Machine learning improved fits of the sound horizon at the baryon drag epoch

Andoni Aizpuru,[*] Rubén Arjona,[†] and Savvas Nesseris[‡]
Instituto de Física Teórica UAM-CSIC, Universidad Autonóma de Madrid, Cantoblanco, 28049 Madrid, Spain

(Received 4 June 2021; accepted 26 July 2021; published 16 August 2021)

The baryon acoustic oscillations (BAO) have proven to be an invaluable tool in constraining the expansion history of the Universe at late times and are characterized by the comoving sound horizon at the baryon drag epoch $r_s(z_d)$. The latter quantity can be calculated either numerically using recombination codes or via fitting functions, such as the one by Eisenstein and Hu, made via grids of parameters of the recombination history. Here we quantify the accuracy of these expressions and show that they can strongly bias the derived constraints on the cosmological parameters using BAO data. Then, using a machine learning approach, called the genetic algorithms, we proceed to derive new analytic expressions for $r_s(z_d)$, which are accurate at the ∼0.003% level in a range of 10σ around the Planck 2018 best fit or ∼0.018% in a much broader range, compared to ∼2–4% for the Eisenstein and Hu expression, thus obtaining an improvement of two to three orders of magnitude. Moreover, we also provide fits that include the effects of massive neutrinos and an extension to the concordance cosmological model assuming variations of the fine structure constant. Finally, we note that our expressions can be used to ease the computational cost required to compute $r_s(z_d)$ with a Boltzmann code when deriving cosmological constraints using BAO data from current and upcoming surveys.

DOI: 10.1103/PhysRevD.104.043521

I. INTRODUCTION

Some of the strongest constraints on the expansion of the Universe at late times come from baryon acoustic oscillations (BAO) data. The BAO were formed in the early Universe, while it was very homogeneous [as probed today by the cosmic microwave background (CMB)] except for tiny fluctuations, and the photons and baryons were tightly coupled [1]. As the Universe expanded, it became cooler and less dense, while the fluctuations grew due to gravity. Acoustic waves were generated as the photon-baryon fluid was attracted and fell onto the overdensities producing compressions and rare factions due to the gravitational collapse and radiation pressure.

These acoustic waves propagated until the Universe became cool enough for the electrons and protons to recombine and then the baryons and photons decoupled. The time when the baryons were released from the drag of the photons is known as the drag epoch, z_d [2]. From then on, photons expanded freely while the acoustic waves froze in the baryons in a scale given by the size of the sound horizon at the drag epoch, dubbed $r_s(z_d)$. Progressively, baryons fell into dark matter potential wells, but dark matter was also attracted to baryon overdensities. Neutrinos did not interact, so they streamed away while dark matter responded to gravity and fell onto the overdensity.

The perturbations were dominated by photons and baryons as they were coupled, resulting in overdensities and overpressure that tried to equalize with the surrounding resulting in an expanding sound wave moving at the speed of sound, approximately $c_s^2 \sim 1/3$. The perturbation in photons and baryons was carried outward and the photons and baryons continued to expand whereas neutrinos spread out. Dark matter continued to fall into perturbations, which kept growing.

As the expanding Universe continued to cool down, it reached a point when the electrons and protons began to combine. Since photons did not scatter as efficiently they started to decouple. The sound speed dropped and the pressure wave slowed down. The process continued until the photons were completely decoupled, and then the perturbations smoothed out.[1] In fact, the sound speed of the baryon perturbation dropped so much that the pressure wave stalled. Thus, the original dark matter perturbation was left surrounded by a baryon perturbation in a shell. The two components attracted each other and the perturbations started to mix.[2]

[*] andoni.aizpuru@estudiante.uam.es
[†] ruben.arjona@uam.es
[‡] savvas.nesseris@csic.es

[1] http://mwhite.berkeley.edu/BAO.

[2] https://lweb.cfa.harvard.edu/∼deisenst/acousticpeak/.

The BAO provides a characteristic scale that is frozen in the galaxy distribution providing a standard ruler that can be measured as a function of redshift in either the galaxy correlation function or the galaxy power spectrum. The BAO determination of the geometry of the Universe is quite robust against systematics and has been measured by several surveys, such as the SDSS [3] and 2dFGRS [4]. The BAO signature provides a standard ruler that can be used to measure the geometry of the Universe and it can measure both the angular diameter distance $d_A(z)$ and the expansion rate $H(z)$. Measurements of the BAO only provide the combination of H_0 and $r_\mathrm{s}(z_\mathrm{d})$, which means that the two parameters are fully degenerate. As a result, the constraints obtained from the analysis of the BAO can be influenced significantly on the assumption of $r_\mathrm{s}(z_\mathrm{d})$ [5].

In order to accurately estimate $r_\mathrm{s}(z_\mathrm{d})$, one may use either recombination codes, such as RECFAST [6], CosmoRec [7] or HyRec [8,9], or analytic approximations based on fits of grids of parameters of the recombination history. A prominent example of the latter approach is the formula by Eisenstein and Hu [10], hereafter known as EH, which provides a fit of $r_\mathrm{s}(z_\mathrm{d})$ in terms of the matter and baryon density parameters. This formula has been extensively used in the literature in analyses of the BAO data, see for example Refs. [11–15]. However, as already observed in Ref. [10], this expression is only accurate to the ∼2% level and as a result is not appropriate for deriving cosmological constraints from BAO data in a percent cosmology era with current and upcoming surveys.

