Testing general relativity with gravitational waves: A reality check

Michele Vallisneri

Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA
(Received 20 July 2012; published 1 October 2012)

The observations of gravitational-wave signals from astrophysical sources such as binary inspirals will be used to test general relativity for self-consistency and against alternative theories of gravity. I describe a simple formula that can be used to characterize the prospects of such tests, by estimating the matched-filtering signal-to-noise ratio required to detect non-general-relativistic corrections of a given magnitude. The formula is valid for sufficiently strong signals; it requires the computation of a single number, the *fitting factor* between the general-relativistic and corrected waveform families; and it can be applied to all tests that embed general relativity in a larger theory, including tests of individual theories such as Brans-Dicke gravity, as well as the phenomenological schemes that introduce corrections and extra terms in the post-Newtonian phasing expressions of inspiral waveforms. The formula suggests that the volume-limited gravitational-wave searches performed with second-generation ground-based detectors would detect alternative-gravity corrections to general-relativistic waveforms no smaller than 1%–10% (corresponding to fitting factors of 0.9 to 0.99).

I. INTRODUCTION AND MAIN RESULTS

The possibility of performing high-precision tests of general relativity (GR) in its dynamical, strong-gravity regime [1] is perhaps the most exciting prospect of the budding field of gravitational-wave (GW) astronomy [2]. Several authors have carried out detailed analyses of such tests for both ground-based and space-based GW detectors [3–25]; by and large, the tests proposed so far belong in two classes.

In the first, GR is tested against specific alternative theories, such as scalar-tensor or massive-graviton theories, which recover GR for particular value of one or more additional parameters, such as the Brans-Dicke coupling constant, or the graviton mass [3–18]. Thus, the strength of the tests is characterized by the accuracy with which the alternative-theory parameters can be measured and either found to be consistent with GR, or to deviate from it.

In the second class of tests, GR is tested for self-consistency by treating some of the coefficients in the post-Newtonian (PN) expansion of the phasing as free variables rather than deterministic functions of the source parameters, and verifying whether the recovered values are consistent with GR predictions [19–22]. The strength of these tests is characterized by the amplitude of the deviations from GR that could be discerned in the PN coefficients. More general tests are possible with the parametrized post-Einstein (ppE) formalism [23,26], which, in addition to modifying the PN coefficients, adds extra terms to the PN amplitude and phasing and to the merger and ringdown waveforms, and recovers individual alternative theories for specific forms of the extra terms.

As advocated in [24,25], GR-by-GW tests find a more satisfying formulation in Bayesian model selection [27,28], which compares the *Bayesian evidence*, given

the observed data *s*, for the alternative-theory/modified-GR scenario (henceforth "AG," for "alternative gravity") and for the Einstein-GR hypothesis. Model selection was applied to the PN consistency tests in Refs. [24,29,30], and to ppE inspiral waveforms in [25]. (For a comprehensive discussion of model selection in the context of GW detection, rather than GR tests, see also Refs. [31–34].) To wit, in model selection we compute the Bayesian *odds ratio*,

$$\mathcal{O} = \frac{P(\text{AG}|s)}{P(\text{GR}|s)} = \frac{P(\text{AG}) \int p(s|\theta^{i,a}) p(\theta^{i,a}) d\theta^{i,a}}{P(\text{GR}) \int p(s|\theta^i) p(\theta^i) d\theta^i}, \quad (1)$$

where $P(\text{AG})$ and $P(\text{GR})$ are the prior probabilities assigned to the AG and GR hypotheses; θ^i and θ^a are the source parameters (masses, spins, etc.) and additional AG parameters, respectively; $p(s|\theta)$ is the likelihood of the observed data *s* given θ; and $p(\theta)$ is the prior probability distribution for θ.[1] The odds ratio describes the degree to which we should prefer one hypothesis over the other after having observed the data, and it incorporates the Bayesian law of parsimony (also known as Occam's razor)— although models with additional parameters will always fit the data better, they will be relatively disfavored by the improbability that more parameters assume particular values in their prior ranges [27,28].

