neutrinos on structure formation are somewhat complex; hence, we compute the linear evolution of density perturbations numerically using Boltzmann codes CAMB [87] and CLASS [88], wherever possible. We will now summarize the key physical effects. A more extensive summary and analytic approximations are given in Sec. 2.3 and Appendix A.2; we recommend Refs. [89,90] for detailed study.

![Figure 2.2: No cooling boundary](image)

Figure 2.2: No cooling boundary: if one assumed that observers trace dark matter halos of mass $10^{12} M_\odot$ or greater, one would find a bimodal probability distribution over the neutrino mass sum $m_\nu$. This distribution is shown here for a normal hierarchy (orange/upper curve) and degenerate hierarchy (green curve). The range of values $m_\nu$ consistent with observation ($58$ meV $< m_\nu < 0.23$ eV, shaded in red) is greatly disfavored, ruling out this model.—By contrast, we shall assume here that observers trace galaxies. Crucially, we shall argue that for $m_\nu \gtrsim 10$ eV, galaxies do not form even though halos do. This novel catastrophic boundary excludes the mass range above $10$ eV, leading to a successful anthropic explanation of the neutrino mass (see Fig. 2.3).

After becoming nonrelativistic, neutrinos contribute approximately as pressureless matter to the Friedmann equation. However, they contribute very differently from cold dark matter (CDM) to the growth of perturbations, because neutrinos are light and move fast. This introduces a new physical scale into the problem of structure formation: the free-streaming scale is set by the distance over which neutrinos travel until becoming nonrelativistic. It is roughly given by the horizon scale when they become nonrelativistic, with comoving wavenumber $k_{nr}$ (see Appendix A.2.1 for more details). On this and smaller scales, $k \gtrsim k_{nr}$, neutrinos wipe out their own density perturbations. More importantly, as a nonclustering matter component they change the rate at which CDM perturbations grow, from linear
growth in the scale factor ($\delta \propto a$) on large scales, to sub-linear growth on smaller scales $k \gtrsim k_{nr}$. This suppresses the CDM power spectrum at small scales, see Fig. 2.1a.

The linear quantity most closely related to the abundance of dark matter halos on the galactic scale $R$ is not the dimensionless CDM power spectrum $k^3P_{cc}(k)$. Rather, halo abundance is controlled by the smoothed density contrast $\sigma_R$, which is approximately given by the integrated power, 

$$\sigma_R \sim \int_0^{1/R} \frac{dk}{k} k^3P_{cc}(k),$$

up to the wavenumber corresponding to the relevant scale. (A more precise formula is given in the main text, where we also describe in detail how halo abundance is computed from $\sigma_R$ using the Press-Schechter formalism.) This distinction turns out to be crucial for large neutrino masses.

We see from Fig. 2.1a that for small neutrino masses $m_\nu \lesssim 10$ eV, the integrand $k^3P_{cc}(k)$ increases monotonically. Hence the integral for $\sigma_R$ is dominated by its upper limit, i.e., by the power on the scale $k_{gal} \sim 1/R$. This yields the “bottom-up” scenario of hierarchical structure formation familiar from our own universe: small halos typically form first, and more massive halos virialize later.

However, for $m_\nu \gtrsim 10$ eV, the small scale power becomes so suppressed that the dimensionless power spectrum develops a maximum at the free-streaming scale $k_{nr} < k_{gal}$. In this regime, the smoothed density contrast $\sigma_R$ on galactic scales $R$ is no longer dominated by the power at wavenumber $\sim 1/R$. Instead, the power at larger scales than $R$ contributes dominantly to $\sigma_R$. This results in a top-down scenario, where halos first form on cluster scales, nearly simultaneously with galactic-scale halos.

The transition from bottom-up to top-down structure formation around $m_\nu \approx 10$ eV has not (to our knowledge) been noted in the context of anthropic explanations of the neutrino mass sum. We find here that it is crucial to the analysis, for two reasons. First, it implies that the scales that dominantly contribute to $\sigma_R$ are unaffected by free-streaming for $m_\nu \gtrsim 10$ eV. Therefore, increasing $m_\nu$ beyond $\sim 10$ eV does not suppress CDM structure. In fact, we find that $\sigma_R$ increases in this range (Fig. 2.1b). The second implication works in the opposite direction: in the top-down scenario that arises for $m_\nu \gtrsim 10$ eV, galaxies will not form inside halos at an abundance comparable to our universe.

Let us discuss each of these implications in turn. We begin by pretending that the stellar mass per halo mass is unaffected by $m_\nu$; in particular, let us suppose that there is no dramatic suppression of star formation in the top-down regime, $m_\nu \gtrsim 10$ eV. If so, we would be justified in regarding halos as a fair proxy for observers. Here we consider $10^{12}M_\odot$ halos [91]. From Fig. 2.1b, we see that halo abundance decreases with $m_\nu$ up to $m_\nu \sim 10$ eV; then it begins to increase. Combining this with the assumed prior distribution that favors large $m_\nu$, we would find the probability distribution over $m_\nu$ is bimodal (Fig. 2.2). The first peak is at $m_\nu \approx 1$ eV, followed by a minimum near 10 eV and a second peak at much
CHAPTER 2. ANTHROPIC ORIGIN OF THE NEUTRINO MASS?

35

Figure 2.3: Our main result: the probability distribution over the neutrino mass sum $m_\nu$ for (a) a normal hierarchy and (b) a degenerate hierarchy, assuming that observers require galaxies. The plot is the same as Fig. 2.2 but the mass range is cut off at $7.7 \text{ eV} (10.8 \text{ eV})$ in the normal (degenerate) case. For greater masses, the first halos form late and are of cluster size; we argue that galaxies do not form efficiently in such halos. We use the second observer model described in Sec. 2.2.2 results look nearly identical with the first model. We assume a flat prior over $m_\nu$ (see Fig. 2.4 for other priors).—The central $1\sigma$ and $2\sigma$ regions are shaded. Vertical red lines indicate the lowest possible values for the neutrino mass consistent with available data: $m_{\text{obs}} \approx 58 \text{ meV}$ for a normal hierarchy, and $m_{\text{obs}} \approx 150 \text{ meV}$ for a degenerate hierarchy. We find that these values are within $2\sigma$ of the median. The agreement would further improve with a less conservative treatment of the detrimental effects of neutrinos on gas cooling in halos, and/or the cosmological detection of a neutrino mass sum larger than the minimal value.

greater mass\footnote{Fig. 2.2 does not show the entire peak since CAMB gives results only for $m_\nu \lesssim 40 \text{ eV}$. Absent the earlier catastrophic boundary at $10 \text{ eV}$ that we will assert, a robust effect that would eventually suppress the probability at large neutrino mass is the smallness of the baryon fraction for $m_\nu \gtrsim 100 \text{ eV}$. This would suppress the number of baryons (and hence, observers) in the causal patch \footnote{92}. It would also impose dynamical obstructions to star formation \footnote{93}.}

Therefore, if observers traced dark matter halos with $M \gtrsim 10^{12} M_\odot$, one should conclude that small neutrino masses are greatly disfavored. Such a result would be in significant tension with the current upper bound of 0.23 eV, and it would seem to render an anthropic origin of the neutrino mass sum implausible.

However, our fundamental assumption is that observers trace galaxies, not halos. In some cosmologies including our own, galaxies in turn trace halos; if they do, halos are an equally good proxy. But the change of regime from bottom-up to top-down structure formation for $m_\nu \gtrsim 10 \text{ eV}$ is catastrophic for galaxy formation.

From observation, we know that stars do not form efficiently in bound structures that are much larger than the mass scale of our own galactic halo, $10^{12} M_\odot$. Heuristically, this can
be explained by noting that in halos of this size, the cooling timescale for the baryonic gas is greater than the age of the universe \[93-97\]. In our universe there are galaxies because galactic halos, which produce stars efficiently, formed earlier than these larger halos, which do not. Clusters inherit galaxies that formed in smaller halos, but they do not have significant star formation themselves.

In a top-down scenario due to large neutrino mass, however, galactic halos would form much later. They would typically be embedded in larger halos that virialize roughly at the same time, with masses characteristic of galaxy groups clusters—but without many galaxies to inherit. The virial temperature and dynamical timescale relevant for baryon cooling will be set by the largest of the nested halos. (See Sec. [2.3.4 and Appendix A.3 for details.) Therefore, cooling will not be efficient: the top-down scenario produces star-poor dark matter clumps, with most baryons remaining in hot gas.

As a first approximation for this cooling boundary, we cut off the probability distribution at a value \(m_\nu \sim 10 \text{ eV}\) that corresponds to the onset of the top-down regime. This overestimates the amount of galaxies just below the cutoff and underestimates it just above. In future work, we plan to include explicit models for successful galaxy cooling beyond the crude top-down vs. bottom-up criterion. This should replace the sharp cutoff by a smooth decay of the probability.

