

Testing general relativity using Bayesian model selection: Applications to observations of gravitational waves from compact binary systems

Walter Del Pozzo,^{1,2,*} John Veitch,^{1,3} and Alberto Vecchio¹

¹*School of Physics and Astronomy, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom*

²*Nikhef National Institute for Subatomic Physics, Science Park 105, 1098 XG Amsterdam, The Netherlands*

³*Physics and Astronomy, Cardiff University, Queen's Buildings, The Parade, Cardiff CF24 3AA, United Kingdom*

(Received 27 November 2010; published 29 April 2011)

Second-generation interferometric gravitational-wave detectors, such as Advanced LIGO and Advanced Virgo, are expected to begin operation by 2015. Such instruments plan to reach sensitivities that will offer the unique possibility to test general relativity in the dynamical, strong-field regime and investigate departures from its predictions, in particular, using the signal from coalescing binary systems. We introduce a statistical framework based on Bayesian model selection in which the Bayes factor between two competing hypotheses measures which theory is favored by the data. Probability density functions of the model parameters are then used to quantify the inference on individual parameters. We also develop a method to combine the information coming from multiple independent observations of gravitational waves, and show how much stronger inference could be. As an introduction and illustration of this framework—and a practical numerical implementation through the Monte Carlo integration technique of nested sampling—we apply it to gravitational waves from the inspiral phase of coalescing binary systems as predicted by general relativity and a very simple alternative theory in which the graviton has a nonzero mass. This method can (and should) be extended to more realistic and physically motivated theories.

DOI: 10.1103/PhysRevD.83.082002

PACS numbers: 04.80.Nn, 02.70.Rr, 02.70.Uu

I. INTRODUCTION

General relativity (GR) has so far passed every experimental test with flying colors. Yet, a whole range of efforts is underway to put Einstein's theory under even more intense experimental scrutiny in the coming years [1]. Amongst these, highly sensitive gravitational-wave (GW) experiments are opening new means to probe gravity in the dynamical, strong-field regime [2]. Ground-based gravitational-wave laser interferometers, LIGO [3,4] and Virgo [5], have now reached a sensitivity that could plausibly lead to the first direct detection of gravitational waves. The upgrade of these instruments to the second generation (also known as advanced configuration) is already underway; Advanced LIGO [6] and Advanced Virgo [7] are expected to start science observations by 2015, and to provide a wealth of detections from a variety of sources [8–10]. As soon as a positive detection is achieved, one can surely expect that gravitational-wave data will be used to test the predictions of general relativity; see [1,2] for recent reviews. In the long term future, very-high sensitivity laser interferometers, both space-based—such as the Laser Interferometer Space Antenna (LISA) [11] and Decigo [12]—and ground-based third-generation instruments, such as the Einstein gravitational-wave telescope (ET) [13–15], will increase our ability to test alternative theories of gravity.

Tests of general relativity through gravitational-wave observations have already been discussed in several studies [16–39]. Of particular interest are observations of the coalescence of binary systems, as they allow us to probe the dynamic, highly relativistic and strong-field regime through the precise monitoring of the amplitude and phase evolution of gravitational waves, which can be accurately predicted in general relativity and alternative theories of gravity.

The conceptual approach that most of these studies have employed to investigate the ability to test general relativity is based on the computation of the expected accuracy of the measurements of the unknown parameters that characterize the radiation, including those introduced by alternative theories of gravity; moreover, the statistical errors are estimated using the inverse of the Fisher information matrix [40]. These results are useful as approximate figures of merit for the constraints that can be expected to be derived from observations. However, they suffer from various conceptual (and practical) limitations that we address in this paper. The first is that at the conceptual level, those studies do not actually address whether observations will be able to discriminate an alternative theory of gravity from general relativity. They simply *assume* that the true theory of gravity is different from general relativity and, by computing the expected statistical errors on the unknown signal parameters (including those that encode deviations from GR), make some statements on whether or not the observations have sensitivity to the relevant parameter(s). This

*walterdp@nikhef.nl

leaves aside the issue that the computation of the variance-covariance matrix simply provides a lower bound, the Cramer-Rao bound [41,42], to the variance of the statistical errors, which is a meaningful bound on the accuracy of parameter estimation only in the limit of a high signal-to-noise ratio; see e.g. Refs. [43–46]. Second, those studies ignore what would be the consequences of (small) deviations from GR, such that detections can still be achieved using GR waveforms when the actual theory of gravity differs from GR, but parameter estimation can be affected due to the use of the “wrong” waveform model; this issue has recently been discussed in Ref. [35] and termed “bias in gravitational-wave astronomy.” In addition to this, past studies do not take into account the fact that one can take advantage of the (hopefully) many detections to provide better constraints on alternative theories of gravity by combining the observations. In fact, although each astrophysical source will be characterized by different astrophysical parameters (such as masses and spins, in the case of coalescing binaries), if there is a deviation from GR described by some “global” fundamental parameters of the new theory that are the same for every system—e.g. a graviton mass different from zero will affect the gravitational radiation dispersion velocity—then one can *combine* (see e.g. Refs. [47,48]) all the observations and obtain better constraints. Finally, there has been no actual attempt to provide a method to address these issues in an analysis that can be implemented in practice and therefore move towards using actual gravitational-wave data to place constraints on theories of gravity.

In this paper we tackle these specific issues: we introduce a conceptual and practical approach within the framework of Bayesian inference to discriminate between different theories of gravity by performing model selection, and to estimate the unknown model parameters. The approach based on Bayesian inference is particularly simple and powerful, as it provides both a statement on the relative probabilities of models (a given alternative theory of gravity versus general relativity) and on the distribution of the unknown parameters that characterize the theory, the (marginalized) posterior probability density functions (PDFs). The conceptual simplicity of Bayes’ theorem is balanced by the computationally expensive N -dimensional integral (where N , usually $\gtrsim 10$, is the total number of the unknown parameters) on which this theorem relies for the calculation of the evidence and the marginalized PDFs. Several integration techniques, such as (reversible jump) Markov-chain Monte Carlo methods [49,50], thermodynamic integration [51], and nested sampling [52], have been explored in a range of fields to tackle this computational challenge. For our analysis we use a nested sampling algorithm that some of us have developed for applications in the context of observations of coalescing binaries with ground-based instruments [53], and that has been shown to allow such N -dimensional integrals to be computed in an

efficient and relatively computationally inexpensive way [53–57].

The method that we present here is completely general and can be applied to observations of any gravitational-wave source and any alternative theory of gravity. However, for the purpose of introducing and illustrating the method, in this paper we focus specifically on the case of a “massive-graviton” (MG) theory—i.e. a theory of gravity in which the boson mediating the gravitational interaction is characterized by a rest mass m_g different from zero, and the corresponding Compton wavelength of the graviton λ_g is finite (the GR case corresponds to $m_g = 0$ and $\lambda_g \rightarrow \infty$)—and consider how one would go about testing GR against this alternative theory in a statistically rigorous way, by providing both a conceptual and practical (in the sense that it is readily applicable to gravitational-wave observations) approach to this problem. The reason for choosing a massive-graviton theory for our proof-of-concept analysis is twofold: (i) the gravitational radiation emitted by coalescing compact binaries in a massive-graviton theory takes a particularly simple form characterized by only one additional unknown parameter—the Compton wavelength of the graviton λ_g —and therefore provides an ideal proof-of-concept case to study, and (ii) several studies have already explored the feasibility of placing new limits on the graviton mass using gravitational-wave observations, and it is therefore useful to compare those expectations with actual results from a rigorous statistical analysis performed on mock data sets. However, the analysis that we show here can be applied to any other theory of gravity.

Within the context described above—and reminding the reader that we consider only a specific class of models and gravitational waveforms—we concentrate on the following issues:

- (i) Can we, in principle, discriminate among different theories of gravity using gravitational waves?
- (ii) Furthermore, are standard GR templates used in the day-to-day analysis of gravitational-wave data from ground-based interferometers able to detect non-GR signals and vice versa?
- (iii) What kind of bias on the estimates of the relevant parameters (e.g. masses) is introduced by an *a priori* choice of a theory of gravity?
- (iv) Considering an ensemble of gravitational-wave observations, can we improve our knowledge of the theory of gravity and of its relevant parameters?

The paper is organized as follows. In Sec. II we discuss the statistical method that we employ. In Sec. III we present the models we use as a test case for our study, and introduce the gravitational waveform generated by inspiralling compact binaries in general relativity, and in a massive-graviton theory. Sections IV and V contain the main results of the paper: in Sec. IV we show the results of our analysis for the observation of a single gravitational-

wave signal; in Sec. V we derive a method for combining multiple observations which are expected from advanced interferometers, and we show how effective it can be for the specific example at hand to further constrain λ_g ; finally, in Sec. VI we summarize our work.

II. METHOD

Let us consider a set of theories of gravity $\{\mathcal{H}_j\}$, including general relativity, that we want to test using observations of gravitational waves emitted during the coalescence of binary systems of compact objects, either black holes or neutron stars. Each theory makes a prediction on the gravitational waveform $h(t; \vec{\theta})$ that depends on the specific theory $\mathcal{H}_j$ and a set of unknown parameters $\vec{\theta}$. The statements that we can make on any given theory is based on a data set d (observations) and all the relevant prior information I that we hold.

Within the framework of Bayesian inference, the key quantity that one needs to compute is the posterior probability of a given theory (a “model” or hypothesis) $\mathcal{H}_j$. Applying Bayes’ theorem we obtain

$$P(\mathcal{H}_j|d, I) = \frac{P(\mathcal{H}_j|I)P(d|\mathcal{H}_j, I)}{P(d|I)}, \quad (1)$$

where $P(\mathcal{H}_j|d, I)$ is the posterior probability of the model $\mathcal{H}_j$ given the data, $P(\mathcal{H}_j|I)$ is the prior probability of hypothesis $\mathcal{H}_j$, and $P(d|\mathcal{H}_j, I)$ is the marginal likelihood or evidence for $\mathcal{H}_j$ that can be written as

$$P(d|\mathcal{H}_j, I) = \mathcal{L}(\mathcal{H}_j) = \int d\vec{\theta} p(\vec{\theta}|\mathcal{H}_j, I) p(d|\vec{\theta}, \mathcal{H}_j, I). \quad (2)$$

In the previous expression $p(\vec{\theta}|\mathcal{H}_j, I)$ is the prior probability density of the unknown parameter vector $\vec{\theta}$ within the theory $\mathcal{H}_j$, and $p(d|\vec{\theta}, \mathcal{H}_j, I)$ is the likelihood function of the observation d , assuming a given value of the parameters $\vec{\theta}$ and the theory $\mathcal{H}_j$.

If we want to compare different models—for this paper, we concentrate on general relativity versus an alternative theory of gravity—in light of the observations made, we can compute the relative posterior probabilities, known as the odds ratio,

$$O_{i,j} = \frac{P(\mathcal{H}_i|d)}{P(\mathcal{H}_j|d)} = \frac{P(\mathcal{H}_i) P(d|\mathcal{H}_i)}{P(\mathcal{H}_j) P(d|\mathcal{H}_j)} = \frac{P(\mathcal{H}_i)}{P(\mathcal{H}_j)} B_{i,j}, \quad (3)$$

where $P(\mathcal{H}_j)/P(\mathcal{H}_i)$ is the prior odds of the two hypotheses, the confidence we assign to the models before any observation, and $B_{i,j}$ is the Bayes factor. Here, we are interested in what the data can tell us about the relative probabilities of two models, and so we will not involve the prior odds any further.