A cogent way of understanding the statistical significance of odds ratios is to set up a *decision scheme* based on the value of $\mathcal{O}$ [30,31]. Namely, we declare that we have detected AG whenever $\mathcal{O}$ is greater than a set threshold $\mathcal{O}_{\text{thr}}$. We set $\mathcal{O}_{\text{thr}}$ by requiring a given *false-alarm rate F*: this is the fraction of observations in which the underlying signal is GR, but $\mathcal{O}$ happens to pass the threshold. *F* gets

---

[1]In this paper we forgo annotating probabilities with the customary conditional dependence on "all other" assumptions, usually denoted as *I*.

smaller the more averse we are to falsely claiming AG detection, and its choice in practice should be guided by the prior $P(\text{AG})$. Now, for a given $\mathcal{O}_{\text{thr}}$, the *efficiency E* of detection is the fraction of observations in which the underlying signal is AG, and $\mathcal{O}$ passes the threshold, so AG is detected correctly.[2] A way of understanding the strength of a test of GR is then to choose a reasonably low F (say, 10^{-4}) and ask how strong an AG effect and how loud a GW signal we would need to detect AG with reasonably high E (say, $1/2$, but it turns out in practice that E rises sharply after that).

In Ref. [25], Cornish and colleagues point out that the odds ratio for AG over GR grows with the signal-to-noise ratio (henceforth, SNR) of the *residual* obtained after the best-fit GR waveform has been subtracted from the data; thus, alternative models that are not fit well by varying the GR parameters can be detected more easily than models that are. Indeed, Cornish and colleagues show that in the limit of large signal SNR and small AG deviations the logarithm of the odds ratio scales as $(1 - \text{FF})\text{SNR}^2$, with FF the *fitting factor* [36] between the GR and AG waveforms:

$$\text{FF}(\theta_{\text{AG}}) = \max_{\theta_{\text{GR}}} \frac{(h_{\text{GR}}(\theta_{\text{GR}}), h_{\text{AG}}(\theta_{\text{AG}}))}{|h_{\text{GR}}(\theta_{\text{GR}})||h_{\text{AG}}(\theta_{\text{AG}})|}. \quad (2)$$

Here $h_{\text{GR}}(\theta_{\text{GR}})$ and $h_{\text{AG}}(\theta_{\text{AG}})$ are the GR and AG waveform families (so $\theta_{\text{GR}} \equiv \theta^i$ and $\theta_{\text{AG}} \equiv \theta^{i,a}$), and $(\cdot, \cdot)$ is the standard noise-weighted inner product, such that the sampling probability of a Gaussian-noise realization n is $\propto e^{-(n,n)/2}$, and the optimal matched-filtering SNR of an observed signal h is its norm $|h| \equiv (h, h)^{1/2}$ (see, e.g., [37]). In the FF, the parameters θ_{AG} are fixed by the AG waveform contained in the data, and the inner product is maximized over θ_{GR}. The FF is by definition independent of SNR, and it tends to one when the AG corrections vanish or can be completely reabsorbed by varying θ_{GR}.

In this paper, I formalize and generalize this scaling statement by deriving the full expression of the odds ratio for the AG and GR hypotheses, in the limit of large SNR; the result is valid when AG embeds GR, which is the case for all classes of tests discussed above[3] (see Sec. II). Moreover, I derive the decision-scheme statistics for the resulting $\mathcal{O}$, and show that the efficiency $E(F)$ is a remarkably simple function [Eq. (19), a combination of the error

---

[2]The performance of decision schemes is characterized by their *receiver operating characteristic $E(F)$* [35]. Note that the term "fraction," used above in defining F and E, is ideally the fraction of an infinite number of observations of the same GW signal immersed in different realizations of noise. This characterization of decision schemes is therefore a *frequentist* statement (about the Bayesian statistic $\mathcal{O}$), but one that this Bayesian author finds very reasonable.

