Parameter estimation of binary black holes in the endpoint of the up-down instability

Viola De Renzis,[1,2,*] Davide Gerosa,[1,2,3] Matthew Mould,[3] Riccardo Buscicchio,[1,2] and Lorenzo Zanga[1]

[1]*Dipartimento di Fisica "G. Occhialini", Universitá degli Studi di Milano-Bicocca, Piazza della Scienza 3, 20126 Milano, Italy*

[2]*INFN, Sezione di Milano-Bicocca, Piazza della Scienza 3, 20126 Milano, Italy*

[3]*School of Physics and Astronomy & Institute for Gravitational Wave Astronomy, University of Birmingham, Birmingham, B15 2TT, United Kingdom*

(Received 24 April 2023; accepted 30 May 2023; published 12 July 2023)

Black-hole binary spin precession admits equilibrium solutions corresponding to systems with (anti) aligned spins. Among these, binaries in the up-down configuration, where the spin of the heavier (lighter) black hole is co(counter)aligned with the orbital angular momentum, might be unstable to small perturbations of the spin directions. The occurrence of the up-down instability leads to gravitational-wave sources that formed with aligned spins but are detected with precessing spins. We present a Bayesian procedure based on the Savage-Dickey density ratio to test the up-down origin of gravitational-wave events. This is applied to both simulated signals, which indicate that achieving strong evidence is within the reach of current experiments, and the LIGO/Virgo events released to date, which indicate that current data are not informative enough.

DOI: 10.1103/PhysRevD.108.024024

I. INTRODUCTION

Gravitational-wave (GW) detections provide measurements of the intrinsic properties of astrophysical black holes (BHs), notably their masses and spins. At the time of writing, ground-based interferometers LIGO and Virgo have observed about 70 mergers of stellar-mass BHs with false alarm rates $<1\ \mathrm{yr}^{-1}$ [1–4] and substantially more detections are expected from the upcoming observing runs [5,6].

GWs emitted during the inspiral of BH binaries are mostly beamed along the direction of the orbital angular momentum $\boldsymbol{L}$. If the spins of the two BHs $\boldsymbol{S}_{1,2}$ are misaligned with $\boldsymbol{L}$, couplings between these three momenta cause them to precess [7,8]. The resulting motion imparts characteristic modulations to the amplitude and phase of emitted GWs. From an astrophysical perspective, measuring spin precession is important to elucidate the possible astrophysical formation pathways of BH binaries, with large spin misalignments thought to be indicative of sources formed via dynamical interactions [9,10].

Configurations with spins that are either aligned or antialigned with the orbital angular momentum are equilibrium solutions of the relativistic spin-precession equations. This means that binaries that are *exactly* aligned will remain so. There are four such cases, which we refer to as up-up, down-down, down-up, and up-down, where "up" ("down") indicates spins that are parallel (antiparallel) to the orbital angular momentum and the direction before (after) the hyphen refers to the more (less) massive BH. Crucially, equilibrium does not imply stability. Reference [11] showed that, while up-up, down-down, and down-up binaries are always stable, up-down binaries can be unstable to spin precession. For these sources, infinitesimal perturbations to the spin directions cause large precession cycles. In particular, up-down binaries are stable at early times and turn unstable at the critical orbital separation [11],

$$r_{\mathrm{UD}+} = \frac{(\sqrt{\chi_1} + \sqrt{q\chi_2})^4}{(1-q)^2} M, \tag{1}$$

where $\chi_i = S_i/m_i^2$ are the Kerr parameters of the BHs, $q = m_2/m_1 \leq 1$ is the mass ratio, and $M = m_1 + m_2$ is the total mass of the system.[1] The up-down instability was first derived using a post-Newtonian (PN) approach [11] and then confirmed using both independent PN codes [12,13] and numerical-relativity simulations [14].

Measuring the up-down instability in GW data would provide a direct observation of an exquisite feature of the two-body problem in general relativity. At the same time, the up-down instability might also dilute the effectiveness of the spin orientations in discriminating BH-binary formation channels: GW sources that are observed with

*v.derenzis@campus.unimib.it

[1]Throughout the paper, we use natural units where $c = G = 1$.

precessing spins in the LIGO/Virgo band did not necessarily form with misaligned spins. Rather, the spins used to be (anti) aligned and became misaligned before merger. The flip side of the same coin is that observing unstable binaries will point toward a formation channel that can conceivably explain binaries with up-down spins. Notably, this might include AGN disks surrounding supermassive BHs [15,16], where the spins of embedded stellar-mass BH binaries are expected to either align or antialign with the angular momentum of the disk [17].

The up-down instability provides a testable prediction for GW observations. Reference [18] showed that unstable up-down BHs do not disperse in the available parameter space but converge to a well-defined endpoint late in the inspiral. This is a precessing configuration where all three angular momenta $\boldsymbol{S}_1$, $\boldsymbol{S}_2$, and $\boldsymbol{L}$ are coplanar, and furthermore, the two BH spins are collinear, namely [18],

$$\cos\theta_1 = \frac{\chi_1 - q\chi_2}{\chi_1 + q\chi_2}, \tag{2}$$

$$\cos\theta_2 = \frac{\chi_1 - q\chi_2}{\chi_1 + q\chi_2}, \tag{3}$$

$$\phi_{12} = 0, \tag{4}$$

where θ_i indicate the tilts angles between $\boldsymbol{S}_i$ and $\boldsymbol{L}$, and ϕ_{12} indicates the azimuthal angle between the two BH spins measured in the orbital plane. After the instability is triggered, binaries reach this analytical endpoint after the orbital separation has decreased by only $\lesssim 100M$ [18]. Therefore, binaries that form as up-down and become unstable will appear in our detectors with spin orientations that are well approximated by Eqs. (2)–(4).

In this paper, we perform Bayesian parameter estimation of precessing BH binaries in the endpoint of the up-down instability. Should an unstable up-down binary enter the LIGO band, can we tell that this source was originally stable and aligned? In statistical terms, this is a model-selection problem between a broader hypothesis where binaries are generically precessing and a narrower hypothesis with constraints given by Eqs. (2)–(4). We apply this line of reasoning to both simulated signals and the current catalog of GW events. By employing the Savage-Dickey density ratio, we compute the odds in favor of the up-down hypothesis over that of generically precessing BH binaries. Crucially, this only requires an inference run with the uninformative prior, with the odds computed by post-processing the recovered posterior samples.

In Sec. II, we derive the statistical framework and describe how it can be used to assess whether observed binaries are in the endpoint of the up-down instability. In Sec. III, we present our results for an injection campaign and real sources, and also demonstrate that evolving binary BH spin posteriors backwards in time is a useful diagnostic when investigating the up-down instability. We finish with our conclusions in Sec. IV.

II. METHODS

A. Gravitational-wave signals

We first consider synthetic GW signals from individual binary BH coalescences on quasicircular orbits and target the statistical inference of all 15 parameters of the problem. These are two detector frame masses $m_{1,2}$, 6 spin degrees of freedom (magnitudes $\chi_{1,2}$, tilts $\theta_{1,2}$, azimuthal angles ϕ_{12} and ϕ_{JL}), and seven extrinsic parameters (luminosity distance D_L, sky location α, δ, polar angle θ_{JN}, polarization ψ, coalescence time t_c, and phase ϕ_c).