We believe that our argument for a cooling catastrophe is robust, because the transition to a top-down scenario is a drastic change of regime. However, the underlying physics is complicated, involving shocks, complicated cooling functions, fragmentation, and feedback from stars, black holes, and supernovae. Suppose therefore that we are wrong. That is, suppose that at \(m_\nu \gtrsim 10 \text{ eV}\), some unanticipated combination of processes lead to a stellar mass inside the causal patch that is not much less than in our universe. Then one would find that large neutrino masses are unsuppressed (Fig. [2.2], and the observed value of \(m_\nu\) cannot be explained anthropically. In this sense, the cooling catastrophe we assert can be regarded as a prediction of the anthropic approach to the neutrino mass. To test this prediction, it will be important to investigate galaxy formation for \(m_\nu \gtrsim 10 \text{ eV}\) using simulations that give an adequate treatment of cooling flows and feedback.

**Results:** Our main results, with the cooling cutoff \(m_\nu \lesssim 10 \text{ eV}\) imposed, are shown in Fig. [2.3] We find that the currently allowed range of values for \(m_\nu\) is entirely consistent with an anthropic explanation, at better than \(2\sigma\). Fig. [2.4] shows that that our approach succeeds for a wide range of prior distributions \(dP_{\text{vac}}/dm_\nu \propto m_\nu^{n-1}\): assuming a normal (degenerate) hierarchy, \(m_{\text{obs}}\) lies within \(2\sigma\) of the median if \(0.09 < n \leq 1.0\) (\(0.09 < n < 1.4\)).

Our chief conclusion is that the neutrino mass sum can be anthropically explained, but only if detrimental effects of neutrinos on galaxy and star formation (rather than halo formation) already become significant at or below \(m_\nu \approx 10 \text{ eV}\).

Our results favor larger \(m_\nu\) than the minimum values allowed by the observed mass splittings, and in particular they favor a degenerate over a normal hierarchy. Since the
Figure 2.4: The prior distribution of cosmologically produced vacua is assumed to favor large neutrino mass and to have no special feature near the observed magnitude: \( dP_{\text{vac}}/d\log m_\nu \propto m_\nu^n, \, n > 0 \). One then expects \( n \sim O(1) \), and the previous two figures all show the case \( n = 1 \). This figure shows that the same conclusions obtain for a considerable range of \( n \). (a) The median of the probability distribution as a function of the multiverse force \( n \). (b) The standard deviation of the worst case observed value from the median, as a function of \( n \). The 2\( \sigma \) region is shaded. Orange (the upper curve at large \( n \) in each plot) is for a normal hierarchy; green is for a degenerate hierarchy.

The observed range is consistent within 2\( \sigma \) in either case, these are mild preferences rather than sharp predictions.

There are however two additional reasons why a degenerate hierarchy appears more natural in the context of the anthropic approach. First, with a normal hierarchy one might expect that each neutrino mass scans separately with prior \( n_i \). Each prior would have to be assumed positive and \( O(1) \). The prior for \( m_\nu \) would then be \( n = \sum n_i \), and it becomes less plausible that \( n \) should be small enough to render the anthropic prediction compatible with observation. A degenerate hierarchy, on the other hand, may be the result of some flavor symmetry that links the masses of the individual neutrinos, leaving only a single scanning parameter. Then it is more plausible that \( n \) is small enough to include the observed \( m_\nu \).

The second reason to prefer a degenerate hierarchy is that it eliminates a viable anthropic window where two neutrinos are extremely massive. If each neutrino has mass of order MeV or greater, neutrons would be stable, leading to a (catastrophic) helium-dominated universe \( \text{[98]} \). But neutrons will be unstable and the catastrophe is averted, if one neutrino remains light and only the other two become very heavy. With a normal or inverted hierarchy, one has to explain why the one or two heavy neutrinos did not end up in the extremely large mass range above the MeV scale. This can be resolved by assuming that the prior distributions for the individual neutrino masses do have a feature between the eV and the MeV scale, such that the much larger scale is disfavored. With a degenerate hierarchy, this problem does not arise in the first place, since either all neutrinos are light or all are heavy.
Relation to earlier work: Our analysis builds on the pioneering work of Tegmark, Vilenkin and Pogosian [98, 99] (see also [100,101]), who were the first to argue that the neutrino mass admits an anthropic explanation. We agree with their conclusion, but we claim here that the nature of the relevant catastrophic boundary was not correctly identified. Ref. [98] does not justify its restriction to the region $m_\nu \lesssim 10$ eV. Moreover, it employs an analytic approximation to $\sigma_R$ that greatly underestimates the halo abundance for $m_\nu \gtrsim 5 - 10$ eV. With this approximation, the probability distribution appears to vanish near 10 eV due to a paucity of CDM structure; see Fig. 2.5. Thus, suppression of CDM structure due to massive neutrinos—rather than the obstruction to cooling at $m \gtrsim 10$ eV—would appear to provide the relevant catastrophic boundary underlying the anthropic explanation of the neutrino mass sum.

Here we go further in two respects: our numerical computations show that CDM structure becomes unsuppressed for $m_\nu \gtrsim 10$ eV. Hence, if neutrino masses have an anthropic origin, a different catastrophic boundary is relevant. And we identify a specific physical effect, the transition to a top-down regime, which had not been noted and which supplies a suitable boundary by suppressing galaxy formation.

The analytic approximation in question is Eq. (5) in Ref. [98]. It assumes that massive neutrinos suppress the smoothed density contrast $\sigma_R$ on galactic scales by the same factor by which they suppress the matter power on galactic scales. This is accurate for small neutrino masses, because in a bottom-up scenario the shortest scales dominate the integral for $\sigma_R$. The approximation underestimates the abundance of dark matter halos for $m_\nu \gtrsim 10$ eV, because in this regime $\sigma_R$ is dominated by power at larger scales, which is relatively unsuppressed by free-streaming. More details can be found after Eq. (2.3) and in Sec. 2.3.4.

The discrepancy is revealed by explicit numerical computation of the smoothed density contrast $\sigma_R$ on galactic scales from Boltzmann codes (see Fig. 2.1). One also finds significantly different results for a normal versus degenerate hierarchy, a distinction that was suppressed in the analytical approximation of Ref. [98].

When the halo abundance is correctly computed, the need for a novel catastrophic boundary at or before 10 eV becomes evident (Fig. 2.5). Without it, the probability distribution would strongly disfavor small neutrino masses. It would be in significant tension with the an upper bound of 0.23 eV or even 1 eV, and it would seem to render an anthropic origin of the neutrino mass sum implausible.

Our computation of $\sigma_R$ from Boltzmann codes, and our compensating identification of a novel catastrophic boundary at 10 eV are the main differences to Ref. [98]. Another difference is that we use the causal patch measure [26] to regulate the infinities of eternal inflation. Refs. [98, 99] used a different measure that is no longer viable; see Sec. 2.2.3 for details. This has a visible but comparatively small effect on the probability distribution: by comparing Fig. 2.2 with Fig. 2.5, one sees that the causal patch is somewhat more favorable to an anthropic explanation of $m_\nu$. The causal patch also renders more robust our conjecture
Figure 2.5: The dashed black line shows the probability distribution found by Tegmark et al. [98] for a flat prior over $m_\nu$. This result would seem to remain compatible at about 2$\sigma$ with current observational constraints (red shaded region). However, the analytic fitting function for the density contrast $\sigma_R$ used in [98,99] underestimates $\sigma_R$ above a few eV. The solid curves show the probability distribution that results when $\sigma_R$ is computed numerically from Boltzmann codes: orange/upper=normal hierarchy; green/lower=degenerate hierarchy. They differ slightly from Fig. 2.2 because Ref. [98] used a different measure and observer model. Either way, a successful anthropic explanation of $m_\nu$ requires the identification of a catastrophic boundary at or below 10 eV.

that star formation is ineffective for $m_\nu \gtrsim 10$ eV, as discussed in more detail later.

We also build on the seminal investigation of catastrophic boundaries in cosmology by Tegmark et al. [93] (see also Ref. [102]), who emphasized the crucial role of cooling. We believe that our present work is the first to associate catastrophic cooling failure to a top-down structure formation scenario. Ref. [98] points out a number of distinct catastrophes at very large neutrino mass: For example, neutrinos act as cold dark matter for $m_\nu \gg 100$ eV, which also may be detrimental to star formation. (However, this does not counteract the abundance of CDM structure we find at $m_\nu \gtrsim 10$ eV. Fig. 2.5 illustrates that a cutoff at any scale larger than 10 eV, say at $m_\nu \approx 30$ eV, would make small neutrino masses too improbable for an anthropic explanation to work.)

2.2 Predictions in a Large Landscape

If a theory has a large number of metastable vacua, most predictions will be statistical in nature. We are usually interested in understanding the magnitude of a particular parameter $x$, such as the cosmological constant or in the present case, the neutrino mass; hence we wish to compute a probability distribution $d\mathcal{P} / d\log x$.

Fundamentally, the probability $d\mathcal{P}$ is proportional to the number of observations $dN_{\text{obs}}$. 
that find the parameter to lie in the range \((\log x, \log x + d \log x)\). Thus, our task is to compute \(dN_{\text{obs}}\). This can be done by weighting a prior probability distribution \(f(x)\), which comes from the underlying theory, by the number \(w(x)\) of observations that will be made in a vacuum where \(x\) takes on a particular value:

\[
\frac{dP}{d \log x} = w(x)f(x).
\] (2.4)

We will discuss each factor in turn. Our presentation in each subsection will be general at first, before specializing to the case of the neutrino mass, \(x = m_\nu\).