[3]I thank Curt Cutler for pointing out that this is true also for the PN-coefficient tests.
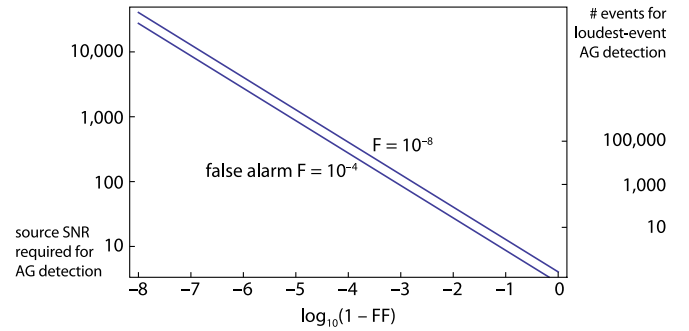

FIG. 1 (color online). SNR required for AG detection with efficiency $E = 1/2$, with false-alarm probability $F = 10^{-4}$ and 10^{-8}, as a function of FF. The right-side vertical axis shows the number of events required in a volume-limited search with detection threshold of 8 to yield a loudest event with the (median) SNR on the left-side vertical axis.

function and its inverse] of the *effective* signal-to-noise ratio $\text{SNR}\sqrt{1 - \text{FF}}$ (see Sec. III). No other information about the waveforms is needed.

Thus, AG detection by model comparison allows us to characterize very generally both kinds of tests discussed above, by computing the *SNR required to positively detect an AG correction as a function of its FF*. Given the sensitivity curve of the detector and the projected detection rates for a source class, we can then derive the magnitude of the AG corrections that we expect to be able to constrain in our observation campaigns. The FF can be computed from the GR Fisher matrix using the formulas of Ref. [38], or directly by maximizing the normalized product (2) over θ_{GR}.

The AG-detection SNR is shown in Fig. 1 for $F = 10^{-8}$–10^{-4}, and it is a rather exacting function of $1 - \text{FF}$. For the typical observations produced by volume-limited searches, which have SNRs at the event-detection threshold ($\simeq 8$), only 10% AG corrections ($1 - \text{FF} = 0.1$) would be detectable, although for most waveform families the strong-signal approximation would not be appropriate at such low SNRs [39]. The required SNR grows roughly threefold for each decade of $1 - \text{FF}$, to SNR $\gtrsim 30$ for 1% effects, SNR $\gtrsim 100$ for one-in-a-thousand effects, and SNR $\gtrsim 1,000$ for one-in-ten-thousand effects.

We can also compute easily the total volume-limited detection rates that would yield one event strong enough (on the median) to detect AG corrections with a given $1 - \text{FF}$ (see Sec. III); these are shown on the right-side vertical axis of Fig. 1. Comparison with the expected binary-inspiral detection rates for second-generation ground-based detectors [40] suggests that precise tests of GR would have to wait for the much higher rates afforded by third-generation detectors [41]. Even pooling together the evidence from all observed events [42] may not help much, reducing the number of required detections by a factor of only a few, because the evidence is dominated by the few loudest sources (see again Sec. III). By contrast, space-based observatories such as the LISA concept [43]

(or its latest incarnation, the European-led eLISA [44]) are not volume limited for some source classes, and would see some events with large SNRs.

The rest of this paper is organized as follows: in Sec. II, I derive the odds ratio in the two cases where the underlying signal is AG and GR; in Sec. III, I study the statistics of the AG decision scheme; in Sec. IV, I discuss the significance and applications of these results.

II. AG-GR ODDS RATIO IN THE HIGH-SNR LIMIT

In the following, we let θ^i be the m-dimensional vector of GR parameters, and $\theta^\mu \equiv (\theta^i, \theta^a)$ the vector of AG parameters, which augments θ^i with the single AG parameter θ^a; the derivation can be extended easily to more AG parameters. We write the true signal as $h_{AG}(\theta^\mu_{true}) = h_0 + \Delta h$, with h_0 a GR signal, and Δh the AG correction, with magnitude proportional to θ^a_{true}.

In a sufficiently small neighborhood of θ^μ_{true}, the signal can be expanded as $h_{AG}(\theta^\mu) = h_0 + \Delta h + \Delta\theta^\mu h_\mu$, with $\Delta\theta^\mu = \theta^\mu - \theta^\mu_{true}$ and $h_\mu \equiv \partial h/\partial\theta^\mu$, evaluated at h_0. If the SNR is sufficiently large, this approximation is valid throughout the region of parameter space that supports most of the likelihood [39].

