

## Bayesian evidence for the tensor-to-scalar ratio $r$ and neutrino masses $m_\nu$ : Effects of uniform versus logarithmic priors

L. T. Hergt<sup>1,2,\*</sup>, W. J. Handley<sup>1,2,†</sup>, M. P. Hobson<sup>1,‡</sup> and A. N. Lasenby<sup>1,2,§</sup>

<sup>1</sup>*Astrophysics Group, Cavendish Laboratory,*

*J. J. Thomson Avenue, Cambridge CB3 0HE, United Kingdom*

<sup>2</sup>*Kavli Institute for Cosmology, Madingley Road, Cambridge CB3 0HA, United Kingdom*



(Received 24 February 2021; accepted 22 April 2021; published 8 June 2021)

We review the effect that the choice of a uniform or logarithmic prior has on the Bayesian evidence and hence on Bayesian model comparisons when data provide only a one-sided bound on a parameter. We investigate two particular examples: the tensor-to-scalar ratio  $r$  of primordial perturbations and the mass of individual neutrinos  $m_\nu$ , using the cosmic microwave background temperature and polarization data from Planck 2018 and the NuFIT 5.0 data from neutrino oscillation experiments. We argue that the Kullback–Leibler divergence, also called the relative entropy, mathematically quantifies the Occam penalty. We further show how the Bayesian evidence stays invariant upon changing the lower prior bound of an upper constrained parameter. While a uniform prior on the tensor-to-scalar ratio disfavors the  $r$  extension compared to the base  $\Lambda$ CDM model with odds of about 1:20, switching to a logarithmic prior renders both models essentially equally likely.  $\Lambda$ CDM with a single massive neutrino is favored over an extension with variable neutrino masses with odds of 20:1 in case of a uniform prior on the lightest neutrino mass, which decreases to roughly 2:1 for a logarithmic prior. For both prior options we get only a very slight preference for the normal over the inverted neutrino hierarchy with Bayesian odds of about 3:2 at most.

DOI: [10.1103/PhysRevD.103.123511](https://doi.org/10.1103/PhysRevD.103.123511)

### I. INTRODUCTION

The “principle of insufficient reason” (Bernoulli [1]) or “principal of indifference” (renamed by Keynes [2]) states that, in the event of multiple, mutually exclusive, possible outcomes and in the absence of any relevant evidence, we should assign the same probability to all outcomes [3]. In a Bayesian analysis, this is generalized to continuous parameters in the form of uninformative priors. Complete prior ignorance about a location parameter is represented by assigning a uniform distribution to the prior. Ignorance about a scale parameter on the other hand is represented by assigning a logarithmic prior, i.e., a uniform distribution on the logarithm of the parameter [3]. However, it is not always clear whether a parameter should be treated as a location or scale parameter. This is quite commonly discussed when faced with a strictly positive parameter such as a mass or an amplitude that is very small, yet still unconstrained. In general, the decision whether to use a uniform or logarithmic prior has effects on credibility bounds and on the Bayesian evidence, i.e., on both levels of Bayesian inference: parameter estimation and model

comparison. Under the reasoning that you can set the lower bound to zero and thus incorporate all possible small values, the uniform prior is often preferred, whereas the logarithmic prior is criticized for a lack of an unambiguous lower bound, and because the ultimate choice of the lower bound might affect a 95% credibility bound and the Bayesian evidence.

In this paper we show that the very last statement is typically not true and that the choice of a lower bound for such a logarithmic prior is less problematic than commonly assumed. To that end we will look at two cosmological examples in particular: the tensor-to-scalar ratio  $r$  of primordial perturbations as well as the neutrino masses  $m_\nu$ , where both uniform and logarithmic priors have been applied historically (for the tensor-to-scalar ratio, see, e.g., [4–9]; and for the neutrino masses see, e.g., [7,10–18]).

The best constraints on the tensor-to-scalar ratio  $r_{0.05} \lesssim 0.06$  come from joint analyses of cosmic microwave background (CMB) data, CMB lensing, and baryon acoustic oscillations (BAO) [7,8], where a uniform prior on  $r$  was adopted. A common goal of upcoming CMB experiments such as the Simons Observatory [19], the LiteBIRD satellite [20] and the next-generation “Stage-4” ground-based CMB experiment (CMB-S4) [21] is to push to a tensor-to-scalar ratio of  $r \sim 10^{-3}$ . In pushing to such small values of  $r$ , the question of whether to adopt a uniform or logarithmic prior in one’s analysis becomes more pertinent.

\*lh561@mrao.cam.ac.uk

†wh260@mrao.cam.ac.uk

‡mph@mrao.cam.ac.uk

§a.n.lasenby@mrao.cam.ac.uk

Since neutrino oscillation experiments measure nonzero mass differences, we can conclude that two or more neutrinos must have mass. However, the absolute scale of the individual neutrino masses  $m_i$  cannot be measured by the oscillation experiments, but only the mass-squared splittings  $\Delta m_{ij}^2 = m_i^2 - m_j^2$ . The strongest bound on the absolute neutrino mass scales is currently provided again by combined CMB and BAO data, limiting the sum of the neutrino masses to  $\sum m_\nu \lesssim 0.12$  eV at 95% confidence [17] (see also [15,18] for other recent analyses).

When investigating the three discrete neutrino mass eigenstates, the question of uniform vs logarithmic priors arises again. Note, however, that given the known mass splittings from oscillation experiments, the three neutrino mass scales are linked. If one mass scale is known, then the others can be inferred from the mass squared splittings. Hence, only one mass scale is truly unknown and assuming scale invariant (i.e., logarithmic) priors on all three neutrino masses simultaneously would unduly favor smaller neutrino masses and thus a normal neutrino hierarchy (NH) with  $m_1 < m_2 \ll m_3$  compared to an inverted neutrino hierarchy (IH) with  $m_3 \ll m_1 < m_2$  (for more on this see also discussions in [10,11]).

This paper is structured as follows: in Sec. II we will start by giving a brief description of our Bayesian analysis framework, including the data and base cosmological model used, as well as the means of computing the Bayesian evidence. In Sec. III we apply this to the tensor-to-scalar ratio  $r$  and compare to a theoretical mock example. In Sec. IV we perform the equivalent analysis for the neutrino masses and contrast the results for the two neutrino hierarchies. We conclude in Sec. V.

## II. METHODS

### A. Bayesian inference

There are two levels to Bayesian inference: parameter estimation and model comparison (see, e.g., [3,22]). Both these levels are based on Bayes's theorem, which relates inference inputs (likelihood and prior) to yielded outputs (posterior and evidence):

$$\begin{aligned} \Pr(\theta|D, M) \times \Pr(D|M) &= \Pr(D|\theta, M) \times \Pr(\theta|M), \\ \text{Posterior} \times \text{Evidence} &= \text{Likelihood} \times \text{Prior}, \\ \mathcal{P}_M(\theta) \times \mathcal{Z}_M &= \mathcal{L}_M(\theta) \times \pi_M(\theta). \end{aligned} \quad (1)$$

The posterior  $\mathcal{P}$  is the main quantity of interest in a parameter estimation, representing our state of knowledge about the parameters  $\theta$  in a given model  $M$ , inferred from our prior information  $\pi$  and the likelihood  $\mathcal{L}$  of the parameters under the data  $D$ . The evidence  $\mathcal{Z}$  is pivotal for model comparisons.

Were we interested only in parameter estimation, then it would be sufficient to care only about the proportionality of

the posterior to the product of likelihood and prior and the Bayesian evidence could be neglected as a mere normalization factor. However, for the comparison of say two models  $A$  and  $B$  the evidence becomes important with the posterior odds ratio of the two models given by

$$\frac{\Pr(B|D)}{\Pr(A|D)} = \frac{\Pr(B)}{\Pr(A)} \times \frac{\mathcal{Z}_B}{\mathcal{Z}_A}. \quad (2)$$

Typically models are assigned the same prior preference such that the first term on the right-hand side becomes unity, leaving simply the evidence ratio  $\mathcal{Z}_B/\mathcal{Z}_A$ , which can be interpreted as betting odds for the two models. We typically quote this in terms of the log-difference of evidences between two models  $\Delta \ln \mathcal{Z} = \ln(\mathcal{Z}_B/\mathcal{Z}_A)$ .

The evidence is the marginal likelihood

$$\mathcal{Z}_M = \int \mathcal{L}_M(\theta) \pi_M(\theta) d\theta = \langle \mathcal{L}_M \rangle_{\pi}, \quad (3)$$

and can be numerically approximated with Laplace's method [23], estimated from a posterior distribution attained e.g., from a Monte Carlo Markov Chain via the Savage–Dickey density ratio [24–27] or via a nearest-neighbor approach [28,29] or computed more directly with nested sampling, which can additionally estimate the corresponding numerical uncertainty [30–36].

If the posterior distribution and the evidence have both been determined, then as a byproduct one can also compute the Kullback–Leibler (KL) divergence, also called the relative entropy:

$$\mathcal{D}_{\text{KL},M} = \int \mathcal{P}_M(\theta) \ln \left( \frac{\mathcal{P}_M(\theta)}{\pi_M(\theta)} \right) d\theta = \left\langle \ln \frac{\mathcal{P}_M}{\pi_M} \right\rangle_{\mathcal{P}}, \quad (4)$$

which quantifies the overall compression from prior to posterior distribution.

### B. Kullback–Leibler divergence and Occam's razor

It should be noted that the Bayesian evidence naturally incorporates the so-called Occam's razor that penalizes models for unnecessary complexity. It can be formulated as the principle to “Accept the simplest explanation that fits the data” [22]. This can be neatly demonstrated using a Gaussian likelihood with mean  $\mu$  and variance  $\sigma^2$  having a single parameter  $x \in [x_{\min}, x_{\max}]$  with a uniform prior (see, e.g., [22,37]). The Bayesian evidence decomposes into two terms:

$$\mathcal{Z} = \mathcal{L}(\mu) \times \frac{\sigma \sqrt{2\pi}}{x_{\max} - x_{\min}}. \quad (5)$$

The first term on the right-hand side is the maximum likelihood point. With additional parameters, this term would only increase and therefore can only favor the given model. The second term incorporates the ratio of posterior to prior uncertainty. Since the posterior uncertainty  $\sigma$  is

generally smaller than the prior uncertainty ( $x_{\max} - x_{\min}$ ), this term penalizes the given model for each of its parameters and thus embodies its Occam penalty. Note that the posterior and prior uncertainties appear inversely in the normalization factor of the actual distributions.