Over the years attempts to improve the EH formula have appeared. For example, the dependence of $r_\mathrm{s}(z_\mathrm{d})$ on various parameters, including massive and massless neutrinos, was examined in Ref. [16]. On the other hand, fits of $r_\mathrm{s}(z_\mathrm{d})$ including neutrinos and relativistic species were found in Refs. [17,18]. Finally, how the fraction of the baryonic mass in Helium Y_P and the relativistic degrees of freedom N_eff affects the sound horizon and how both are degenerate, was studied in Ref. [19].

The main limitation of the aforementioned analyses is that some *ad hoc* parametrizations were fitted to grids of parameters and $r_\mathrm{s}(z_\mathrm{d})$, thus being limited from the start on how accurate they can be. Hence, in our work we use machine learning to provide, in a data driven approach, extremely accurate fits to the comoving sound horizon at the baryon drag epoch $r_\mathrm{s}(z_\mathrm{d})$. We then compare these expressions against both the original formula of EH and the exact numerical estimation of the sound horizon, in order to quantify the amount of bias this expression introduces in the constraints.

In our analysis we also consider separately the effect of massive neutrinos and a varying fine structure constant and we find that our fits provide an improvement of a factor of three compared to other simple parametrizations and can be used in current and upcoming surveys to derive cosmological constraints so as to ease the computational cost that would be required when computing $r_\mathrm{s}(z_\mathrm{d})$ via a Boltzmann code.

The structure of our paper is as follows: in Sec. II we present the theoretical background and main assumptions in our work, while in Sec. III we present some details on our machine learning approach used to improve the sound horizon fits. In Sec. IV we present our main results, while in Sec. V we summarize our conclusions. Finally, in the Appendix we present some complementary fits for the redshift at the drag and recombination epochs.

II. THEORY

The comoving sound horizon at the drag epoch is given by

$$r_\mathrm{s}(z_\mathrm{d}) = \frac{1}{H_0}\int_{z_\mathrm{d}}^{\infty} \frac{c_\mathrm{s}(z)}{H(z)/H_0}\mathrm{d}z, \tag{1}$$

where z_d is the redshift at the drag epoch, see Eq. (4) of Ref. [10], while $c_\mathrm{s}(z)$ is the sound speed in the baryon-photon fluid given by

$$c_\mathrm{s} = \frac{c}{\sqrt{3(1+R)}}, \tag{2}$$

where $R = \frac{3\rho_b}{4\rho_\gamma} = \frac{3\Omega_{b,0}}{4\Omega_{\gamma,0}}a$ and c is the speed of light in vacuum. By definition, the sound horizon at the baryon drag epoch is the comoving distance a wave can travel prior to z_d and it depends on the epoch of recombination, the expansion of the Universe and the baryon-to-photon ratio. The sound horizon is well determined by the CMB measurements of the acoustic peaks.

Regarding the neutrinos, neutrino flavour oscillation experiments have shown that they are massive [20], providing a direct evidence for physics beyond the Standard Model. Cosmology is a very propitious stage to probe neutrino properties since they leave an imprint in the CMB and in the distribution of large-scale structure in the Universe. The energy density of massive neutrinos, $\rho_\nu = \sum m_{\nu,i} n_{\nu,i}$, corresponds to

$$\Omega_\nu h^2 \sim \frac{\sum m_{\nu,i}}{94\ \mathrm{eV}}, \tag{3}$$

where n_ν represents number density of neutrinos.

We also consider variations of fundamental constants, which are usually assumed to be constant over space-time. These constants are defined operationally, meaning that nature by itself does not force it to be constant. They have to be obtained experimentally since they are not given by the theory, see for instance Ref. [21] for a review on the variation of fundamental constants. Here we will examine the interesting case where the fine structure constant, defined in laboratory scales at late times as $\alpha_0 = \frac{e^2}{\hbar c}$, is rescaled and we will express its relative variation over its

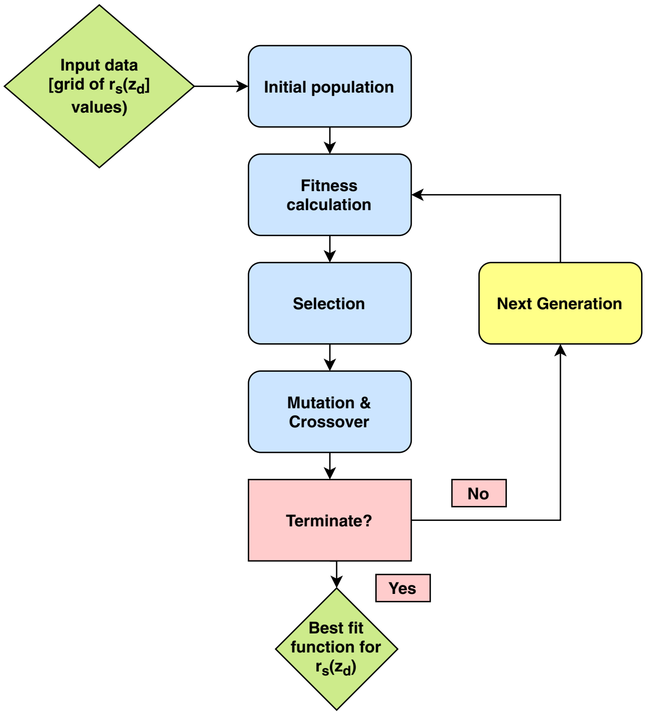

FIG. 1. A flowchart of the list of the steps for the GA reconstruction of $r_s(z_d)$.

standard model value as α/α_0. Thus, we assume that α is the value at early times of the fine structure constant and is rescaled with respect to its laboratory (late time) value α_0, with a sharp transition at intermediate redshifts.