### 2.2.1 Prior as a Multiverse Probability Force

The prior is defined by

\[
f(x) = \frac{dN_{\text{vac}}}{d \log x},
\] (2.5)

Here \(x\) will be a parameter in the effective theory at low energies whose scale \(\log x\) one would like to predict or explain; \(dN_{\text{vac}} = f(x)d \log x\) is the number of long-lived metastable vacua\(^3\) in which the parameter takes on values in the range \((\log x, \log x + d \log x)\).

With the notable exception of the cosmological constant, the prior distribution for most parameters is not well known. This is a technical problem: in the string landscape, \(f(x)\) should in principle be computable. In practice, it is difficult to derive phenomena far below the fundamental scale (the Planck or string scale) directly. However, this need not be an obstruction to progress, any more than the fact that we cannot derive the Standard Model from a more fundamental theory prevented us from discovering it.

Consider an arbitrary low-energy parameter \(x\). In any large landscape the prior distribution \(f(x)\) should admit an effective description\(^7\)[75,102,103]. To avoid putting in the answer, one may assume that \(f(x)\) has no special features (such as a maximum) in a wide logarithmic range of values. This range should include but be much larger than the range compatible with observation. One can then parametrize the prior distribution by a “statistical pressure” or “multiverse probability force” towards large or smaller values,

\[
n \equiv \frac{d \log f}{d \log x},
\] (2.6)

where \(n\) is approximately constant. For example, a flat prior distribution over \(\log x\) corresponds to \(n = 0\). If the prior is flat over \(x\), \(dN_{\text{vac}}/dx = \text{const.}\), then \(n = 1\).

\(^3\)Strictly, what matters is not the abundance of such vacua in the effective potential but in the multiverse: cosmological dynamics could favor the production of some vacua over others. For most low-energy parameters one expects that such selection effects are uncorrelated with \(x\) in the range of interest. In any case, we shall take the prior \(f\) to be an effective distribution that incorporates cosmological dynamics.
CHAPTER 2. ANTHROPIC ORIGIN OF THE NEUTRINO MASS?

Suppose that there is a regime change sufficiently near the observed value \( \log x_0 \), such that the number of observers (or at least, of observers like us) \( w(x) \) drops dramatically above or below a critical value \( \log x_c \). Suppose for definiteness that \( x_0 \lesssim x_c \). If the probability force favors large values of \( x \), but not too strongly \([n > 0, n \sim O(1)]\), then the observed value can be explained. Similarly, with a negative probability force, one can explain the proximity of \( x \) to nearby catastrophic boundary at some smaller \( x_c \lesssim x_0 \).

Recent successful examples of this approach include an explanation of the coincidence that dark and baryonic densities are comparable [92], the fine-tuning of the weak scale [92, 104, 105], and the comparability of several large, a priori unrelated timescales in cosmology [102]. In each case, the required assumption about the probability force is weak and qualitative: \( n \sim \pm O(1) \). Thus phenomenological models of the landscape have significant explanatory value, while constraining the underlying prior distributions through the sign (and roughly the strength) of the probability force \( n \). It is particularly instructive to keep track of the combination of (and possible conflict between) forces \( n_i \) needed to simultaneously explain multiple parameters \( x_i \) [102, 105, 106].

Now let us turn to the prior for the total neutrino mass, \( P_{\text{vac}}(m_\nu) \). We know of no physical reason why a minimum neutrino mass should be necessary for observers. Hence, to obtain a normalizable probability distribution \( f \), we must assume that the effective prior distribution favors large \( m_\nu \):

\[
\frac{dP_{\text{vac}}}{d\log m_\nu} \propto m_\nu^n, \quad n > 0,
\]

in some large logarithmic neighborhood of the observed value, \( \sim 0.1 \text{ eV} \). A natural and simple choice is \( n = 1 \), and we will use this value for definiteness in most plots. More generally, we will find that a comfortable range of values \( 0 < n \approx O(1) \) is consistent with an anthropic explanation of the neutrino mass, but not a value much greater than 1 (see Fig. 2.4).

2.2.2 Anthropic Weighting

The probability distribution over \( \log x \) relevant for comparing the theory with observation is obtained by conditioning \( \tilde{p} \) on the presence of observers. More quantitatively, one weights by the number of observations

\[
w(x) = \frac{dN_{\text{obs}}}{dN_{\text{vac}}}
\]

that are made in a vacuum where \( x \) takes on a specific value. Generically, \( w(x) \) will be unsuppressed in a large region either above or below the observed value, or both. Thus, the anthropic factor is not doing all the work; the prior distribution is crucial for comparing the theory to observation.

In this paper we will consider two different models for the number of observations \( w(x) \) that are performed in the universe. Both are based on the assumption that observers require galaxies, say of halo mass comparable to the Milky Way’s, \( 10^{12} \) solar masses. The first model...
assumes that the rate at which observations occur in a given spatial region per unit proper time, $\dot{w}(x)$, is proportional to the total mass $M_{\text{gal}}$ of such galaxies, at every instant; hence

$$w(x) = \int dt \ M_{\text{gal}}(t) \quad (\text{Observer Model 1}) .$$  \hspace{1cm} (2.9)

The second model (which reduces to the choice made in Ref. [98]) assumes instead that the rate of observation is proportional to the rate $\dot{M}$ at which the above total galaxy mass grows:

$$w(x) = \int dt \ \dot{M}_{\text{gal}}(t) \quad (\text{Observer Model 2}) .$$  \hspace{1cm} (2.10)

The two models can be thought of as two different approximations taken to an extreme. In the first, observations would be made continuously in the galaxy, at fixed rate per unit stellar mass, no matter how old the stars become. In the second model, observations would occur instantaneously as baryons cool and form stars; no observations would be assigned to a galaxy that is not growing. (The second model was used in Ref. [98]; note that in the context of the measure used there the integral over time is trivial, yielding the collapse fraction $F_{R_\star}$.) The truth is likely somewhere in between the two models. However, we will find that our results depend only weakly on the model, so we expect our conclusions to be robust.

\subsection{2.2.3 Measure}

A cosmology with at least one long-lived de Sitter vacuum gives rise to eternal inflation: the universe will grow without bound and remain at finite temperature in arbitrarily large volumes at late times. Hence, all possible events will occur infinitely many times. This applies in particular to observations. Thus a regulator or “measure” must be introduced to obtain a finite anthropic factor $w(x)$. For this problem to exist, it is not necessary that the theory predict a large landscape; one de Sitter vacuum (such as, apparently, ours [3,4]) is enough. But the measure problem becomes particularly glaring in the landscape context: globally, every type of vacuum bubble is produced infinitely many times, and each bubble universe contains an infinite comoving volume.

Existing analyses of the anthropic origin of neutrino masses preceded a period of significant progress on the measure problem of eternal inflation. Following Weinberg [50], Refs. [98,99] regulate the divergences of the cosmological dynamics by estimating the number of observers per baryon. This measure can no longer be considered viable [107,108]. Note, however, that our choice of measure is not responsible for the main differences between our results and those of Refs. [98], as described at the end of Sec. 2.1.

In this paper, we will use the causal patch measure [26], which regulates eternal inflation by considering a single causally connected region and averaging over its possible histories.
This proposal is very generally defined, requiring only causal structure. It is also well motivated: it merely applies to cosmology an existing restriction that was already needed for the unitary evaporation of black holes \[30\]. Though proposed on formal grounds, the causal patch has met with phenomenological success; two examples are described in Appendix \[A.1\]. We take this as evidence that it approximates the correct measure well (at least in regions with positive cosmological constant \[109\]).

A potential landscape is consistent with the observed cosmological history only if it is multi-dimensional with large energy differences between neighboring vacua \[5\]. String theory gives rise to such a structure upon compactification to three spatial dimensions \[28\], with \(\Delta \Lambda\) not much below unity.

The causal patch will contain a particular decay chain through de Sitter vacua in the landscape, ending with a big crunch in a vacuum with negative cosmological constant; each such chain is weighted by its probability, i.e., by the product of branching ratios \[26\]. For a typical decay chain, none of the vacua will have anomalously small cosmological constant \(\Lambda \ll \Delta \Lambda\). Thus, after conditioning on observers, there will be one vacuum with small cosmological constant in the causal patch, and we need only be concerned with how the causal patch regulates the volume of the corresponding bubble universe.

Here we focus on the variation of the neutrino mass only, so we shall take this vacuum to be otherwise like ours. In particular we set the cosmological constant to the observed value, \(\Lambda \sim 10^{-123}\), and we take the spatial geometry to be flat. The metric is of the Friedman-Robertson-Walker (FRW) type:

\[
ds^2 = -dt^2 + a(t)^2 (dr^2 + r^2 d\Omega^2) ,
\]

where \(a\) is the scale factor, \(r\) is the comoving radius, \(t\) is proper time, and \(d\Omega^2\) is the metric on the unit two-sphere.