More generally, the KL divergence can actually be used as an estimator of the Occam penalty, which becomes clearer when rewriting the log-evidence according to<sup>1</sup>:

$$\ln \left( \int \mathcal{L}_M \pi_M d\theta \right) = \int \mathcal{P}_M \ln \mathcal{L}_M d\theta - \int \mathcal{P}_M \ln \left( \frac{\mathcal{P}_M}{\pi_M} \right) d\theta,$$

$$(\text{log-})\text{evidence} = \text{parameter fit} - \text{Occam penalty},$$

$$\ln \mathcal{Z}_M = \langle \ln \mathcal{L}_M \rangle_{\mathcal{P}} - \mathcal{D}_{\text{KL},M}, \quad (6)$$

where we have dropped the dependence on  $\theta$  to save space. Analogous to the example from Eq. (5), the first term on the right-hand side encapsulates the fit of the model, while the KL divergence is the average log-ratio of posterior to prior distribution [see also the last equality in Eq. (4)], thus identifying it as the Occam penalty. This equation appears in passing in the Appendix of [38] in the calculation of tension metrics and it has been intuitively applied, e.g., in the third figure in [39], but as far as the authors are aware Eq. (6) is the first time this analytical form is explicitly connected to the trade-off between parameter fit and model complexity.

While known to experts, a widely unappreciated fact is that the evidence stays unaffected by an unconstrained parameter, i.e., when the data provide no information for that parameter [40]. In terms of Eq. (6) this is reflected in an invariant likelihood  $\langle \ln \mathcal{L} \rangle_{\mathcal{P}}$  and a zero KL divergence  $\mathcal{D}_{\text{KL}}(\pi = \mathcal{P}) = 0$ . Using the alternative labeling we can rephrase this: adding an unconstrained parameter does not affect the fit, but it also does not incur an additional Occam penalty and hence also leaves the evidence unaffected.

A popular measure for an effective number of constrained parameters is the Bayesian model complexity [41]. However, this quantity relies on the use of a point estimator such as the posterior mean or mode, which is why we prefer using the Bayesian model dimensionality  $d$  in the following sections (see [42] for a more detailed discussion on Bayesian complexities/dimensionalities). The Bayesian model dimensionality can be computed straightforwardly from the posterior distribution as the posterior variance of the log-likelihood:

$$\frac{d_M}{2} = \int \mathcal{P}_M(\theta) \left( \ln \frac{\mathcal{P}_M(\theta)}{\pi_M(\theta)} - \mathcal{D}_{\text{KL},M} \right)^2 d\theta, \quad (7)$$

$$= \langle (\ln \mathcal{L}_M)^2 \rangle_{\mathcal{P}} - \langle \ln \mathcal{L}_M \rangle_{\mathcal{P}}^2. \quad (8)$$

<sup>1</sup>Note that proving Eq. (6) becomes surprisingly straightforward when going from right to left and making use of Bayes's theorem (1).

TABLE I. Cosmological parameters of the base  $\Lambda$ CDM cosmology the way they are sampled in our Bayesian analysis. The second column shows their corresponding prior ranges. The third column lists their mean and 68% limits from our base  $\Lambda$ CDM nested sampling run with TT, TE, EE + lowE data from Planck 2018 and is in almost perfect agreement with Table 2 in [7].

Parameter	Prior range	68% limits
$\omega_b = h^2 \Omega_b$	[0.019, 0.025]	$0.02236 \pm 0.00015$
$\omega_c = h^2 \Omega_c$	[0.025, 0.471]	$0.1199 \pm 0.0014$
$100\theta_s$	[1.03, 1.05]	$1.04191 \pm 0.00029$
$\tau_{\text{reio}}$	[0.01, 0.40]	$0.0540_{-0.0084}^{+0.0073}$
$\ln(10^{10} A_s)$	[2.5, 3.7]	$3.043_{-0.016}^{+0.015}$
$n_s$	[0.885, 1.040]	$0.9641 \pm 0.0042$

Note the connection to Eq. (6), where we used the posterior average of the log-likelihood. As such, these two quantities provide an interesting additional perspective to that of the  $(\ln \mathcal{Z}, \mathcal{D}_{\text{KL}})$  pair. The posterior average of the log-likelihood informs us about the parameter fit and the posterior variance of the log-likelihood measures the models' complexity in the form of the number of constrained parameters.

### C. Cosmological models

In the following sections we perform Bayesian model comparisons on one-parameter extensions to the  $\Lambda$ CDM model (Universe dominated today by a cosmological constant  $\Lambda$  and by cold dark matter), which we parametrize with the standard six cosmological parameters listed in Table I with their corresponding prior ranges.

In Sec. III we extend the  $\Lambda$ CDM model by the tensor-to-scalar ratio  $r$  of primordial perturbations, which is set to  $r = 0$  in  $\Lambda$ CDM. In Sec. IV we extend the base model by allowing for three distinct neutrino masses. In the  $\Lambda$ CDM model these are typically fixed to two massless neutrinos and a single massive neutrino with  $m_\nu = 0.06$  eV.

### D. Data

We use the 2018 temperature and polarization data from the Planck satellite [43], which we abbreviate as "TT, TE, EE + lowE". Note that this is the same abbreviation as in the corresponding Planck publication itself. The specific use of "lowE" but lack of "lowT" might lead to the conclusion that only  $E$ -mode and no temperature data were used at low multipoles. However, this is not the case. Both high- $\ell$  and low- $\ell$  temperature autocorrelation data are implied in that abbreviation.

In Sec. IV we additionally use the NuFIT 5.0 (2020) data from neutrino oscillation experiments [44–46] to set Gaussian priors on the mass squared splittings  $\delta m^2$  and  $\Delta m^2$ .

### E. Statistical and cosmological software

We explore the posterior distributions of cosmological and nuisance parameters using COBAYA [47], which

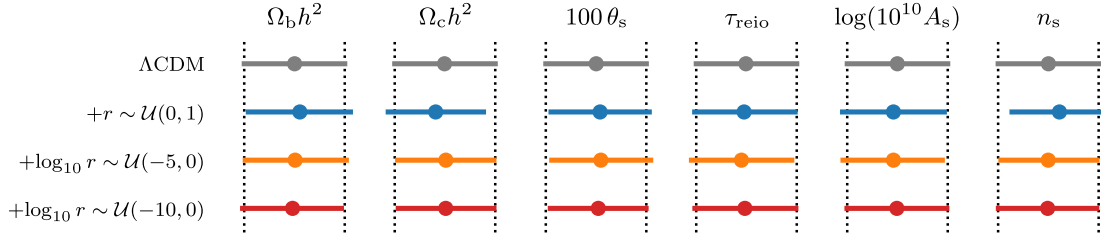


FIG. 1. Stability of the cosmological parameters for the tensor-to-scalar ratio extension of the base  $\Lambda$ CDM cosmology with different priors on  $r$ : uniform in blue, logarithmic with lower bound  $-5$  in orange and logarithmic with lower bound  $-10$  in red. For each parameter we show the mean and the extent from quantile 0.16 to 0.84, i.e., the inner 68% limits.

provides both the Monte Carlo Markov Chain sampler developed for COSMOMC [48,49] with a “fast dragging” procedure described in [50] and also the nested sampling code POLYCHORD [35,36], tailored for high-dimensional parameter spaces, which can simultaneously determine the Bayesian evidence alongside its numerical uncertainty. Both samplers are interfaced with the cosmological Boltzmann code CLASS [51–53], which computes the theoretical CMB power spectra for temperature and polarization modes.

We use GETDIST [54] to generate the data tables of marginalized parameter values. The postprocessing of the nested sampling output for the computation of Bayesian evidence, KL divergence, and Bayesian model dimensionality, as well as the plotting functionality for posterior contours is performed using the PYTHON module ANESTHETIC [55].

All inference products required to compute the results presented in this paper are available for download from ZENODO [56].

### III. TENSOR-TO-SCALAR RATIO

The tensor-to-scalar ratio  $r$  quantifies what fraction of primordial perturbations is in the form of gravitational waves, produced, e.g., during cosmic inflation and potentially detectable in their contribution to CMB  $B$ -modes.

So far, the major experiments probing the contribution of tensor modes to the CMB power spectrum have adopted a uniform prior on  $r$  [7,8]. However, the common target of  $r \sim 10^{-3}$  for many upcoming CMB experiments, such as the Simons Observatory, the LiteBIRD satellite, or CMB-S4, warrants the question as to whether a scale invariant prior might be better to handle such low values. This question frequently brings up arguments of the ambiguity of the lower bound to a logarithmic prior and its potential effect on the Bayesian evidence.

#### A. Tensor-to-scalar ratio: Posteriors

Figure 1 gives an overview of the stability of the cosmological base parameters across different priors for  $r$  and compares them to the  $\Lambda$ CDM base model by showing their mean and 68% ranges. In addition to the  $\Lambda$ CDM base run, we have taken nested sampling runs with both a

uniform prior on the tensor-to-scalar ratio  $r \sim \mathcal{U}(0, 1)$  and with two logarithmic priors with different lower bounds,  $\log_{10} r \sim \mathcal{U}(-5, 0)$  and  $\log_{10} r \sim \mathcal{U}(-10, 0)$ . The near perfect alignment across different setups reflects how little the tensor-to-scalar ratio correlates with the other parameters.

In Fig. 2 we focus on the spectral index  $n_s$  and the tensor-to-scalar ratio  $r$  (or  $\log_{10} r$ ) in particular by showing their one-dimensional marginalized posterior distributions. Figure 3 shows the corresponding two-dimensional joint probability contours of the 68% and 95% levels for  $n_s$  and  $r$  (or  $\log_{10} r$ ). We have included shaded histograms in the 1D plots and scatter points in the 2D plots to give a notion of the prior distributions.

As already expected from Fig. 1, the marginalized posterior for the spectral index is near identical, irrespective of the prior on  $r$ . The tensor-to-scalar ratio in the right panel

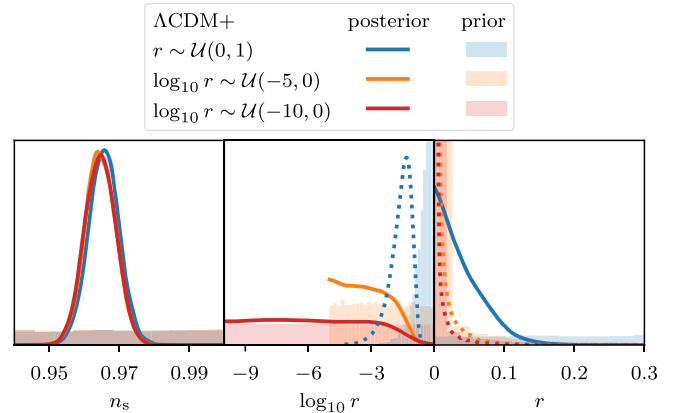


FIG. 2. Normalized one-dimensional posterior distributions for Planck 2018 TT, TE, EE + lowE data for the spectral index  $n_s$  and the tensor-to-scalar ratio  $r$  of primordial perturbations, contrasting the difference between using a uniform (blue) or logarithmic (orange and red) prior on  $r$ . The shaded histograms illustrate the prior distributions. Note that the dotted lines show the inferred parameters  $r$  and  $\log_{10} r$  in the respective opposite domain. This is done only to provide a more direct visual comparison. However, these dotted contours are not data-driven parameter constraints. In particular the blue dotted line results purely from a lack of small prior samples when sampling uniformly over  $r$ , and does not in fact constitute a lower bound on the tensor-to-scalar ratio.