In addition to computing the relative probabilities of different theories, one usually wants to make inference

on the unknown parameters, and therefore one needs to compute the joint posterior probability density function

$$p(\vec{\theta}|d, \mathcal{H}_j, I) = \frac{p(\vec{\theta}|\mathcal{H}_j, I) p(d|\vec{\theta}, \mathcal{H}_j, I)}{p(d|\mathcal{H}_j, I)}. \quad (4)$$

From the previous expression it is simple to compute the marginalized PDF on any given parameter, say θ_1 , within a given theory of gravity $\mathcal{H}_j$,

$$p(\theta_1|d, \mathcal{H}_j, I) = \int d\theta_2 \dots \int d\theta_N p(\vec{\theta}|d, \mathcal{H}_j, I). \quad (5)$$

The key quantities for Bayesian inference in Eqs. (3)–(5) can be efficiently computed using e.g. a nested sampling algorithm [52]. The basic idea of nested sampling is to use a collection of n objects, called live points, randomly sampled from the prior distributions, but subject to a constraint over the value of their likelihood. A live point $\vec{\xi}$ is a point in the multidimensional parameter space. At each iteration, the live point $\vec{\xi}^*$ having the lowest likelihood $\mathcal{L}(\vec{\xi}^*)$ is replaced with a new point $\vec{\xi}$ sampled from the prior distribution. To be accepted, the new point must obey the condition

$$\mathcal{L}(\vec{\xi}) > \mathcal{L}(\vec{\xi}^*). \quad (6)$$

The above condition ensures that regions of progressively increasing likelihood are explored, and the evidence integral, Eq. (2), is calculated using those points as the computation progresses. In this paper we use a specific implementation of this technique that we have developed for ground-based observations of coalescing binaries. The interested reader is referred to Ref. [53] for technical details.

III. MODELS

In this section we introduce and review the gravitational waveform approximations that we use, and spell out the models that we consider in the proof-of-concept analyses, whose results are presented in Secs. IV and V.

A. Gravitational waveforms

In general relativity, gravitational waves generated during the inspiral of compact binary systems are accurately modeled using the post-Newtonian approach; see e.g. Ref. [58] for a review. Here we use the standard restricted post-Newtonian approximation to the radiation, computing directly the waveform in the frequency domain by taking advantage of the stationary phase approximation. The amplitude of the radiation therefore contains the leading order Newtonian contribution, and higher order post-Newtonian terms are retained only in the phase. We further consider the expansion to post²-Newtonian order, and we assume that the compact objects have no spins. In summary, we will consider the frequency domain GW signal in the general relativity case to be

$$h_{\text{GR}}(f) = \mathcal{A} f^{-7/6} e^{i\Psi_{\text{GR}}(f)}, \quad (7)$$

where

$$\mathcal{A} = \frac{1}{\sqrt{30}\pi^{2/3}} \frac{\mathcal{M}^{5/6}}{D_L} \quad (8)$$

is the amplitude of the signal from a source at luminosity distance D_L , and the phase $\Psi_{\text{GR}}(f)$ at the post²-Newtonian order is given by [59]

$$\begin{aligned} \Psi_{\text{GR}}(f) = & 2\pi f t_c - \Phi_c + \frac{3}{128}(\pi M f)^{-5/3} \eta^{-1} \\ & \times \left[1 + \left(\frac{3715}{756} + \frac{55}{9} \eta \right) (\pi M f)^{2/3} - 16\pi(\pi M f) \right. \\ & \left. + \left(\frac{15293365}{508032} + \frac{27145}{504} \eta + \frac{3085}{72} \eta^2 \right) (\pi M f)^{4/3} \right]. \end{aligned} \quad (9)$$

In the previous expressions f is the GW frequency, and

$$M = m_1 + m_2, \quad (10)$$

$$\eta = \frac{m_1 m_2}{M^2}, \quad (11)$$

$$\mathcal{M} = \eta^{3/5} M \quad (12)$$

are the total mass, the symmetric mass ratio, and the “chirp” mass, respectively, of a binary of components masses m_1 and m_2 . Lastly, Φ_c is the constant phase at some (arbitrary) reference time t_c .

In a theory of gravity in which the propagation velocity of gravitational radiation is different from the speed of light, $v_g \neq c$, and depends on the frequency f as $(v_g/c)^2 = 1 - (c/f\lambda_g)^2$, there is a very simple imprint on the phase of the radiation, as shown by Will in Ref. [18]. This is equivalent to the effect that a graviton (the particle mediating the gravitational interactions) with nonzero mass (or a finite size graviton Compton wavelength) would induce. Following the accepted convention in the literature, we will call this a MG theory. In this paper, we will consider the result derived in Ref. [18] for the expression of the waveform emitted by a binary system in a MG theory:

$$h_{\text{MG}}(f) = \mathcal{A} f^{-7/6} e^{i\Psi_{\text{MG}}(f)}. \quad (13)$$

The amplitude is the same as for the general relativity case, Eq. (8), and the gravitational-wave phase becomes

$$\Psi_{\text{MG}}(f) = \Psi_{\text{GR}}(f) - \frac{\pi^2 D M}{\lambda_g^2 (1+z)} (\pi M f)^{-1}, \quad (14)$$

where z is the cosmological redshift of the source. It is useful to notice that the frequency dependency of the massive-graviton term is proportional to $(\pi M f)^{-1}$, which is the same as the first post-Newtonian contribution; see Eq. (9). The distance D that appears in the massive-graviton phase term (14) is given by

$$D \equiv \frac{1+z}{H_0} \int_0^z \frac{dz'}{(1+z') \sqrt{\Omega_M(1+z'^3) + \Omega_\Lambda}}, \quad (15)$$

where we have assumed a flat universe, and H_0 , Ω_M , and Ω_Λ are the Hubble parameter, and the matter and cosmological constant density parameters, today; in this paper we will use values consistent with the standard Λ CDM cosmology, $H_0 = 70 \text{ km s}^{-1} \text{ Mpc}^{-1}$, $\Omega_M = 0.3$, and $\Omega_\Lambda = 0.7$ [60]. We note that D , in general, differs from the luminosity distance D_L that enters into the GW amplitude, Eq. (8), by a factor $(1+z)^{-1}$. However, in this paper we concentrate on sources within 150 Mpc, $z < 0.04$, observed with ground-based laser interferometers with sensitivity typical of second generation. With these assumptions, D differs from D_L by less than 5% and we set the two numbers to be the same, and ignore the difference; therefore, we present the results directly as a function of λ_g . In real observations, z cannot be measured from gravitational-wave observations alone, and therefore one cannot measure independently D and λ_g : one can only refer to the combination $\propto D/[\lambda_g^2(1+z)]$ [61]. None of these assumptions have an impact on the conceptual approach that we propose in this paper. However, some of the results on the specific cases considered here—an alternative theory characterized a massive graviton—are affected by the approximation. During the actual analysis of the data, one would use the most accurate expression of the waveform available.

B. Hypotheses

In this paper we consider tests of general relativity versus a massive-graviton theory, according to the assumptions that we have discussed in the previous section. Here we explicitly state the models (or, equivalently, hypotheses) that we consider which are codified with the following notation.

- (i) $\mathcal{H}_{\text{GR}}$: The data consist of (zero mean) Gaussian and stationary noise of known spectral density plus an inspiral signal of the form described by Eqs. (7)–(9) with $\lambda_g^{-1} = 0$.
- (ii) $\mathcal{H}_{\text{MG}}$: The data consist of (zero mean) Gaussian and stationary noise of known spectral density plus an inspiral signal of the form described by Eqs. (13) and (14), but with λ_g as an additional unknown free parameter.

In our analysis, for both models $\mathcal{H}_{\text{GR}}$ and $\mathcal{H}_{\text{MG}}$ we actually compute the Bayes factors between the hypothesis that the data contain noise and signal, according to a given theory, versus the hypothesis that the data contain only noise. We call these Bayes factors $B_{\text{GR,noise}}$, $B_{\text{MG,noise}}$, for a signal vs noise within general relativity and a massive-graviton theory, respectively (see Sec. II, and, in particular, II.D of Ref. [53] for more details). The Bayes factor

between the MG and the GR models $B_{\text{MG,GR}}$ is then calculated simply as

$$B_{\text{MG,GR}} = \frac{B_{\text{MG,noise}}}{B_{\text{GR,noise}}}; \quad (16)$$

see Eqs. (2) and (3).

IV. RESULTS

In this section we present the results pertinent to the analysis of simulated gravitational-wave observations using individual detections. In Sec. V we shall then generalize our approach to the combination of multiple gravitational-wave events to produce more constraining statements on alternative theories of gravity and, in the specific case considered in this paper, a massive graviton. In this exemplifying study, we shall use the reference sensitivity of second-generation ground-based instruments, such as Advanced LIGO [62].

We begin in Sec. IV B by discussing how to distinguish between the MG and GR models using Bayesian model selection. We consider data sets in which coalescing binary waveforms are present and are generated using MG waveforms—we inject these waveforms into Gaussian and stationary noise for a fixed massive-graviton wavelength and binary system parameters—and analyze the data using both MG and GR waveforms to calculate the value of the Bayes factor $B_{\text{MG,GR}}$, Eq. (16). In Sec. IV C we then discuss the effect of analyzing the data with GR waveforms—focusing specifically on the issue of parameter estimation—in the case in which the actual correct theory of gravity is represented by a massive-graviton theory; we therefore address the issue of bias (in particular) as well as statistical errors in parameter estimation which, except for the pseudoanalytical approach in Ref. [63], cannot be addressed (and has not been addressed) solely with Fisher information matrix calculations. Finally, as the value of the graviton wavelength λ_g is already tightly bound by Solar System tests and measurements of galaxy motions in clusters of galaxies [1], we consider the case in which the correct theory of gravity is general relativity ($\lambda_g \rightarrow \infty$) and investigate how to set bounds on λ_g and how tight these constraints are expected to be. In order to do so, we inject GR signals and analyze the data sets using massive-graviton waveforms to compute bounds on λ_g from the marginalized PDF of λ_g that the analysis yields.

Before discussing the results in Secs. IV B, IV C, IV D, and V we provide details about the simulations on which our analyses are based.

A. Details of the simulations

Our results are derived from a set of simulations on synthetic data sets analyzed using a nested sampling algorithm that we have developed for ground-based observations of coalescing binaries [53]. Here we provide details

on the parameters of the simulations that are used throughout the paper.

We generate Gaussian and stationary noise in the frequency domain with noise spectral density representative of second-generation instruments [62] and superimpose (“inject,” in the jargon of the LIGO Scientific Collaboration) frequency domain gravitational-wave inspiral signals corresponding to either the general relativity model, Eqs. (7)–(9), or the massive-graviton model, Eqs. (13) and (14). Observations to gain information on the graviton dispersion relation require the use of at least three instruments at geographically separated locations to fully break the distance-source position determination; see Eq. (14). However, in order to reduce the computational costs of the simulations that are, in general, significant, we actually use data sets from a single instrument and assume the source sky location is known. In practice, we simulate the effect of a network of interferometers by constraining the sky position parameters, longitude and latitude, to vary only within a square with a width of 0.1 radians centered around the injection value.