We can now compute the value $P(AG|s_{AG})$ of the evidence for the AG hypothesis when the data contain an AG signal, $s_{AG} = h_{AG}(\theta^\mu_{true}) + n$. The likelihood can be written as

$$p(s_{AG}|\Delta\theta^\mu) = \mathcal{N} e^{-|s_{AG}-h(\theta^\mu)|^2/2} = \mathcal{N} e^{-|n-\Delta\theta^\mu h_\mu|^2/2}, \tag{3}$$

and it is maximized by $\Delta\theta^\mu_{ML} = (G^{-1})^{\mu\nu}(n, h_\nu)$, with $G_{\mu\nu} = (h_\mu, h_\nu)$ the $(m+1)$-dimensional AG *Fisher matrix*. Switching to parameters $\delta\theta^\mu = \Delta\theta^\mu - \Delta\theta^\mu_{ML}$ that describe displacements around the maximum, we resum the exponential as

$$p(s_{AG}|\delta\theta^\mu) = \mathcal{N} e^{-|n|^2/2+(G^{-1})^{\mu\nu}(n,h_\mu)(n,h_\nu)/2-G_{\mu\nu}\delta\theta^\mu\delta\theta^\nu/2}. \tag{4}$$

The evidence follows by integrating out the $\delta\theta^\mu$, which we do under the assumptions of flat priors $p(\theta^\mu) = 1/\Delta\theta^\mu_{prior}$ in the relevant region of parameter space, large enough to encompass the Fisher-matrix exponential:

$$P(AG|s_{AG}) = P(AG) \int p(\theta^\mu)p(s_{AG}|\delta\theta^\mu)$$
$$= P(AG)\frac{(2\pi)^{(m+1)/2}\sqrt{|G^{-1}|}}{\prod_\mu \Delta\theta^\mu_{prior}}$$
$$\times \mathcal{N} e^{-|n|^2/2+(G^{-1})^{\mu\nu}(n,h_\mu)(n,h_\nu)/2}. \tag{5}$$

This expression can be understood as the product of the maximum likelihood (the normalized exponential) with the prior $P(AG)$ and the Bayesian Occam factor (the fraction), which weighs (by volume) the region of uncertainty for the AG parameters after the observation with the region allowed by their priors. In the high-SNR limit, the posterior region of uncertainty is just the Fisher 1-σ ellipsoid, which has volume proportional to $\sqrt{|G^{-1}|}$. The second term in the exponential is the enhancement of likelihood due to overfitting noise: this is a random variable (a function of the noise realization) with expectation value[4] equal to $m + 1$.

We repeat this computation for the GR hypothesis, expanding the signal as $h_{GR}(\theta^i) = h_0 + \Delta\theta^i h_i$, with $\Delta\theta^i = \theta^i - \theta^i_{true}$, and integrating over $\delta\theta^i = \Delta\theta^i - (F^{-1})^{ij}(n + \Delta h, h_j)$, with F_{ij} the m-dimensional GR Fisher matrix. From the point of view of GR waveforms, Δh behaves as an additional noise component. Thus,

$$P(GR|s_{AG}) = P(AG)\frac{(2\pi)^{m/2}\sqrt{|F^{-1}|}}{\prod_i \Delta\theta^i_{prior}}$$
$$\times \mathcal{N} e^{-|n+\Delta h|^2/2+(F^{-1})^{ij}(n+\Delta h,h_i)(n+\Delta h,h_j)/2}, \tag{6}$$

where $F_{ij} \equiv (h_i, h_j)$ is the m-dimensional Fisher matrix.