By definition, the causal patch is the causal past of the future endpoint of a geodesic; thus its boundary consists of the past light-cone of such a point. We are interested in the boundaries of the causal patch during the time when a long-lived de Sitter vacuum still contains matter. A future decay has an exponentially small effect on the location of the patch boundary at much earlier times, so the patch can be computed by treating the vacuum as completely stable. The patch boundary is thus the cosmological event horizon. Its comoving radius at FRW time \(t\) is obtained by tracing a light-ray back from future de Sitter infinity:

\[
r_{\text{patch}}(t) = \int_t^\infty \frac{dt'}{a(t')} .
\]
The physical volume of the patch is

\[ V_{\text{phys}}(t) = \frac{4\pi}{3}a(t)^3 r_{\text{patch}}(t)^3. \] (2.13)

As described in the previous subsection, we estimate the rate of observations per unit time as proportional to the total mass of all galaxies in the physical volume of the patch (for observer model 1), or to the rate of increase of this mass (for observer model 2). We can write this quantity as

\[ M_{\text{gal}}(t) = \rho_{bc}(t)V_{\text{phys}}(t)F_R(t)G_R(t). \] (2.14)

The first two factors give the total mass \( M_{bc} \) of baryons and cold dark matter in the patch at the time \( t \). The collapse fraction \( F_R \) is the fraction of this mass that is contained in halos of mass greater than \( 10^{12} M_\odot \), corresponding to a comoving distance scale \( R \): \( M_{\text{halo}} = M_{bc}F_R \). The galaxy fraction \( G_R \) is the fraction of this latter mass that represents baryons in galaxies, \( M_{\text{gal}} = M_{\text{halo}}G_R \).

Combining this with Eqs. (2.9), (2.7), and (2.4), the (unnormalized) probability distribution over the neutrino mass is given by

\[ \frac{dp}{d\log m_\nu} \propto m_\nu^n \int dt (r_{\text{patch}}a)^3 \rho_{bc} F_R G_R. \] (2.15)

in the first observer model; we replace \( F_R G_R \) by \( \frac{d}{dt}(F_R G_R) \) for the second observer model.\(^5\) Factors in the integrand may in general depend on both \( m_\nu \) and \( t \).

### 2.3 Calculation of \( dP/d\log m_\nu \)

#### 2.3.1 Fixed, Variable, and Time-dependent Parameters

We will consider a one-parameter family\(^6\) of cosmologies, differing from our universe only in the total mass of active neutrinos. More precisely, we consider two such families, since we treat the cases of normal and degenerate neutrino hierarchy separately. Thus, we hold fixed all fundamental parameters other than \( m_\nu \). In particular, we fix the vacuum energy

---

\(^5\)Note that the time derivative should not be taken of the entire integrand, for this model. The loss of mass across the horizon due to the shrinking comoving volume of the patch does not produce “negative galaxies” inside the patch. At some cost in readability, we could have made this more explicit by defining the integrand in Eq. (2.10) as the causal patch volume times the rate of change of the average physical density contributed by galaxies.\(^6\)It would clearly be of interest to compute the probability distribution over several parameters including the neutrino mass; for examples of multivariate probability distributions in the landscape, see e.g. [99]. Each additional scanning parameter is an additional opportunity to falsify the model. But already with one parameter scanning, one can falsify a model, in the usual way: by computing a probability distribution from the theory. If one finds that the observed value is several standard deviations from the mean, the model is ruled out at the corresponding level of confidence.
density, $\rho_\Lambda = \Lambda / 8\pi G$, and the spatially flat geometry of the universe (imposed, presumably, by a mechanism like inflation that is uncorrelated with $m_\nu$). We also hold fixed $\chi_b \equiv \rho_b / n_\gamma$ and $\chi_c \equiv \rho_c / n_\gamma$, the masses per photon of baryons and CDM. These quantities remain invariant under changes of $m_\nu$, since we hold fixed the fundamental processes that produced the observed baryon and CDM abundances.

For the actual values of these parameters, we use the Planck TT+lowP+lensing+ext best fit cosmological parameters [85]; see Table 2.1. The best fit assumes a neutrino mass of about 0.06 eV [85], whereas strictly, one should use a best fit marginalized over $m_\nu$ for the purposes of our paper. However, this has virtually no effect on the fixed cosmological parameters such as $\rho_\Lambda$, $\chi_b$, and $\chi_c$, because neutrinos are already constrained to contribute a very small fraction to the total density. For example, the best-fit for the Hubble parameter (Planck TT+lowP+lensing+ext [85]) shifts from $67.9 \pm 0.55$ ($m_\nu \approx 0.06 \text{ eV}$) to $67.7 \pm 0.6$ (marginalized over $m_\nu$). This difference is negligible compared to current error bars and the discrepancies between different cosmological datasets.

When considering entire cosmological histories, as we do, it is best to specify each cosmology in terms of time-independent parameters such as $\Lambda$, $\chi_b$, $\chi_c$, and $m_\nu$. However, we use Boltzmann codes such as CAMB and CLASS to compute power spectra wherever possible (i.e., for $z \geq 0$). These codes expect input parameters that specify the cosmological model in terms of their present values, at redshift $z = 0$. It is not clear what one would mean by the “present” time in an alternate cosmology, but for the purposes of CAMB and CLASS, $z = 0$ is defined to be the time at which the CMB temperature takes the observed value, $T_{\text{CMB}} \approx 2.7 \text{ K}$.

Thus we must derive the values of various time-dependent quantities at the time when the universe reaches this temperature, as a function of $m_\nu$, with other time-independent parameters fixed as described above. One finds for the Hubble parameter and the density parameters

\begin{align}
H(m_\nu; z = 0) &= H_0 \left( \frac{\chi_{b\Lambda\nu}}{\chi_{b\Lambda\nu}(0)} \right)^{1/2}, \\
\Omega_X(m_\nu; z = 0) &= \frac{\chi_X}{\chi_{b\Lambda\nu}}, \quad X \in \{b, c, \Lambda, \nu\}.
\end{align}

(2.16) (2.17)

Here multiple indices imply summation, for example $\chi_{bc} \equiv \chi_b + \chi_c$. The fixed parameters $\chi_b$ and $\chi_c$ were defined above. The fixed parameter $\chi_\Lambda \equiv \rho_\Lambda / n_\gamma(z = 0)$ is defined for notational convenience as the observed vacuum energy per photon at the present observed CMB temperature. The $m_\nu$-dependent parameter $\chi_\nu(m_\nu) = \frac{3}{11} m_\nu$ is the neutrino mass per photon. $H_0 \equiv H(0.06 \text{ eV}; z = 0)$ and $\Omega_\nu = \chi_\nu(0.06 \text{ eV})$ are observed values, corresponding to the Planck best fit baseline model.

\footnote{Unless otherwise specified, we quote the Hubble parameter in units km s$^{-1}$ Mpc$^{-1}$ throughout.}
Table 2.1: The cosmological parameters used in our calculation, as well as the resulting mass per photon of baryons and CDM, $\chi_b$ and $\chi_c$. $T_{\text{CMB}}$ is a Planck TT+lowP+BAO fit, while all others are from Planck TT+lowP+lensing+ext best fit values. We take $k_{\text{pivot}} = 0.05 \text{ Mpc}^{-1}$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{\text{CMB}}$</td>
<td>2.722 K</td>
</tr>
<tr>
<td>$H_0$</td>
<td>67.90</td>
</tr>
<tr>
<td>$\Omega_m$</td>
<td>0.3065</td>
</tr>
<tr>
<td>$\Omega_\Lambda$</td>
<td>0.6935</td>
</tr>
<tr>
<td>$\Omega_b h^2$</td>
<td>0.02227</td>
</tr>
<tr>
<td>$\Omega_c h^2$</td>
<td>0.1184</td>
</tr>
<tr>
<td>$10^9 A_s$</td>
<td>2.143</td>
</tr>
<tr>
<td>$n_s$</td>
<td>0.9681</td>
</tr>
<tr>
<td>$\chi_b$</td>
<td>0.5745 eV</td>
</tr>
<tr>
<td>$\chi_c$</td>
<td>3.054 eV</td>
</tr>
</tbody>
</table>

### 2.3.2 Homogeneous Evolution

For computing the volume of the causal patch, the factor $(r_{\text{patch}} a)^3$ in Eq. (2.15), we will need to know the scale factor. Unless structure is present, the integrand will be suppressed by the Press-Schechter factor $F_R$; hence it suffices to use an analytic solution valid to excellent approximation in the matter and vacuum eras:

$$a(m_\nu; t) = \left[ \cot \lambda \sinh \left( \frac{3 \sin \lambda}{2} H_0 t \right) \right]^{2/3},$$

(2.18)

The solution depends on $m_\nu$ through $\Omega_\Lambda(m_\nu; z = 0) \equiv \sin^2 \lambda$.

Since $\chi_{bc}$ does not depend on $m_\nu$ and $\rho_{bc} = n_\gamma \chi_{bc}$, $\rho_{bc}(z = 0)$ does not depend on $m_\nu$. Moreover, since the scale factor in Eq. (2.18) is normalized so that $a = 1$ at $z = 0$, we have $\rho_{bc}(t) = \rho_{bc}(z = 0)/a(t)^3$ for all values of $m_\nu$. Thus Eq. (2.15) simplifies to

$$\frac{d\mathcal{P}}{d \log m_\nu} \propto m_\nu^3 \int dt r_{\text{patch}}^3 F_R G_R ,$$

(2.19)

where $r_{\text{patch}}$ is given by Eq. (2.12), and we are dropping $m_\nu$-independent normalization factors as usual.