We inject inspiral signals from binary systems with chirp mass $\mathcal{M} = 5M_\odot$ and symmetric mass ratio $\eta = 0.15$ —the individual component masses are $m_1 = 2.9M_\odot$ and $m_2 = 12.7M_\odot$ —and fixed, but random polarization, inclination, and sky location. In order to explore how the results depend on the signal-to-noise ratio (SNR), we repeat the injections considering different distances. Note that if one increases the distance to the source, the gravitational waveforms are affected in two ways: the amplitude decreases (and so does the SNR), and the contribution of the massive-graviton term to the phase increases; see Eq. (14). Unless otherwise stated, we generated MG waveforms using a value of the parameter $\lambda_g = 10^{15}$ m. The actual present lower bound on λ_g , which is inferred from Solar System experiments, is $\lambda_g \geq 2.8 \times 10^{15}$ m [18]. The purpose of this section is the investigation of our method’s behavior and its feasibility as a rigorous but practical way of performing tests of GR, and so we have ignored the existing physical bound of the graviton Compton wavelength here. The magnitude of the non-GR term in Eq. (14) is in fact $\propto \lambda_g^{-2}$; thus, to have a significant contribution from this term, we need to have small values of λ_g .

In the analysis, the prior distributions on the parameters are chosen as follows. We use prior probability distributions that are uniform on the chirp mass in the range $1 \leq \mathcal{M}/M_\odot \leq 15$, in η over the range 0.1–0.25, and in D_L over the range 1–1500 Mpc. The prior probability distributions on the polarization, and initial phase angles are flat and cover the whole range. For the inclination angle ι , the prior is proportional to $|\sin \iota|$ over the whole range. The prior on the signal coalescence time is flat within a range ± 100 msec centered on the true value. As stated above, the prior on the sky position is flat in the angles and has

a width of 0.1 radians around the actual value of the injection.

When we analyze the data considering the MG model, one of the parameters of the signal is of course λ_g . The functional form that we choose for the prior on λ_g is the scale invariant prior

$$p(\lambda_g | \mathcal{H}_{\text{MG}}) \propto \begin{cases} 1/\lambda_g & \text{if } \lambda_g^{(\min)} \leq \lambda_g \leq \lambda_g^{(\max)} \\ 0 & \text{elsewhere.} \end{cases} \quad (17)$$

We choose the lower bound of the prior range as $\lambda_g^{(\min)} = 10^{14.5}$ m and we analyze the data using different upper limits in the range $10^{15.5} \text{ m} \leq \lambda_g \leq 10^{20.5} \text{ m}$. The prior (17) is effectively noninformative on the scale of λ_g between its bounds [64], and also implies that $p(\log \lambda_g | \mathcal{H}_{\text{MG}}) = \text{const}$. In the numerical integration, we actually consider $\log \lambda_g$ as a massive-graviton parameter, as it is more easily sampled over this large of a range.

The analysis of the simulated data sets is carried out using the nested sampling algorithm that we have developed for ground-based observations of coalescing binary systems, with a total of 1 000 “live points” for each analysis. Technical details about the method are available in Ref. [53] and the software is released as part of the LALAPPS_INSPNEST program, which can be found in the LIGO Scientific Collaboration LALAPPS software distribution [65].

The stochastic nature of the noise affects the actual value of the Bayes factor that is recovered by the analysis, which varies from noise realization to noise realization for fixed values of the signal injection parameters. In several instances, in the results that we provide below we average over the noise realization; that is, we inject the same signal on many different and statistically independent noise realizations and we report (some measure of) the spread of the Bayes factors. We also note that we have chosen the tunable parameters of the nested sampling algorithm used in the analysis such that the fluctuations due to the noise dominate any fluctuation of the Bayes factor values due to the Monte Carlo nature of the technique.

B. Selecting a theory

We first consider the problem of discriminating between a massive-graviton theory and general relativity. In order to do so, we generate signals from binary systems with mass parameters $\mathcal{M} = 5M_\odot$ and $\eta = 0.15$, as we have described in the previous section. We fix the value of the Compton wavelength to $\lambda_g = 10^{15}$ m, and in order to explore the dependency of the results on the SNR, we repeat the injections using a different value of the luminosity distances, to span the range $5 \lesssim \text{SNR} \lesssim 20$. We also generate and analyze 10 data sets with independent noise realizations for fixed injection parameters. In the analyses, the prior on λ_g is given by Eq. (17); however, for each data set the analysis is repeated for six different

ranges of $p(\lambda_g | \mathcal{H}_{\text{MG}})$, with $\lambda_g^{(\min)} = 10^{14.5}$ m (which is kept constant), and $\lambda_g^{(\max)}$ is changed by an order of magnitude each time to span the range $\lambda_g^{(\max)} = 10^{15.5} \text{ m} - 10^{20.5} \text{ m}$.

The results of the analysis are summarized in Fig. 1, where we plot the Bayes factors $B_{\text{MG,GR}}$ [see Eq. (3)], as a function of the optimal SNR of the injection and the upper bound of the prior on the Compton wavelength. For each value of the prior range and the SNR, $B_{\text{MG,GR}}$ shows a range of values; in fact, we plot the median and 95% confidence interval of $\log B_{\text{MG,GR}}$ to average over the ef-

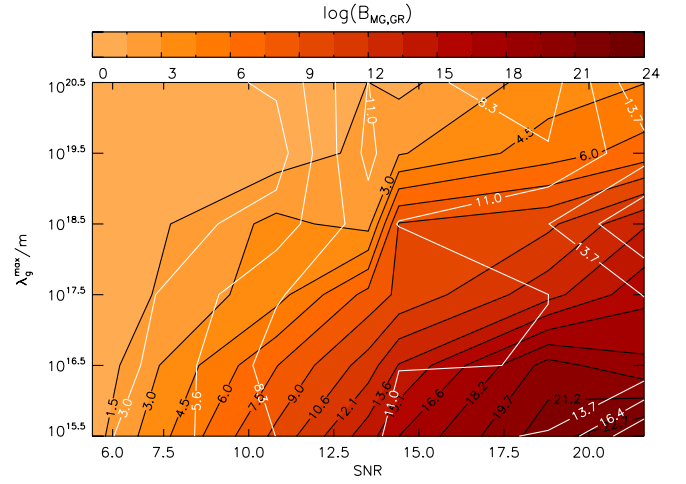


FIG. 1 (color online). The Bayes factor of the massive-graviton vs general relativity model $B_{\text{MG,GR}}$ [see Eq. (16)] obtained by analyzing data sets containing a massive-graviton waveform as a function of the SNR of the injection and the prior range on λ_g . For each value of the injection and analysis parameters, 10 data sets with independent noise realizations are generated and processed in order to quantify the variation of $B_{\text{MG,GR}}$ produced by the stochastic nature of the noise. The black contours show the median value, while the white contours show the 95% confidence interval of $\log B_{\text{MG,GR}}$. For all the injections, the mass parameters are $\mathcal{M} = 5M_\odot$, $\eta = 0.15$, and $\lambda_g = 10^{15}$ m. The location and orientation of the source in the sky were drawn randomly and kept fixed in each run, and the distance changed in order to produce different signal-to-noise ratios. The lower bound of the prior massive-graviton wavelength is kept constant in all the analyses and set to $\log_{10}(\lambda_g^{(\min)}/\text{m}) = 14.5$; as an upper bound we consider, in turn, six values (spaced by an order of magnitude), from $\log_{10}(\lambda_g^{(\max)}/\text{m}) = 15.5$ to 20.5 , which are shown on the vertical axis. The Bayes factor decreases with the increase of the prior width because the evidence, Eq. (2), is smaller when integrated over a larger volume in the parameter space. This plot shows that, even for λ_g smaller than the Solar System bound, the odds are in favor of the MG theory only for high SNR, greater than 10, and for prior widths spanning at most 2 orders of magnitude in λ_g . As we have no physical reason to prefer a small prior over a wide one, we expect that Advanced LIGO will not be sensitive enough to favor the MG hypothesis over GR.

fects of different realizations of the noise. The results indicate that, for any given prior range on λ_g , the Bayes factor $B_{\text{MG,GR}}$ is, on average, a monotonic increasing function of the SNR. At the same time, for a given SNR, $B_{\text{MG,GR}}$ is a monotonic *decreasing* function of the prior range on λ_g . To understand this last behavior, consider that the integral in Eq. (2) is performed over the whole parameter space. For a given likelihood, the effect of enlarging the prior on λ_g is then to decrease the evidence. Nevertheless,

we note that, in our simulations, $\log B_{\text{MG,GR}}$ favors the MG model for $\text{SNR} \gtrsim 15$ and for priors on λ_g that span at most 2 orders of magnitude.

A by-product of our analysis are the marginalized PDFs [see Eq. (5)], of the parameters characterizing the relevant model. Figure 2 shows the marginalized PDFs for selected parameters for a specific injection (and noise realization) with $\text{SNR} = 14.4$. For such a system the total number of wave cycles produced by the massive-graviton term in