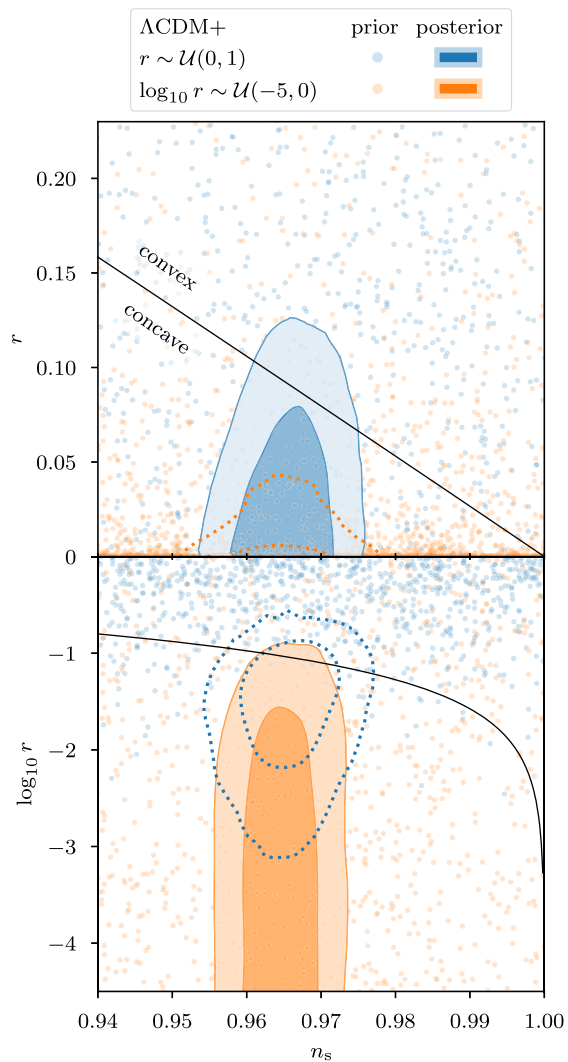


FIG. 3. Two-dimensional version of Fig. 2 showing the 68% and 95% levels of the posterior contours for Planck 2018 TT, TE, EE + lowE data for spectral index  $n_s$  and tensor-to-scalar ratio  $r$ , where again a uniform (blue) or logarithmic (orange) prior on  $r$  was used. The scattered dots give a notion of that prior distribution. Note that the dotted lines are not true constraints as explained in Fig. 2. The thin black line divides the  $n_s$ - $r$  parameter space into regions of convex and concave inflationary potentials.

of Fig. 2 drops off exponentially from  $r = 0$  to larger values, thereby significantly compressing the prior, which spans up to unity. When sampling logarithmically the posterior levels off towards small scales and shows a steplike behavior at the upper bound.

We have included the kernel density estimate from the uniform  $r$  samples in the  $\log_{10} r$  plot and vice versa (dotted lines). This allows us to compare more directly what sort of numerical values were actually used in those two cases. At a first naïve glance one might be concerned that the dotted blue line actually indicates a lower bound, however, looking at the blue shaded histogram in the 1D plot or the blue scatter points in the 2D plot it becomes clear that

this is entirely prior driven and reflects that uniform sampling of  $r$  does not reach such low values (see also [57] on a related discussion about the importance of adjusting the density when setting the  $x$  scale to “log”). With a target of  $r \sim 10^{-3}$  this highlights how the parameter space is sampled rather inefficiently at those low values of interest when applying a uniform prior, which would be an argument for adopting a logarithmic prior in the future.

One problem to be aware of with the unconstrained posteriors from a logarithmic prior is that upper bounds in form of, e.g., 95% limits will change with the lower bound on the logarithmic parameter: the smaller the lower prior bound, the smaller also the upper posterior bound. This lack of a stable posterior bound is a result of the definition via percentiles, a notion inspired by a normal distribution. For other types of distributions, such as the steplike posteriors seen in the middle panel of Fig. 2, percentiles of that sort are not the ideal measure for an upper bound. For such a steplike posterior a better alternative would be to quantify the position of the step directly, e.g., where the posterior drops to some fraction of its plateau value. In the case that an exponential distribution provides a good fit to the nonlogarithmic parameter (see the mock example in the following Sec. III C), the parameter value where the posterior is  $1/e$  times its maximum turns out to be a stable choice, which corresponds to the mean of the exponential distribution. Indeed, using that  $1/e$  measure for the step position we get roughly the same upper bound on the tensor-to-scalar ratio for all prior options:

$$\begin{aligned} r &< 0.06, \\ \log_{10} r &< -1.2. \end{aligned} \quad (9)$$

Note, that these are *not* the habitually quoted 95% upper bounds on the tensor-to-scalar ratio. For the uniform sampling run of  $r$ , this limit in this case is closer to roughly an 80% upper bound. Note further that the choice of the  $1/e$  fraction provides a particularly stable bound, because of the connection to the mean of the exponential distribution.

### B. Tensor-to-scalar ratio: Evidence and Kullback–Leibler divergence

Nested sampling provides us with distributions for log-evidence  $\ln \mathcal{Z}$ , KL divergence  $\mathcal{D}_{\text{KL}}$  and Bayesian model dimensionality  $d$  in the same way as for the posterior of free model parameters, which can be calculated straightforwardly using ANESTHETIC’s analysis tools for nested sampling output [55]. Figure 4 shows the contours for those quantities in a triangle plot. We have normalized all quantities with respect to the base  $\Lambda\text{CDM}$  model, such that, e.g., for the log-evidence we have

$$\Delta \ln \mathcal{Z} = \ln \mathcal{Z} - \ln \mathcal{Z}_{\Lambda\text{CDM}}. \quad (10)$$

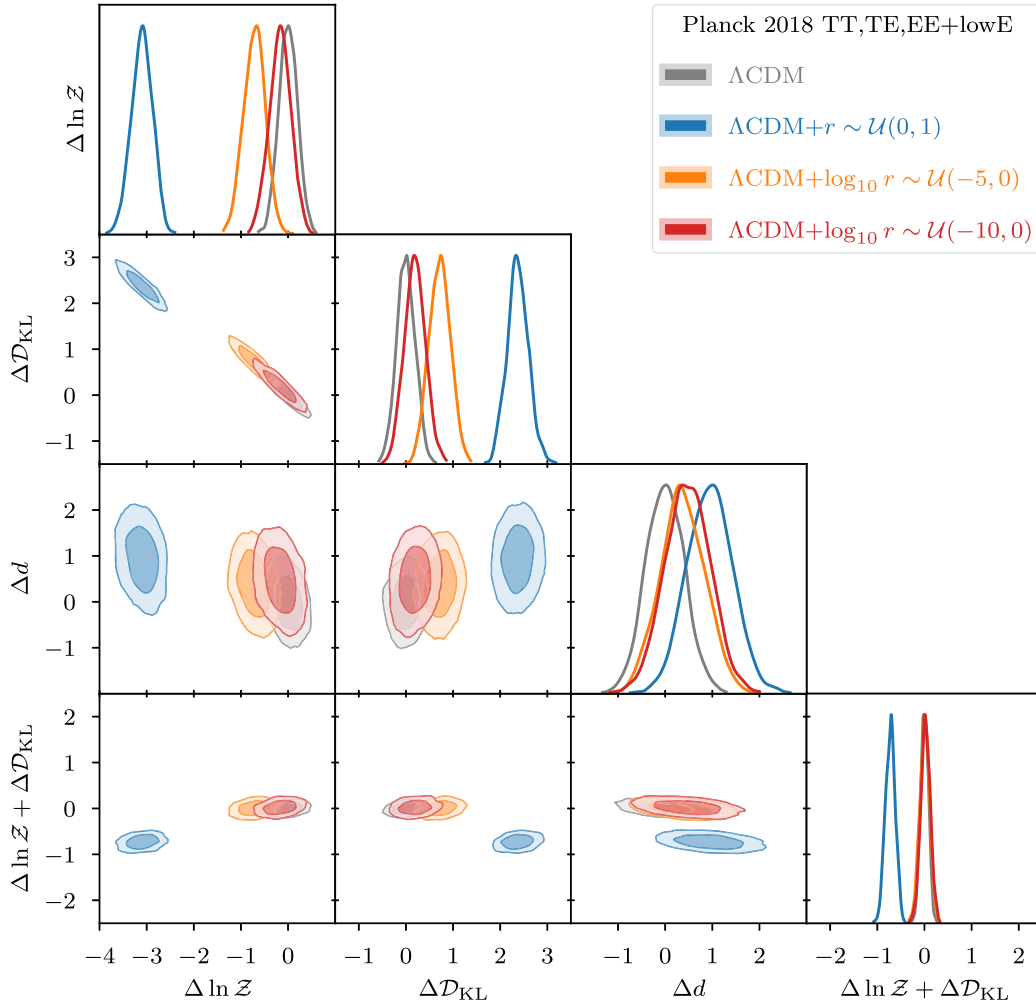


FIG. 4. Effect of uniform vs logarithmic priors on Bayesian model comparison for the tensor-to-scalar ratio  $r$ : log-evidence  $\Delta \ln \mathcal{Z}$ , Kullback–Leibler divergence  $\mathcal{D}_{\text{KL}}$  (in nats), Bayesian model dimensionality  $d$ , and posterior average of the log-likelihood  $\langle \ln \mathcal{L} \rangle_{\mathcal{P}} = \ln \mathcal{Z} + \mathcal{D}_{\text{KL}}$ . The probability distributions represent errors arising from the nested sampling process. In the limit of infinite life points these distributions would become point statistics, in contrast to posterior distributions. We normalize with respect to the  $\Lambda\text{CDM}$  model without  $r$  (i.e., with  $r = 0$ ). Note, how switching from uniform to logarithmic sampling of  $r$  (i.e., from blue to orange/red) moves the contours along their  $\ln \mathcal{Z}$ ,  $\mathcal{D}_{\text{KL}}$  degeneracy line, i.e., relative entropy is traded in for evidence. Note further by comparison of the orange and red lines, how changing the lower bound of the logarithmic sampling interval (by five log-units) barely affects the contours (bar some expected statistical fluctuation due to the sampling error).

Table II lists the summary statistics for the quantities from Fig. 4.

The marginalized plot for the difference in log-evidence (topmost panel) with  $\Delta \ln \mathcal{Z} = -3.10 \pm 0.23$  for the  $r$  extension of  $\Lambda\text{CDM}$  shows that it is considerably disfavored when applying a uniform prior. However, switching from a uniform to a logarithmic prior negates the difference in log-evidence completely, such that the log  $r$  extension ends up almost on par with the base  $\Lambda\text{CDM}$  model.