The comoving volume of the causal patch is shown in Fig. 2.6. We note that already at the homogeneous level, a nonzero neutrino mass is slightly disfavored because it decreases
the size of the causal patch. We also note that the patch size is maximal at early times and decreases rapidly. Hence galaxies that form very late effectively fail to contribute to the probability for a given parameter value.

![Figure 2.6: The comoving volume of the causal patch for $m_\nu = 0$ (black), $m_\nu = 4$ (blue), and $m_\nu = 8$ (red).]

2.3.3 Halo Formation

The next factor in Eq. (2.19) is the collapse fraction $F_R(m_\nu, t)$. It captures the effects of neutrinos on structure formation: recall that $F_R$ is defined as the fraction of baryonic and cold dark matter that is contained in halos of mass $10^{12} M_\odot$ or greater. It captures the effects of neutrinos on structure formation. Recall that $F_R$ is defined as the fraction of baryonic and cold dark matter in virialized halos of mass scale $10^{12} M_\odot$ or greater. This corresponds to a comoving distance scale $R \sim 1.8 \text{ Mpc}^8$.

The collapse fraction can be determined using the Press-Schechter formalism \[110\]. Before nonlinearities are important, the density contrast \[9\] $\delta(x, t)$ smoothed on a scale $R$ has a Gaussian distribution,

$$\mathcal{P}(\delta, t) \, d\delta \sim \exp\left(-\frac{\delta^2}{2\sigma_R^2}\right) \, d\delta,$$

(2.20)

with standard deviation $\sigma_R(t)$. Fluctuations that exceed a certain threshold $\delta_* \sim O(1)$ in the linear analysis will have become gravitationally bound. Hence,

$$F_R(t) = \int_{\delta_*}^{\infty} \mathcal{P}(\delta, t) \, d\delta = \text{erfc}\left(\frac{\delta_*}{\sqrt{2\sigma_R(t)}}\right).$$

(2.21)

---

8The comoving scale $R$ is independent of $m_\nu$ because $\rho_{bc}(z = 0)$ is. However, when expressed in units of Mpc/$h$ it depends on $m_\nu$ through Eq. (2.16).

9We use the CDM density contrast and power spectrum to compute the Press-Schechter factor $F$. This matches $N$-body simulations better than using the full matter density contrast including neutrinos \[111\]. It is also a conservative choice, since the total matter power spectrum is further suppressed at large $m_\nu$, by a factor $(1 - f_\nu)^2$ below the free streaming scale.
We use the canonical value $\delta^* = 1.69$, which is obtained by comparing the linear perturbation to a spherical collapse model.\footnote{For structure that forms in the vacuum era, the collapse threshold is slightly lowered, whereas in the presence of an appreciable neutrino fraction $\delta^*$ should be slightly increased. If we adapted $\delta^*$ accordingly, the net effect would be to further suppress structure at large $m_\nu$, in favor of an anthropic origin of the neutrino mass. However, appropriate values of $\delta^*$ have so far been estimated only for rather small neutrino masses. Ultimately, it would be preferable to sidestep the Press-Schechter approximation altogether. Our analysis could be dramatically improved by using proper $N$-body simulations to compute structure formation, including an adequate treatment of baryonic physics.}

The standard deviation of the smoothed density contrast is given by \footnote{Available Boltzmann codes do not return power spectra for negative redshifts, that is, for times when the CMB temperature is below 2.7 K. In this regime only, we estimate $\sigma_R$ by extrapolating our numerical results to negative redshifts semi-analytically as described in Appendix A.2.2. This regime is not a dominant contributor to the overall probability distribution, due to the smallness of the causal patch at late times, and since vacuum domination terminates structure formation in any case. We compute $F_R$ and $\dot{F}_R$ from Eq. (2.21); the results are shown in Fig. 2.7.\footnote{We use the CAMB code to compute the CDM power spectrum $P_{cc}(k)$ as a function of time, in models with different neutrino mass. We evaluate the integral in Eq. (2.24) numerically. We have also checked our results using the CLASS code. We noticed a small discrepancy in the output of $k^3 P_{cc}$ at the largest neutrino masses we consider, $m_\nu \sim 10$ eV, where CLASS gives a slightly larger amplitude for the free-streaming peak. By lowering the cutoff on $m_\nu$ described in Sec. 2.3.4, the CLASS output would only strengthen the anthropic explanation of the observed neutrino mass range.}}

$$\sigma_R^2 \equiv \langle \delta_R^2(x) \rangle, \quad (2.22)$$

with

$$\delta_R \equiv \int d^3 x' \delta(x) W_R(|x - x'|), \quad (2.23)$$

where $\delta(x) = \delta \rho_c / \rho_c$ is the fractional overdensity of cold dark matter. We use the top hat window function, $W_R(x) = 1$ for $|x| \leq R$ and $W_R(x) = 0$ otherwise.

Equivalently, the smoothed density contrast can be computed from Fourier-transformed quantities:

$$\sigma_R^2 = \int_0^{\infty} dk \frac{k^3}{k^2} \langle |W_R(k)|^2 \rangle, \quad (2.24)$$

where $W_R(k) = \frac{3}{(kR)^2} \left( \sin kR - kR \cos kR \right)$. The CDM power spectrum is defined by

$$\langle \tilde{\delta}(k) \tilde{\delta}(k') \rangle = \left(2\pi\right)^3 P_{cc}(k) \delta^3(k - k'), \quad (2.25)$$

where $\tilde{\delta}(k)$ is the Fourier transform of $\delta(x)$ and $\delta^3$ is the Dirac delta function.
CHAPTER 2. ANTHROPOIC ORIGIN OF THE NEUTRINO MASS?

2.3.4 Galaxy Formation: Neutrino-Induced Cooling Catastrophe

The final factor $G_R(m_\nu, t)$ in Eq. (2.19) is the fraction of the halo mass in baryons within galaxies. To approximate this, we must first investigate the effect of a top-down structure scenario (present at $m_\nu \gtrsim 8 - 10$ eV, as discussed in Sec. 2.1) on galaxy formation.

In our universe galaxies form in halos with masses between $10^7 M_\odot$ and $10^{12} M_\odot$. Larger halos can inherit galaxies from mergers, resulting in galaxy groups and clusters, with masses ranging from $10^{13} M_\odot$ to $10^{15} M_\odot$. However, halos in the latter mass range do not themselves produce a significant amount of stars, relative to their total mass.

This fact can be understood as a consequence of the ability, or failure, of baryons to cool rapidly inside newly formed dark matter halos. (For more detail, see Appendix A.3 and references given there.) Baryons are shock-heated to a virial temperature $T_{\text{vir}}$ when they fall into a large dark matter halo. In order to condense into a galaxy at the center of the halo, the baryons must first shed their thermal energy. Cooling can occur by bremsstrahlung at temperatures large enough to ionize hydrogen, or by atomic and molecular line cooling at the lower temperatures attained in smaller halos.

Analytically, one can estimate the time it takes baryons to cool, $t_{\text{cool}}$. The cooling time grows with the mass of the halo (for large masses), and with the time of its formation. It is also easy to compute the gravitational timescale of the halo, $t_{\text{grav}}$, which is somewhat shorter than the time of its formation.

A good match to observation is obtained by the following criterion. If $t_{\text{cool}} < t_{\text{grav}}$, then cooling is efficient. A significant fraction of baryons (up to 10%) is converted into stars. This process occurs rapidly, on a timescale that can be treated as instantaneous compared to the

---

Figure 2.7: The Press-Schechter factor (solid lines) and its derivative (dashed lines) at the galaxy scale, $10^{12} M_\odot$, for a normal hierarchy with $m_\nu = 0$ eV (black), $m_\nu = 2$ eV (brown), $m_\nu = 4$ eV (blue), $m_\nu = 6$ eV (purple) and $m_\nu = 8$ eV (red). Each is used to define either of the two observer models of Sec. 2.2.2. Note that massive neutrinos suppress structure at all times, but much more so at early times [114–119].
age of the universe when the halo formed.

On the other hand, if $t_{\text{cool}} > t_{\text{grav}}$, then star formation is limited by the cooling time. In this regime, one would still expect a certain amount of rapid star formation at the dense core of the halo, but this is not seen in observations. (This is known as the cooling flow problem.) Observations do not constrain the possibility that a significant portion of baryons will form stars in the distant future, on a timescale much greater than the age of the universe. This time would greatly exceed $t_\Lambda$. Since the causal patch is of a fixed physical size of order the de Sitter horizon scale, there will be exponentially few halos left in it at late times. Thus, star formation at very late times does not contribute to the probability of a particular universe. (This sensitivity to the matter content inside the cosmological horizon is a key feature distinguishing the causal patch from other interesting measures, such as the fat geodesic or scale factor time cutoff [19], and it is responsible for several of the chief successes of the causal patch, e.g. [92, 102, 120–125].) Thus, we may take $t_{\text{cool}} < t_{\text{grav}}$ as a robust condition for galaxy formation to occur in a newly formed halo.