Signals are analyzed using the parallel version of the BILBY inference code [19,20]. We use the IMRPhenomXPHM approximant [21] for both injection and recovery. We consider a three-detector network made of LIGO Livingston, LIGO Hanford, and Virgo at the sensitivity expected for the upcoming O4 run. We use data segments of 4 s, a sampling frequency of 2048 Hz, a low-frequency cutoff of 20 Hz, and zero noise. Spin orientations are quoted at a reference frequency of 20 Hz. We use the DYNESTY sampler [22] with 2048 live points, a random walk sampling method, a number of autocorrelation equal to 50, and a likelihood that is marginalized over time and distance.

Our priors are those commonly used in the standard LIGO/Virgo analyses [1–4]. In particular, detector-frame component masses are distributed uniformly in $m_{1,2} \in [5, 100]M_\odot$ with bounds in mass ratio $q \in [1/8, 1]$ and detector-frame chirp mass $\mathcal{M} \in [10, 60]M_\odot$ while spins are distributed uniformly in magnitude $\chi_{1,2} \in [0, 0.99]$ and isotropically in directions.

In the following, we also postprocess GW data using publicly available posterior samples for the GWTC-2.1 [3] and the GWTC-3 [4] data releases. Among the available datasets, we use results from the IMRPhenomXPHM waveform model where the merger rate is uniform in comoving volume and source-frame time. We consider binary BH mergers with false alarm rates $<1\ \mathrm{yr}^{-1}$ in at least one of the detection pipelines. From these, we exclude all the events that potentially contain a neutron star. The resulting list of 69 events is reported in Table I.

When needed, we covert between PN orbital separation r and GW frequency f_{ref} using the 2PN expressions from Ref. [8].

B. Savage-Dickey density ratio

Given the data d associated with a measurement and model hypothesis $\mathcal{H}$ characterized by parameters θ, the Bayesian evidence is defined as

$$\mathcal{Z}(d|\mathcal{H}) = \int \mathcal{L}(d|\theta, \mathcal{H})\pi(\theta|\mathcal{H})\mathrm{d}\theta, \tag{5}$$

TABLE I. Current GW events and their Bayes factors in favor of the up-down hypothesis over generic spin precession. We select events with false alarm rates $<1\ \mathrm{yr}^{-1}$ in at least one of the LIGO/Virgo searches, excluding those that can potentially include a neutron star.

Event	$\ln \mathcal{B}$	Event	$\ln \mathcal{B}$
GW150914	0.14	GW190731_140936	0.11
GW151012	0.54	GW190803_022701	0.11
GW151226	0.50	GW190805_211137	0.61
GW170104	−0.02	GW190828_063405	0.3
GW170608	0.18	GW190828_065509	0.15
GW170729	0.47	GW190910_112807	−0.06
GW170809	0.26	GW190915_235702	0.29
GW170814	−0.06	GW190924_021846	0.31
GW170818	0.58	GW190925_232845	0.24
GW170823	0.26	GW190929_012149	−0.15
GW190408_181802	0.02	GW190930_133541	0.59
GW190412	0.6	GW191103_012549	0.58
GW190413_052954	−0.01	GW191105_143521	0.06
GW190413_134308	0.07	GW191109_010717	−0.83
GW190421_213856	0.09	GW191127_050227	0.31
GW190503_185404	−0.04	GW191129_134029	0.33
GW190512_180714	0.33	GW191204_171526	0.79
GW190513_205428	0.48	GW191215_223052	0.11
GW190514_065416	−0.01	GW191216_213338	0.27
GW190517_055101	0.53	GW191222_033537	−0.16
GW190519_153544	0.35	GW191230_180458	0.25
GW190521	−0.26	GW200112_155838	0.07
GW190521_074359	−0.42	GW200128_022011	0.46
GW190527_092055	0.23	GW200129_065458	0.63
GW190602_175927	0.44	GW200202_154313	0.1
GW190620_030421	0.52	GW200208_130117	−0.04
GW190630_185205	−0.15	GW200209_085452	0.21
GW190701_203306	0.05	GW200216_220804	0.26
GW190706_222641	0.8	GW200219_094415	0.05
GW190707_093326	0.04	GW200224_222234	0.2
GW190708_232457	0.15	GW200225_060421	−0.11
GW190720_000836	0.58	GW200302_015811	0.05
GW190725_174728	0.39	GW200311_115853	0.32
GW190727_060333	0.44	GW200316_215756	0.57
GW190728_064510	0.32		

where $\mathcal{L}$ is the likelihood and π is the prior distribution. Model selection in favor of, say, a "narrow" model $\mathcal{H}_{\rm N}$ over a "broad" model $\mathcal{H}_{\rm B}$ requires computing the posterior odds,

$$\mathcal{O} = \frac{\mathcal{Z}(d|\mathcal{H}_{\rm N})}{\mathcal{Z}(d|\mathcal{H}_{\rm B})}\frac{\pi(\mathcal{H}_{\rm N})}{\pi(\mathcal{H}_{\rm B})}, \tag{6}$$

where the first term (ratio of the evidences) is the Bayes factor $\mathcal{B}$. Values of the posterior odds are often associated to descriptive terms using the so-called Jeffrey scale [23], where $|\ln \mathcal{O}| < 1$ is classified as "inconclusive," $1 < |\ln \mathcal{O}| < 2.5$ is classified as "weak" evidence, $2.5 < |\ln \mathcal{O}| < 5$ is classified as "moderate" evidence, and $|\ln \mathcal{O}| > 5$ is classified as "strong" evidence. The sign of the log Bayes factor indicates which of the two models is statistically favored, with $\ln \mathcal{O} > 0$ signaling a preference for $\mathcal{H}_{\rm N}$ over $\mathcal{H}_{\rm B}$. In the following, we consider equal model priors such that $\mathcal{O} = \mathcal{B}$.

Let us now assume that model $\mathcal{H}_{\rm N}$ is nested within $\mathcal{H}_{\rm B}$. That is, among the parameters $\theta = \{\varphi, \gamma\}$, a subset of parameters φ is common to both models, while the other parameters γ are constrained to $\gamma_{\rm N}(\varphi)$ in the narrow model. Let us also assume that the prior on φ is the same for the two models. In symbols, this is

$$\pi(\varphi|\mathcal{H}_{\rm N}) = \pi(\varphi|\gamma = \gamma_{\rm N}(\varphi), \mathcal{H}_{\rm B}). \tag{7}$$

Within these assumptions, the Bayes factor in favor of the narrow model reduces to

$$\mathcal{B} = \int \frac{p(\varphi, \gamma = \gamma_{\rm N}(\varphi)|d, \mathcal{H}_{\rm B})}{\int \pi(\varphi', \gamma = \gamma_{\rm N}(\varphi)|\mathcal{H}_{\rm B})\mathrm{d}\varphi'}\mathrm{d}\varphi. \tag{8}$$

A formal proof of Eq. (8) is presented in the Appendix. For the specific case where γ_N does not depend on φ, one has

$$\mathcal{B} = \frac{p(\gamma = \gamma_{\rm N}|d, \mathcal{H}_{\rm B})}{\pi(\gamma = \gamma_{\rm N}|\mathcal{H}_{\rm B})}, \tag{9}$$

where the numerator (denominator) corresponds to the posterior (prior) marginalized over the common parameters φ. Equation (9) is the so-called Savage-Dickey density ratio [24]. The key, practical advantage of both these expressions is that they only depend on the broad model $\mathcal{H}_{\rm B}$. One does *not* need to perform inference in the narrow model $\mathcal{H}_{\rm N}$, which can be challenging for nontrivial submanifolds $\gamma_{\rm N}(\varphi)$. It is sufficient to sample the broad model $\mathcal{H}_{\rm B}$ and then evaluate the resulting posterior and prior probability densities at the location prescribed by the narrow model.