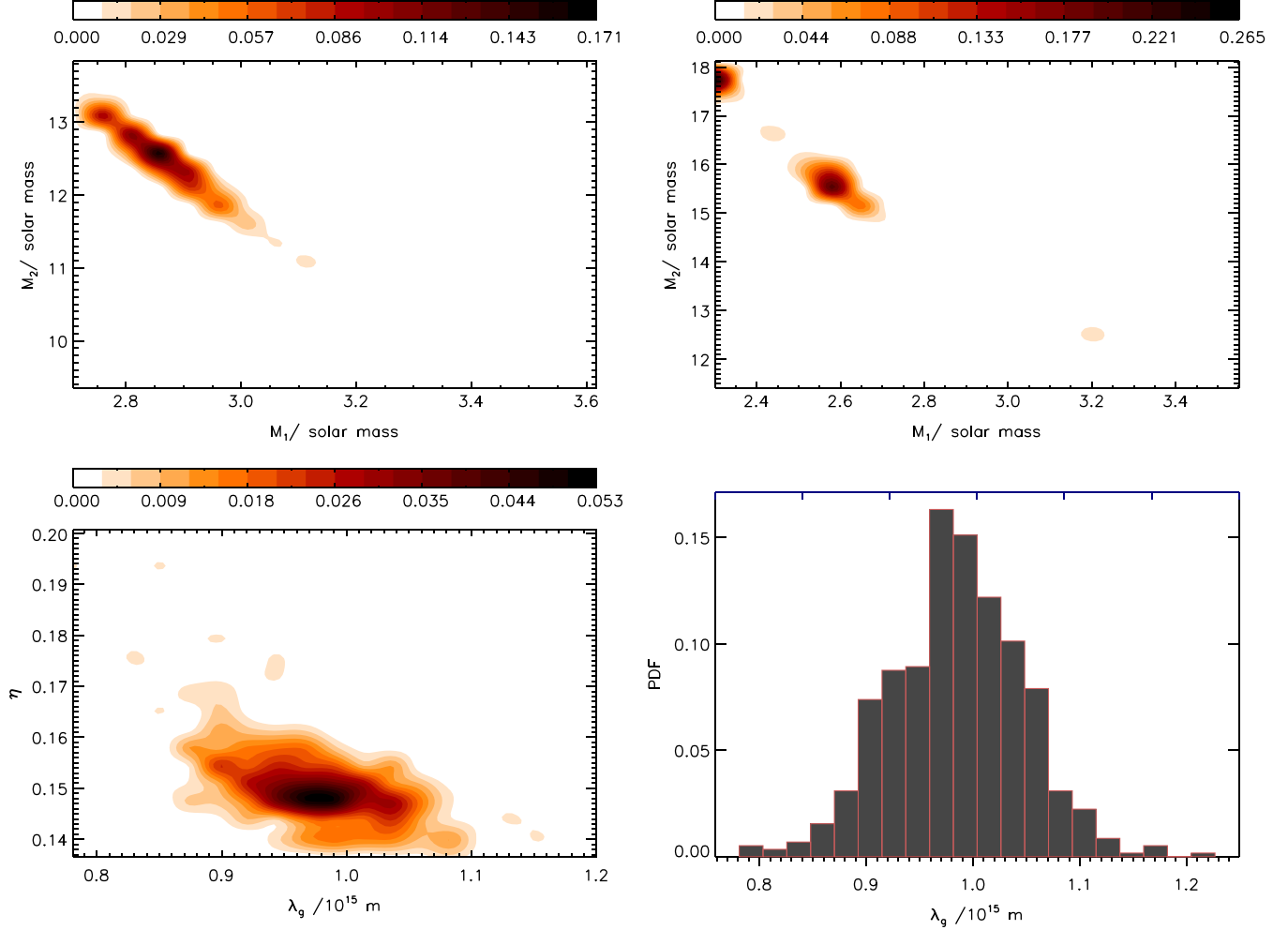


FIG. 2 (color online). Marginalized posterior density functions of selected parameters obtained by analyzing a data set in which a massive-graviton signal was injected with parameters $\mathcal{M} = 5.0M_\odot$, $\eta = 0.15$ (giving $m_1 = 2.9M_\odot$ and $m_2 = 12.7M_\odot$), and $\lambda_g = 10^{15}$ m at a distance of $D_L = 60$ Mpc. Sky position and orientation of the orbital angular momentum are randomly chosen and yield an optimal SNR of 14.4. The corresponding number of wave cycles in the sensitivity band associated with the massive-graviton term is 27. The prior on the massive graviton is flat in $\log_{10}(\lambda_g)$ and nonzero for $\log_{10}(\lambda_g/\text{m}) \in [14.5, 20.5]$. The analyses yield $\log B_{\text{MG,noise}} = 83$, $\log B_{\text{GR,noise}} = 65$, and therefore $\log B_{\text{MG,GR}} = 18$. Clearly, the signal hypothesis is preferred to the noise-only hypothesis using both GR and MG waveforms. However, the Bayes factor is clearly in favor of the MG model rather than the GR model. *Top-left panel*: The marginalized posterior PDFs of m_1 and m_2 obtained by analysis of the data with the MG model, $p(m_1, m_2|d, \mathcal{H}_{\text{MG}})$. Note that the distribution is peaked at the true value of the injection parameters. *Top-right panel*: The marginalized posterior PDF of m_1 and m_2 obtained by analysis of the data with the GR model, $p(m_1, m_2|d, \mathcal{H}_{\text{GR}})$. The posterior PDF now is offset from the true value of the parameters and shows a bias in the recovery of the parameters. *Bottom-left panel*: The marginalized posterior PDF of η and λ_g obtained by analysis of the data with the MG model, $p(\eta, \lambda_g|d, \mathcal{H}_{\text{MG}})$. *Bottom-right panel*: The marginalized posterior distribution of λ_g using the MG model, $p(\lambda_g|d, \mathcal{H}_{\text{MG}})$. When the data are analyzed with the MG model (which corresponds to the injected waveform model) the signal parameters are correctly recovered.

Eq. (14) within the sensitivity band is 27. The prior range on λ_g is $10^{14.5} \text{ m} \leq \lambda_g \leq 10^{17.5} \text{ m}$. For this specific noise realization, we obtained $\log B_{\text{MG,noise}} = 81.7$ and $\log B_{\text{GR,noise}} = 65$, giving $\log B_{\text{MG,GR}} = 18$. The signal is clearly detected using both the MG and the GR model. However, for the particular value of $\lambda_g = 10^{15} \text{ m}$ that we consider in this set of simulations, the MG model is favored over GR. Using the MG model, the parameters are correctly recovered, which is not the case for the GR model, where a significant bias is observed (see top-right panel of Fig. 2). This is an issue that has not been considered so far by other studies and will be discussed in detail in Sec. IV C.

When the width of the prior on λ_g is large enough to dominate the integral (2) and therefore the Bayes factor does not favor the MG model, the posterior density function on λ_g , *if one assumes the MG model*, shows a characteristic behavior, which is shown in Fig. 3: the analysis excludes the region of the $\log_{10}(\lambda_g)$ prior that would give rise to an appreciable effect, ruling out the lower parts of the prior distribution.

C. Parameter estimation and bias

The assumption that general relativity is the correct theory of gravity may affect parameter estimation (and therefore inference) if the actual theory is different. The impact is clearly not only restricted to the statistical

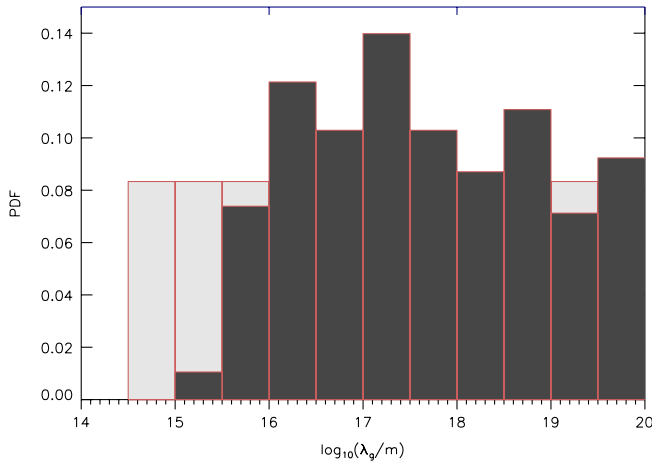


FIG. 3 (color online). The prior and marginalized posterior probability density function of $\log_{10}(\lambda_g)$ for the analysis of the same injection parameters used for Fig. 2. In this analysis, however, the prior distribution is nonzero over the interval $14.5 \leq \log_{10}(\lambda_g/\text{m}) \leq 20.5$. For this particular noise realization $\log B_{\text{MG,noise}} = 82$ while $\log B_{\text{GR,noise}} = 81$. The charcoal colored histogram is an example PDF for $\log_{10}(\lambda_g)$ when $\log B_{\text{MG,GR}}$ does not favor the MG model because of the prior width. The injected waveform has $\mathcal{M} = 5M_\odot$, $\eta = 0.15$, and $\lambda_g = 10^{15} \text{ m}$, with optimal SNR of 14.4. The light grey histogram is the prior distribution of $\log_{10}(\lambda_g)$.

errors with which the parameters are recovered but may also lead to bias. One might think that if a non-GR signal is detected with GR templates, then the effects on the estimation of the parameters is negligible. However, this is not necessarily true, as we have already shown in Fig. 2 (top-right panel): the inspiral binary signal is clearly detected ($\log B_{\text{GR,noise}} = 65$), but the mass estimates are unambiguously biased; in fact, the signal present in the data follows the massive-graviton model.

The particular aspect of bias is very important, as it introduces a conceptually different source of error—a systematic offset from the true value which does not fade away as the signal-to-noise ratio increases—that has completely been ignored so far. In fact, all the work on so-called parameter estimation in the context of other theories of gravity has been carried out so far using the Fisher information matrix formalism that *assumes* that there is no bias and deals only with statistical errors. The result that we have just highlighted in Fig. 2 shows that this is a key issue. Using the MG model as an example of deviation from general relativity, we show in this section the possible effects. Of course every theory of gravity would lead to different results; in the future, it is therefore important to study each one of them individually. For the specific case of a massive-graviton theory, we note that the phase contribution of λ_g is proportional to $(\pi M f)^{-1}$ [see Eq. (14)]: it scales as the general relativity post¹-Newtonian contribution, which is controlled by η . Therefore we expect that the parameter that will be mostly affected by processing the data with a GR waveform will be the symmetric mass ratio η (and in turn the estimate of the individual mass components). The bottom-left panel of Fig. 2 shows precisely the degeneracy (or correlation) between λ_g and η (clearly in the case in which the data are analyzed assuming the MG model).

In order to explore these effects, we perform a set of numerical experiments in which we inject MG waveforms according to Eqs. (13) and (14) at fixed SNR and mass parameters ($\mathcal{M} = 5M_\odot$, $\eta = 0.15$, and $\text{SNR} = 21$) and vary the value of the graviton Compton wavelength of the signal in the interval $10^{14.5} \text{ m} \leq \lambda_g \leq 10^{17} \text{ m}$, using the values reported in Table I. For a fixed signal-to-noise ratio (or D_L), it is useful to notice that for $\lambda_g \gtrsim 3.2 \times 10^{15} \text{ m}$ (and for the parameters of the injections) the number of wave cycles $\mathcal{N}_{\text{MG}}$ to which the massive-graviton term contributes in the detection band (i.e. for frequencies above 20 Hz, in this case) drops below 1. As usual, for each set of injection parameters we generate 10 independent noise realizations. In addition, the data are analyzed three times, using three different prior ranges for λ_g [see Eq. (17)]: the lower bound of the range is kept fixed to $\lambda_g^{(\min)} = 10^{14.5} \text{ m}$ and the upper bound is set to $\lambda_g^{(\max)} = 10^{15.5} \text{ m}$, $10^{18.5} \text{ m}$, and $10^{20.5} \text{ m}$, in turn.

We explore the dependency of the Bayes factor $\log B_{\text{MG,GR}}$ on the value of λ_g , and we investigate whether,

TABLE I. Total number of wave cycles from the massive-graviton phase term as a function of λ_g for the mass parameter values used in the injections ($\mathcal{M} = 5M_\odot$ and $\eta = 0.15$) and a low-frequency cutoff $f_{\text{low}} = 20$ Hz, to which results in Fig. 4 refer.

λ_g (m)	$\mathcal{N}_{\text{MG}}$
3.16×10^{14}	171
5.62×10^{14}	54
10^{15}	17
1.78×10^{15}	5.4
3.16×10^{15}	1.7
5.62×10^{15}	0.5
10^{16}	0.2
1.78×10^{16}	0.05
3.16×10^{16}	0.02
5.62×10^{16}	0.005
10^{17}	0.001

in analyzing the data assuming the GR model, the estimates of the unknown parameters are affected.