Changing the lower bound for the logarithmic prior, on the other hand, barely affects the evidence value at all. We have performed a run with a lower bound of  $\log_{10} r = -5$  and another with  $\log_{10} r = -10$ , i.e., five orders of magnitude difference in  $r$ . Despite this large difference in the lower bound the corresponding log-evidence  $\ln \mathcal{Z}$  changes

only very little such that the distributions significantly overlap one another. As explained in Sec. II A, this is due to log  $r$  being unconstrained below a certain threshold and the Bayesian evidence picking up only on constrained parameters. This can seem counterintuitive, since the Bayesian evidence is generally understood to automatically penalize additional parameters. The key point is that the Occam penalty essentially enters into the Bayesian evidence in the form of the ratio of posterior to prior volume. If both volumes are the same, then they divide out and do not contribute to the Occam penalty.

The last point becomes clearer by also taking into account the KL divergence and recalling Eq. (6), where we identified  $\mathcal{D}_{\text{KL}}$  as a measure for the Occam penalty. Looking at the correlation plot between log-evidence and KL divergence

TABLE II. Mean and standard deviation of the log-evidence  $\ln \mathcal{Z}$ , Kullback–Leibler divergence  $\mathcal{D}_{\text{KL}}$  and Bayesian model dimensionality  $d$  of the base  $\Lambda$ CDM cosmology and its  $r$  extension from Planck 2018 TT, TE, EE + lowE data [43]. The  $\Delta$  indicates normalization with respect to the base  $\Lambda$ CDM model.

Model	$\ln \mathcal{Z}$	$\mathcal{D}_{\text{KL}}$	$d$	$\Delta \ln \mathcal{Z}$	$\Delta \mathcal{D}_{\text{KL}}$	$\Delta d$
$\Lambda$ CDM	$-1431.05 \pm 0.20$	$38.57 \pm 0.20$	$17.10 \pm 0.40$	$-0.00 \pm 0.20$	$0.00 \pm 0.20$	$0.00 \pm 0.40$
$\Lambda$ CDM + $r \sim \mathcal{U}(0, 1)$	$-1434.15 \pm 0.23$	$40.94 \pm 0.23$	$18.05 \pm 0.48$	$-3.10 \pm 0.23$	$2.37 \pm 0.23$	$0.95 \pm 0.48$
$\Lambda$ CDM + $\log_{10} r \sim \mathcal{U}(-5, 0)$	$-1431.76 \pm 0.23$	$39.29 \pm 0.23$	$17.47 \pm 0.47$	$-0.71 \pm 0.23$	$0.72 \pm 0.23$	$0.37 \pm 0.47$
$\Lambda$ CDM + $\log_{10} r \sim \mathcal{U}(-10, 0)$	$-1431.22 \pm 0.23$	$38.76 \pm 0.22$	$17.57 \pm 0.48$	$-0.17 \pm 0.23$	$0.19 \pm 0.22$	$0.47 \pm 0.48$

makes it clear that there is a trade-off happening between those two quantities when switching between uniform and logarithmic priors. While the evidence increases for the logarithmic prior, the KL divergence decreases, as expected from the posterior plots in Fig. 2, which shows how the change from prior to posterior happens only at about  $\log_{10} r \gtrsim -2$ . This is further reflected in the Bayesian model dimensionality  $d$ , which shows a clear growing trend from about  $d = 17$  for the base  $\Lambda$ CDM model via a  $\log r$  extension to about  $d = 18$  for the  $r$  extension reflecting the one additional sampling parameter. Note that the total number of sampled parameters consists of six base cosmological parameters (+1 for the  $r$  extension) and 21 nuisance parameters from the Planck likelihood.

Because of the trade-off between log-evidence and KL divergence it is interesting also to look at their sum, which from Eq. (6) we know turns out to be the posterior average of the log-likelihood:

$$\ln \mathcal{Z} + \mathcal{D}_{\text{KL}} = \langle \ln \mathcal{L} \rangle_{\mathcal{P}}. \quad (11)$$

This makes for an interesting pairing with the Bayesian model dimensionality, since  $d/2$  is the posterior variance of the log-likelihood. As such, these two quantities provide an alternative perspective to that of the evidence and KL divergence. The posterior average and variance of the log-likelihood are a measure of the fit and complexity, respectively.  $\langle \ln \mathcal{L} \rangle_{\mathcal{P}}$  is shown in the last panel in Fig. 4, where we indeed see that the line for uniform sampling of  $r$  has moved much closer to the other lines, which is to be expected, since  $r$  and  $\log r$  are fundamentally the same parameter and therefore lead to a similar goodness of fit.

This behavior can also be understood analytically, which we explore in the following section in a one-dimensional mock example, simulating the  $r$  vs  $\log r$  result.

### C. Mock example

To illustrate further the role of a uniform vs a logarithmic prior on a Bayesian model comparison, we propose the following mock example, which is loosely based on the pedagogical example by Sivia and Skilling [37] explaining the effect of an additional (although in that case constrained) parameter, which we already outlined in Sec. II A.

Here, we will not assume a Gaussian likelihood that ultimately fully constrains a parameter, but rather we will

assume an exponential distribution as our likelihood on a strictly positive parameter (which is the maximum-entropy distribution when only a mean is known):

$$\mathcal{L}(a) = P_0 e^{-a/\mu}, \quad (12)$$

where  $P_0 = \Pr(D|a=0)$  is the maximum likelihood value for the data  $D$  at  $a=0$  and where  $\mu$  is the mean of the likelihood distribution describing the data. Thus, the likelihood is constrained only on one side, providing an upper bound, as shown in the left panel of Fig. 5.

We will assume a model  $A$ , where we sample the parameter  $a$  uniformly in the interval  $[a_1, a_2]$ . Furthermore, we will assume a model  $B$ , where we uniformly sample the parameter  $b = \log_{10} a$  in the interval  $[b_1, b_2]$ , corresponding to logarithmically sampling the parameter  $a$ . Since both models are fundamentally governed by the same quantity and will use the same likelihood, any difference in Bayesian inference quantities will be purely prior driven.

We will make the following assumptions on the ordering of the prior limits:

$$0 = a_1 < 10^{b_1} \ll \mu \ll 10^{b_2} = a_2 = 1. \quad (13)$$

This ordering is motivated as follows: for the upper limit we require that the likelihood has essentially dropped to zero. Hence, without loss of generality, we can set the upper limit to one and require  $\mu \ll 1$ . The lower limit for the positive parameter  $a$  can be explicitly set to zero when sampling uniformly. However, when sampling logarithmically we need to pick some finite lower limit, which we require to be in the region  $10^{b_1} \ll \mu$ , where the likelihood has essentially saturated with respect to  $b$  (see right panel in Fig. 5). The dependence of Bayesian quantities such as the evidence  $\mathcal{Z}$  or the Kullback–Leibler divergence  $\mathcal{D}_{\text{KL}}$  on the prior choice on the one hand and on this lower limit  $b_1$  on the other is the goal of this mock example.

The corresponding priors for models  $A$  and  $B$  can thus be written as

$$\pi_A(a) = \frac{1}{a_2 - a_1} \Theta(a - a_1) \Theta(a_2 - a), \quad (14)$$

$$\pi_B(b) = \frac{1}{b_2 - b_1} \Theta(b - b_1) \Theta(b_2 - b), \quad (15)$$

where  $\Theta(x)$  is the Heaviside step function.

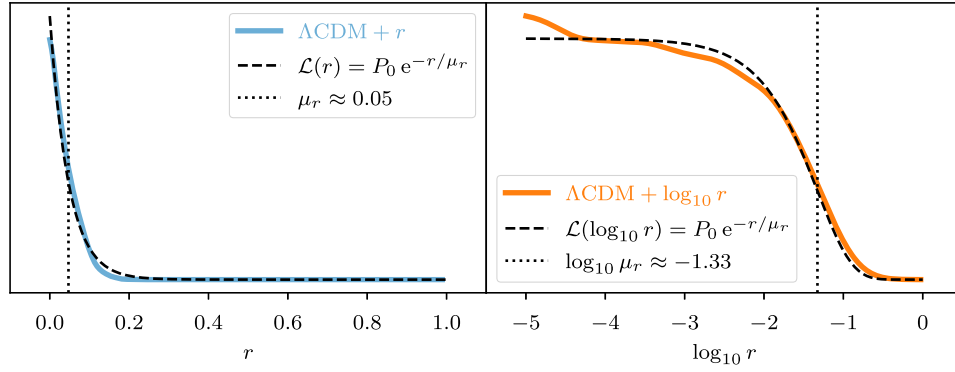


FIG. 5. Exponential likelihood distribution from our mock example in Eq. (12) compared to Planck 2018 temperature and polarization data (TT, TE, EE + lowE) on the tensor-to-scalar ratio  $r$  with uniform sampling of  $r$  on the left and logarithmic sampling of  $r$  on the right. Note how the mean  $\mu_r$  fulfills the ordering required by Eq. (13) and how the lower limit on  $\log_{10} r$  is well into the saturation plateau of posterior/likelihood.

We can compute the evidence and Kullback–Leibler divergence for models  $A$  and  $B$  as

$$\mathcal{Z}_A = \int \mathcal{L}(a)\pi_A(a)da = \frac{P_0\mu}{a_2 - a_1} (e^{-a_1/\mu} - e^{-a_2/\mu}), \quad (16)$$

$$\mathcal{Z}_B = \int \mathcal{L}(10^b)\pi_B(b)db = \frac{P_0}{b_2 - b_1 \ln(10)} \left[ \text{Ei}\left(-\frac{10^{b_2}}{\mu}\right) - \text{Ei}\left(-\frac{10^{b_1}}{\mu}\right) \right], \quad (17)$$

$$\mathcal{D}_{\text{KL},A} = \int \frac{\mathcal{L}(a)\pi_A(a)}{\mathcal{Z}_A} \ln\left(\frac{\mathcal{L}(a)}{\mathcal{Z}_A}\right) da = \ln\frac{P_0}{\mathcal{Z}_A} - 1 - \frac{P_0}{\mathcal{Z}_A} \frac{1}{a_2 - a_1} \left[ a_1 \exp\left(-\frac{a_1}{\mu}\right) - a_2 \exp\left(-\frac{a_2}{\mu}\right) \right], \quad (18)$$

$$\mathcal{D}_{\text{KL},B} = \int \frac{\mathcal{L}(10^b)\pi_B(b)}{\mathcal{Z}_B} \ln\left(\frac{\mathcal{L}(10^b)}{\mathcal{Z}_B}\right) db = \ln\frac{P_0}{\mathcal{Z}_B} - \frac{P_0}{\mathcal{Z}_B} \frac{1}{b_2 - b_1 \ln(10)} \left[ \exp\left(-\frac{10^{b_1}}{\mu}\right) - \exp\left(-\frac{10^{b_2}}{\mu}\right) \right], \quad (19)$$

where Ei refers to the exponential integral. For reasons of brevity and readability, we omit the similar but lengthier terms for the Bayesian model dimensionality here. With the ordering from Eq. (13), i.e., for small  $\mu$  and for a sufficiently low lower bound of the logarithmic prior  $b_1$ , we can then approximate these terms to give:

$$\Delta \ln \mathcal{Z}_A \approx \ln \mu \sim -3, \quad (20)$$

$$\Delta \ln \mathcal{Z}_B \approx \ln\left(1 + \frac{\gamma - \ln \mu}{b_1 \ln 10}\right) \sim 0, \quad (21)$$

$$\Delta \mathcal{D}_{\text{KL},A} \approx -1 - \ln \mu \sim 2, \quad (22)$$

$$\Delta \mathcal{D}_{\text{KL},B} \approx \frac{1}{\gamma - \ln \mu + b_1 \ln 10} - \ln\left(1 + \frac{\gamma - \ln \mu}{b_1 \ln 10}\right) \sim 0, \quad (23)$$

$$\Delta d_A \approx 2 + \frac{1}{\mu^2(1 - \cosh \frac{1}{\mu})} \sim 2, \quad (24)$$

$$\Delta d_B \approx -\frac{2}{\gamma - \ln \mu + b_1 \ln 10} - \frac{2}{(\gamma - \ln \mu + b_1 \ln 10)^2} \sim 0, \quad (25)$$

where  $\gamma$  is Euler's constant and where the  $\Delta$  indicates normalization with respect to a base model  $O$  with  $a = 10^b = 0$  fixed, such that  $\mathcal{Z}_O = P_0$  and  $\mathcal{D}_{\text{KL},O} = 0$ .