C. Application to up-down binaries

For the specific case we are addressing here, the broad model $\mathcal{H}_{\rm B}$ is that of generically precessing BH binaries described in Sec. II A. The narrow model $\mathcal{H}_{\rm N}$ consists of binaries in the endpoint of the up-down instability, which are subject to the three constraints of Eqs. (2)–(4). From these, we define the parameters $\gamma = \{\gamma_1, \gamma_2, \gamma_3\}$, where

$$\gamma_1 = \frac{\cos\theta_1 - \cos\theta_{\rm UD}(q, \chi_1, \chi_2)}{2}, \tag{10}$$

$$\gamma_2 = \frac{\cos\theta_2 - \cos\theta_{\rm UD}(q, \chi_1, \chi_2)}{2}, \tag{11}$$

$$\gamma_3 = \frac{1}{\pi}\arctan\left(\frac{\sin\phi_{12}}{\cos\phi_{12}}\right), \tag{12}$$

and

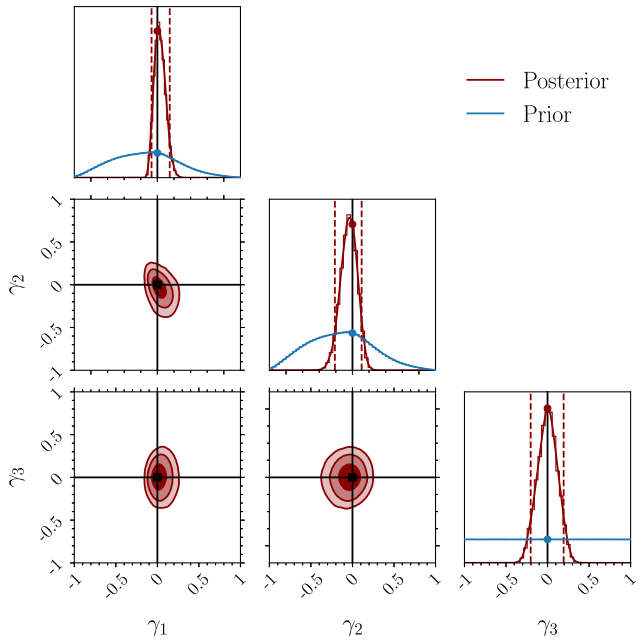

FIG. 1. Joint posterior distributions of the rescaled parameters $\gamma = \{\gamma_1, \gamma_2, \gamma_3\}$ defined in Eqs. (10)–(12). Contour levels correspond to 50%, 90%, and 99% credible regions. Red dashed lines in the 1D marginals indicate the 90% credible intervals. Solid black lines mark the location of the narrow model $\gamma = 0$, i.e., the endpoint of the up-down instability. Black scatter points indicate the value of the posterior (red) and prior (blue) distributions at the endpoint, which are the key ingredients entering the Savage-Dickey evaluation of the Bayesian odds.

$$\cos\theta_{\rm UD}(q,\chi_1,\chi_2) = \frac{\chi_1 - q\chi_2}{\chi_1 + q\chi_2}. \tag{13}$$

While not unique, we find this parametrization convenient because all the γ_i are defined[2] in $[-1,1]$ and the up-down endpoint is mapped to $\gamma = (0,0,0)$. We apply the transformations of Eqs. (10)–(12) to both prior and posterior samples, estimate the corresponding probability density functions using three-dimensional kernel density estimation (KDE), and evaluate the Bayes factor from Eq. (9). We use Gaussian kernels and a bandwidth of 0.2 [25].

An example of this procedure is shown in Fig. 1. We consider a synthetic source in the endpoint of the up-down instability with tilt angles $\cos\theta_1 = \cos\theta_2 = \cos\theta_{\rm UD} = 0.103$ and $\phi_{12} = 0$. The injected system has $m_1 = 49.5M_\odot$ $m_2 = 39.4M_\odot$, $\chi_1 = 0.92$, $\chi_2 = 0.94$, $D_L = 845$ Mpc, $\theta_{JN} = 0.37$, $\phi_{JL} = 5.71$, $\alpha = 6.11$, $\delta = 0.24$, $\psi = 2.28$, $t_c = -0.069$ s (in GPS time), and $\phi_c = 5.12$. The prior and posterior KDEs are evaluated at the origin of the $\{\gamma_1, \gamma_2, \gamma_3\}$ cube (black lines in Fig. 1). The Savage-Dickey estimate of the Bayes factor is $\ln \mathcal{B} = 5.11$. For equal priors, this corresponds to strong evidence that the source is indeed in the up-down endpoint. Figure 1 also shows that the posteriors of the rescaled parameters γ_i are somewhat close to a multivarate Gaussian distribution; this not a generic feature but rather a consequence of the relatively high signal-to-noise ratio (SNR), which for this specific injection is 60.

[2]The trigonometric manipulation in Eq. (12) is necessary because $\phi_{12} \in [0, 2\pi]$.

III. RESULTS

A. Comparing posteriors

Before reporting Bayes factors, it is informative to compare posterior distributions against the predictions of Eqs. (2)–(4). This a preliminary step which is often used to identify promising candidates for a model-selection analysis.

We consider six synthetic signals describing binary BHs that are in the endpoint of the up-down instability when entering the LIGO band at the reference frequency of 20 Hz. We use the same set of source parameters as in Fig. 1. In particular, we fix the detector-frame masses and inject source waveforms with SNR $= 150, 100, 80, 60, 40, 20$, corresponding to luminosity distances $D_L = 338, 508, 634, 845, 1268, 2538$ Mpc. The PN orbital separation of the binary at $f_{\rm ref} = 20$ Hz is $r_{20\,\rm Hz} \simeq 10M$, while the critical separation for the instability is $r_{\rm UD+} = 266M$. The condition $r_{\rm UD+} - r_{20\,\rm Hz} > 100M$ ensures that the predicted endpoint well describes these unstable up-down sources [18].

Our results are shown in Fig. 2, where each panel correspond to a different source. The upper subpanels compare the posterior distributions of $\cos\theta_{1,2}$ (as obtained from our parameter-estimation analysis) against that of $\cos\theta_{\rm UD}$ [as obtained from substituting the posterior samples of q,χ_1,χ_2 into Eq. (13)]. Note how the prior distribution of $\cos\theta_{\rm ud}$ peaks toward positive values, while those of $\cos\theta_{1,2}$ are flat. Close agreement between the posteriors of $\cos\theta_1$, $\cos\theta_2$, and $\cos\theta_{\rm UD}$ provide a qualitative (but not quantitative) indication that the theoretical prediction of the up-down instability is a reasonable description of the data. The lower subpanels report the posterior distribution of $\sin\phi_{12}$, where values close to zero 0 indicate a preference for the up-down hypothesis.

As expected, posteriors for the lowest SNRs tend to cover a large portion of prior range. As the SNR increases, the recovered posteriors approach the injected values that define the endpoint of the up-down instability. In particular, for the case of the highest SNR $= 150$, we find $\cos\theta_{\rm UD} = 0.122^{+0.068}_{-0.061}$ and $\phi_{12} = 0.004^{+0.460}_{-0.485}$ (where we quote the median and 90% credible interval), compared to the injected values $\cos\theta_{\rm UD} = 0.103$ and $\phi_{12} = 0$.