If there are eventually signatures of a variation it would have imprints in different physical mechanisms such as the CMB anisotropies [22]. Constraints on this variation, both temporal and spatial, have been performed already [23–29], and this variation can be produced for example through an evolving scalar field that is coupled to the electromagnetic Lagrangian [24,30–32]. This will give rise to variations of the fine structure constant, a violation of the weak equivalence principle and violations of the standard $T_{\rm CMB}(z)$ law, as the number of photons is no longer conserved. These kinds of models can in principle be constrained by future large scale structure surveys using high-resolution spectroscopic data in combination with local astrophysical data, see Ref. [33] for updated constraints with current data and Ref. [34] for recent forecasts with upcoming surveys.

Another class of models where this occurs is the Bekenstein-Sanvik-Barrow-Magueijo model [35], where the electric charge is allowed to vary. Although such theories preserve the local gauge and Lorentz invariance, the fine structure constant will vary during the matter dominated era.

III. THE GENETIC ALGORITHMS

In this section we will describe the genetic algorithms (GAs) that will be used in our analysis to improve the sound horizon fits. The GA have been successfully used in cosmology for several reconstructions on a wide range of data, see for example Refs. [36–47]. Other applications of the GA cover other areas such as particle physics [48–50] and astronomy and astrophysics [51–53]. Other symbolic regression methods implemented in physics and cosmology can be found at [54–61].

The GA can be regarded as a machine learning technique constructed to carry out unsupervised regression of data; i.e., it performs nonparametric reconstructions that find an analytic function of one or more variables (like in our case here) that describes the data extremely well. The GA emulates the concept of biological evolution through the principle of natural selection, as brought by the genetic operations of mutation and crossover.

In essence, a set of trial functions evolves as time passes by through the effect of the stochastic operators of crossover, i.e., the joining of two or more candidate functions to form another one, and mutation, i.e., a random alteration of a candidate function. This process is then repeated thousands of times with different random seeds to ensure convergence and explore properly the functional space. In Fig. 1 we present a flowchart of the steps the GA goes through when reconstructing a function.

Since the GA is constructed as a stochastic approach, the probability that a population of functions will bring about offspring is principally assumed to be proportional to its fitness to the data, where in our analysis is given by a χ^2 statistic and give the information on how good every individual agrees with the data. For the simulated data in our analysis we are assuming that the likelihoods are sufficiently Gaussian that we use the χ^2 in our GA approach. Then, the probability to have offspring and the fitness of each individual is proportional to the likelihood causing an evolutionary pressure that favors the best-fitting functions in every population, hence directing the fit towards the minimum in a few generations.

In our analysis we reconstruct the $r_s(z_d)$ function considering that it depends on the following variables: $\{\Omega_m h^2, \Omega_b h^2\}$, $\{\Omega_m h^2, \Omega_b h^2, \Omega_\nu h^2\}$ and $\{\Omega_m h^2, \Omega_b h^2, \alpha/\alpha_0\}$, respectively. To calculate the sound horizon we use the code CLASS by Ref. [62] and the HyRec-2 recombination module Hy Rec2020 [8,9]. We then make grids of parameters and $r_s(z_d)$ and fit the values with the genetic algorithms. For example, when $\{\Omega_m h^2 = 0.13, \Omega_b h^2 = 0.0214\}$ we have that $r_s(z_d) = 151.365$ Mpc. Our reconstruction procedure is as follows. First, our predefined grammar was constructed on the following functions: exp, log, polynomials etc. and a set of operations $+, -, \times, \div$, see Table I for the complete list.

Once the initial population has been constructed, the fitness of each member, which indicates how accurately each individual of the population fits the data, is computed by a χ^2 statistic using the $r_s(z_d)$ data points directly as input, i.e.,

TABLE I. The grammars used in the GA analysis. Other complex forms are automatically produced by the mutation and crossover operations as described in the text.

Grammar type	Functions
Polynomials	c, x, $1+x$
Fractions	$\frac{x}{1+x}$
Trigonometric	$\sin(x)$, $\cos(x)$, $\tan(x)$
Exponentials	e^x, x^x, $(1+x)^{1+x}$
Logarithms	$\ln(x)$, $\ln(1+x)$

$$\chi^2 = \sum_{i=1}^{N} [r_{\mathrm{s,i}}(z_\mathrm{d}) - r_{s,\mathrm{GA}}(z_\mathrm{d})]^2, \tag{4}$$

where N represents the number of data points, which in our case was around 4000, and $r_\mathrm{s}(z_\mathrm{d})_{GA}$ is the fitting function derived by the GA. Notice that in Eq. (4) we are not considering uncertainties in each data point since we are taking directly the output value derived with the code CLASS.

Then, through a tournament selection process, see Ref. [36] for more details, the best-fitting functions in each generation are chosen and the two stochastic operations of crossover and mutation are used. The final output of the code is a mathematical function of $r_\mathrm{s}(z_\mathrm{d})$ that describes the sound horizon at the drag epoch in terms of the various cosmological parameters of interest.