Figure 4 summarizes the results, as a function of λ_g . The analyses of the data yield $\log B_{\text{MG,noise}}$ and $\log B_{\text{GR,noise}}$ always greater than unity and, in general, $\gg 1$ showing that the signal hypothesis is clearly favored over the

noise hypothesis, and therefore the signal is “detected” (left panel of Fig. 4). This happens regardless of the choice of the prior on λ_g , and of the model (MG or GR). Furthermore, $\log B_{\text{MG,noise}}$ is found in the interval $200 \leq \log B_{\text{MG,noise}} \leq 250$, whereas $\log B_{\text{GR,noise}}$ increases from ≈ 70 for $\lambda_g = 3.16 \times 10^{14}$ m to the same values of $\log B_{\text{MG,noise}}$ for $\lambda_g \gtrsim 10^{15}$ m. In fact, when λ_g exceeds $\approx 10^{15}$ m (for this specific choice of masses and signal-to-noise ratio), $\log B_{\text{MG,GR}} \approx 0$ and there is no conclusive evidence from the data that the MG model should be preferred to the GR model. This behavior is consistent with what one would expect: as λ_g increases, the MG waveform looks progressively more like a GR waveform, and this is reflected in the behavior of the Bayes factors. Eventually, the Occam’s razor takes over, and naturally the simplest theory is favored (if there are no strong reasons to prefer the more complex one).

The recovered median value of η varies with the two models used to analyze the data, and within the MG model it varies with the prior chosen on λ_g . For the largest prior size, $10^{14.5} \text{ m} \leq \lambda_g \leq 10^{20.5} \text{ m}$, MG behaves similarly to GR (right panel of Fig. 4). When the size of the phase terms that are neglected in the model template is large, therefore for $\lambda_g \leq 10^{15}$ m (cf. Table I for our choice of parameters),

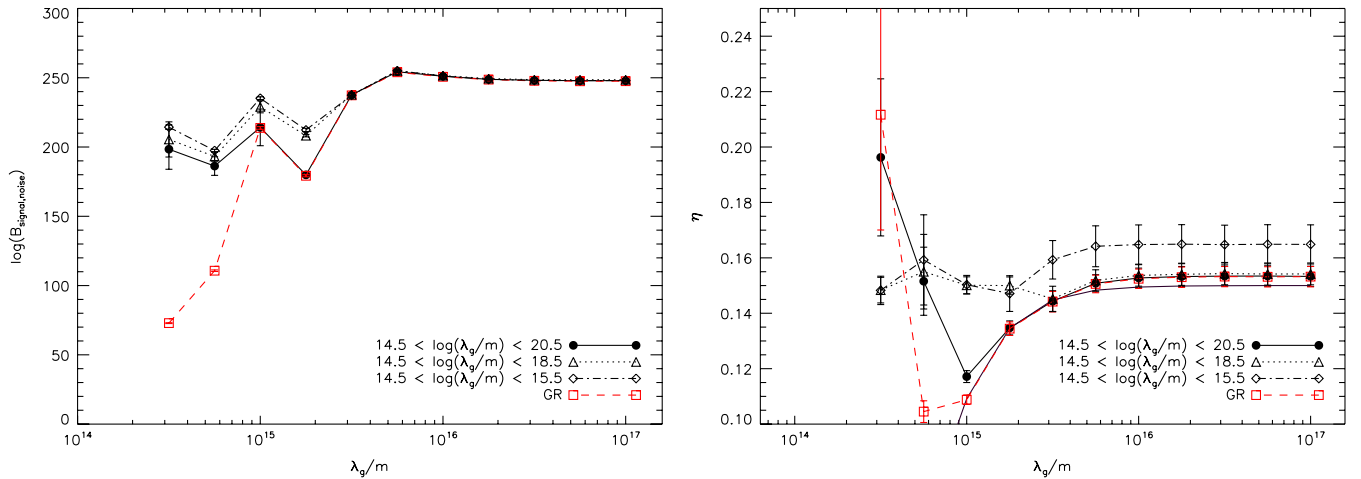


FIG. 4 (color online). *Left panel:* The Bayes factor of the signal over noise-only hypotheses assuming the MG model (black filled circles, open diamonds, and open triangles) and GR (red open squares) as a function of the value of λ_g (in meters) of the injected signal. In the MG case each symbol corresponds to a different prior range on λ_g shown in the bottom-right corner of the plots. All the injections are carried out for a signal with $\mathcal{M} = 5M_\odot$ and $\eta = 0.15$, and distance and angular parameters are selected in such a way as to produce an optimal signal-to-noise ratio of 21.6. It is clear from the fact that $\log B_{\text{MG,noise}}$ and $\log B_{\text{GR,noise}}$ are $\gg 1$ that the signals are unambiguously detected. As expected, the MG theory is (strongly) favored for small values of λ_g , but for $\lambda_g \gtrsim 10^{15}$ m (for this specific choice of masses, distance, and signal-to-noise ratio) $\log B_{\text{MG,GR}} \approx 0$; thus, the data do not provide any conclusive evidence in favor of the MG theory. *Right panel:* The median of the maximum likelihood value of the symmetric mass ratio η as recovered by the MG model and the GR model. The symbols are the same as the left panel. Each point and respective error bar are the sample mean and 95% probability intervals obtained from 10 independent nested sampling runs [66]. The results described by the solid line correspond to a value of eta in the GR model that gives a number of cycles that is as close as possible to the MG (for the signal’s injection parameters). Note that although $\log B_{\text{MG,GR}} \approx 0$ for $10^{15} \lesssim \lambda_g/m \lesssim 10^{16}$, and therefore the MG model is not favored over GR, the value of η recovered by the GR model is systematically biased to compensate for the additional phase shift due to the mass of the graviton that the GR model cannot account for properly.

the best estimate of η is severely biased (although the signal is clearly detected using just GR templates). Consider the contribution to the total number of cycles in the detector band from the massive-graviton phase term contribution $\mathcal{N}_{\text{MG}}$. When $\mathcal{N}_{\text{MG}} \lesssim 1$, the GR waveform does not differ significantly from a MG waveform, so small or no systematic effects are introduced in the parameter recovery. When $\mathcal{N}_{\text{MG}} \gtrsim 1$, the GR template has no way to distinguish the phase shift due to the massive-graviton term from the post¹-Newtonian term; see Eq. (9). Consequently, the only way in which the GR model template can match the observed MG signal is to lower the value of η to account for the phase shift. Needless to say, even if the η PDF is strongly peaked, the best estimate of η is actually quite different from the injected value (notice that $\log \mathcal{B}_{\text{GR,noise}} \gg 1$). Without the knowledge of the “true” model describing the observed data, even in the case of a successful detection, inference on the parameters may be affected. For the intermediate prior, $10^{14.5} \text{ m} \leq \lambda_g \leq 10^{18.5} \text{ m}$ (empty triangles in Fig. 4), the MG model manages to recover the correct value of the parameter η . In this case, no bias is ever introduced. When we consider the smallest prior, $10^{14.5} \text{ m} \leq \lambda_g \leq 10^{15.5} \text{ m}$ (open diamonds in Fig. 4), η is correctly estimated for $\lambda_g < 3 \times 10^{15} \text{ m}$, corresponding approximately to the upper boundary of the prior itself. When the injected value of λ_g is greater than the upper bound of the prior probability distribution function, η is instead overestimated. Our estimate is biased, but this time this is in the opposite direction compared to what we observed before.

We further illustrate the subtleties associated with the bias in Fig. 5. For this we generate a single noise realization and inject a MG waveform with $\mathcal{M} = 5M_\odot$, $\eta = 0.15$, and $\lambda_g = 10^{16} \text{ m}$. The data set is then analyzed using three different priors over the graviton Compton wavelength [same functional form as in Eq. (17) but different intervals over which the prior is nonzero]: $10^{14.5} \text{ m} \leq \lambda_g \leq 10^{15} \text{ m}$, $10^{15} \text{ m} \leq \lambda_g \leq 10^{15.5} \text{ m}$, and $10^{15.75} \text{ m} \leq \lambda_g \leq 10^{16.25} \text{ m}$ (note that only the last interval contains the value of the injected signal). In Fig. 5 we show the two-dimensional marginalized posterior PDFs on the mass parameters and the graviton Compton’s wavelength. The bias on the recovered values of $\mathcal{M}$ and η is a clear function of the distance between the injected value and the upper bound of the prior chosen to analyze the data. For $\mathcal{M}$ the recovered value differs from the real one only by a few percent. $\mathcal{M}$ is in fact mainly determined by the leading Newtonian term in Eq. (14). η and λ_g are instead anticorrelated in (14); therefore, when λ_g is forced to be smaller than the actual injection value, η is pushed toward the highest acceptable values set by its own prior. It is important to notice that in the three cases, $\log \mathcal{B}_{\text{MG,noise}}$ is essentially the same; therefore it cannot be used to prefer one choice of the prior over another. Nevertheless, it is possible to infer the inadequacy of a prior by checking for signs of railing against the prior boundary.

The simple examples presented in this section are merely indicative of a possible wider problem when gravitational-wave astronomy begins to take place.

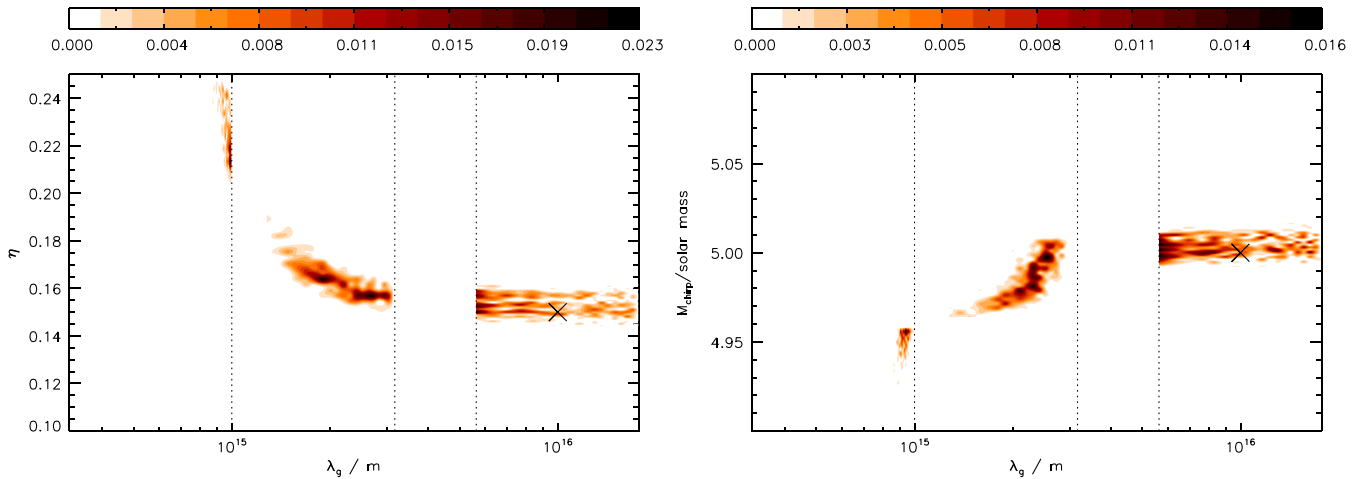


FIG. 5 (color online). The effect of the priors on the marginalized posterior density functions of selected parameters. A massive-graviton signal was injected with $\mathcal{M} = 5M_\odot$, $\eta = 0.15$, and $\lambda_g = 10^{16} \text{ m}$, producing an optimal SNR of 21.6. The same data set is analyzed three times, changing only the prior range of the graviton’s Compton wavelength (everything else being the same): $10^{14.5} \leq \lambda_g/\text{m} \leq 10^{15}$, $10^{15} \leq \lambda_g/\text{m} \leq 10^{15.5}$, and $10^{15.75} \leq \lambda_g/\text{m} \leq 10^{16.25}$. *Left panel*: The marginalized posterior PDF of η and λ_g . *Right panel*: The marginalized posterior PDF of $\mathcal{M}$ and λ_g . In both panels, the vertical dotted lines show the limits of the priors used, and the cross marks show the injected values of the parameters. The bias in parameter estimation is clearly a function of the distance between the “real” (injected) value and the upper bound of the prior on λ_g .