Note that systematic effects are not captured in both these results and the rest of the paper because we perform zero-noise runs and use the same waveform model for both injection and recovery. Waveform systematics in the specific region of parameter space where the up-down instability take place still need to be investigated.

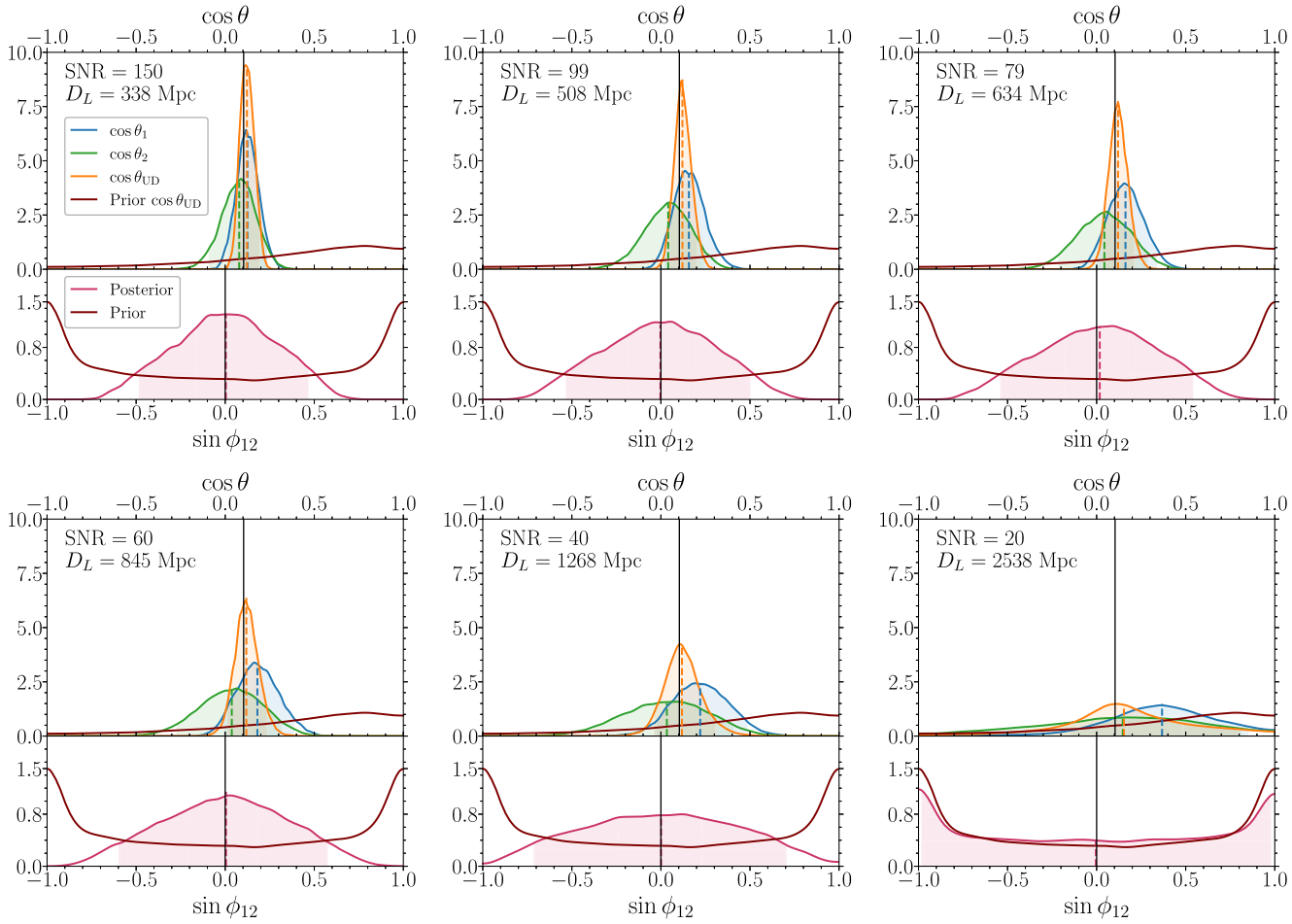

FIG. 2. Panels from left to right and from top to bottom show parameter-estimation results for the same GW source injected at decreasing (increasing) values of the SNR (luminosity distance D_L). The upper subpanels show posterior distributions of $\cos\theta_1$ (blue), $\cos\theta_2$ (green), $\cos\theta_{\rm UD}$ (orange), and the prior distribution of $\cos\theta_{\rm UD}$ (dark red); the prior distributions of $\cos\theta_{1,2}$ are flat. The lower subpanels show posterior (pink) and prior (dark red) distributions of $\sin\phi_{12}$. Black vertical lines indicate the injected values. Dashed vertical lines mark the medians of each distribution while shaded areas indicate the 90% credible intervals.

We further note a common feature that characterize all cases shown in Fig. 2, including those at low SNR. While the recovered values of $\cos\theta_{1,2}$ depart from the injected values as the SNR decreases, the medians of $\cos\theta_{\rm UD}$ tend to remain closer to that of the injected endpoint. This seems to indicate that, if the source is truly in the endpoint of the up-down instability, the estimator $\cos\theta_{\rm UD}$ might be more accurate than $\cos\theta_{1,2}$. We interpret this as a consequence of more accurate measurements of q and $\chi_{1,2}$ compared those of the spin tilts. This implies we can measure what the endpoint of a binary *would* be from the $q-\chi$ posteriors. However, inferring that the given source is in fact in its endpoint requires computation of posterior odds.

B. Model selection

While comparing posteriors as in Fig. 2 provides a useful indication of a potential up-down signature, this statement needs to be quantified with a full Bayesian model selection. For the same series of six injections, Fig. 3 shows the Bayes factor in favor of the up-down hypothesis over that of generic BH binaries computed using the Savage-Dickey density ratio (orange points). The Bayes factor increases from $\ln\mathcal{B}\sim 1.96$ for SNR $= 20$ (weak evidence) to $\ln\mathcal{B}\sim 6.89$ for SNR $=$ 150 (strong evidence). While this is a controlled experiment where the true source parameters are injected in the up-down configuration, the successful recovery of a large value of $\mathcal{B}$ indicates that data are informative about this property in a concrete measurement setting.