The approximate numerical results for the Bayesian model dimensionality can be compared to Fig. 3 in [42] which links posterior shapes to values for the model dimensionality and lists  $d = 0$  for a top hat (matching model  $B$ ),  $d = 1$  for a Gaussian, and  $d = 2$  for a Laplace distribution (matching model  $A$ ).

The approximate numerical values assume  $\mu \sim 0.06$ , which is roughly the posterior mean of the tensor-to-scalar ratio under uniform sampling in the preceding section. Hence, we can compare these zeroth-order numerical approximations to the results in Fig. 4 and Table II, which indeed match. Figure 6 makes this comparison more thoroughly, comparing the results from our one-dimensional mock example in Eqs. (16) and (17) with the nested sampling results from Table II for a variable lower bound  $b_{\min}$  of the logarithmic prior. The mean  $\ln \mathcal{Z}$  of the base model with  $r = 0$  for both the mock example and for the base  $\Lambda$ CDM nested sampling run are zero by definition of our normalization. They serve only as calibration for the models with



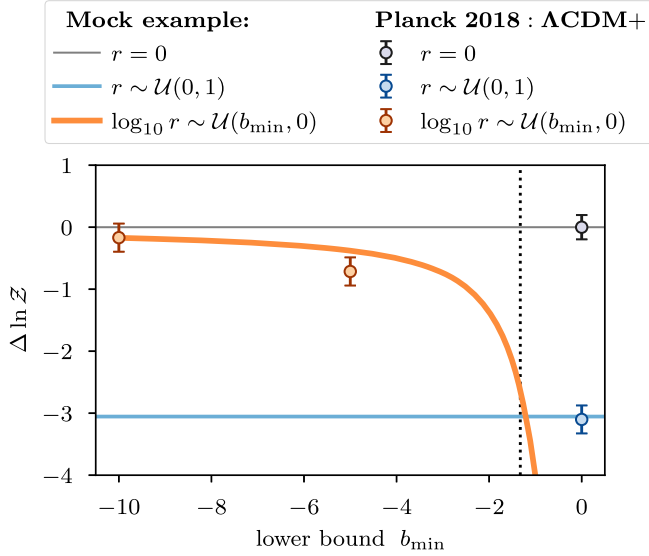


FIG. 6. Dependence of the log-evidence on the lower prior bound  $b_{\min}$ . Comparison of the results in Eqs. (16) and (17) for the one-dimensional mock example (solid lines) to the nested sampling results from Table II (dots with error bars). The vertical dotted line corresponds to the mean used in the mock likelihood distribution (cf. Fig. 5).

uniform (blue) and logarithmic (orange) priors. All three nested sampling runs agree well with the prediction from the mock example within their margins of errors.

Figure 6 illustrates how the evidence levels off with regards to the choice of the lower bound of the logarithmic prior (orange line) also reflected in the near equal evidences of the nested sampling runs with lower prior bounds of  $-5$  and  $-10$ , respectively. Note that the good agreement between mock example and data in Fig. 6 is due to the fact that the tensor-to-scalar ratio is almost completely uncorrelated with the other cosmological parameters, with the biggest (yet still small) correlation coming from the spectral index  $n_s$  (cf. Fig. 3).

#### IV. NEUTRINO MASSES

In Planck’s baseline cosmology, the neutrinos are assumed to be comprised of two massless neutrinos and one massive neutrino with mass  $m_\nu = 0.06$  eV with the effective number of neutrino species set slightly larger than three to  $N_{\text{eff}} = 3.046$  [7,58,59].

Upcoming CMB experiments such as the Simons Observatory, LiteBIRD, or CMB-S4 and large scale structure experiments such as Euclid will allow us to fully constrain the sum of neutrino masses  $\sum m_\nu$ . However, even under the most optimistic assumptions, it will not be possible to disentangle the individual contributions of the three neutrino flavors with cosmological data alone [16]. To achieve that, we need additional data from solar, atmospheric, reactor, and accelerator experiments as

summarized in NuFIT 5.0 (2020) [45,46] that provide us with the mass square splittings:

$$\delta m^2 = 7.42_{-0.20}^{+0.21} \times 10^{-5} \text{ eV}^2 \quad (\text{NH \& IH}), \quad (26)$$

$$\Delta m^2 = \begin{cases} 2.517_{-0.028}^{+0.026} \times 10^{-3} \text{ eV}^2 & (\text{NH}), \\ 2.498_{-0.028}^{+0.028} \times 10^{-3} \text{ eV}^2 & (\text{IH}), \end{cases} \quad (27)$$

where  $\delta m^2$  is the smaller squared mass splitting between the light and the medium neutrino mass for the normal neutrino hierarchy (NH) and between the medium and the heavy neutrino mass for the inverted neutrino hierarchy (IH), and  $\Delta m^2$  is the larger squared mass splitting between the light and the heavy neutrino mass in both cases.

With the knowledge of the two squared mass splittings, the remaining uncertainty lies mostly with the scale of the lightest neutrino. In the following Bayesian analysis we therefore apply Gaussian priors according to Eqs. (26) and (27) and vary over the lightest neutrino mass.

#### A. Neutrino masses: Posteriors

We have taken nested sampling runs for an extension of the base  $\Lambda$ CDM cosmology with three individual neutrino masses, where we have used both a uniform prior  $m_{\text{light}} \sim \mathcal{U}(0, 1)$  and logarithmic priors with different lower bounds,  $\log_{10} m_{\text{light}} \sim \mathcal{U}(-5, 0)$  and  $\log_{10} m_{\text{light}} \sim \mathcal{U}(-10, 0)$ , on the lightest neutrino mass. The other two neutrino masses are then derived from  $m_{\text{light}}$  together with  $\delta m^2$  and  $\Delta m^2$  from Eqs. (26) and (27):

$$m_{\text{medium}}^2 = \begin{cases} m_{\text{light}}^2 + \delta m^2 & (\text{NH}), \\ m_{\text{light}}^2 + \Delta m^2 - \delta m^2 & (\text{IH}), \end{cases} \quad (28)$$

$$m_{\text{heavy}}^2 = m_{\text{light}}^2 + \Delta m^2. \quad (29)$$

Figure 7 gives an overview of the stability of the cosmological base parameters across the different priors for  $m_{\text{light}}$  and compares them to the  $\Lambda$ CDM base model by showing their mean and 68% ranges. Compared to Fig. 1 for the tensor-to-scalar ratio there are some small parameter shifts visible in relation to the base  $\Lambda$ CDM model, but all shifts stay well within the 68% bounds.

Figure 8 shows the one-dimensional marginalized posterior distributions for the three individual neutrino masses  $m_{\text{light}}$ ,  $m_{\text{medium}}$ , and  $m_{\text{heavy}}$ , as well as the sum of all three  $\sum m_\nu$  for both the normal and the inverted hierarchy. We have included shaded histograms to give a notion of the prior distributions. The vertical black dotted lines indicate roughly the lower bound for the medium and heavy neutrino mass as determined from the mass squared splittings under the assumption where the light neutrino mass is zero.

When looking at the lightest neutrino mass in the first row, the picture is very similar to that for the

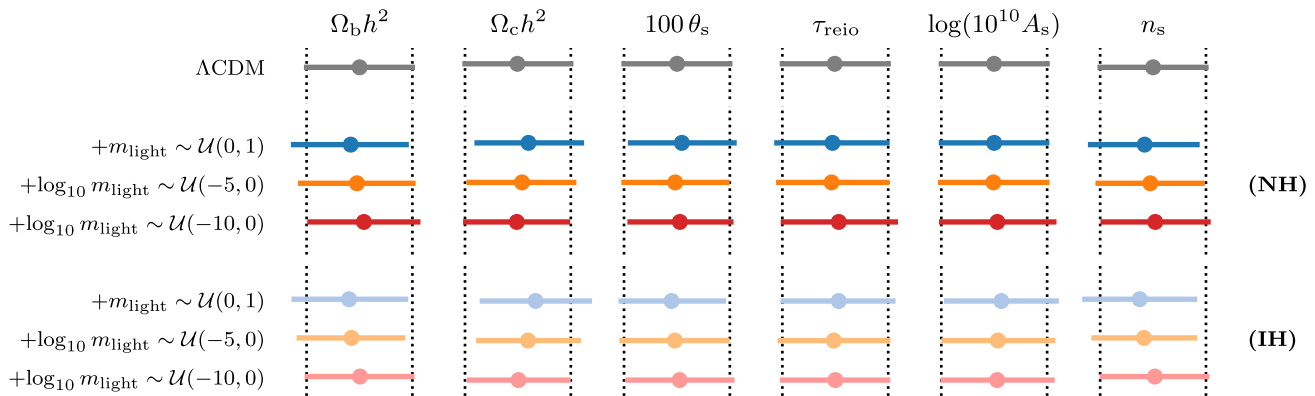


FIG. 7. Stability of the cosmological parameters for the three-neutrino extension of the base  $\Lambda$ CDM cosmology for different priors on  $m_{\text{light}}$ : uniform in blues, logarithmic with lower bound of  $-5$  in oranges and logarithmic with lower bound of  $-10$  in reds. The darker set of colors corresponds to the normal normal neutrino hierarchy (NH) and the lighter set to the inverted hierarchy (IH). For each parameter we show the mean and the extent from quantile 0.16 to 0.84, i.e., the inner 68% limits.

tensor-to-scalar ratio before, and most of what we have said in Sec. III A applies here, too. One has an almost exponential dropoff from zero when sampling uniformly over the mass (left column), significantly compressing the prior, which turns into a more steplike behavior with respect to the logarithm of the mass when sampling the mass logarithmically (right column).

Note that the medium and heavy mass from rows 2 and 3 as well as the sum of all masses in the bottom row are *derived* quantities and therefore do not show the same prior behavior visible for the light neutrino mass. This is not so apparent for the derived masses, when sampling uniformly over the light neutrino mass, although one can see a slight step in the histogram of the prior for the heavy mass in the NH case and for both medium and heavy mass in the IH case. However, when sampling logarithmically over the light neutrino mass, then the picture is much clearer. The probability density for medium and heavy neutrino mass bulks up around their rough lower minimum set by the smaller and larger mass square splitting, respectively.