The cooling function that determines the rate of heat dissipation has a complicated form in the relevant halo mass range (see [126] and references therein). Appendix A.3 describes two different approximations to $t_{\text{cool}}$ and $t_{\text{grav}}$ that capture different cooling regimes that halos in our analysis might explore. One finds in either regime that at late times, cooling is inefficient for halo masses above the scale of the Milky Way halo:

$$M_{\text{vir}} > 10^{12} M_\odot, \quad t_{\text{vir}} \gtrsim O(\text{Gyr}) \implies \text{No Galaxy} \quad (2.26)$$

Importantly, the boundary is consistent with the observation that in our universe, there are no galaxies much larger than the Milky Way.

It would be interesting to implement a more precise version of the above boundary as a cutoff on the time until galaxy formation is efficient, at any value of $m_\nu$. Massive neutrinos delay structure formation more dramatically than they suppress it (Fig. 2.7), so such a cutoff would exclude an appreciable fraction of halos from contributing to galaxy formation even at rather small $m_\nu$. Thus it would lead to a greater suppression of intermediate neutrino masses between 1 and 10 eV, and thus would favor the anthropic approach. Instead, we will argue more conservatively for a cooling cutoff on $m_\nu$ around 10 eV. We will now identify a change of regime for $m_\nu > 10 \text{ eV}$. As we shall see, this transition places the dominant halo population so far into the regime of inefficient cooling, that the above rough estimate suffices to conclude that galaxy formation is highly suppressed.

For $m_\nu \lesssim 8 – 10 \text{ eV}$, recall that the dimensionless matter power spectrum $k^3 P_{\text{cc}}(k)$ increases monotonically with $k$ (see Fig. 2.1a), and the integral for the smoothed density contrast $\sigma_R$ in Eq. (2.24) is dominated by the power at the small galactic scale $R$. In this range, the power spectrum preserves the standard hierarchical structure formation we see in our universe, where low mass halos generally form earlier than more massive ones. Thus, it is not likely for a $10^{12} M_\odot$ halo to be nested inside a more massive overdensity that collapses at the same time.
Above $m_\nu \approx 8 - 10 \text{ eV}$, neutrinos suppress small scale power so much that the dimensionless power spectrum $k^3 P_{cc}(k)$ develops a maximum near the scale associated with free streaming $k_{nr}$ (Fig. 2.1b). This corresponds to a mass of order $5 - 100$ times the scale of the Milky Way halo, roughly the scale of galaxy clusters. It implies that the smoothed density contrast on small scales such as $10^{12} M_\odot$ is no longer dominated by the power at the corresponding wavenumber $k$. Instead, the integral in Eq. (2.24) is dominated by the maximum of the integrand, near $k_{nr}$.

This implies that $10^{12} M_\odot$ overdensities become gravitationally bound at the same time as overdensities on larger scales: a top-down scenario. The virial temperature and cooling time will be set by the largest scale that the $10^{12} M_\odot$ overdensity is embedded in, $M_{\text{vir}} \gg 10^{12} M_\odot$. Moreover, for such large halos virialization will occur quite late (see Fig. 2.7), $t_{\text{vir}} \gg 5.3$ Gyr. Hence, for $m_\nu \gtrsim 8 - 10 \text{ eV}$, the cooling condition in Eq. (2.26) becomes violated, by a substantial margin.

Note that this conclusion is insensitive to the halo mass scale we associate with observers. Whether we require $10^{10} M_\odot$ or $10^{12} M_\odot$ halos: if the power spectrum peaks at larger scales, the putative galactic halos will be embedded in and virialize together with perturbations on a mass scale well above $10^{12} M_\odot$, leading to a cooling problem.

Let us summarize these considerations and formulate our cooling cutoff on the neutrino mass. If there exists some large scale $k_* < k_{\text{gal}}$ such that $k_*^3 P_{cc}(k_*) > k_{\text{gal}}^3 P_{cc}(k_{\text{gal}})$, we interpret this as indicating top-down structure formation. Let $m_{\nu}^{\text{max}}$ be the greatest neutrino mass sum for which this criterion is not met, i.e., the largest neutrino mass compatible with bottom-up structure formation. From Boltzmann codes we find $m_{\nu}^{\text{max}} = 7.7 \text{ eV}$ for the normal hierarchy and $m_{\nu}^{\text{max}} = 10.8 \text{ eV}$ for the degenerate hierarchy. We have argued that cooling fails substantially in the top-down regime, because the first virialized halos are large and form late. Hence, we treat $m_{\nu}^{\text{max}}$ as a sharp catastrophic boundary. We approximate $G_R$ as a step function that vanishes past this critical mass:

$$G_R(m_\nu, t) = \begin{cases} 
1 & m_\nu < m_{\nu}^{\text{max}} \\
0 & m_\nu \geq m_{\nu}^{\text{max}}.
\end{cases} \quad (2.27)$$

We evaluate the integral in Eq. (2.19) numerically using Mathematica. The integration is started before structure begins to form, at redshift $z = 12$, when $F_R$ is negligible. The integration is terminated deep in the vacuum era when $r_{\text{patch}}$ becomes exponentially small. Our final result is described in Sec. 2.1; see Figures 2.3 and 2.4.

---

11 The peak (the free streaming scale) moves to smaller scales as $m_\nu$ is increased. Eventually it crosses the galaxy scale: for $m_\nu \gtrsim 100 \text{ eV}$ neutrinos act as cold dark matter. But this does not yield an anthropically allowed region, because the dark matter to baryon density ratio $\zeta$ will be too large. This may be detrimental to disk fragmentation [93,98]. If the causal patch is used, $\zeta \gg 1$ is robustly suppressed independently of any effects on galaxy and star formation, because the total mass of baryons (and thus of observers) in the patch scales like $(1 + \zeta)^{-1}$ [92].
Bibliography


Appendix A

A.1 Cosmological Constant and the Causal Patch

The cosmological constant offers a nice example of the predictive power of a large landscape, and it also illustrates the advantages of the causal patch measure over competing proposals. In this appendix we review Weinberg’s 1987 prediction of a positive cosmological constant \[50\], which has since been confirmed by observation \[3,4\]. We then turn to the more recent success of the causal patch measure in improving the quantitative agreement with the observed magnitude of \(\Lambda > 0\) (particularly in settings where the primordial density contrast is also allowed to vary), while eliminating specific anthropic assumptions. The goal is to make contact between an example many readers will be familiar with, and the more general formalism for making predictions in the landscape described in Sec. 2.2.

A.1.1 Weinberg’s Prediction: \(\Lambda \sim t_{\text{vir}}^{-2}\)

Because \(\Lambda = 0\) is not a special value from the point of view of particle physics, the prior distribution over the cosmological constant \(\Lambda\) should have no sharp feature near \(\Lambda = 0\); hence to leading order in a Taylor expansion, \(dN_{\text{vac}}/d\Lambda \approx \text{const.}\) for \(|\Lambda| \ll 1\). Hence we have

\[P_{\text{vac}}(\Lambda) \propto \Lambda = \exp(\log \Lambda) : \quad (A.1)\]

the prior favors large magnitude of the cosmological constant. So far, this is just a restatement of the cosmological constant problem in a landscape setting: among many (nonsupersymmetric) vacua, most will tend to have large \(\Lambda\), since precise cancellations between the positive and negative contributions to \(\Lambda\) are unlikely.

For \(\Lambda > 0\), structure formation would be severely diminished if \(\Lambda\) was large enough to dominate over the matter density of the universe before the time \(t_{\text{gal}}\) when density perturbations on the scale of galactic haloes would otherwise become nonlinear. (For negative \(\Lambda\) of sufficient magnitude, the universe recollapses too soon.) Crudely, the weighting factor \(w(x)\)

\[w(x) \approx \exp(-x^2)\]

\[\Rightarrow \quad P_{\text{vac}}(\Lambda) \propto \Lambda = \exp(\log \Lambda) : \quad (A.1)\]
may be approximated as vanishing for $\Lambda > \rho_{\text{NL}}$ and constant for $\Lambda < \rho_{\text{NL}}$, where $\rho_{\text{vir}} \sim t_{\text{gal}}^{-2}$ is the energy density at that time \[^{50}\]. A refinement \[^{91}\] models $w(x)$ as the fraction of baryons that enter structure of a specified minimum mass.

Thus, the resulting distribution $P(\log \Lambda) = w f$ peaks around $x \sim -2 \log t_{\text{gal}}$. $P$ is suppressed at larger values of $x$ due to the anthropic factor $w$, and at smaller values of $x$ because the prior probability $f$ is low. The model, proposed by Weinberg in 1987, thus predicted a nonzero cosmological constant not much smaller than $\rho_{\text{NL}}$. Just such a value has since been discovered \[^{3,4}\]. The model could have been ruled out at any level of confidence if, instead of a detection, the observational upper bound on $\Lambda$ had continued to improve, moving ever deeper into the region suppressed by the prior.

Weinberg’s argument had a few shortcomings, which we list here. First, the approach actually favors a somewhat larger value of $\Lambda$; the observed value is small at $2-3 \sigma$ depending on the assumptions made about the size of galaxies required by observers. More concerningly, the approach would not appear to be robust against variations of the initial density contrast $Q$. It strongly favors vacua in which both $Q$ and $\Lambda$ are larger than the observed values, unless the prior for $Q$ favors a small magnitude, or unless there is a catastrophic boundary very close to the observed values of $Q$. Neither of these arguments are easy to make.