We repeat the same study for six additional series of BH binaries in the up-down endpoint with different parameters θ (gray points), which are part of the broader set of injections described in Sec. III D. As expected, the Bayes factor increases with the SNR in all cases, though the overall normalization depends on the other source parameters. For the case discussed above and shown with orange scatter points, a strong evidence in favor of the nested model is achieved at SNR $\gtrsim 60$—values within

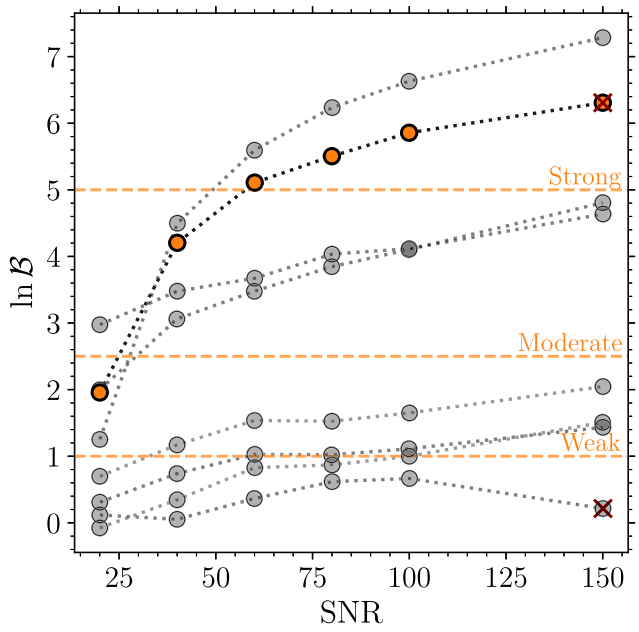

FIG. 3. Natural logarithm of the Bayes factor in favor of the up-down hypothesis as a function of the SNR. We consider the same sources as in Fig. 2 (orange scatter points) as well as six other series of BH binaries in the up-down endpoint (gray scatter points). Horizontal dashed lines indicate the threshold values of the Bayes factor in the Jeffrey scale. Crosses indicate the sources shown in Fig. 4.

reach of next LIGO-Virgo observing run [6]. However, this is not generic. We find that the distinguishability power critically depends on the source parameters. Even among this limited set, there are cases that provide only weak or even inconclusive evidence even at SNR = 150.

C. Backpropagation

We can further visualize the up-down signature of BH binaries by backpropagating posteriors samples [13,26]. If a detected source is truly an unstable up-down binary, evolving it backward in time should allow us to see it in the up-down spin configuration instead of the particular precessing configuration as observed. For a given injection, we numerically evolve each posterior sample backward from detection at $f_{\rm ref} = 20$ Hz to past-time infinity at $f_{\rm ref} = 0$ Hz using precession-averaged PN equations as implemented in Refs. [27,28]. This procedure requires q, $\chi_{1,2}$, $\theta_{1,2}$, ϕ_{12}, and r at $f_{\rm ref} = 20$ Hz as inputs and returns the values of the tilt angles θ_{12} at 0 Hz (ϕ_{12} does not enter the dynamics at infinitely large orbital separations [28,29]).

Figure 4 shows two examples which were selected from those of Fig. 3. Both sources have SNR = 150; one provides strong evidence in favor of the up-down endpoint (left panel, ln $\mathcal{B} = 6.31$) while the other returns an inconclusive result (right panel, ln $\mathcal{B} = 0.22$). The parameters of the former are listed in Sec. II C, while those of the latter are $m_1 = 26M_\odot$, $m_2 = 26M_\odot$, $\chi_1 = 0.17$, $\chi_2 = 0.57$, $\theta_{12} = 2.15$, $\phi_{12} = 0$, $D_L = 190.06$ Mpc, $\psi = 2.89$, $\phi = 3.33$, $\alpha = 3.78$, $\delta = -0.081$, $\theta_{JN} = 0.41$, $\phi_{JL} = 3.71$, and $t_c = -0.01$ s.

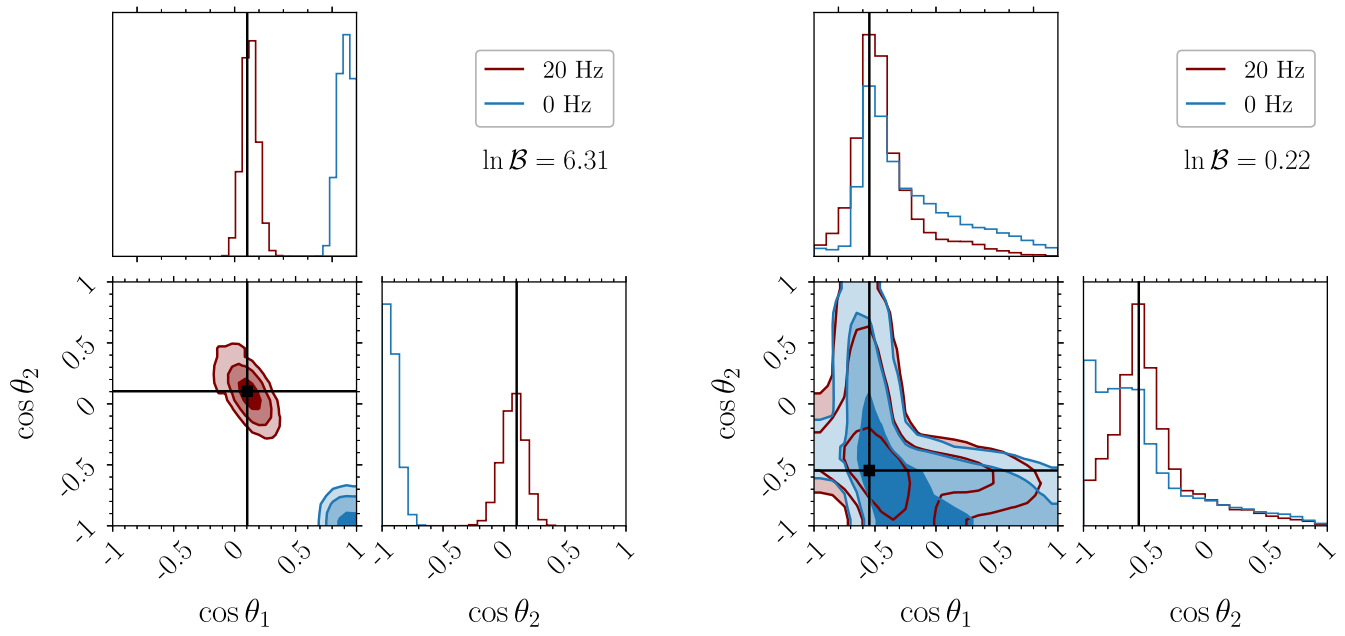

FIG. 4. Joint posterior distribution of the tilt angles θ_1 and θ_2 for the sources described in Sec. III C and marked with crosses in Fig. 3. The left (right) panel shows a case that presents strong (inconclusive) evidence in favor of the up-down hypothesis. Posterior samples are evolved numerically from $f_{\rm ref} = 20$ Hz (red) to 0 Hz (blue). Solid black lines indicate the injected values. Contour levels mark the 50%, 90%, and 99% credible regions.

For the binary with large $\mathcal{B}$ (left panel in Fig. 4), the posterior distribution at 0 Hz is constrained to be close to an aligned binary with up-down spins. In particular, we find $\cos\theta_1 > 0.80$ and $\cos\theta_2 < -0.99$ at 90% confidence. This result is an additional, visual indication that data taken at ∼20 Hz are well described by a BH binary that *used to be* aligned but is being observed precessing.

On the other hand, for the inconclusive case (right panel in Fig. 4), the joint distribution of $\cos\theta_1$ and $\cos\theta_2$ at $f_{\rm ref} = 0$ Hz occupies a much broader region of the prior volume ($\cos\theta_1 > -0.65$ and $\cos\theta_2 < 0.41$ at 90% confidence). As indicated by the Bayes factor, this is a source where data are compatible with a variety of precessing configurations, some that did and some that did not form with up-down spin directions.