There are two perspectives that one can adopt here. On one hand, one could criticize the choice of a logarithmic prior for being ultimately too prior (or theory) driven and not reflective of the data. On the other hand, one could say that this is the natural result of our state of knowledge of the mass square splittings and our true ignorance about the scale of the lightest neutrino mass.

We wonder whether this very last statement could be contested, e.g., could we say that we would expect the lightest neutrino mass to be of a magnitude similar to that of the medium neutrino mass in the NH? However, this is not the case, when checking for precedence by looking at the other set of leptons, the electron, muon, and tauon, where we have roughly around two orders of magnitude between their masses [60].

Comparing the two hierarchies with one another, we can see that the major difference lies in the medium neutrino

mass (and therefore also the sum of all neutrino masses), which is restricted to larger masses in the inverted hierarchy compared to the normal hierarchy, as expected from the mass square splitting (black dotted lines).

## B. Neutrino masses: Evidence and Kullback–Leibler divergence

In Fig. 9 we show the results from our nested sampling runs for the log-evidence  $\ln \mathcal{Z}$ , KL divergence  $\mathcal{D}_{\text{KL}}$ , Bayesian model dimensionality  $d$  and posterior average of the log-likelihood ( $\ln \mathcal{L}$ ). We again normalize with respect to the base  $\Lambda$ CDM model. Table III lists the summary statistics for these quantities. As already the case for the posterior, the picture here is again similar to the one for the tensor-to-scalar ratio in Sec. III B.

Looking at the distributions for the log-evidence (top-most diagonal panel) shows that the addition of the neutrino parameters with uniform sampling over the light neutrino mass (either hierarchy) is disfavored with over three log-units compared to the base  $\Lambda$ CDM model with a single massive neutrino of fixed mass (and two massless). Since the mass square splittings enter on the prior level in our analysis and remain essentially unconstrained by the cosmological data, any change to the evidence is almost entirely driven by the light neutrino mass parameter. Hence, it is not surprising that upon switching to a logarithmic prior on  $m_{\text{light}}$  the log-evidence increases again while the KL divergence drops close to the level of the  $\Lambda$ CDM model. We need to keep in mind that since this is an extension to the  $\Lambda$ CDM model, it has in principle a better chance of fitting the data, such that any difference in the Bayesian evidence can be attributed to an Occam penalty, which the shift between uniform and logarithmic sampling confirms.

As expected from our investigations for the tensor-to-scalar ratio and especially with regards to our mock example from Sec. III C, changing the lower bound for

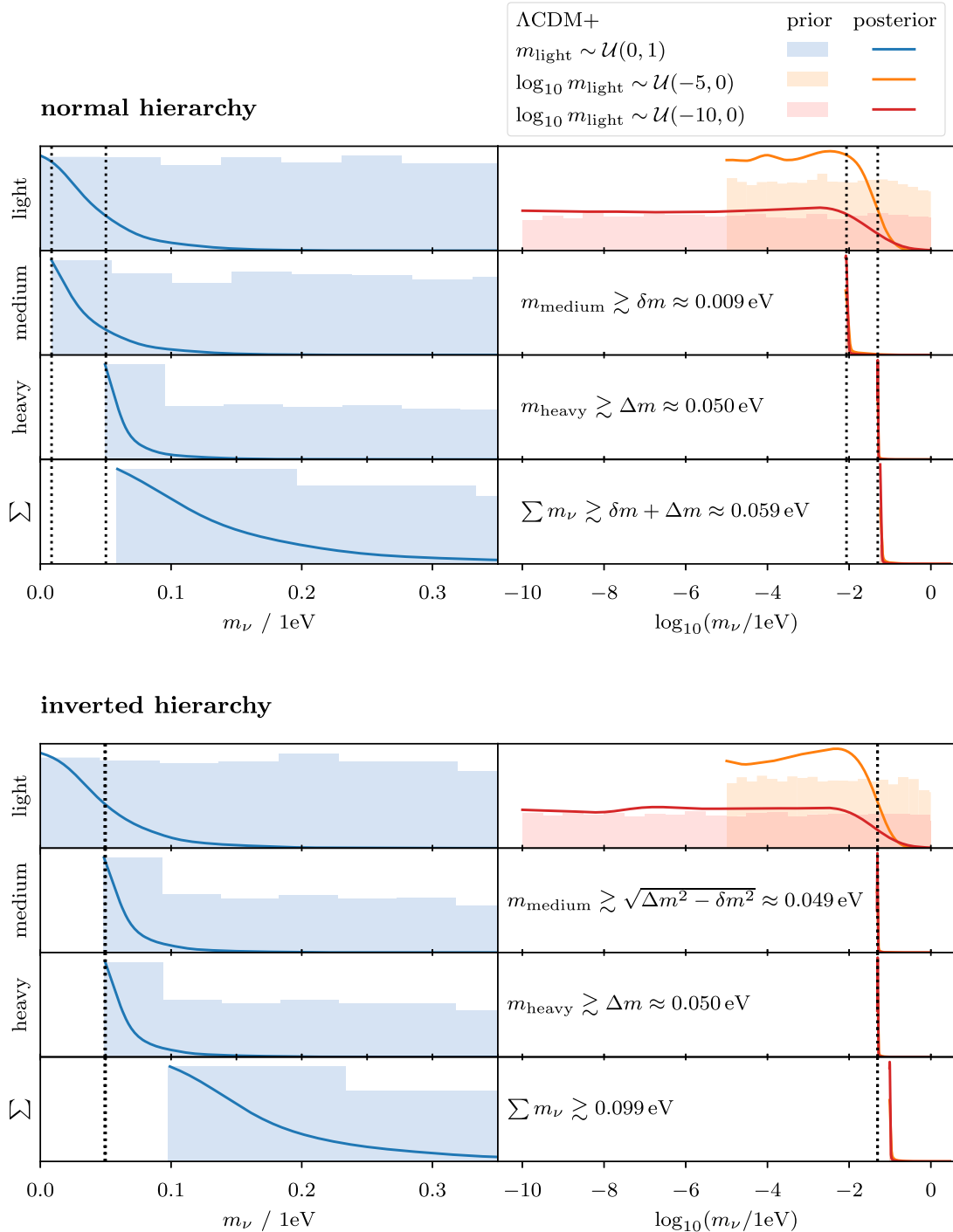


FIG. 8. One-dimensional posterior distributions of neutrino masses with normal neutrino hierarchy (NH) in the top panel and with inverted hierarchy (IH) in the bottom panel for TT, TE, EE + lowE data from Planck 2018 and neutrino oscillation data on the mass squared splittings from NuFIT 5.0 (2020). The vertical black dotted lines give the rough lower limit on medium and heavy mass that is set by the mass squared splittings  $\delta m^2$  and  $\Delta m^2$ . For the inverted hierarchy these dotted lines appear almost on top of each other. The rows show the posteriors for the light, medium, and heavy neutrino mass and sum of all neutrino masses, respectively. The columns contrast the difference between using a uniform (blue, left) or logarithmic (orange and red, right) prior on the light neutrino mass  $m_{\text{light}}$ . The shaded histograms give a notion of that prior distribution.

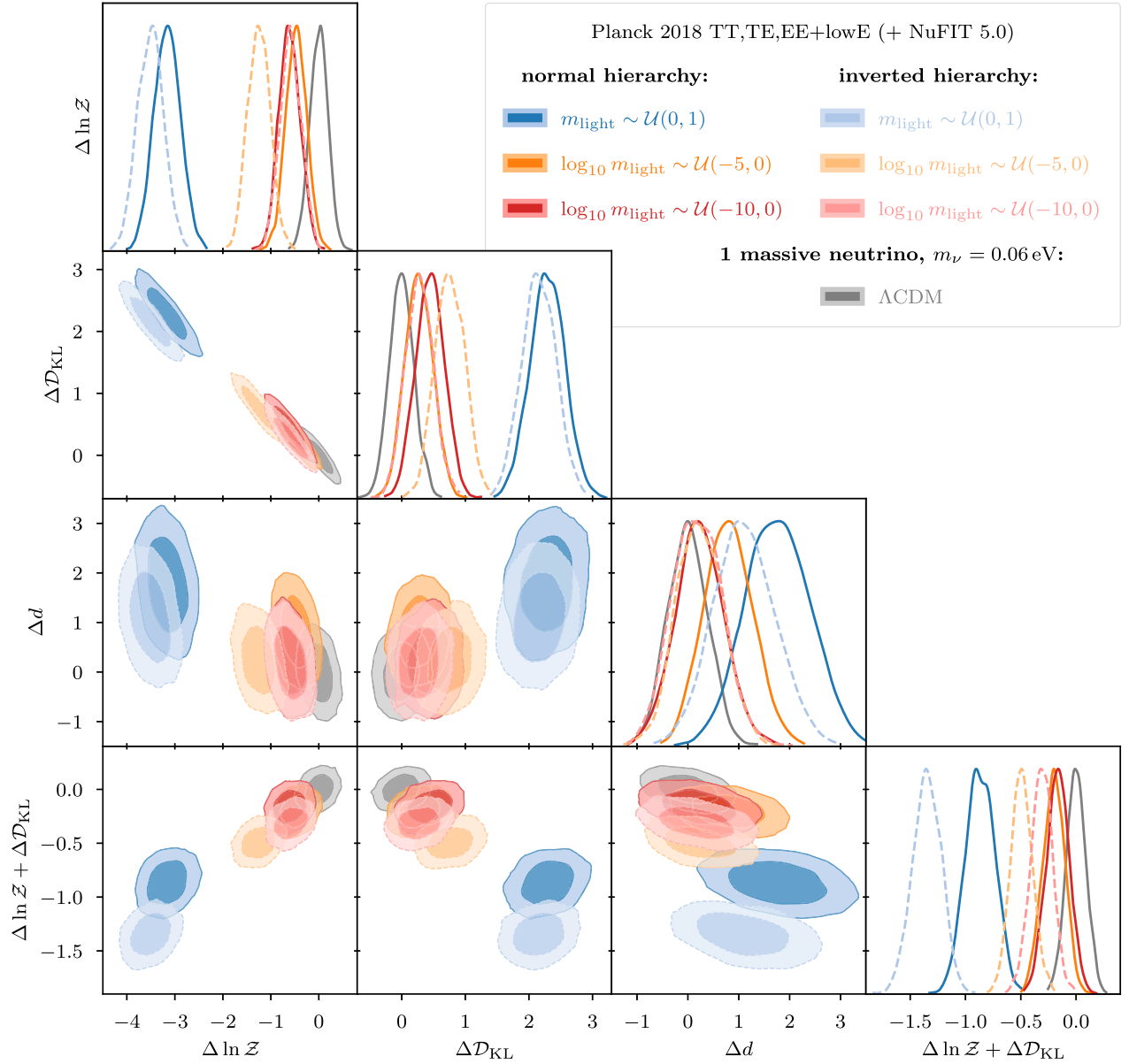


FIG. 9. Effect of uniform vs logarithmic priors on the light neutrino mass  $m_{\text{light}}$  for Bayesian model comparison: log-evidence  $\Delta \ln \mathcal{Z}$ , Kullback–Leibler divergence  $\mathcal{D}_{\text{KL}}$ , Bayesian model dimensionality  $d$ , and posterior average of the log-likelihood  $\langle \ln \mathcal{L} \rangle_{\mathcal{P}} = \ln \mathcal{Z} + \mathcal{D}_{\text{KL}}$ . The probability distributions represent errors arising from the nested sampling process. In the limit of infinite life points these distributions would become point statistics, in contrast to posterior distributions. We normalize with respect to the  $\Lambda\text{CDM}$  model with two massless and only one massive neutrino with  $m_{\nu} = 0.06$  eV. Note how switching from uniform to logarithmic sampling of  $m_{\text{light}}$  moves the contours along their  $\ln \mathcal{Z}$ ,  $\mathcal{D}_{\text{KL}}$  degeneracy line, i.e., relative entropy is traded in for evidence. Note further by comparison of the orange and red lines, how changing the lower bound of the logarithmic sampling interval (by five log-units!) barely affects the contours (bar some expected statistical fluctuation due to the sampling error).