We can now form the odds ratio $\mathcal{O}_{AG} = P(AG|s_{AG})/P(GR|s_{AG})$, using the shorthand $X_\mu \equiv (X, h_\mu)$:

$$\mathcal{O}_{AG} = \frac{p(AG)}{p(GR)}\frac{(2\pi)^{1/2}\sqrt{|G^{-1}|/|F^{-1}|}}{\Delta\theta^{AG}_{prior}} e^{[|\Delta h|^2-(F^{-1})^{ij}\Delta h_i\Delta h_j]/2+[(\Delta h,n)-(F^{-1})^{ij}\Delta h_i n_j]+[(G^{-1})^{\mu\nu}n_\mu n_\nu-(F^{-1})^{ij}n_i n_j]/2}; \tag{7}$$

this expression can be simplified considerably by noting that $(F^{-1})^{ij}h_i(h_j, \cdot)$ acts as the linear projector P_{GR} onto the local tangent space of signal derivatives taken with respect to GR parameters, so

$$|\Delta h|^2 - (F^{-1})^{ij}\Delta h_i\Delta h_j = |(1-P_{GR})\Delta h|^2,$$
$$(\Delta h, n) - (F^{-1})^{ij}\Delta h_i n_j = ((1-P_{GR})\Delta h, n); \tag{8}$$

thus it is only the component $\Delta h_\perp \equiv (1 - P_{GR})\Delta h$ of the AG correction that enters the odds ratio; this is indeed the *residual* that cannot be reabsorbed by shifting the

estimated values of the GR parameters, and the larger the $\Delta h_\perp$, the more evidence there is for the AG hypothesis.

The Occam factor and noise-overfitting contributions to the maximum likelihood also bear some simplification: using the block-matrix decomposition of $G_{\mu\nu}$ and its inverse,

---

[4]From the definition of inner product as $(a, b) = 4\,\mathrm{Re}\int a^*(f)b(f)/S_n(f)df$ and the definition of noise spectral density S_n from $\langle n^*(f)n(f')\rangle = S_n(f)\delta(f - f')/2$, it follows in general that $\langle (n, a)(n, b)\rangle_n = (a, b)$. Then $\langle (G^{-1})^{\mu\nu}(n, h_\mu) \times (n, h_\nu)\rangle = (G^{-1})^{\mu\nu}G_{\nu\mu} = I^\mu_\mu = m + 1$.

$$G_{\mu\nu} = \begin{pmatrix} F_{ij} & b_i \\ b_j & c \end{pmatrix}, \qquad (G^{-1})^{\mu\nu} = \begin{pmatrix} (F^{-1})^{ij} + (F^{-1})^{ik}b_kb_l(F^{-1})^{lj}/k & -(F^{-1})^{ik}b_k/k \\ -b_k(F^{-1})^{kj}/k & 1/k \end{pmatrix}, \qquad (9)$$

where $b_i = (h_i, h_a)$, $c = (h_a, h_a)$, and $k = c - b_ib_j(F^{-1})^{ij}$, we can show that

$$|G_{ij}| = |cF_{ij} - b_ib_j| = |F_{ij}|k,$$

$$(G^{-1})^{\mu\nu}n_\mu n_\nu - (F^{-1})^{ij}n_in_j = (\Delta h_\perp, n)^2/|\Delta h_\perp|^2, \quad (10)$$

so

$$\mathcal{O}_{AG} = \frac{p(AG)}{p(GR)}\frac{(2\pi)^{1/2}\Delta\theta_{est}^a}{\Delta\theta_{prior}^a}e^{|\Delta h_\perp|^2/2 + x|\Delta h_\perp| + x^2/2}, \quad (11)$$

where $x = (\Delta h_\perp, n)/|\Delta h_\perp|$ is a normal random variable with zero mean and unit variance (see again footnote 4), and $\Delta\theta_{est}^a = k^{-1/2}$ is the estimation error for the AG parameter, as given by the corresponding diagonal element of the inverse Fisher matrix G^{-1}. Remarkably (if logically), the odds ratio turns out to be a function of the posterior uncertainty and prior range for the additional AG parameter alone.