### A.1.2 Causal Patch Prediction: $\Lambda \sim t_{\text{obs}}^{-2}$

In much of the older literature, the divergences of eternal inflation were regulated by computing the number of observers per baryon. (See the beginning of Sec. 2.2.3 for a brief discussion of the measure problem, and Ref. \[^{6}\] for a review.) This was a reasonable first guess, particularly in the context of a landscape where only the cosmological constant varies. However, it is no longer viable in light of more recent insights \[^{107,108}\].

The ratio is not well-defined in a landscape where some vacua may not contain any baryons. Worse, it does not actually regulate all infinities, since a long-lived metastable vacuum with positive cosmological constant (such as ours) will have infinite four-volume in any comoving volume; hence, an infinite number of observers “per baryon” will be produced by thermal fluctuations at late times. The number of measures that are well-defined and not clearly ruled out is surprisingly small, and the causal patch measure has had the greatest quantitative success so far (at least \[^{109}\] when we are interested in relative probabilities for events in vacua with positive cosmological constant, as we are here). Here we give two examples.

First let us recompute the probability distribution over the cosmological constant $dP / d \log \Lambda$, with $\Lambda > 0$, using the causal patch. We consider a class of observers that live at the (arbitrary but fixed) time $t_{\text{obs}}$; for comparing with out observations, we will choose $t_{\text{obs}} = 13.8 \text{ Gyr}$. But the causal patch at late times coincides with the interior of the cosmological horizon. Because of the exponential expansion, the average density decreases like $e^{-3t_{\text{obs}}/t_{\Lambda}}$. If $t_{\text{obs}} \gg t_{\Lambda} \sim \Lambda^{-1/2} \sim O(10) \text{ Gyr}$, no observers will be present in the patch, no matter
whether or not galaxies form. This is a much more stringent cutoff than the suppression of galaxy formation which only sets in for a larger value of $\Lambda$, such that $t_{\text{gal}} \gg t_{\Lambda}$. It agrees very well with the observed value of $\Lambda$, resolving the mild $(2 - 3\sigma)$ tension with Weinberg’s estimate. It is unaffected by any increase in the primordial density contrast, since $t_{\text{obs}}$ contains Gyr time scales that are not shortened by hastening structure formation. It solves the “Why Now” problem directly. And it does all this without making any specific assumptions about the nature of observers, except that they are made of stuff that redshifts faster than vacuum energy. (However, in the present paper we do assume that observers require galaxies.)

The causal patch can also explain why dark and baryonic matter have comparable abundances: the “Why Comparable” coincidence. One makes the qualitative assumption that the dark-to-baryonic density ratio $\zeta$ favors large values. But when $\zeta \gg 1$, the causal patch suppresses baryonic observers by a factor $1/(1 - \zeta)$, which counteracts the prior distribution, leading to the prediction that $\zeta \sim O(1)$ \[92\].

### A.2 Structure Formation with Neutrinos

Our calculation was done almost entirely using Boltzmann codes, not analytic approximations. However, for completeness we summarize here the physical origin of the effects of neutrinos on structure formation. In the final subsection \[A.2\] we explain the semi-analytic extrapolation formula we have used to extend the code output to negative redshifts. For excellent in-depth treatments of neutrino cosmology, see Refs. \[89\] [90].

#### A.2.1 Neutrino Cosmology

Around a second after the big bang at the time of decoupling, neutrinos are frozen out with a Fermi-Dirac distribution whose temperature is set by the primordial plasma. Due to $e^\pm$ annihilations that heat up the plasma soon after neutrino decoupling, this temperature differs from the temperature of the CMB, which decouples from the plasma much later: $T_{\nu,0} = (4/11)^{1/3} T_{\text{CMB}} = 1.95$ K.

The energy density and pressure of a single neutrino with mass $m$ at a fixed time since decoupling is thus approximately given by

\[
\rho_{\nu} = 2 \int \frac{d^3p}{(2\pi)^3} \frac{\sqrt{p^2 + m^2}}{e^{p/T_{\nu}(z)} + 1},
\]

\[
P_{\nu} = 2 \int \frac{d^3p}{(2\pi)^3} \frac{p^2}{3 \sqrt{p^2 + m^2} e^{p/T_{\nu}(z)} + 1},
\]

where $T_{\nu}(z) = T_{\nu,0}(1 + z)$ is the neutrino temperature as it redshifts from the value set at decoupling.
At early times, neutrinos contribute as radiation and add to the total radiation density as
\[ \rho_R = \left[ 1 + \frac{7}{8} \left( \frac{4}{11} \right)^{4/3} N_{\text{eff}} \right] \rho_\gamma, \tag{A.4} \]
where
\[ \rho_\gamma = \frac{\pi^2}{15} T_{\text{CMB}}^4, \tag{A.5} \]
and where \( N_{\text{eff}} = 3.046 \) is the effective number of neutrino species, with a slight deviation from 3 due to non-thermal spectral distortions from the \( e^\pm \) annihilations.

Similarly, the number density of neutrinos per species is set by the CMB number density:
\[ n_\nu = \frac{3}{11} n_\gamma, \tag{A.6} \]
where
\[ n_\gamma = \frac{2\zeta(3)}{\pi^2} T_{\text{CMB}}^3. \tag{A.7} \]

Neutrinos become approximately non-relativistic once their thermal energy drops below the relativistic kinetic energy, \( 3T_\nu(z) < m_\nu \), which occurs at a redshift \( z_{\text{nr}} \) of
\[ 1 + z_{\text{nr}} = 1991 \left( \frac{m_\nu}{1 \text{ eV}} \right). \tag{A.8} \]

Well after this transition, the density of non-relativistic neutrinos asymptotes to
\[ \rho_\nu = m_\nu n_\nu, \tag{A.9} \]
where \( m_\nu \) is the sum of masses of all non-relativistic neutrino species. In terms of this, the neutrino density parameter counting only massive neutrinos is
\[ \Omega_\nu = \frac{\rho_\nu}{\rho_\ast}, \tag{A.10} \]
where \( \rho_\ast \) is the critical density defined by \( H^2 = 8\pi G \rho_\ast/3 \), which gives
\[ \Omega_\nu h^2 = \left( \frac{m_\nu}{94.5 \text{ eV}} \right). \tag{A.11} \]

The neutrino free streaming scale is set by the typical distance neutrinos travel thermally up to a given time. Roughly, it is given by the horizon scale at early times and stops growing soon after the neutrinos become nonrelativistic; hence it can be crudely approximated by the horizon scale at the nonrelativistic transition, \( k_{\text{nr}} \).

\footnote{The non-relativistic transition is far from sudden. The neutrino pressure Eq. (A.3) has a non-negligible tail long after the redshift Eq. (A.8), which smears out the transition. We thank J. Lesgourgues for explaining this point to us.}
APPENDIX A. APPENDIX

On small scales, there are two effects by which neutrinos suppress structure. The most obvious is that density perturbations will be washed out. Thus, free streaming eliminates the contribution of neutrinos to structure, and thus suppresses the total matter power by a factor $\sim (1 - f_\nu)^2$, where

$$f_\nu = \frac{\Omega_\nu}{\Omega_m}.$$  \hspace{1cm} (A.12)

defines the massive neutrino fraction. Conversely, on larger scales neutrinos will remain confined to the over-dense regions and will behave like cold dark matter.

A secondary but more important effect is that the density of massive neutrinos contribute via the Friedmann equation to the Hubble parameter, which controls the friction term in the growth of matter perturbations. But on short scales, they do not contribute to the source term (the density contrast). Therefore, CDM perturbations grow more slowly in the presence of a nonclustering matter component on short scales [127]:

$$\delta_c \propto a^p, \quad k > k_{nr},$$

$$\delta_c \propto a, \quad k \lesssim k_{nr},$$  \hspace{1cm} (A.13)

where

$$p = -1 + \frac{\sqrt{1 + 24(1 - f_\nu)}}{4} \approx 1 - \frac{3}{5} f_\nu < 1,$$  \hspace{1cm} (A.14)

with the last approximation valid in the limit of small neutrino masses.

![Figure A.1](image)

Figure A.1: The growth factor Eq. (A.15) (solid line), which behaves like $x^{1/3}$ (dashed line) during the matter era, and asymptotes to a constant value well above $x(t_\Lambda) = 1$.

A.2.2 Late-Time Extrapolation of Numerical Results

Available Boltzmann codes do not offer output for negative redshifts. In order to estimate the smoothed density contrast $\sigma_R$ in this regime, we extrapolate our numerical results for
σ_R(z) from positive to negative z, i.e., from $a < 1$ to $a > 1$. The most straightforward approach would be a linear extrapolation in some time variable, fitting both the value and the derivative of $\sigma_R$ at $z = 0$. However, there is a physical effect that we must incorporate analytically: vacuum domination turns off structure growth on all scales. This effect is not strong enough at $z = 0$ to have a significant imprint on the value or time derivative of $\sigma_R$. However, the effect is also rather simple, and thus easy to incorporate analytically.