the logarithmic prior does not affect the Bayesian evidence. We have again performed runs with two different lower bounds of  $\log m_{\text{light}} = -5$  and  $\log m_{\text{light}} = -10$ , i.e., five orders of magnitude apart. With both of these bounds well into the area of the posterior (see top right panel of Fig. 8) where it has leveled off, we do not expect much change to the evidence value. This is clearly confirmed in Table III and Fig. 9 for the normal hierarchy. For the inverted

hierarchy it is not as clear but still reasonable in light of the uncertainties.

Looking at the normalized posterior average of the log-likelihood  $\langle \ln \mathcal{L} \rangle_{\mathcal{P}} = \ln \mathcal{Z} + \mathcal{D}_{\text{KL}}$  we again roughly confirm

$$\Delta \ln \mathcal{Z}_{\text{uni}} + \Delta \mathcal{D}_{\text{KL,uni}} \approx -1, \quad (30)$$

TABLE III. Mean and standard deviation of the log-evidence  $\ln \mathcal{Z}$ , Kullback–Leibler divergence  $\mathcal{D}_{\text{KL}}$  and Bayesian model dimensionality  $d$  of the base  $\Lambda$ CDM cosmology and its three-neutrino extension from Planck 2018 TT, TE, EE + lowE data [43]. The second block of rows shows the results from the normal neutrino hierarchy and the third block for the inverted hierarchy. The  $\Delta$  indicate normalization with respect to the base  $\Lambda$ CDM model.

	Model	$\ln \mathcal{Z}$	$\mathcal{D}_{\text{KL}}$	$d$	$\Delta \ln \mathcal{Z}$	$\Delta \mathcal{D}_{\text{KL}}$	$\Delta d$
	$\Lambda$ CDM	$-1431.04 \pm 0.19$	$38.56 \pm 0.19$	$17.08 \pm 0.39$	$-0.00 \pm 0.19$	$0.00 \pm 0.19$	$0.00 \pm 0.39$
Normal	$\Lambda$ CDM + $m_{\text{light}} \sim \mathcal{U}(0, 1)$	$-1434.20 \pm 0.27$	$40.85 \pm 0.27$	$18.85 \pm 0.63$	$-3.16 \pm 0.27$	$2.29 \pm 0.27$	$1.77 \pm 0.63$
	$\Lambda$ CDM + $\log_{10} m_{\text{light}} \sim \mathcal{U}(-5, 0)$	$-1431.51 \pm 0.22$	$38.83 \pm 0.22$	$17.88 \pm 0.48$	$-0.47 \pm 0.22$	$0.27 \pm 0.22$	$0.80 \pm 0.48$
	$\Lambda$ CDM + $\log_{10} m_{\text{light}} \sim \mathcal{U}(-10, 0)$	$-1431.65 \pm 0.22$	$39.01 \pm 0.22$	$17.31 \pm 0.48$	$-0.61 \pm 0.22$	$0.45 \pm 0.22$	$0.23 \pm 0.48$
Inverted	$\Lambda$ CDM + $m_{\text{light}} \sim \mathcal{U}(0, 1)$	$-1434.55 \pm 0.27$	$40.72 \pm 0.27$	$18.19 \pm 0.60$	$-3.51 \pm 0.27$	$2.16 \pm 0.27$	$1.11 \pm 0.60$
	$\Lambda$ CDM + $\log_{10} m_{\text{light}} \sim \mathcal{U}(-5, 0)$	$-1432.29 \pm 0.23$	$39.32 \pm 0.23$	$17.31 \pm 0.46$	$-1.25 \pm 0.23$	$0.76 \pm 0.23$	$0.23 \pm 0.46$
	$\Lambda$ CDM + $\log_{10} m_{\text{light}} \sim \mathcal{U}(-10, 0)$	$-1431.63 \pm 0.22$	$38.84 \pm 0.22$	$17.27 \pm 0.49$	$-0.59 \pm 0.22$	$0.28 \pm 0.22$	$0.19 \pm 0.49$

$$\Delta \ln \mathcal{Z}_{\log} + \Delta \mathcal{D}_{\text{KL}, \log} \approx 0, \quad (31)$$

matching our mock results from Eqs. (20) to (23), independent from the mock parameter  $\mu$ .

### C. Neutrino hierarchy

A Bayesian model comparison of the normal vs the inverted neutrino hierarchy is beyond the scope of this paper and has been done before with more stringent data [17,61,62]. However, with posteriors and evidences at hand, we shall briefly discuss the situation here.

There have been claims to a strong preference of the normal over the inverted neutrino hierarchy [10], however, such strong evidence can typically be traced back to prior volume effects [11], i.e., the effect of a reduced sampling space for the inverted hierarchy. In other words, we need to watch out and properly distinguish to what extent any Bayesian preference is assigned already on the prior level and to what extent is that preference indeed driven by the data.

In our analysis both hierarchies start out on an equal footing. With the same prior on the light neutrino mass and equivalent Gaussian priors on the mass squared splittings from neutrino oscillation experiments, the prior volume for both hierarchies is essentially the same. Note that although the means for the larger mass squared splitting  $\Delta m^2$  are slightly different in the two hierarchies, its standard deviations are essentially the same.

There is a slight tendency for all prior options of a better fit of the normal compared to the inverted neutrino hierarchy. However, with an evidence difference of less than one log-unit (odds of maximally 2:1) any preference for the normal hierarchy is meager at best, especially when also accounting for the sampling error (see Fig. 9). It should be noted, though, that we have used only CMB temperature and polarization data here. Adding data from CMB lensing or baryon acoustic oscillations would have further shrunk the constraints on the sum of neutrino

masses and thereby possibly strengthened the case for the normal hierarchy.

## V. DISCUSSION

We demonstrate how switching between a uniform and a logarithmic prior on some single-bounded model parameter results in a trade-off between Bayesian evidence and Kullback–Leibler divergence (or relative entropy). The common scenario is that of insufficient data sensitivity, leading to a one-sided bound on a parameter. For a location parameter this typically causes an exponential dropoff, which translates to a steplike behavior when turned into the corresponding scale parameter. We show that the ambiguity of the lower bound of the scale parameter does not affect a Bayesian model comparison, provided the lower bound is chosen sufficiently far into the likelihood plateau.

We demonstrate this behavior for two cases of parameter extensions to the  $\Lambda$ CDM model of cosmology, namely for the tensor-to-scalar ratio of primordial perturbations and for the case of three nondegenerate neutrino masses. In both cases we confirm that switching from a uniform prior to a logarithmic prior will get rid of (most of) the Occam penalty associated with that parameter, since unconstrained parameters do not affect the Bayesian evidence. Thus the Bayesian evidence is roughly on par with the unextended (base) model, with the only difference in the form of an uninformative parameter. Furthermore and for the same reason, the exact choice of the lower bound for the logarithmic prior does not change the Bayesian evidence. When the likelihood levels off, e.g., due to insufficient sensitivity in the data, then so does the Bayesian evidence.

## ACKNOWLEDGMENTS

L. T. H. would like to thank the Isaac Newton Trust, the STFC, and the Cavendish Laboratory for their support. W. J. H. was supported by a Gonville & Caius Research Fellowship. This work was performed using the resources

provided by the Cambridge Service for Data Driven Discovery (CSD3) operated by the University of Cambridge Research Computing Service ([www.csd3.cam.ac.uk](http://www.csd3.cam.ac.uk)), provided by Dell EMC and Intel using Tier-2 funding from the Engineering and Physical Sciences Research Council (capital Grant No. EP/P020259/1), and DiRAC ([www.dirac.ac.uk](http://www.dirac.ac.uk)) funding from the Science and Technology Facilities Council (STFC) (capital Grants No. ST/P002307/1 and No. ST/R002452/1 and operations

Grant No. ST/R00689X/1). DiRAC is part of the National e-Infrastructure. This work was also performed using the DiRAC Data Intensive service at Leicester (DiaL), operated by the University of Leicester IT Services, which forms part of the STFC DiRAC HPC Facility ([www.dirac.ac.uk](http://www.dirac.ac.uk)). The equipment was funded by BEIS capital funding via STFC (capital Grants No. ST/K000373/1 and No. ST/R002363/1 and operations Grant No. ST/R001014/1). DiRAC is part of the National e-Infrastructure.