We can link $\Delta h_\perp$ to the fitting factor FF by finding the $\Delta\theta^i$ that maximizes the normalized match,

$$FF = \max_{\Delta\theta^i} \frac{(h_0 + \Delta h, h_0 + \Delta\theta^ih_i)}{|h_0 + \Delta h| \cdot |h_0 + \Delta\theta^ih_i|}, \quad (12)$$

which is given (unsurprisingly) by $\Delta\theta^i = (F^{-1})^{ij}(\Delta h, h_j)$, and replacing it in Eq. (12), yielding

$$1 - FF = \frac{1}{2}\frac{|\Delta h_\perp|^2}{|h_0|^2} = \frac{1}{2}\frac{|\Delta h_\perp|^2}{SNR^2}, \quad (13)$$

which is valid to $O(SNR^{-4})$. Thus, for fixed FF the odds ratio scales as SNR^2, just as it does in the Bayesian decision scheme for the (non)detection of a known signal in noise; for fixed SNR the odds ratio scales as $1 - FF$, so the odds ratio is larger with stronger and less reabsorbable AG deviations. The effects of detector noise add some statistical fluctuations through the random variable x.

This derivation can be repeated with small changes to yield the odds ratio when the data contain a GR signal, $s_{GR} = h_{GR}(\theta_{true}^i) + n$, with $h_{GR}(\theta_{true}^i) = h_0$, leading to

$$\mathcal{O}_{GR} = \frac{p(AG)}{p(GR)}\frac{(2\pi)^{1/2}\Delta\theta_{est}^a}{\Delta\theta_{prior}^a}e^{x^2/2}, \quad (14)$$

where again x is a normal random variable with zero mean and unity variance. Equations (11), (13), and (14) comprise the main novel result of this paper, and in the next section we use them to characterize the statistics of our decision scheme.

III. AG-GR DECISION SCHEME

The distribution of $\mathcal{O}_{GR}$, as implied by the distribution of x through Eq. (14), determines the *background* of false AG detections for a chosen threshold $\mathcal{O}_{thr}$, quantified by the

false-alarm probability $F = P(\mathcal{O}_{GR} > \mathcal{O}_{thr})$. We choose $\mathcal{O}_{thr}$ to yield the desired F, and evaluate the corresponding efficiency $E = P(\mathcal{O}_{AG} > \mathcal{O}_{thr})$ from Eq. (11). Surprisingly, because the ratios of priors $P(AG)/P(GR)$ and the Occam factors are the same in $\mathcal{O}_{GR}$ and $\mathcal{O}_{AG}$, their only effect is to rescale $\mathcal{O}_{thr}$, and they cancel out when we compute E as a function of F. We can then work with the renormalized odds ratios

$$\mathcal{O}'_{GR} = e^{x^2/2}, \qquad \mathcal{O}'_{AG} = e^{x^2/2 + \sqrt{2}xSNR_{AG} + SNR_{AG}^2}, \quad (15)$$

where $SNR_{AG} \equiv SNR\sqrt{1 - FF}$ plays the role of an *effective* SNR for AG detection.

This is not to say that the priors $P(AG)$ and $P(GR)$ are unimportant. Indeed, our prior degree of belief in AG sets our requirements for F [31]. From basic Bayesian reasoning, the probability that AG is true when it is "detected" the odds-ratio decision scheme is

$$P(AG|detected) = \frac{E \times P(AG)}{E \times P(AG) + F \times P(GR)}$$

$$= \left(1 + \frac{F}{E}\frac{P(GR)}{P(AG)}\right)^{-1}; \quad (16)$$

since GR is so well tested, it seems reasonable that $P(AG) \ll P(GR)$; then F must be $\ll P(AG)$ if we are to believe that we have truly detected AG, because a false alarm is *a priori* much more probable than a true detection.

Combining Eq. (14) with the definition of F and the sampling distribution $p(x) = e^{-x^2/2}/\sqrt{2\pi}$, we obtain

$$F = \text{erfc}\left(\sqrt{\log\mathcal{O}'_{thr}}\right), \quad (17)$$

with $\text{erfc}(z) = 1 - \text{erf}(z)$ the *complementary error function*, defined from the error function[5] $\text{erf}(z) = (2/\sqrt{\pi}) \times \int_0^z e^{-t^2}dt$. Likewise, combining Eq. (11) with the definition of E and $p(x)$, we find

$$E = \frac{1}{2}(\text{erf}(-SNR_{AG} + \sqrt{\log\mathcal{O}'_{thr}})$$

$$- \text{erf}(-SNR_{AG} - \sqrt{\log\mathcal{O}'_{thr}})). \quad (18)$$

Next, we solve Eq. (17) for $\mathcal{O}'_{thr}$ and replace it in Eq. (18):

$$E = 1 - \frac{1}{2}(\text{erf}(-SNR_{AG} + \text{erfc}^{-1}(F))$$

$$- \text{erf}(-SNR_{AG} - \text{erfc}^{-1}(F))), \quad (19)$$

where $z = \text{erfc}^{-1}(P)$ is the solution of $\text{erfc}(z) = P$. Solving $E(SNR_{AG}) = 1/2$ yields the SNR_{AG} required