Severe bias in the parameter(s) recovery may be introduced by the choice of the model, cf. Fig. 4, and priors, cf. Fig. 5.

D. Bounds

The results of the previous section demonstrate that if a MG theory represents the behavior of gravity, it is unlikely that in the near future gravitational-wave observations will be able to provide sufficiently stringent observational results to tip the odds in favor of such a theory. On the other hand, even if the correct theory of gravity is characterized by a massless graviton, it is interesting to investigate what limit on λ_g one could place experimentally. This can be addressed in a straightforward way in Bayesian inference; it simply requires the evaluation of the marginalized posterior density function $p(\lambda_g|d, \mathcal{H}_{\text{MG}})$, from which one can compute a lower limit on λ_g corresponding to a given probability P :

$$\int_{\lambda_g^{(P)}}^{\infty} d\lambda_g p(\lambda_g|\vec{d}, \mathcal{H}_{\text{MG}}) = P. \quad (18)$$

In our case we decide to set (arbitrarily) $P = 0.95$ and therefore compute the 95% lower limit on the graviton Compton's wavelength that we label $\lambda_g^{95\%}$. In order to explore this point, we first assume that GR is the correct theory of gravity. We then produce 100 independent injections for sources with the same physical parameters ($\mathcal{M} = 5M_\odot$ and $\eta = 0.15$) and luminosity distance but different location/orientation in the sky (drawn uniformly on the two-sphere) so as to produce a range of optimal

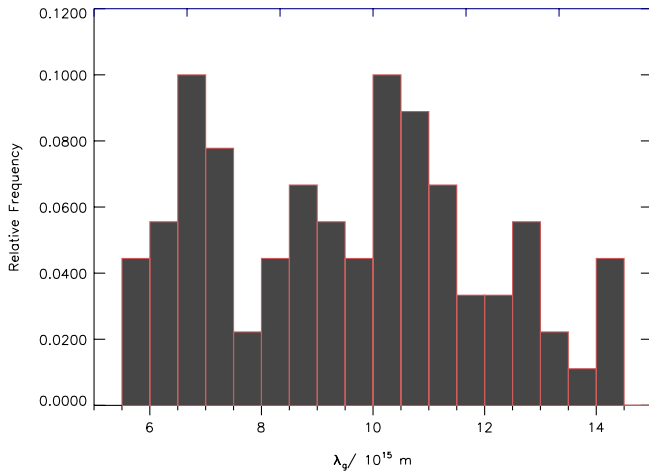


FIG. 6 (color online). The 95% lower limit on the graviton Compton wavelength $\lambda_g^{95\%}$ [see Eq. (18)] in observations with second-generation ground-based instruments of inspiral signals from binaries with chirp mass $\mathcal{M} = 5M_\odot$ and $\eta = 0.15$ modeled using general relativity waveforms. The histogram shows the relative frequency of the lower limit from 100 injections with an optimal signal-to-noise ratio in the range $5 \leq \text{SNR} \leq 25$. By comparison, the Solar System bound on the graviton Compton wavelength is $\lambda_g \geq 2.8 \times 10^{15} \text{ m}$ [18].

signal-to-noise ratios. The inspiral waveform used to generate the injection is the GR waveform described by Eqs. (7) and (9), and we analyzed the data using the massive-graviton model, Eqs. (13) and (14), with a prior $p(\lambda_g|\mathcal{H}_{\text{MG}})$ with $10^{14.5} \text{ m} \leq \lambda_g \leq 10^{20.5} \text{ m}$. From the marginalized PDF of λ_g we compute the 95% lower limit, by setting $P = 0.95$ in Eq. (18). The results are shown in Fig. 6. Advanced LIGO could therefore put a tighter constraint on the value of λ_g than the one currently inferred from the observation of the orbit of Mars. Also, Bayesian model selection allows slightly tighter bounds on λ_g compared to Fisher information matrix based studies. For values of the parameters similar to the ones we used, the typical lower limit from second-generation instruments such as Advanced LIGO is of the order of $\text{few} \times 10^{15} \text{ m}$ [18,32].

V. COMBINING MULTIPLE DETECTIONS

Beyond the present challenge of directly detecting gravitational waves for the first time, second- and third-generation instruments are expected to detect an increasing number of signals in the coming years [8–10]. Therefore, it is imperative that we exploit the information that multiple observations can bring in a statistical way. Bayes' theorem offers the possibility of combining the results from each individual observation in a conceptually simple way. In the context of testing theories of gravity, this may be particularly powerful if a given theory is characterized by some “global parameters”—e.g. the Compton wavelength of gravitons—that are independent of the actual gravitational-wave signal at hand. In this case one can construct posterior density functions that take into account all the data available and therefore strengthen the inference process. For the specific case considered in this paper, we will consider how one can set more stringent lower limits on the graviton's Compton wavelength using observations of a number of coalescing binaries, each with different parameters. Specifically, in our case we consider the example of inferring λ_g from the combined probability distribution from multiple, independent, observations.

Let us assume that we have a set of N independent observations, $d_1, \dots, d_N$, in which a gravitational-wave signal from a coalescing binary is detected (here N should not be confused with the number of dimensions of the signal's parameter vector $\vec{\theta}$ introduced in Sec. II). We want to estimate the marginal PDF of λ_g from the joint set of observations. From Bayes' theorem we can write

$$p(\lambda_g|d_1, \dots, d_N) \propto p(\lambda_g)p(d_1, \dots, d_N|\lambda_g), \quad (19)$$

and from the chain rule,

$$p(d_1, \dots, d_N|\lambda_g) = p(d_1|\lambda_g, d_2, \dots, d_N)p(d_2, \dots, d_N|\lambda_g). \quad (20)$$

Since the observations are independent, Eq. (20) simplifies

to

$$p(d_1|\lambda_g, d_2, \dots, d_N) \propto p(d_1|\lambda_g), \quad (21)$$

and, in general, we can write

$$p(\lambda_g|d_1, \dots, d_N) \propto p(\lambda_g) \prod_{i=1}^N p(d_i|\lambda_g), \quad (22)$$

where

$$p(d_i|\lambda_g) = \int d\vec{\theta} p(\vec{\theta}) p(d_i|\vec{\theta}, \lambda_g) \quad (23)$$

is the marginalized likelihood for the i th observation.

A. A pedagogical example

One can first develop some intuition about the benefits and power of this approach by considering a simple case, in which the relevant functions that enter the computation of $p(\lambda_g|d_1, \dots, d_N)$, Eq. (22), have simple analytical forms; this case is also useful to disentangle conceptual issues from practical ones related to the discrete nature of the probability density functions with which one deals in practice, and that we shall address in the next section.

The results of Sec. IV B show that in the case in which the correct theory of gravity is general relativity, the posterior PDF $p(\log_{10}(\lambda_g)|d)$ is well approximated by a sigmoid function; see Fig. 3. Let us therefore consider a set of $i = 1, \dots, N$ observations, each of which yields a posterior PDF of the form

$$p(\log_{10}(\lambda_g)|d_i) \propto \frac{1}{1 + a_i e^{-b_i \log_{10}(\lambda_g)}}, \quad (24)$$

where a_i and b_i are real numbers that differ from observation to observation. We can simulate the outcome of N observations on a range of binary systems by simply selecting the values of a_i and b_i , and then combine the results by using Eq. (22). In fact, for the specific choice of prior on λ_g that we consider here, we have

$$p(d_i|\lambda_g) \propto \frac{p(\log_{10}(\lambda_g)|d_i)}{p(\lambda_g)} = p(\log_{10}(\lambda_g)|d_i), \quad (25)$$

where, differently from everywhere else in the paper, we assumed a flat prior distribution on λ_g . As an example, we can assume that we have, say, 50 detections of coalescing binaries, each of which comes from a different source and a different SNR, that leads to a different 95% lower limit on the graviton's Compton wavelength $\lambda_g^{95\%}$, by randomly drawing values of a and b to construct the PDFs (24). Following this procedure, we have generated 50 distributions according to Eq. (24). For each of them we compute the 95% lower limit on λ_g , based on a single observation, and the same quantity using the combining set of observations, from Eq. (22). Figure 7 summarizes the results. It shows a monotonic increase in the value of the 95% lower

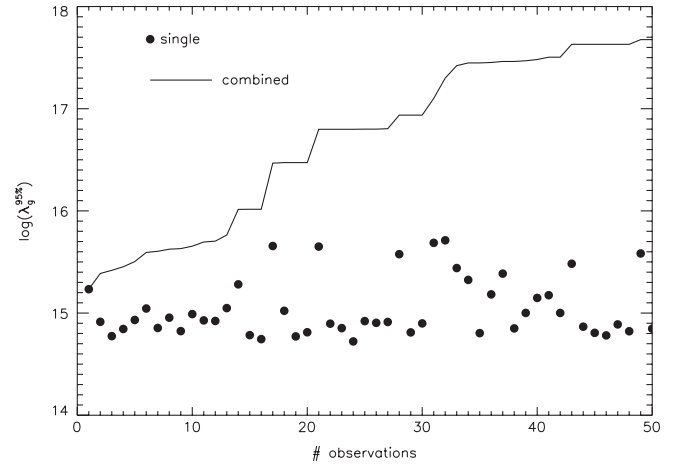


FIG. 7. The 95% lower limit on the graviton Compton wavelength $\lambda_g^{95\%}$ for an example of individual and combined (independent) observations of inspiral binaries. The plot shows $\lambda_g^{95\%}$ for each individual observation (solid circle) and for the combined set of observations (solid line), following Eq. (22), as a function of the number of observations. The posterior PDF on λ_g from which the single and combined results are obtained is assumed to be of the form (24), with the coefficients a and b randomly drawn from uniform distributions in the range $a \in [2, 10]$ and $b \in [1, 10]$.

limit calculated from the combined PDF, compared to each single 95% lower limit. It is interesting to note that how much we know about λ_g (here quantified by $\lambda_g^{95\%}$) never decreases as more information (data) is collected. If a “bad” observation is recorded (with an individual limit on λ_g significantly worse than the rest of the data), this leaves the overall 95% limit constant.

B. Combining observations in practice

In practice, we are dealing with a finite number of samples from the posterior distributions relevant to each detection, rather than with an analytical function. This introduces certain complications when producing the combined PDF on the λ_g parameter, which we have tackled using the procedure that we describe below. The technique that we propose here is applied to the specific example of the graviton's Compton wavelength but can be applied to any parameter in any theory.