IV. RESULTS

In this section we now present our machine learning fits to the sound horizon at the baryon drag epoch $r_\mathrm{s}(z_\mathrm{d})$. First, we will only include the dependence on the matter and baryon density parameters $\{\Omega_m h^2, \Omega_b h^2\}$, while later we will also consider the effect of massive neutrinos and a varying fine structure constant, i.e., the parameter vectors will be $\{\Omega_m h^2, \Omega_b h^2, \Omega_\nu h^2\}$ and $\{\Omega_m h^2, \Omega_b h^2, \alpha/\alpha_0\}$, respectively.

The computation of the sound horizon is described in Sec. III and we fit the values with both traditional minimization approaches and with the genetic algorithms. To simplify our notation we make the following definitions that will be used throughout the text: $\omega_b = \Omega_b h^2$, $\omega_m = \Omega_m h^2$, and $\omega_\nu = \Omega_\nu h^2$. In what follows, we will now describe our approach in more detail and present the results for the various cases.

A. Matter and baryons only

First, we consider the standard case of matter and baryons, as was also studied in Ref. [10] (hereafter denoted as EH). This case was obtained by simulating values for $\Omega_m h^2 \in [0.025, 0.5]$ and $\Omega_b h^2 \geq 0.0125$ and is given by [10]

$$r_\mathrm{s}(z_\mathrm{d}) \simeq \frac{44.5 \ln(\frac{9.83}{\omega_m})}{\sqrt{1 + 10\,\omega_b^{3/4}}}\ \mathrm{Mpc}, \tag{5}$$

which is accurate up to ~2%. Since now the recombination codes have more improved physics (for example an improved post-Saha expansion at early phases of hydrogen recombination, see Refs. [8,63] for a discussion), we have considered the same parametrization as in EH but with the coefficients as free parameters. By fitting the parametrization to a grid of values for $r_\mathrm{s}(z_\mathrm{d})$ for the range $\Omega_m h^2 \in [0.13, 0.15]$ and $\Omega_b h^2 \in [0.0214, 0.0234]$, which is around 10σ from the Planck best fit, we find the following improved expression

$$r_\mathrm{s}(z_\mathrm{d}) = \frac{45.5337 \ln(\frac{7.20376}{\omega_m})}{\sqrt{1 + 9.98592\,\omega_b^{0.801347}}}\ \mathrm{Mpc}, \tag{6}$$

which is accurate up to ~0.009%. Using the same grid of values with the GA we find the following fit that is even better

$$r_\mathrm{s}(z_\mathrm{d}) = \frac{1}{a_1 \omega_b^{a_2} + a_3 \omega_m^{a_4} + a_5 \omega_b^{a_6} \omega_m^{a_7}}\ \mathrm{Mpc}, \tag{7}$$

where

$$a_1 = 0.00785436,\quad a_2 = 0.177084,\quad a_3 = 0.00912388,$$
$$a_4 = 0.618711,\quad a_5 = 11.9611,\quad a_6 = 2.81343,$$
$$a_7 = 0.784719.$$

In this case, our GA improved expression given by Eq. (7) is accurate up to ~0.003%.

Next, we also consider a broader range of values for the parameter grid in order to allow for the fitting function to be used in BAO analyses without compromising its accuracy. In particular, we consider the range $\Omega_m h^2 \in [0.05, 0.25]$ and $\Omega_b h^2 \in [0.016, 0.03]$ and we find with the GA the following fit

$$r_\mathrm{s}(z_\mathrm{d}) = \left[\frac{1}{a_1 \omega_b^{a_2} + a_3 \omega_b^{a_4} \omega_m^{a_5} + a_6 \omega_m^{a_7}} - \frac{a_8}{\omega_m^{a_9}}\right]\ \mathrm{Mpc}, \tag{8}$$

where

$$a_1 = 0.00257366,\quad a_2 = 0.05032,\quad a_3 = 0.013,$$
$$a_4 = 0.7720642,\quad a_5 = 0.24346362,\quad a_6 = 0.00641072,$$
$$a_7 = 0.5350899,\quad a_8 = 32.7525,\quad a_9 = 0.315473.$$

which is accurate up to ~0.018%, i.e., a two orders of magnitude improvement from the EH expression of Eq. (5).

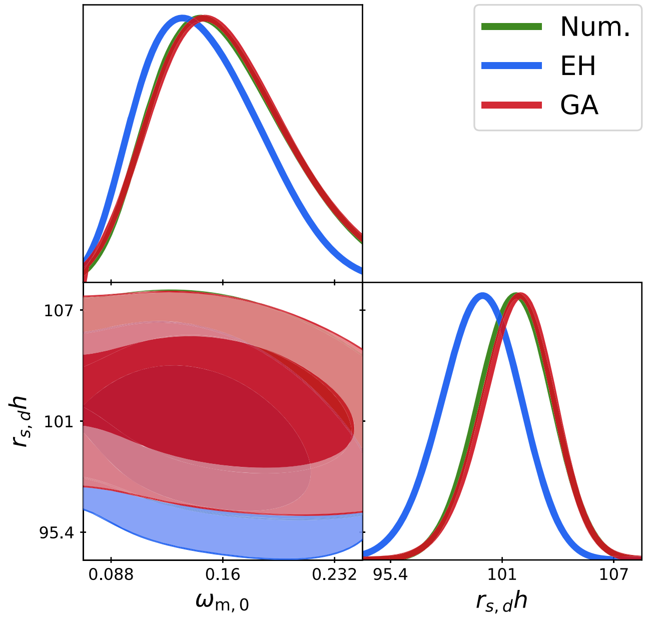

FIG. 2. A comparison of the confidence contours for the expression by EH for the sound horizon given by Eq. (5) (blue contour) against the improved expression found by the machine learning approach (GA) given by Eq. (8) (red contours) and the exact numerical approach (Num.) calculated via HyRec2020 (green contour), using the current BAO data as described in Ref. [42].