D. Injection campaign

We now investigate the distinguishability of up-down sources in a wider region of the parameter space. We construct a set of injections by drawing binaries from the standard uninformative priors; we sample q and $\chi_{1,2}$ and enforce $\cos\theta_{1,2}$ and ϕ_{12} from Eqs. (2)–(4). We then impose the following constraints:

(i) We only consider binaries with $r_{\rm UD+} - r_{20\,\rm Hz} > 200M$, which is a conservative condition to ensure that the analytical instability endpoint well describes binaries that formed in the up-down configuration.

(ii) We further require sources to have SNR > 20, thus adopting a threshold that is about twice the current detection limit [1–4]. Spin effects are known to be challenging to measure [30–32], and the model-selection problem tackled here inevitably requires loud signals.

Our results are shown in Fig. 5, where we report the Bayes factor as a function of the mass ratio q, the critical separation $r_{\rm UD+}$, and the SNR. It is immediate to note that all injections have mass ratios $q \gtrsim 0.8$; this is a direct consequence of selecting binaries with a large value of $r_{\rm UD+} \propto (1-q)^{-2}$ [cf. Eq. (1)] and is largely independent of the total mass M which only enters the source-frame/detector-frame conversion of the frequency.

Among the 151 sources we select, we find that 31 present inconclusive evidence in favor of the up-down origin, 45 sources present weak evidence, 73 present moderate evidence, and 2 present strong evidence (recall that we are assuming equal model priors such that the posterior odds and the Bayes factor coincide).

We find a broad trend indicating that binaries with more unequal masses tend to have larger Bayes factors, while binaries with close-to-equal masses cover a larger range of Bayes factors. The value of q is closely correlated with

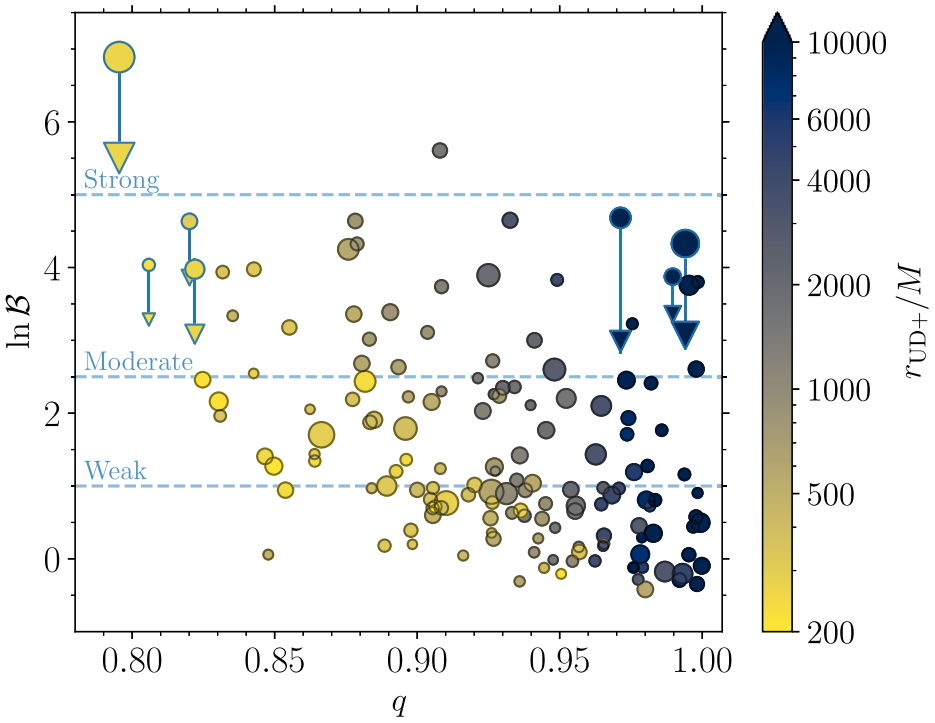

FIG. 5. Natural logarithm of the Bayes factor $\mathcal{B}$ as a function of the mass ratio q for a set of 151 GW signals injected in the endpoint of the up-down instability. The critical orbital separation $r_{\rm UD+}$ is reported on the color bar, and the size of each scatter point is directly proportional to the three-detector SNR. Horizontal dashed blue lines correspond to the threshold values of the Jeffrey scale for weak, moderate, and strong evidence. The scatter points connected by vertical lines are sources that were injected and recovered both with (upper markers, circles) and without (lower markers, triangles) higher-order modes.

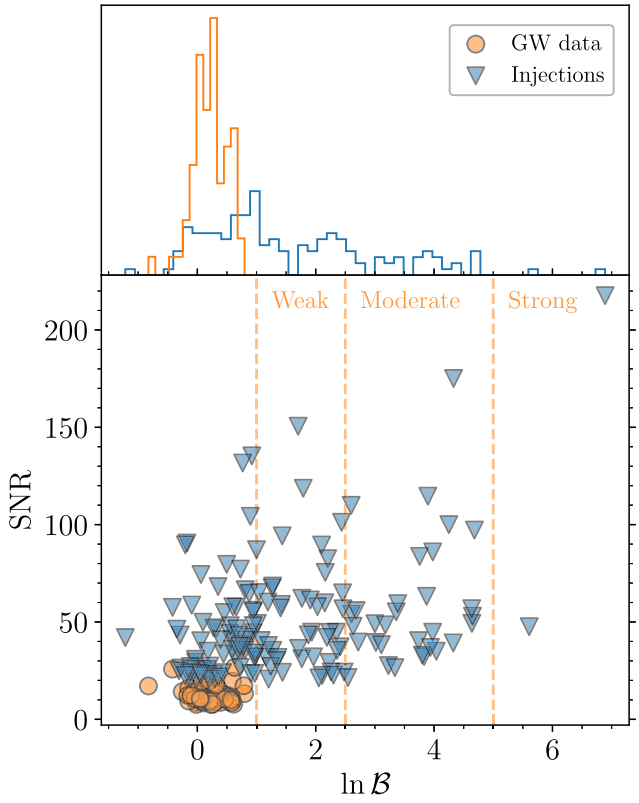

FIG. 6. Natural logarithm of the Bayes factor as a function of the SNR for 151 simulated sources (blue triangles) and 69 GW events from GWTC-3 (orange circles). Vertical dashed orange lines indicate the threshold values of the Jeffrey scale for weak, moderate, and strong evidence. The upper panel shows an histogram of the Bayes factors.

$r_{\mathrm{UD}+}$ from Eq. (1), which implies that pinpointing the up-down origin of binaries with lower values of the critical separation $r_{\mathrm{UD}+}$ is going to be somewhat easier (as long as $r_{\mathrm{UD}+}$ is still sufficiently large that the analytical endpoint provides a reasonable prediction, see above).

Figure 6 shows Bayes factors and SNRs for the same set of injections (blue triangles). As expected the two are positively correlated (cf. Fig. 3), though with a large dispersion, including several loud sources that still return an inconclusive model selection. Even SNRs as large as $\sim$200 do not guarantee a decisive model selection result since the value of $\mathcal{B}$ strongly depends on the specific parameters of the source.

A key ingredient to this analysis is the inclusion of higher-order emission modes in the adopted waveform model. Higher harmonics can break degeneracies between the mass and spin parameters [21,33–35], thus aiding our model selection problem. We further investigate this point by considering seven sources among those with the smaller and larger values of q from our set and repeat their analysis without higher-order modes. As expected, we find that the Bayes factor decreases, with differences (in logarithmic scale) that are up to $\sim$1.5.