To state the problem, we wish to find an appropriate approximation to the posterior PDF for each observation, which we can multiply together in order to get the combined PDF, Eq. (22), through Eq. (25). A typical approach is to categorize the samples into a series of bins, creating a histogram which approximates the underlying PDF. A histogram $\mathbf{m}$ is a set of k integers, $\mathbf{m} = (m_1, \dots, m_k)$, which register the number of samples falling into k independent categories. As a histogram approximates a probability distribution, which we know is normalized, we can approximate the probability in each bin by dividing each

count by $n = \sum_{i=1}^k m_i$. This process can introduce a large amount of noise to the distributions, as there are likely to be very few posterior samples falling into the bins of low probability density, which occurs near the region of the 95% limit (or any other large probability interval) in which we are interested. A naïve approach to combining these histograms would be to multiply the (normalized) results in each bin, but if one of the histograms contains zero samples in a bin, then all combined results produced from this distribution will also register zero in that bin no matter how many samples may appear there in the other histograms used in the combination. This problem arises because of the treatment of the normalized count in each bin, as if it were the actual probability in that bin. In fact, given a particular histogram, we can calculate the actual probability distribution for the probability in each bin, using the Dirichlet distribution. This allows us to avoid the probability in any one bin going to zero, although that may still be the most likely value given the histograms.

The probability of obtaining a given histogram of k bins from a set of n items is governed by a multinomial distribution:

$$p(\mathbf{m}; \mathbf{p}) = \frac{n!}{m_1! \cdots m_k!} \prod_{i=1}^k p_i^{m_i}, \quad (26)$$

$$n = \sum_{i=1}^k m_i. \quad (27)$$

Thus we can alternatively describe a histogram using the probabilities $\mathbf{p} = (p_1, \dots, p_k)$ associated with each bin of the histogram itself. The distribution of the probabilities entering the definition of the multinomial distribution is described by the Dirichlet distribution. In fact, the probability density for the variables $\mathbf{p} = (p_1, \dots, p_k)$ given the histogram $\mathbf{m} = (m_1, \dots, m_k)$ is

$$p(\mathbf{p}; \mathbf{m}) = \frac{1}{Z(\mathbf{m})} \prod_{i=1}^k p_i^{m_i-1}, \quad (28)$$

where $p_1, \dots, p_k \geq 0$, $\sum_{i=1}^k p_i = 1$, and $m_1, \dots, m_i > 0$. In this context, the parameters m_i can be interpreted as “prior observation counts” for events governed by p_i . Furthermore, in the limit $m_i = 0$ the Dirichlet distribution is a conjugate noninformative prior for the multinomial distribution. The normalization constant $Z(\mathbf{m})$ is given by

$$Z(\mathbf{m}) = \frac{\prod_{i=1}^k \Gamma(m_i)}{\Gamma(\sum_{i=1}^k m_i)}, \quad (29)$$

where Γ is the gamma function. We use the Dirichlet distribution to calculate the probability $p_1, \dots, p_k$ associated with each of the k bins given the current histogram $\mathbf{m}$. Thanks to this procedure, no bin is ever assigned a zero probability. Therefore, to calculate the combined PDFs,

Eq. (22), we multiply “probability histograms,” i.e. the Dirichlet distributions, describing the probabilities associated with each bin given each histogram, instead of the multinomial distributions corresponding to each histogram. The combined PDF can also be thought of as the joint probability distribution of the probability that each bin has to receive a future point, given all the previous observations.

An alternative approach to a very similar problem is presented in Ref. [48]. However, the method presented there may be unsuitable to estimate the probability density functions of the parameters in regions of the parameter space where they are close to zero. Given the typical behavior of $p(\lambda_g|d)$ (e.g. Fig. 3), these regions are critical for our goal; see Eq. (18). We stress again that the technique that we implement here does not rely on any form of approximation or smoothing of the original, discrete PDF.

We applied the method described above to a set of 50 observations of gravitational waves modeled using general relativity and added, in the same way as reported in the previous section, to Gaussian and stationary noise following the spectral density of second-generation instruments. The parameters of the injections were chosen randomly to create a distribution of signal-to-noise ratios, and we have analyzed the data using the MG model. In the analysis the prior range for $p(\lambda_g|\mathcal{H}_{\text{MG}})$ [see Eq. (17)] was $14.5 \leq \log_{10}(\lambda_g/m) \leq 20.5$. For this specific simulation we used a higher number of live points in the nested sampling algorithm than before, specifically 10^4 , in order to generate more accurate PDFs, in particular, in regions in which the PDF is close to zero. Figure 8 shows the result of the analysis, where both the 95% lower limit on λ_g from each individual observation and the combined result, ob-

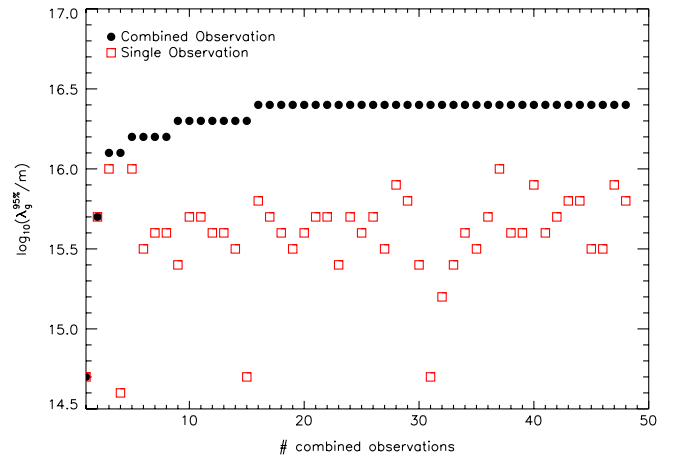


FIG. 8 (color online). The 95% lower limit on $\log_{10}(\lambda_g)$ using the detection of 50 inspiral signals. The plot shows $\lambda_g^{95\%}$ for each individual observation (red squares) and for the combined observations (black solid circles). Notice that the combined lower limit does not change much after 10 detections of approximately the same quality (i.e. same lower limit on λ_g) are combined.

tained by applying Eq. (22), are reported. The results confirm the behavior that one expects and that we have shown using the proof-of-concept example described in Sec. VA and summarized in Fig. 7. Given the discrete nature of the process, the sampling of the original PDFs might influence the results presented in Fig. 8. We tested the robustness of our combination method against this effect by studying the scatter introduced on $\lambda_g^{95\%}$ by randomly resampling the nested sampling output chains which yield the marginalized posterior PDFs. For a given single observation, we randomly resample the output chains of the nested sampling algorithm to produce 100 marginalized PDFs of λ_g . Using the procedure that we have described earlier in the section, for each of the marginalized PDFs we evaluate $\log_{10}(\lambda_g^{95\%})$, and then compute the sample mean of the single observation $\log_{10}(\lambda_g^{95\%})$'s and the spread of these values. In particular, we consider the interval over which 68% of the $\log_{10}(\lambda_g^{95\%})$'s fall; with slight abuse of terminology, we call this the “ 1σ interval.” We repeat this procedure for each of the individual detections. For a fixed order of the observations, we then draw randomly a value of $\lambda_g^{95\%}$ coming from the single observation and produce the combined 95% limit on λ_g ; we repeat the procedure 100 times to compute a sample mean and a 1σ interval in the same way as above. The results are summarized in Fig. 9.

The results of this test confirm that our procedure is robust. We do observe some variations in the combined value of $\log_{10}(\lambda_g^{95\%})$, but this is restricted to ≈ 0.1 . Moreover, the trend of the mean is consistent with that presented in both Figs. 7 and 8, and shows a clear improvement (well beyond the uncertainty region) with respect to single observations. We are therefore confident in the

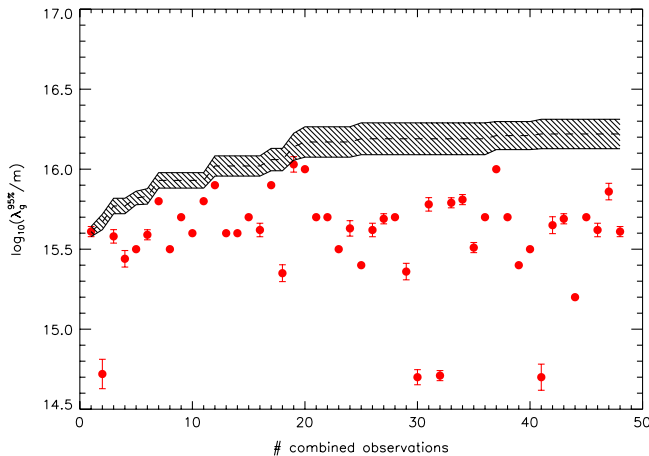


FIG. 9 (color online). The shaded region is the 1σ confidence region over the 95% lower limit on $\log_{10}(\lambda_g)$ obtained by combining 100 realizations of the single observation PDFs. The order of the observations is held fixed. The circles are the 95% lower limits on $\log_{10}(\lambda_g)$ for each single observation. The error bars are the 1σ confidence interval of each point.

histogram method of combining multiple independent observations of a constant underlying parameter; it can be applied to any experiment detecting even a small number of gravitational-wave events.

Our study already shows a significant improvement on the value of $\lambda_g^{95\%}$ after the combination of only five PDFs of λ_g , which turns out to be larger (by a factor of a few) than the value that any single experiment yields. It is therefore clear that the ability of collectively using the results from many detections shall provide a powerful tool to tighten the observational constraints on the relevant parameters and physical quantities.

VI. CONCLUSIONS

We have developed a statistical framework and an analysis approach to discriminate among theories of gravity using gravitational-wave observations of binary systems. This is a general framework and analysis pipeline that can be straightforwardly extended to any source. Our analysis strategy is based on Bayesian inference; using a nested sampling algorithm, we can compute the evidence of each theory and then the Bayes factors between them. As a by-product of the analysis, we are also able to compute marginalized posterior probability functions of each model parameter. This approach not only provides a rigorous mathematical quantification of the relative probability of two (or more) theories of gravity given prior information and observations, but also allows the exploration of correlations and bias on the process of parameter estimation in addition to purely statistical errors. We are able to quantify the bias introduced in the analysis if the assumed model does not actually correspond to the “real world” one.

In this framework, we have also introduced the combination of the results from observations of an arbitrary number of sources as a natural extension of the method that enables more powerful inference and therefore better constraints. We consider the fundamental conceptual issues, and the practical ones that derive from dealing with manipulating sampling distributions with few data points.

As a proof of principle, we have applied this analysis strategy to simulated data sets from second-generation instruments containing gravitational waves generated during the inspiral phase of nonspinning compact binary systems in general relativity and in a massive-graviton theory of gravity. The focus of this paper has been the methodology, rather than the ability of future experiments to test massive-graviton theories, due to the simplifications that we have made on the gravitational waveform. However, we have derived results that complement past studies that are solely concerned with estimating the expected statistical errors using approximations of the variance-covariance matrix. We have shown that Advanced LIGO (and other instruments of comparable sensitivity) will be capable of putting a limit on λ_g that is a few times

more stringent than current bounds. Furthermore, the combination of multiple detections can lead to significant improvements in such a bound.