In order to quantify the bias introduced in deriving constraints on the cosmological parameters by using the less accurate expression of Eq. (5), we will now present the confidence contours and parameter constraints obtained via a Markov chain Monte Carlo with the code MONTEPYTHON3 of Ref. [64], using the currently available BAO data as described in Ref. [42] and the aforementioned $r_s(z_d)$ expressions. As mentioned earlier, $r_s(z_d)$ and $h \equiv H_0/100$ are degenerate, we in what follows we will consider the combination $r_{s,d}h = r_s(z_d)h$.

In particular, in Fig. 2 we show a comparison of the confidence contours for the EH expression for the sound horizon given by Eq. (5) (blue contour) against the machine learning improved expression (GA) given by Eq. (8) (red contours) and the exact numerical approach (Num.) calculated via HyRec2020 (green contour). Furthermore, in Table II we show the best fit, mean, and 95% limits for $(\omega_{m,0}, r_{s,d}h)$ obtained from the Markov chain Monte Carlo runs.

As can be seen, using the older and less accurate expression biases strongly the constraints for both $\omega_{m,0}$ and $r_s(z_d)h$ by almost half a σ and shifts the best-fit $\omega_{m,0}$ by ~9.3% from its true value, which is obtained using the full numerical approach. This implies that any analysis, e.g., Refs. [11–15], using the simple EH formula of Eq. (5) will be biased by about half a σ and should be interpreted with some care.

B. Matter, baryons, and massive neutrinos

Next, we also include massive neutrinos and this time we compare with the expression of Ref. [17], where the following fit was presented

$$r_s(z_d) \approx \frac{55.154 \exp\left[-72.3(\omega_\nu + 0.0006)^2\right]}{\omega_m^{0.25351}\omega_b^{0.12807}} \text{ Mpc}, \qquad (9)$$

which is accurate up to 0.29% within our range of values considered. Notice that this expression is accurate up to 0.021% if we limit to the range within 3σ of values derived by Planck and that $\omega_\nu = 0.0107\,(\sum m_\nu/1.0 \text{ eV})$.

In our case we consider the parameters in the range $\Omega_m h^2 \in [0.13, 0.15]$, $\Omega_b h^2 \in [0.0214, 0.0234]$, which is around 10σ from Planck, and for the massive neutrinos in the range $0 < \sum m_\nu < 0.6$ eV. Then, with the GA we find the improved fit that reads as follows:

$$r_s(z_d) = \frac{a_1 e^{a_2(a_3+\omega_\nu)^2}}{a_4\omega_b^{a_5} + a_6\omega_m^{a_7} + a_8(\omega_b\omega_m)^{a_9}} \text{ Mpc}, \qquad (10)$$

where the coefficients take the following values

$$\begin{aligned} a_1 &= 0.0034917, & a_2 &= -19.972694, & a_3 &= 0.000336186, \\ a_4 &= 0.0000305, & a_5 &= 0.22752, & a_6 &= 0.00003142567, \\ a_7 &= 0.5453798, & a_8 &= 374.14994, & a_9 &= 4.022356899, \end{aligned}$$

which is accurate up to 0.0076%, i.e., roughly a factor of three improvement over Eq. (9) in the range within 3σ of Planck and a factor of ~30 in the broader range.

C. Matter, baryons, and the fine structure constant

Finally, we also consider the effects of a varying fine structure constant on the sound horizon at the drag redshift. The fine structure constant α is already included in the recombination code HyRec2020 [8,9], thus the only modification in the code that was needed in this case was passing an extra parameter to CLASS.

Then, we simulate values of the $r_s(z_d)$ for the range $\Omega_m h^2 \in [0.13, 0.15]$, $\Omega_b h^2 \in [0.0214, 0.0234]$, and $\alpha/\alpha_0 \in [0.98, 1.02]$. The range for α/α_0 might seem

TABLE II. The best fit, mean, and 95% limits for $(\omega_{\mathrm{m},0}, r_{s,d}h)$ as discussed in the text. As seen, the older EH approach biases the estimated mean values for the parameters by almost half a σ, even though they share the same value of the χ^2 at the minimum $\chi^2_{\mathrm{min}} = 10.95$. The contours are shown in Fig. 2.