E. Current gravitational-wave data

Finally, we apply our model-selection analysis as described in Sec. II B to current GW events reported up to GWTC-3. We analyze the 69 binary BH coalescences listed in Table I (see Sec. II A).

Figure 6 (orange circles) compares the Bayes factor and the source SNR (estimated using the median of the optimal network SNR posterior samples).

The Bayes factor in favor of the up-down hypothesis for current GW signals lies within the range $\ln \mathcal{B} \in [-0.8, 0.8]$, which is inconclusive. None of the current events support the up-down endpoint model, but they do not allows us to exclude it either. This is somewhat expected given that SNRs of current event are $\lesssim$30, which is unlikely to provide meaningful constraints (cf. Figs. 3 and 6). Our finding agrees with previous analyses [1–4] indicating that current data provide loose constraints on the orientations of individual BH spins, which in turn are key ingredients in the up-down model selection problem. We conclude that the current catalog of GW events does not contain promising up-down candidates.

At the same time, we note that the Bayes factor for the entire observed catalog $\sum_i \ln B_i \simeq 15$ shows a preference for the narrow hypothesis $\mathcal{H}_{\mathrm{N}}$. Properly quantifying the astrophysical relevance of this finding requires a deeper investigation on the systematics of the single-event B_i's as well as additional population modeling to include selection effects.

IV. CONCLUSIONS

In this paper, we performed parameter estimation of BH binaries that have encountered the up-down instability [11]. Binaries that are formed with the spin of the heavier (lighter) BH aligned (antialigned) with the orbital angular momentum might enter the LIGO/Virgo band with significant spin precession. Their final configuration (i.e., the endpoint of the up-down instability) can be computed in closed form [18] and allows us to test the up-down origin of precessing binary BHs. More ambitiously, one could also target up-down binaries as they become unstable (i.e., $r = r_{\mathrm{UD}+}$) and start precessing. While worthy of further investigation, the rate of these events is presumably very low.

We presented a statistical approach based on the Savage-Dickey density ratio for the calculation of the Bayes factor and applied it to both simulated signals (which act as a control set) and current GW events. The identification of unstable up-down binaries depends on the source SNR, with higher-order emission modes providing an important contribution. At least within the limited set of injections performed here, we find that SNRs greater than $\sim$100 are required. However, this is a necessary but not sufficient condition for the up-down origin to be distinguishable, as the resulting posterior odds strongly depends on the source parameters. Our model selection analysis is slightly more discriminative for sources

with unequal masses and, consequently, with smaller values of $r_{\mathrm{UD}+}$. Posterior samples for all the injections presented in this paper are publicly available at https://github.com/ViolaDeRenzis/updowninjections [36].

Among the current LIGO/Virgo events, we do not find promising candidates that could be interpreted as binary systems that were originally aligned in the up-down configuration. This result is not surprising, given the present SNRs which are $\lesssim 30$.

Future LIGO/Virgo upgrades as well as new facilities will largely increase the available statistical sample [5,6]. The methodology developed in this paper provides a straightforward, postprocessing operation that can be performed on posterior samples from future GW catalogs. Looking ahead, testing the up-down hypothesis is particularly relevant in the context of supermassive BH binaries observed by LISA. Some of those sources are expected to have SNRs as large as ~ 3000 [37] and their spins might be brought to the up-down configuration by interactions with galactic-scale accretion disks [17,38,39].

A future detection of the up-down instability presents the opportunity to confirm this prediction of the general-relativistic two-body problem.

ACKNOWLEDGMENTS

We thank Colm Talbot, Isobel Romero-Shaw, Chris Moore, Francesco Iacovelli, Salvatore Vitale, Neil Cornish, Sylvia Biscoveanu, Vijay Varma, and Max Isi for discussions. V. D. R., D. G., and M. M. are supported by ERC Starting Grant No. 945155–GWmining, Cariplo Foundation Grant No. 2021-0555, MUR PRIN Grant No. 2022-Z9X4XS, and the ICSC National Research Centre funded by NextGenerationEU. D. G. is supported by Leverhulme Trust Grant No. RPG-2019-350. R. B. is supported by Italian Space Agency Grant No. 2017-29-H.0. Computational work was performed at CINECA with allocations through INFN, Bicocca, and ISCRA project HP10BEQ9JB.

APPENDIX: SAVAGE-DICKEY DENSITY RATIO

Following the notation introduced in Sec. II A, let us assume that we have some observed data d and two hypotheses such that

$$\mathcal{H}_{\mathrm{N}}: \mathcal{H}_{\mathrm{B}} \wedge \gamma = \gamma_{\mathrm{N}}(\varphi). \tag{A1}$$

With this definition, the evidence of the narrow model is

$$\begin{aligned}\mathcal{Z}(d|\mathcal{H}_N) &= \int \mathcal{L}(d|\varphi,\mathcal{H}_{\mathrm{N}})\pi(\varphi|\mathcal{H}_{\mathrm{N}})\mathrm{d}\varphi \\ &= \int \mathcal{L}(d|\varphi,\gamma=\gamma_{\mathrm{N}}(\varphi),\mathcal{H}_{\mathrm{B}})\pi(\varphi|\gamma=\gamma_{\mathrm{N}}(\varphi),\mathcal{H}_{\mathrm{B}})\mathrm{d}\varphi.\end{aligned} \tag{A2}$$

One can manipulate the first term in the integrand using Bayes' theorem,

$$\mathcal{L}(d|\varphi,\gamma=\gamma_{\mathrm{N}}(\varphi),\mathcal{H}_{\mathrm{B}}) = \frac{p(\varphi,\gamma=\gamma_{\mathrm{N}}(\varphi)|d,\mathcal{H}_{\mathrm{B}})\mathcal{Z}(d|\mathcal{H}_{\mathrm{B}})}{\pi(\varphi,\gamma=\gamma_{\mathrm{N}}(\varphi)|\mathcal{H}_{\mathrm{B}})}, \tag{A3}$$

and write the Bayes factor in favor of the narrow model as

$$\begin{aligned}\mathcal{B} &= \frac{\mathcal{Z}(d|\mathcal{H}_{\mathrm{N}})}{\mathcal{Z}(d|\mathcal{H}_{\mathrm{B}})} \\ &= \int \mathrm{d}\varphi\, p(\varphi,\gamma=\gamma_{\mathrm{N}}(\varphi)|d,\mathcal{H}_{\mathrm{B}})\frac{\pi(\varphi|\gamma=\gamma_{\mathrm{N}}(\varphi),\mathcal{H}_{\mathrm{B}})}{\pi(\varphi,\gamma=\gamma_{\mathrm{N}}(\varphi)|\mathcal{H}_{\mathrm{B}})}.\end{aligned} \tag{A4}$$

The rule of conditional probability implies

$$\begin{aligned}\frac{\pi(\varphi,\gamma=\gamma_{\mathrm{N}}(\varphi)|\mathcal{H}_{\mathrm{B}})}{\pi(\varphi|\gamma=\gamma_{\mathrm{N}}(\varphi),\mathcal{H}_{\mathrm{B}})} &= \pi(\gamma=\gamma_{\mathrm{N}}(\varphi)|\mathcal{H}_{\mathrm{B}}) \\ &= \int \pi(\varphi',\gamma=\gamma_{\mathrm{N}}(\varphi)|\mathcal{H}_{\mathrm{B}})\mathrm{d}\varphi',\end{aligned} \tag{A5}$$

where in the second equality, we have explicitly indicated the marginalization over the common parameters. This yields

$$\mathcal{B} = \int \frac{p(\varphi,\gamma=\gamma_{\mathrm{N}}(\varphi)|d,\mathcal{H}_{\mathrm{B}})}{\int \pi(\varphi',\gamma=\gamma_{\mathrm{N}}(\varphi)|\mathcal{H}_{\mathrm{B}})\mathrm{d}\varphi'}\mathrm{d}\varphi, \tag{A6}$$

which is equal to Eq. (8).