- 
- [1] J. Bernoulli, *Ars conjectandi* (Thurnisiorum, Basiliae, 1713).
- [2] J. M. Keynes, *A Treatise on Probability* (Macmillan And Co., London, 1921).
- [3] D. S. Sivia and J. Skilling, *Data Analysis: A Bayesian Tutorial* (Oxford University Press, New York, 2006), pp. 1–246.
- [4] K. Lau, J.-Y. Tang, and M.-C. Chu, Cosmic microwave background constraints on the tensor-to-scalar ratio, *Res. Astron. Astrophys.* **14**, 635 (2014).
- [5] G. Barenboim and W.-I. Park, On the tensor-to-scalar ratio in large single-field inflation models, [arXiv:1509.07132](https://arxiv.org/abs/1509.07132).
- [6] P. Creminelli, S. Dubovsky, D. López Nacir, M. Simonović, G. Trevisan, G. Villadoro, and M. Zaldarriaga, Implications of the scalar tilt for the tensor-to-scalar ratio, *Phys. Rev. D* **92**, 123528 (2015).
- [7] Planck Collaboration, Planck 2018 results. VI. Cosmological parameters, *Astron. Astrophys.* **641**, A6 (2020).
- [8] The Keck Array and BICEP2 Collaborations, Constraints on Primordial Gravitational Waves Using Planck, WMAP, and New BICEP2/Keck Observations Through the 2015 Season, *Phys. Rev. Lett.* **121**, 221301 (2018).
- [9] K. Hirano, Inflation with very small tensor-to-scalar ratio, [arXiv:1912.12515](https://arxiv.org/abs/1912.12515).
- [10] F. Simpson, R. Jimenez, C. Pena-Garay, and L. Verde, Strong Bayesian evidence for the normal neutrino hierarchy, *J. Cosmol. Astropart. Phys.* **06** (2017) 029.
- [11] T. Schwetz, K. Freese, M. Gerbino, E. Giusarma, S. Hannestad, M. Lattanzi, O. Mena, and S. Vagnozzi, Comment on “Strong Evidence for the Normal Neutrino Hierarchy”, [arXiv:1703.04585](https://arxiv.org/abs/1703.04585).
- [12] A. Caldwell, A. Merle, O. Schulz, and M. Totzauer, Global Bayesian analysis of neutrino mass data, *Phys. Rev. D* **96**, 073001 (2017).
- [13] F. Capozzi, E. Di Valentino, E. Lisi, A. Marrone, A. Melchiorri, and A. Palazzo, Global constraints on absolute neutrino masses and their ordering, *Phys. Rev. D* **95**, 096014 (2017).
- [14] A. Loureiro, A. Cuceu, F. B. Abdalla, B. Moraes, L. Whiteway, M. McLeod, S. T. Balan, O. Lahav, A. Benoit-Lévy, M. Manera, R. P. Rollins, and H. S. Xavier, Upper Bound of Neutrino Masses from Combined Cosmological Observations and Particle Physics Experiments, *Phys. Rev. Lett.* **123**, 081301 (2019).
- [15] F. Capozzi, E. Di Valentino, E. Lisi, A. Marrone, A. Melchiorri, and A. Palazzo, Addendum to “Global constraints on absolute neutrino masses and their ordering”, *Phys. Rev. D* **101**, 116013(A) (2020).
- [16] M. Archidiacono, S. Hannestad, and J. Lesgourgues, What will it take to measure individual neutrino mass states using cosmology?, *J. Cosmol. Astropart. Phys.* **09** (2020) 021.
- [17] S. R. Choudhury and S. Hannestad, Updated results on neutrino mass and mass hierarchy from cosmology with Planck 2018 likelihoods, *J. Cosmol. Astropart. Phys.* **07** (2020) 037.
- [18] P. Stöcker, C. Balázs, S. Bloor, T. Bringmann, T. E. Gonzalo, W. Handley, S. Hotinli, C. Howlett, F. Kahlhoefer, J. J. Renk, P. Scott, A. C. Vincent, and M. White (The GAMBIT Cosmology Workgroup), Strengthening the bound on the mass of the lightest neutrino with terrestrial and cosmological experiments, [arXiv:2009.03287](https://arxiv.org/abs/2009.03287).
- [19] The Simons Observatory Collaboration, The Simons Observatory: Science goals and forecasts, *J. Cosmol. Astropart. Phys.* **02** (2019) 056.
- [20] The LiteBIRD Collaboration, Astro2020 APC White Paper: LiteBIRD: An all-sky cosmic microwave background probe of inflation, *Bull. Am. Astron. Soc.* **51** (2019), <https://baas.aas.org/pub/2020n7i286/release/1>.
- [21] CMB-S4 Collaboration, CMB-S4 science case, reference design, and project plan, [arXiv:1907.04473](https://arxiv.org/abs/1907.04473).
- [22] D. J. C. MacKay, Model comparison and Occam’s Razor, in *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, England, 2003), Chap. 28, pp. 343–355.
- [23] D. J. C. MacKay, Laplace’s method, in *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, England, 2003), Chap. 28, pp. 341–342.
- [24] J. M. Dickey, The weighted likelihood ratio, linear hypotheses on normal location parameters, *Ann. Math. Stat.* **42**, 204 (1971).
- [25] R. Trotta, Applications of Bayesian model selection to cosmological parameters, *Mon. Not. R. Astron. Soc.* **378**, 72 (2007).
- [26] R. Trotta, Forecasting the Bayes factor of a future observation, *Mon. Not. R. Astron. Soc.* **378**, 819 (2007).
- [27] L. Verde, S. M. Feeney, D. J. Mortlock, and H. V. Peiris, (Lack of) Cosmological evidence for dark radiation after Planck, *J. Cosmol. Astropart. Phys.* **09** (2013) 013.

- [28] A. Heavens, Y. Fantaye, A. Mootoovaloo, H. Eggers, Z. Hosenie, S. Kroon, and E. Sellentin, Marginal Likelihoods from Monte Carlo Markov Chains, [arXiv:1704.03472](https://arxiv.org/abs/1704.03472).
- [29] A. Heavens, Y. Fantaye, E. Sellentin, H. Eggers, Z. Hosenie, S. Kroon, and A. Mootoovaloo, No Evidence for Extensions to the Standard Cosmological Model, *Phys. Rev. Lett.* **119**, 101301 (2017).
- [30] J. Skilling, Nested sampling for general Bayesian computation, *Bayesian Analysis* **1**, 833 (2006).
- [31] D. S. Sivia and J. Skilling, Nested sampling, in *Data Analysis: A Bayesian Tutorial* (Oxford University Press, New York, 2006), Chap. 9, pp. 181–208.
- [32] F. Feroz and M. P. Hobson, Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses, *Mon. Not. R. Astron. Soc.* **384**, 449 (2008).
- [33] F. Feroz, M. P. Hobson, and M. Bridges, MultiNest: An efficient and robust Bayesian inference tool for cosmology and particle physics, *Mon. Not. R. Astron. Soc.* **398**, 1601 (2009).
- [34] F. Feroz, M. P. Hobson, E. Cameron, and A. N. Pettitt, Importance nested sampling and the MultiNest algorithm, *Open J. Astrophys.* **2**, 10 (2019).
- [35] W. J. Handley, M. P. Hobson, and A. N. Lasenby, PolyChord: Nested sampling for cosmology, *Mon. Not. R. Astron. Soc.: Lett.* **450**, L61 (2015).
- [36] W. J. Handley, M. P. Hobson, and A. N. Lasenby, PolyChord: Next-generation nested sampling, *Mon. Not. R. Astron. Soc.* **453**, 4385 (2015).
- [37] D. S. Sivia and J. Skilling, Model selection, in *Data Analysis: A Bayesian Tutorial* (Oxford University Press, New York, 2006), Chap. 4, pp. 78–102.
- [38] C. Heymans *et al.*, KiDS-1000 Cosmology: Multi-probe weak gravitational lensing and spectroscopic galaxy clustering constraints, *Astron. Astrophys.* **646**, A140 (2021).
- [39] W. Handley, Curvature tension: Evidence for a closed universe, *Phys. Rev. D* **103**, L041301 (2021).
- [40] W. Handley and P. Lemos, Quantifying tensions in cosmological parameters: Interpreting the DES evidence ratio, *Phys. Rev. D* **100**, 043504 (2019).
- [41] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, Bayesian measures of model complexity and fit, *J. R. Stat. Soc.: Ser. B (Statistical Methodology)* **64**, 583 (2002).
- [42] W. Handley and P. Lemos, Quantifying dimensionality: Bayesian cosmological model complexities, *Phys. Rev. D* **100**, 023512 (2019).
- [43] Planck Collaboration, Planck 2018 results. V. CMB power spectra and likelihoods, *Astron. Astrophys.* **641**, A5 (2020).
- [44] I. Esteban, M. C. Gonzalez-Garcia, A. Hernandez-Cabezudo, M. Maltoni, and T. Schwetz, Global analysis of three-flavor neutrino oscillations: Synergies and tensions in the determination of  $\theta_{23}$ ,  $\delta CP$ , and the mass ordering, *J. High Energy Phys.* **2019** (2019) 106.
- [45] I. Esteban, M. C. Gonzalez-Garcia, M. Maltoni, T. Schwetz, and A. Zhou, The fate of hints: Updated global analysis of three-flavor neutrino oscillations, *J. High Energy Phys.* **09** (2020) 178.
- [46] NuFIT | NuFIT 5.0, <http://www.nu-fit.org/> (2020), accessed: 2020-08-13.
- [47] J. Torrado and A. Lewis, Cobaya: Code for Bayesian Analysis of hierarchical physical models, [arXiv:2005.05290](https://arxiv.org/abs/2005.05290).
- [48] A. Lewis and S. Bridle, Cosmological parameters from CMB and other data: A Monte Carlo approach, *Phys. Rev. D* **66**, 103511 (2002).
- [49] A. Lewis, Efficient sampling of fast and slow cosmological parameters, *Phys. Rev. D* **87**, 103529 (2013).
- [50] R. M. Neal, Taking bigger metropolis steps by dragging fast variables, [arXiv:math/0502099](https://arxiv.org/abs/math/0502099).
- [51] J. Lesgourgues, The cosmic linear anisotropy solving system (CLASS) I: Overview, [arXiv:1104.2932](https://arxiv.org/abs/1104.2932).
- [52] D. Blas, J. Lesgourgues, and T. Tram, The cosmic linear anisotropy solving system (CLASS). Part II: Approximation schemes, *J. Cosmol. Astropart. Phys.* **07** (2011) 034.
- [53] J. Lesgourgues and T. Tram, The cosmic linear anisotropy solving system (CLASS) IV: Efficient implementation of non-cold relics, *J. Cosmol. Astropart. Phys.* **09** (2011) 032.
- [54] A. Lewis, GetDist: A Python package for analysing Monte Carlo samples, [arXiv:1910.13970](https://arxiv.org/abs/1910.13970).
- [55] W. Handley, Anesthetic: Nested sampling visualization, *J. Open Source Software* **4**, 1414 (2019).
- [56] L. T. Hergt, Bayesian evidence for the tensor-to-scalar ratio  $r$  and neutrino masses  $m_\nu$ : Effects of uniform vs logarithmic priors (supplementary inference products) (Version 1.0.0) (2021), [http://doi.org/10.5281/zenodo.4556360](https://doi.org/10.5281/zenodo.4556360).
- [57] J. C. Forbes, A PDF PSA, or Never gonna set\_xscale again —guilty feats with logarithms, [arXiv:2003.14327](https://arxiv.org/abs/2003.14327).
- [58] G. Mangano, G. Miele, S. Pastor, T. Pinto, O. Pisanti, and P. D. Serpico, Relic neutrino decoupling including flavor oscillations, *Nucl. Phys.* **B729**, 221 (2005).
- [59] P. F. de Salas and S. Pastor, Relic neutrino decoupling with flavor oscillations revisited, *J. Cosmol. Astropart. Phys.* **07** (2016) 051.
- [60] Particle Data Group, Review of particle physics, *Prog. Theor. Exp. Phys.* **2020**, 083C01 (2020).
- [61] S. Hannestad and T. Schwetz, Cosmology and the neutrino mass ordering, *J. Cosmol. Astropart. Phys.* **11** (2016) 035.
- [62] A. F. Heavens and E. Sellentin, Objective Bayesian analysis of neutrino masses and hierarchy, *J. Cosmol. Astropart. Phys.* **04** (2018) 047.