Method	Param	Best fit	mean ± σ	95% lower	95% upper
Numerical	$\omega_{\mathrm{m},0}$	0.1968	$0.1641^{+0.04}_{-0.051}$	0.0788	0.251
	$r_{s,d}h$	102.1	$101.7^{+1.9}_{-1.8}$	97.91	105.4
EH	$\omega_{\mathrm{m},0}$	0.1816	$0.1488^{+0.036}_{-0.044}$	0.07544	0.2222
	$r_{s,d}h$	100.3	$99.9^{+2.2}_{-1.9}$	95.74	103.9
GA	$\omega_{\mathrm{m},0}$	0.1959	$0.1645^{+0.04}_{-0.054}$	0.07738	0.2535
	$r_{s,d}h$	102.3	$101.7^{+1.9}_{-1.8}$	97.94	105.5

restrictive, but in Ref. [65] it was shown that with current data any variations are constrained to $\Delta\alpha/\alpha_0 \sim 10^{-3}$, while with future large scale structure data and local astrophysical measurements the constraints can be further reduced to $\Delta\alpha/\alpha_0 \sim 10^{-6}$. Following the same procedure as before we find the following fitting formula using an EH-like parametrization

$$r_\mathrm{s}(z_\mathrm{d}) = \frac{a_1 \ln(\frac{a_2}{\omega_m})}{\sqrt{1 + a_3\omega_b^{a4}}}(\alpha/\alpha_0)^{a_5}\ \mathrm{Mpc}, \tag{11}$$

which is accurate up to ~0.047% and the parameters are given by

$$a_1 = 45.417, \quad a_2 = 7.15466, \quad a_3 = 10.1167,$$
$$a_4 = 0.811586, \quad a_5 = -1.254537.$$

On the other hand, with the GA we have found an improved fit which reads as follows

$$r_\mathrm{s}(z_\mathrm{d}) = \frac{1}{a_1\omega_b^{a_2}\omega_m^{a_3}[(\alpha/\alpha_0)^{a_4} + \omega_b^{a_5}\omega_m^{a_6}] + a_7\omega_m^{a_8}}\ \mathrm{Mpc}, \tag{12}$$

where the coefficients take the following values

$$a_1 = 0.00730258, \quad a_2 = 0.088182, \quad a_3 = 0.099958,$$
$$a_4 = 1.97913, \quad a_5 = 0.346626, \quad a_6 = 0.0092295,$$
$$a_7 = 0.0074056, \quad a_8 = 0.8659935,$$

which is accurate up to 0.0077% and is roughly a factor of six improvement over the EH-like parametrization of Eq. (11).

V. CONCLUSIONS

In summary, we have presented extremely accurate machine learning fits to the comoving sound horizon at the baryon drag epoch $r_\mathrm{s}(z_\mathrm{d})$ as a function of cosmological parameters and we compared our results with other expressions found in the literature. In particular, we considered the widely used Eisenstein-Hu fitting formula given by Eq. (5), which is accurate to the ~2% level, and showed how it may strongly bias any constraints on the matter density parameter obtained by using the current BAO data as described in Ref. [42].

In particular, we found that the confidence contours are biased by roughly half a sigma, while the matter density parameter $\omega_{\mathrm{m},0}$ is shifted at a ~9.3% level from its correct value, which is obtained using the full numerical analysis. On the other hand, our machine learning fits given by Eq. (7) do not suffer from this issue, as they are accurate to within ~0.003%. Furthermore, in our analysis we also considered the effect of massive neutrinos, see Eq. (10) and a varying fine structure constant, see Eq. (12), finding that our fits have an improvement of a factor of three to four compared to other simple EH-like parametrizations.

It should be noted though that according to Ref. [9], HyRec2020 achieves an accuracy of the order of $\sim 10^{-4}$, which is comparable to the precision of the GA results. On the other hand, many forthcoming surveys like Euclid, see Ref. [66], expected to measure the cosmological parameters to about 1% precision, which is about two orders of magnitude larger than the precision of the GA results. As a result, the latter are not expected to bias any analyses with data products from forthcoming surveys in the near term, such as Euclid.

To conclude, we presented machine learning improved expressions for the sound horizon at the drag redshift, which are more accurate in some cases even by two orders of magnitude compared to other similar expressions already found in the literature. The advantage of our approach is that the new expressions do not bias the parameter constraints obtained from BAO data, thus they can be used in BAO analyses coming from current and upcoming surveys to derive cosmological constraints and ease the computational cost that would be required when computing $r_\mathrm{s}(z_\mathrm{d})$ with a full Boltzmann code.

Numerical Analysis Files.—The genetic algorithm code used by the authors in the analysis of the paper and the expressions of the fits can be found at [67].

ACKNOWLEDGMENTS

The authors acknowledge use of the codes: CLASS version 3.0.1, MONTEPYTHON3, and the HyRec-2 recombination module HyRec2020. They also acknowledge support from the Research Project No. PGC2018-094773-B-C32 and the Centro de Excelencia Severo Ochoa Program No. SEV-2016-0597. S. N. also acknowledges support from the Ramón y Cajal program through Grant No. RYC-2014-15843. The authors also acknowledge use of the Hydra cluster at the Instituto de Física Teórica (IFT), on which some of the numerical computations for this paper took place.

APPENDIX: FITS FOR THE REDSHIFT OF THE DRAG EPOCH AND THE PHOTON-DECOUPLING SURFACE

Here we provide some fits for the redshift at the drag epoch $z_{\rm d}$, which can be used in Eq. (1) as a complementary fit instead of the analytic fit of $r_{\rm s}(z_{\rm d})$ and also a fit to the redshift at the photon-decoupling surface z_*.

1. The drag redshift $z_{\rm d}$

The fit for the drag redshift from Ref. [10] is given by

$$z_d = \frac{1291(\omega_m)^{0.251}}{1+0.659(\omega_m)^{0.828}}[1+b_1(\omega_b)^{b2}], \tag{A1}$$

where

$$b_1 = 0.313(\omega_m)^{-0.419}[1+0.607(\omega_m)^{0.674}],$$
$$b_2 = 0.238(\omega_m)^{0.223},$$

and which is accurate up to ∼3.7%.