The Savage-Dickey density ratio is recovered by a suitable change of variables,

$$\{\varphi,\gamma\} \to \{\bar{\varphi}=\varphi, \bar{\gamma}=\gamma-\gamma_{\mathrm{N}}(\varphi)\}. \tag{A7}$$

The determinant of the resulting Jacobian is

$$\det\begin{pmatrix}\partial\bar{\varphi}/\partial\varphi & \partial\bar{\varphi}/\partial\gamma \\ \partial\bar{\gamma}/\partial\varphi & \partial\bar{\gamma}/\partial\gamma\end{pmatrix} = \det\begin{pmatrix}1 & 0 \\ -\mathrm{d}\gamma_{\mathrm{N}}/\mathrm{d}\varphi & 1\end{pmatrix} = 1, \tag{A8}$$

such that, for any probability distribution P, one can simply write

$$P(\varphi,\gamma=\gamma_{\mathrm{N}}(\varphi)) = P(\varphi,\bar{\gamma}=0). \tag{A9}$$

With this transformation, Eq. (A6) reduces to

$$\mathcal{B} = \frac{\int p(\varphi,\bar{\gamma}=0|d,\mathcal{H}_{\mathrm{B}})\mathrm{d}\varphi}{\int \pi(\varphi',\bar{\gamma}=0|\mathcal{H}_{\mathrm{B}})\mathrm{d}\varphi'} = \frac{p(\bar{\gamma}=0|d,\mathcal{H}_{\mathrm{B}})}{\pi(\bar{\gamma}=0|\mathcal{H}_{\mathrm{B}})}, \tag{A10}$$

as reported in Eq. (9); see also Ref. [40].

[1] B. P. Abbott *et al.*, Phys. Rev. X **9**, 031040 (2019).
[2] R. Abbott *et al.*, Phys. Rev. X **11**, 021053 (2021).
[3] R. Abbott *et al.*, arXiv:2108.01045.
[4] R. Abbott *et al.*, arXiv:2111.03606.
[5] V. Baibhav, E. Berti, D. Gerosa, M. Mapelli, N. Giacobbo, Y. Bouffanais, and U. N. Di Carlo, Phys. Rev. D **100**, 064060 (2019).
[6] B. P. Abbott, R. Abbott, T. D. Abbott, S. Abraham, F. Acernese, K. Ackley, C. Adams, V. B. Adya, C. Affeldt, M. Agathos *et al.*, Living Rev. Relativity **23**, 3 (2020).
[7] T. A. Apostolatos, C. Cutler, G. J. Sussman, and K. S. Thorne, Phys. Rev. D **49**, 6274 (1994).
[8] L. E. Kidder, Phys. Rev. D **52**, 821 (1995).
[9] I. Mandel and A. Farmer, Phys. Rep. **955**, 1 (2022).
[10] M. Mapelli, in *Handbook of Gravitational Wave Astronomy* (Springer, New York, 2021), p. 16.
[11] D. Gerosa, M. Kesden, R. O'Shaughnessy, A. Klein, E. Berti, U. Sperhake, and D. Trifirò, Phys. Rev. Lett. **115**, 141102 (2015).
[12] C. O. Lousto and J. Healy, Phys. Rev. D **93**, 124074 (2016).
[13] N. K. Johnson-McDaniel, S. Kulkarni, and A. Gupta, Phys. Rev. D **106**, 023001 (2022).
[14] V. Varma, M. Mould, D. Gerosa, M. A. Scheel, L. E. Kidder, and H. P. Pfeiffer, Phys. Rev. D **103**, 064003 (2021).
[15] B. McKernan, K. E. S. Ford, R. O'Shaugnessy, and D. Wysocki, Mon. Not. R. Astron. Soc. **494**, 1203 (2020).
[16] D. Gerosa and M. Fishbach, Nat. Astron. **5**, 749 (2021).
[17] J. M. Bardeen and J. A. Petterson, Astrophys. J. Lett. **195**, L65 (1975).
[18] M. Mould and D. Gerosa, Phys. Rev. D **101**, 124037 (2020).
[19] G. Ashton *et al.*, Astrophys. J. Suppl. Ser. **241**, 27 (2019).
[20] R. J. E. Smith, G. Ashton, A. Vajpeyi, and C. Talbot, Mon. Not. R. Astron. Soc. **498**, 4492 (2020).
[21] G. Pratten, C. García-Quirós, M. Colleoni, A. Ramos-Buades, H. Estellés, M. Mateu-Lucena, R. Jaume, M. Haney, D. Keitel, J. E. Thompson, and S. Husa, Phys. Rev. D **103**, 104056 (2021).
[22] J. S. Speagle, Mon. Not. R. Astron. Soc. **493**, 3132 (2020).
[23] H. Jeffreys and R. B. Lindsay, Phys. Today **16**, No. 3, 68 (1963).
[24] W. D. Penny and G. R. Ridgway, PLoS One **8**, e59655 (2013).
[25] L. Kelley, J. Open Source Software **6**, 2784 (2021).
[26] M. Mould and D. Gerosa, Phys. Rev. D **105**, 024076 (2022).
[27] D. Gerosa and M. Kesden, Phys. Rev. D **93**, 124066 (2016).
[28] D. Gerosa, G. Fumagalli, M. Mould, G. Cavallotto, D. Padilla Monroy, D. Gangardt, and V. De Renzis, arXiv:2304.04801.
[29] D. Gerosa, M. Kesden, U. Sperhake, E. Berti, and R. O'Shaughnessy, Phys. Rev. D **92**, 064016 (2015).
[30] S. Vitale, R. Lynch, J. Veitch, V. Raymond, and R. Sturani, Phys. Rev. Lett. **112**, 251101 (2014).
[31] M. Pürrer, M. Hannam, and F. Ohme, Phys. Rev. D **93**, 084042 (2016).
[32] V. De Renzis, D. Gerosa, G. Pratten, P. Schmidt, and M. Mould, Phys. Rev. D **106**, 084040 (2022).
[33] E. Payne, C. Talbot, and E. Thrane, Phys. Rev. D **100**, 123017 (2019).
[34] C. Mills and S. Fairhurst, Phys. Rev. D **103**, 024042 (2021).
[35] R. Cotesta, A. Buonanno, A. Bohé, A. Taracchini, I. Hinder, and S. Ossokine, Phys. Rev. D **98**, 084028 (2018).
[36] V. De Renzis and D. Gerosa, https://github.com/ViolaDeRenzis/updowninjections.
[37] P. Amaro-Seoane *et al.*, arXiv:1702.00786.
[38] M. C. Miller and J. H. Krolik, Astrophys. J. **774**, 43 (2013).
[39] N. Steinle and D. Gerosa, Mon. Not. R. Astron. Soc. **519**, 5031 (2023).
[40] K. Chatziioannou, N. Cornish, A. Klein, and N. Yunes, Phys. Rev. D **89**, 104023 (2014).