Figure A.2: The parameters $C_{\text{eff}}$ and $p_{\text{eff}}$, for normal (orange, top) and degenerate (green, bottom) hierarchies, obtained by fitting Eq. (A.23) to CAMB output for $\sigma_R$ and its derivative at $z = 0$. The resulting fitting function for $\sigma_R(z)$ is used to compute the Press-Schechter factor at negative redshift only. Note that $p_{\text{eff}} \approx 1$ throughout. This may seem surprising, but it is consistent with our earlier finding that at large neutrino masses, the scales whose power contributes dominantly to $\sigma_R$ are precisely the ones on which free-streaming is not effective. This is closely related to the discrepancy we find with Ref. [98], whose estimate $p_{\text{eff}} \approx p(k_{\text{gal}}) \approx 1 - 8 f_{\nu}$ would yield a monotonically decreasing curve in (b).

In a universe with negligible neutrino mass, the CDM density contrast grows as

$$\delta \propto G_\Lambda(x) \equiv \frac{5}{6} \sqrt{1 + \frac{1}{x} \int_0^x \frac{dy}{y^{1/6}(1+y)^{3/2}}} ,$$

where

$$x \equiv \frac{\rho_\Lambda}{\rho_m} = \left. \frac{\Omega_\Lambda}{\Omega_m} \right|_{z=0} (1 + z)^{-3} .$$

As seen in Fig. [A.1], density perturbations grow like the scale factor during the matter dominated era; they asymptote to a constant value at times $t > t_\Lambda$.

With nonzero neutrino mass, a reasonable approximation is obtained by combining the analytic result for the matter era, Eq. (A.13), with the $m_\nu = 0$ transition to the vacuum
dominated era:

\[ \delta \propto G^\Lambda(x) \quad (k < k_{nr}), \quad (A.17) \]

\[ \delta \propto G^\Lambda(x)^p \quad (k > k_{nr}). \quad (A.18) \]

Recall that \( P_{cc}(k) \propto \delta_c(k)^2 \) by Eq. (2.25).

In order to improve on this result, we can incorporate the information gained from the use of Boltzmann codes. Instead of computing \( p \) and \( k_{FS} \) analytically as described in the previous subsection, we can read off a slope \( p(k) \) from the numerical output near \( z = 0 \):

\[ p(k) \equiv \frac{1}{2} \frac{d \log P_{cc}(x)}{d \log G^\Lambda(x)} \quad (A.19) \]

We can also fix the constant of proportionality \( C \) by matching the magnitude of \( P_{cc} \) obtained from CAMB at \( z = 0 \). This yields a semi-analytic power spectrum as a function of time, for any fixed \( k \) and fixed neutrino mass:

\[ P_{cc}(x) = CG^\Lambda(x)^{2p(k)}. \quad (A.20) \]

In practice, it is cumbersome to extrapolate the power at each wave number only to integrate over scales to obtain the smoothed density contrast. By the late time corresponding to \( z = 0 \), for any neutrino mass, we expect that the integral in Eq. (2.24) is dominated by the power at some scale \( k \) and will remain dominated by the same scale in the future \( (z < 0) \). For small neutrino masses, this scale will be set by the galaxy scale; for large \( m_\nu \) it will be the scale of the peak of the spectrum \( k^3P(k) \). We incorporate this by matching the analytic growth for \( z < 0 \) directly to the numerical results for \( \sigma_R(x) \) at \( z = 0 \). For every \( m_\nu \), we compute

\[ p_{\text{eff}} \equiv \frac{d \log \sigma_R(x)}{d \log G^\Lambda(x)} \bigg|_{z=0}, \quad (A.21) \]

\[ C_{\text{eff}} \equiv \frac{\sigma_R}{G^\Lambda(x)p_{\text{eff}}} \bigg|_{z=0} \quad (A.22) \]

from the CAMB output for small nonnegative redshifts. The results are shown in Fig. A.2.

As our semi-analytic approximation entering the Press-Schechter factor \( F \) for \( z < 0 \) we use

\[ \sigma_R(z) = C_{\text{eff}}G^\Lambda(x(z))^{p_{\text{eff}}} \quad [\text{used for } z < 0 \text{ only}] \quad (A.23) \]

with \( G^\Lambda \) given by Eq. (A.15). We have checked that the same formula provides an excellent fit to the numerical results at \( z > 0 \), as one would expect. However, we stress again that we use the output from the CAMB code in this regime, not the fitting function. Moreover, the regime \( z > 0 \) dominates in our calculation because the comoving volume of the causal patch decreases rapidly below \( z = 2 \).
A.3 Cooling and Galaxy Formation

In this Appendix, we review the basic time scales that are believed to control cooling flows in dark matter halos. Our discussion closely follows Ref. [102], where further details and references can be found.

Baryonic gas will fall into the gravitational well of newly formed dark matter halos. The baryons are thus shock-heated to high temperatures. In order for stars to form, the baryonic gas must cool and condense. The initial temperature of the baryons is called the virial temperature. By the virial theorem,

\[
\frac{GM_{\text{vir}}\mu}{5R_{\text{vir}}} = T_{\text{vir}},
\]

where \(M_{\text{vir}}\) is the mass of the halo and \(R_{\text{vir}}\) is its virial radius. In the regime of interest for us, \(T_{\text{vir}}\) is large enough to ionize hydrogen. Then one can take the average molecular mass \(\mu\) to be \(m_p/2\), where \(m_p\) is the mass of the proton. With \(M_{\text{vir}} = \frac{4\pi}{3}\rho_{\text{vir}}R_{\text{vir}}^3\) one finds

\[
T_{\text{vir}} \propto M_{\text{vir}}^{2/3} \rho_{\text{vir}}^{1/3},
\]

where the “constants” of proportionality depend negligibly on \(M_{\text{vir}}\).

The timescale for cooling by bremsstrahlung is

\[
t_{\text{brems}} \propto \frac{T_{\text{vir}}^{1/2}}{\rho_{\text{vir}}} \propto \frac{M_{\text{vir}}^{1/3}}{\rho_{\text{vir}}^{5/6}}.
\]

We will be interested in how this timescale compares to the age of the universe when the halo virializes,

\[
t_{\text{vir}} \propto \rho_{\text{vir}}^{-1/2}.
\]

If \(t_{\text{brems}} \lesssim t_{\text{vir}}\), then galaxy formation can be treated as instantaneous, i.e., as occurring nearly simultaneously with halo formation. Keeping track of all constants [102], one finds that this case corresponds to

\[
M_{\text{vir}} t_{\text{vir}}^2 \lesssim (10^{12} M_\odot)(2.2\text{Gyr})^2.
\]

In the opposite case, \(M_{\text{vir}} t_{\text{vir}}^2 \gg (10^{12} M_\odot)(2.2\text{Gyr})^2\), we have \(t_{\text{brems}} \gg t_{\text{vir}}\). In halos with these mass and virialization time combinations, galaxy formation cannot be treated as instantaneous. Instead, it takes a much greater time \(t_{\text{brems}} \gg t_{\text{vir}}\) to convert a comparable fraction of baryons into stars. (If feedback or major mergers disrupt the cooling flow, the contrast would be even more drastic, but we will not assume this here.)

The above analysis assumed cooling of unbound charged particles by bremsstrahlung. This approximation is best for virial temperatures above \(10^7\) K. At lower temperatures the
cooling function is quite complicated, but one can get an estimate by treating it as independent of $T_{\text{vir}}$ in some range $T_{\text{vir}}$. With this approximation, one obtains that the cooling condition is satisfied for

$$M_{\text{vir}}^2 t_{\text{vir}} < (10^{12} M_\odot)^2 (5.3 \text{ Gyr}) .$$

(A.29)

With either scaling, one finds again that cooling is inefficient if $M_{\text{vir}} > 10^{12} M_\odot$, particularly for late virialization $t_{\text{vir}} \gtrsim 10 \text{ Gyr}$.

So far, we have neglected the effects of the cosmological constant. For halos that form deep in the vacuum dominated era, one should use $\rho_{\text{vir}} \sim \rho_\Lambda$ instead of Eq. (A.27). But such halos contribute negligibly in the causal patch because they will be exponentially dilute.

We have also neglected neutrinos. However, Eq. (A.28) is sufficiently general to capture their main effect, which is to change the relation between $M_{\text{vir}}$ and $t_{\text{vir}}$. In a universe with $m_\nu \ll 8 \text{ eV}$, $t_{\text{vir}}$ grows logarithmically with $M_{\text{vir}}$ for overdensities of a fixed relative amplitude. For $10^{12} M_\odot$ halos forming from $1 \sigma$ ($2 \sigma$) overdensities, $t_{\text{vir}} \approx 3.6 \text{ Gyr}$ ($t_{\text{vir}} \approx 1.3 \text{ Gyr}$) and by Eq. (A.28), cooling fails (succeeds).

In a universe with $m_\nu \gtrsim 8 \text{ eV}$, however, small scale power is so suppressed that structure formation proceeds in a top-down manner. (This is shown in detail in the main text.) Then structure on all scales forms much later than 2.4 Gyr. Moreover, smaller structure is embedded in larger halos, which set the virial mass that enters Eq. (A.28). Hence, the timescale for a significant fraction of baryons to form stars is at least $t_{\text{brms}} \gg t_{\text{vir}} \gg O(\text{Gyr})$. 