Any statement regarding the value of the parameters of a model is dependent both on the data *and* on the model itself. We highlight that our ignorance of the true model produces a systematic uncertainty (bias) in the physical parameters which we infer from the data. By recognizing this, one should be encouraged to perform a more rigorous analysis, where the model as well as its parameters are scrutinized. We have demonstrated that Bayesian model selection is ideal for this task, giving us quantitative confidence in one model over another.

We can briefly summarize our findings as follows:

- (i) The calculation of the evidence suggests that second-generation instruments, in general, will not have sufficient sensitivity to favor a massive-graviton theory over general relativity for values of the Compton wavelength of the graviton larger than the existing constraint $\lambda_g \geq 2.8 \times 10^{15}$ m.
- (ii) For all values of λ_g not already excluded, see Sec. IV B, both GR and MG templates will be able to detect gravitational waves, if either of these hypotheses is assumed to describe the real signal.
- (iii) The wrong choice of a theory when analyzing the data significantly affects the estimation of the component masses (the symmetric mass ratio η)

through its sensitivity to the post¹-Newtonian order, where the MG correction enters in the expansion, Eq. (14). Nevertheless, the possibility to evaluate and compare competing models provides a tool to recognize this effect.

- (iv) The combination of multiple detections will lead to significant improvements, a factor of ~ 3 , in the lower bound of λ_g ; see Sec. V B.

This study should be regarded as the starting point for more comprehensive investigations that extend beyond the simple case of nonspinning inspiral waveforms and tackle, in a comprehensive fashion, the theories of gravity that are considered viable alternatives to Einstein's theory. It is essential that these aspects (including an end-to-end analysis pipeline to be used by the time advanced instruments are online) are fully addressed, and we plan to come back to some of them in the future.

ACKNOWLEDGMENTS

We would like to thank F. Pretorius, N. Yunes, and I. Mandel for useful discussions. W. D. P. would like to thank B. Aylott for help in setting up the code. This work has been supported by the Science and Technology Facilities Council of the United Kingdom. The numerical simulations were performed on the Tsunami cluster of the University of Birmingham.

-
- [1] C.M. Will, *Living Rev. Relativity* **9**, 3 (2006), <http://www.livingreviews.org/lrr-2006-3>.
 - [2] B.S. Sathyaprakash and B.F. Schutz, *Living Rev. Relativity* **12**, 2 (2009), <http://relativity.livingreviews.org/Articles/lrr-2009-2>.
 - [3] B.C. Barish and R. Weiss, *Phys. Today* **52**, 44 (1999).
 - [4] B. Abbott *et al.* (LIGO Scientific Collaboration), *Rep. Prog. Phys.* **72**, 076901 (2009).
 - [5] F. Acernese *et al.*, *Classical Quantum Gravity* **21**, S385 (2004).
 - [6] <http://www.ligo.caltech.edu/advLIGO/>.
 - [7] <http://www.cascina.virgo.infn.it/advirgo/>.
 - [8] C. Cutler and K.S. Thorne, in *General Relativity and Gravitation*, edited by N.T. Bishop and D.M. Sunil (World Scientific, Singapore, 2002), p. 72.
 - [9] K. Kokkotas, *Rev. Mod. Astrophys.* **20**, 140 (2008).
 - [10] J. Abadie *et al.* (LIGO Scientific Collaboration and Virgo Collaboration), *Classical Quantum Gravity* **27**, 173001 (2010).
 - [11] P. Bender *et al.*, LISA Pre-Phase A Report, Second Edition, MPQ 233 (1998).
 - [12] S. Kawamura *et al.*, *Classical Quantum Gravity* **23**, S125 (2006).
 - [13] A. Freise, S. Chelkowski, S. Hild, W. Del Pozzo, A. Perreca, and A. Vecchio, *Classical Quantum Gravity* **26**, 085012 (2009).
 - [14] S. Hild, S. Chelkowski, A. Freise, J. Franc, N. Morgado, R. Flaminio, and R. DeSalvo, *Classical Quantum Gravity* **27**, 015003 (2010).
 - [15] M. Punturo *et al.*, *Classical Quantum Gravity* **27**, 084007 (2010).
 - [16] C.M. Will, *Phys. Rev. D* **50**, 6058 (1994).
 - [17] F.D. Ryan, *Phys. Rev. D* **56**, 1845 (1997).
 - [18] C.M. Will, *Phys. Rev. D* **57**, 2061 (1998).
 - [19] C.M. Will, *Classical Quantum Gravity* **20**, S219 (2003).
 - [20] C.M. Will and N. Yunes, *Classical Quantum Gravity* **21**, 4367 (2004).
 - [21] E. Berti, A. Buonanno, and C.M. Will, *Phys. Rev. D* **71**, 084025 (2005).
 - [22] K. Glampedakis and S. Babak, *Classical Quantum Gravity* **23**, 4167 (2006).
 - [23] M. Kesden, J. Gair, and M. Kamionkowski, *Phys. Rev. D* **71**, 044015 (2005).
 - [24] S.A. Hughes, *AIP Conf. Proc.* **873**, 233 (2006).
 - [25] K.G. Arun, B.R. Iyer, M.S.S. Qusailah, and B.S. Sathyaprakash, *Phys. Rev. D* **74**, 024006 (2006).
 - [26] K.G. Arun, B.R. Iyer, M.S.S. Qusailah, and B.S. Sathyaprakash, *Classical Quantum Gravity* **23**, L37 (2006).
 - [27] E. Berti, V. Cardoso, and C.M. Will, *Phys. Rev. D* **73**,

- 064030 (2006).
- [28] L. Barack and C. Cutler, *Phys. Rev. D* **75**, 042003 (2007).
 - [29] S. Alexander, L. S. Finn, and N. Yunes, *Phys. Rev. D* **78**, 066005 (2008).
 - [30] J. R. Gair, C. Li, and I. Mandel, *Phys. Rev. D* **77**, 024035 (2008).
 - [31] N. Yunes and L. S. Finn, *J. Phys. Conf. Ser.* **154**, 012041 (2009).
 - [32] K. G. Arun and C. M. Will, *Classical Quantum Gravity* **26**, 155002 (2009).
 - [33] A. Stavridis and C. M. Will, *Phys. Rev. D* **80**, 044002 (2009).
 - [34] N. Yunes and C. F. Sopuerta, *J. Phys. Conf. Ser.* **228**, 012051 (2010).
 - [35] N. Yunes and F. Pretorius, *Phys. Rev. D* **80**, 122003 (2009).
 - [36] T. A. Apostolatos, G. Lukes-Gerakopoulos, and G. Contopoulos, *Phys. Rev. Lett.* **103**, 111101 (2009).
 - [37] B. F. Schutz, *Classical Quantum Gravity* **26**, 094020 (2009).
 - [38] D. Keppel and P. Ajith, *Phys. Rev. D* **82**, 122001 (2010).
 - [39] C. K. Mishra, K. G. Arun, B. R. Iyer, and B. S. Sathyaprakash, *Phys. Rev. D* **82**, 064010 (2010).
 - [40] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science* (Cambridge University Press, Cambridge, England, 2003).
 - [41] C. Rao, *Bull. Calcutta Math. Soc.* **37**, 81 (1945).
 - [42] H. Cramer, *Mathematical Methods of Statistics* (Pergamon Press, Princeton University Press, NJ, 1946).
 - [43] D. Nicholson and A. Vecchio, *Phys. Rev. D* **57**, 4588 (1998).
 - [44] R. Balasubramanian and S. V. Dhurandhar, *Phys. Rev. D* **57**, 3408 (1998).
 - [45] M. Vallisneri, *Phys. Rev. D* **77**, 042001 (2008).
 - [46] M. Zanolin, S. Vitale, and N. Makris, *Phys. Rev. D* **81**, 124048 (2010).
 - [47] B. F. Schutz, *Nature (London)* **323**, 310 (1986).
 - [48] I. Mandel, *Phys. Rev. D* **81**, 084029 (2010).
 - [49] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, *Markov Chain Monte Carlo in Practice* (Chapman & Hall/CRC, Boca Raton, 1996).
 - [50] P. J. Green, *Biometrika* **82**, 711 (1995).
 - [51] A. Gelman and X.-L. Meng, *Stat. Sci.* **13**, 163 (1998).
 - [52] J. Skilling, *AIP Conf. Proc.* **735**, 395 (2004).
 - [53] J. Veitch and A. Vecchio, *Phys. Rev. D* **81**, 062003 (2010).
 - [54] J. Veitch and A. Vecchio, *Phys. Rev. D* **78**, 022001 (2008).
 - [55] J. Veitch and A. Vecchio, *Classical Quantum Gravity* **25**, 184010 (2008).
 - [56] B. Aylott *et al.*, *Classical Quantum Gravity* **26**, 165008 (2009).
 - [57] B. Aylott, J. Veitch, and A. Vecchio, *Classical Quantum Gravity* **26**, 114011 (2009).
 - [58] L. Blanchet, *Living Rev. Relativity* **9**, 4 (2006), <http://relativity.livingreviews.org/Articles/lrr-2006-4>.
 - [59] L. Blanchet, T. Damour, B. R. Iyer, C. M. Will, and A. G. Wiseman, *Phys. Rev. Lett.* **74**, 3515 (1995).
 - [60] E. Komatsu *et al.*, *Astrophys. J. Suppl. Ser.* **192**, 18 (2011).
 - [61] Disentangling D from λ_g assumes that the cosmological parameters are known, as well as the source redshift. In reality, a statement on λ_g requires that one folds into the analysis the uncertainty about both the cosmology and the redshift, which, if no electromagnetic counterpart is observed, can be reconstructed from the luminosity distance and the cosmological parameters. For the purpose of this study, as we limit ourselves to distances for which $z \ll 1$, we can ignore these effects and we set $D = D_L$ and $1 + z = 1$.
 - [62] We have used the analytical expression of the Advanced LIGO noise curve as included in the LSC Analysis Library, which is given by the expression of the noise spectral density $S(f) = S_0[x^{-4.14} - 5x^2 + 111(1 - x^2 + 0.5x^4)/(1 + 0.5x^2)]$, where $S_0 = 10^{-49} \text{ Hz}^{-1}$ and $x \equiv f/215 \text{ Hz}$. Notice that a range of possible interferometer configurations will be possible with advanced LIGO, and they would yield different noise curves, which in turn differ from the one considered here. Equally, Advanced Virgo is expected to have a noise performance comparable to that of Advanced LIGO, although slightly different in the details. As the study here is not about the relative performance of different instruments, but rather about the actual method used to perform experimental tests on general relativity using observations *given* a specific instrument, the specific choice of the noise curve does not affect the qualitative behavior of the results presented here, although it does affect the quantitative details.
 - [63] C. Cutler and M. Vallisneri, *Phys. Rev. D* **76**, 104018 (2007).
 - [64] H. Jeffreys, *Theory of Probability* (Oxford University Press, New York, 1961).
 - [65] <https://www.lsc-group.phys.uwm.edu/daswg/projects/lalapps.html>.
 - [66] For each injection we perform 10 parallel runs using 10 independent nested sampling chains. We then compute the mean of the value of η recovered by each run and compute the 95% confidence interval of the mean. We do not include the statistical uncertainty on η arising from each run.