To improve this fit, we simulate values for $z_{\rm d}$ in the range $\Omega_m h^2 \in [0.13, 0.15]$ and $\Omega_b h^2 \in [0.0214, 0.0234]$, which is around 10σ from Planck. Then, with the GA we find

$$z_{\rm d} = \frac{1+428.169\omega_b^{0.256459}\omega_m^{0.616388}+925.56\omega_m^{0.751615}}{\omega_m^{0.714129}}, \tag{A2}$$

which is accurate up to ∼0.001%.

2. The redshift at recombination z_*

The fit for the redshift to the photon-decoupling surface z_* from Ref. [68] is given by

$$z_* = 1048[1+0.00124(\Omega_b h^2)^{-0.738}][1+g_1(\Omega_m h^2)^{g_2}], \tag{A3}$$

where

$$g_1 = \frac{0.0783(\Omega_b h^2)^{-0.238}}{1+39.5(\Omega_b h^2)^{0.763}},$$
$$g_2 = \frac{0.560}{1+21.1(\Omega_b h^2)^{1.81}},$$

and which is accurate up to ∼0.3%.

To improve this fit, we simulate values for z_* for the range $\Omega_m h^2 \in [0.13, 0.15]$ and $\Omega_b h^2 \in [0.0214, 0.0234]$, which is around 10σ from Planck. Then, as before, with the GA we find

$$z_* = \frac{391.672\omega_m^{-0.372296}+937.422\omega_b^{-0.97966}}{\omega_m^{-0.0192951}\omega_b^{-0.93681}}+\omega_m^{-0.731631}, \tag{A4}$$

which is accurate up to ∼0.0005%.

[1] S. Dodelson, *Modern Cosmology* (Elsevier, New York, 2003).

[2] S. Weinberg, *Cosmology* (Oxford University Press, New York, 2008).

[3] A. de Mattia *et al.*, Mon. Not. R. Astron. Soc. **501**, 5616 (2021).

[4] W. J. Percival, S. Cole, D. J. Eisenstein, R. C. Nichol, J. A. Peacock, A. C. Pope, and A. S. Szalay, Mon. Not. R. Astron. Soc. **381**, 1053 (2007).

[5] A. Cuceu, J. Farr, P. Lemos, and A. Font-Ribera, J. Cosmol. Astropart. Phys. 10 (2019) 044.

[6] S. Seager, D. D. Sasselov, and D. Scott, Astrophys. J. Lett. **523**, L1 (1999).

[7] J. Chluba and R. M. Thomas, Mon. Not. R. Astron. Soc. **412**, 748 (2011).

[8] N. Lee and Y. Ali-Haïmoud, Phys. Rev. D **102**, 083517 (2020).

[9] Y. Ali-Haimoud and C. M. Hirata, Phys. Rev. D **83**, 043513 (2011).

[10] D. J. Eisenstein and W. Hu, Astrophys. J. **496**, 605 (1998).

[11] F. Beutler, C. Blake, M. Colless, D. H. Jones, L. Staveley-Smith, L. Campbell, Q. Parker, W. Saunders, and F. Watson, Mon. Not. R. Astron. Soc. **416**, 3017 (2011).

[12] E. Komatsu *et al.* (WMAP Collaboration), Astrophys. J. Suppl. **192**, 18 (2011).

[13] K. Bamba, S. Capozziello, S. Nojiri, and S. D. Odintsov, Astrophys. Space Sci. **342**, 155 (2012).

[14] Z. Zhai and Y. Wang, J. Cosmol. Astropart. Phys. 07 (2019) 005.

[15] M. Martinelli *et al.* (EUCLID Collaboration), Astron. Astrophys. **644**, A80 (2020).

[16] K. Thepsuriya and A. Lewis, J. Cosmol. Astropart. Phys. 01 (2015) 034.

[17] E. Aubourg *et al.*, Phys. Rev. D **92**, 123516 (2015).

[18] L. Anderson *et al.* (BOSS Collaboration), Mon. Not. R. Astron. Soc. **441**, 24 (2014).

[19] Z. Hou, R. Keisler, L. Knox, M. Millea, and C. Reichardt, Phys. Rev. D **87**, 083008 (2013).

[20] S. Bilenky, J. Phys. Conf. Ser. **718**, 062005 (2016).

[21] S. J. Landau, in *Proceedings of the IAU Symposium 357: White Dwarfs as Probes of Fundamental Physics and Tracers of Planetary, Stellar & Galactic Evolution Hilo, Big Island, Hawaii, USA, 2019* (2020), arXiv:2002.00095.

[22] J.-P. Uzan, Living Rev. Relativity **14**, 2 (2011).

[23] M. T. Clara and C. J. A. P. Martins, Astron. Astrophys. **633**, L11 (2020).

[24] I. De Martino, C. J. A. P. Martins, H. Ebeling, and D. Kocevski, Universe **2**, 34 (2016).

[25] I. de Martino, C. J. A. P. Martins, H. Ebeling, and D. Kocevski, Phys. Rev. D **94**, 083008 (2016).

[26] A. Hees, O. Minazzoli, and J. Larena, Phys. Rev. D **90**, 124064 (2014).

[27] L. Colao, R. Holanda, and R. Silva, arXiv:2004.08484.

[28] L. Lopez-Honorez, O. Mena, S. Palomares-Ruiz, P. Villanueva-Domingo, and S. J. Witte, J. Cosmol. Astropart. Phys. 06 (2020) 026.

[29] M. R. Wilczynska *et al.*, Sci. Adv. **6**, eaay9672 (2020).

[30] T. R. Taylor and G. Veneziano, Phys. Lett. B **213**, 450 (1988).

[31] J. A. Casas, J. Garcia-Bellido, and M. Quiros, Nucl. Phys. **B361**, 713 (1991).

[32] J. A. Casas, J. Garcia-Bellido, and M. Quiros, Classical Quantum Gravity **9**, 1371 (1992).

[33] L. Hart and J. Chluba, Mon. Not. R. Astron. Soc. **493**, 3255 (2020).

[34] M. Martinelli *et al.*, arXiv:2105.09746.

[35] H. B. Sandvik, J. D. Barrow, and J. Magueijo, Phys. Rev. Lett. **88**, 031302 (2002).

[36] C. Bogdanos and S. Nesseris, J. Cosmol. Astropart. Phys. 05 (2009) 006.

[37] S. Nesseris and A. Shafieloo, Mon. Not. R. Astron. Soc. **408**, 1879 (2010).

[38] S. Nesseris and J. Garcia-Bellido, J. Cosmol. Astropart. Phys. 11 (2012) 033.

[39] S. Nesseris and J. García-Bellido, Phys. Rev. D **88**, 063521 (2013).

[40] D. Sapone, E. Majerotto, and S. Nesseris, Phys. Rev. D **90**, 023012 (2014).

[41] R. Arjona, J. Cosmol. Astropart. Phys. 08 (2020) 009.

[42] R. Arjona and S. Nesseris, J. Cosmol. Astropart. Phys. 11 (2020) 042.

[43] R. Arjona and S. Nesseris, Phys. Rev. D **101**, 123525 (2020).

[44] R. Arjona and S. Nesseris, Phys. Rev. D **103**, 103539 (2021).

[45] R. Arjona and S. Nesseris, Phys. Rev. D **103**, 063537 (2021).

[46] R. Arjona, H.-N. Lin, S. Nesseris, and L. Tang, Phys. Rev. D **103**, 103513 (2021).

[47] R. Arjona and S. Nesseris, arXiv:2105.09049.

[48] S. Abel, D. G. Cerdeo, and S. Robles, arXiv:1805.03615.

[49] B. C. Allanach, D. Grellscheid, and F. Quevedo, J. High Energy Phys. 07 (2004) 069.

[50] Y. Akrami, P. Scott, J. Edsjo, J. Conrad, and L. Bergstrom, J. High Energy Phys. 04 (2010) 057.

[51] M. Wahde and K. Donner, Astron. Astrophys. **379**, 115 (2001).

[52] V. Rajpaul, in *Proceedings of the 56th Annual Conference of the South African Institute of Physics (SAIP 2011), Gauteng, South Africa, 2011* (2012), pp. 519–524, arXiv:1202.1643.

[53] M. Ho, M. M. Rau, M. Ntampaka, A. Farahi, H. Trac, and B. Poczos, Astrophys. J. **887**, 25 (2019).

[54] S.-M. Udrescu and M. Tegmark, Sci. Adv. **6**, eaay2631 (2020).

[55] Y. Setyawati, M. Prrer, and F. Ohme, Classical Quantum Gravity **37**, 075012 (2020).

[56] H. Vaddireddy, A. Rasheed, A. E. Staples, and O. San, arXiv:1911.05254.

[57] K. Liao, A. Shafieloo, R. E. Keeley, and E. V. Linder, Astrophys. J. Lett. **886**, L23 (2019).

[58] E. Belgacem, S. Foffa, M. Maggiore, and T. Yang, Phys. Rev. D **101**, 063505 (2020).

[59] Y. Li, R. S. Rainer, W. Cheng, and Y. Hao, in *Proceedings of the 3rd North American Particle Accelerator Conference (NAPAC2019)* (2019), arXiv:1904.05683.

[60] M. Bernardini, L. Mayer, D. Reed, and R. Feldmann, Mon. Not. R. Astron. Soc. **496**, 5116 (2020).

[61] A. Gmez-Valent and L. Amendola, in *Proceedings of the 15th Marcel Grossmann Meeting on Recent Developments in Theoretical and Experimental General Relativity, Astrophysics, and Relativistic Field Theories* (2019), arXiv:1905.04052.

[62] D. Blas, J. Lesgourgues, and T. Tram, J. Cosmol. Astropart. Phys. 07 (2011) 034.

[63] J. A. Rubino-Martin, J. Chluba, W. A. Fendt, and B. D. Wandelt, Mon. Not. R. Astron. Soc. **403**, 439 (2010).

[64] T. Brinckmann and J. Lesgourgues, Phys. Dark Universe **24**, 100260 (2019).

[65] P. A. R. Ade *et al.* (Planck Collaboration), Astron. Astrophys. **580**, A22 (2015).

[66] R. Laureijs *et al.* (EUCLID Collaboration), arXiv:1110.3193.

[67] https://github.com/RubenArjona.

[68] W. Hu and N. Sugiyama, Astrophys. J. **471**, 542 (1996).