---

[5]With this definition, the cumulative density function (cdf) of a normal variable x with zero mean and unit norm is $\text{cdf}(x) = 1/2(1 + \text{erf}(x/\sqrt{2}))$.

for confident AG detection as a function of F, ranging from 2.75 to 4.05 for $F = 10^{-4}$ down to 10^{-8}. The GW-detection SNR required for AG detection is just $\mathrm{SNR_{AG}}(F)/\sqrt{1 - \mathrm{FF}}$, and it is plotted in Fig. 1 for $F = 10^{-4}$ and 10^{-8}. We already discussed the meaning of these curves in Sec. I.

An interesting question to ask is what detection rates would be needed in a volume-limited search so that the *loudest* observed signal could be used to detect AG corrections of given FF. In such a search, neglecting cosmological effects for simplicity, source distances are distributed as $p(D) = 3/D_{\mathrm{hor}}(D/D_{\mathrm{hor}})^2$, out to the horizon distance D_{hor} where sources are detected at the threshold $\mathrm{SNR_{thr}}$. For N GW detections, the minimum distance is distributed[6] as $p(D_{\min}) = 3N/D_{\mathrm{hor}}(D/D_{\mathrm{hor}})^2(1 - (D/D_{\mathrm{hor}})^3)^{N-1}$, which has median $D_{\mathrm{hor}}(1 - 2^{-1/N})^{1/3}$. If follows that the median maximum SNR is $\mathrm{SNR_{thr}}(1 - 2^{-1/N})^{-1/3}$. Setting this equal to $\mathrm{SNR_{AG}}(F)/\sqrt{1 - \mathrm{FF}}$ and solving for N, we obtain the required number of detections, which scales as $(1 - \mathrm{FF})^{-3/2}$, and is shown in Fig. 1 on the right-side vertical axis for $\mathrm{SNR_{thr}} = 8$.

Figuring out what happens if we pool together the evidence from a number of observed events [29,42] of the same kind is a little harder computationally. The odds ratios take forms similar to the one-signal case:

$$\mathcal{O}'_{\mathrm{GR}} = e^{\sum_i x_i^2/2}, \qquad \mathcal{O}'_{\mathrm{AG}} = e^{\sum_i x_i^2/2 + \sqrt{2}\sum_i x_i \mathrm{SNR}_{\mathrm{AG},i} + \mathrm{SNR}_{\mathrm{AG},i}^2},$$

(20)

where the x_i are independently distributed normal variables with zero mean and unit variance, and the $\mathrm{SNR}_{\mathrm{AG},i}$ are the effective AG-detection SNRs for the individual observations. Here I limit myself to a small Monte Carlo exploration: assuming for simplicity that the FF is the same for all the sources, and taking the median over all $\{\mathrm{SNR}_{\mathrm{AG},i}\}$ realizations in a volume-limited search with $\mathrm{SNR_{thr}} = 8$, I find that with $F = 10^{-4}$ we need $\sim 9/200/4500$ observations to detect AG with $1 - \mathrm{FF} = 10^{-2}/10^{-3}/10^{-4}$, to be compared with $\sim 28/900/30\,000$ using evidence from the loudest source alone. Essentially, because SNRs are distributed as $1/\mathrm{SNR}^4$, the Bayesian-inference problem is dominated by a few very loud events, and there are not very many of those for moderate detection rates. (However, this conclusion differs from the findings of Ref. [29], and it would be interesting to understand why.)

---

[6]Why? Consider first the minimum $x_{\min}$ among N variables independently and uniformly distributed in $[0, 1]$. Its distribution is $p(x_{\min}) = N(1 - x_{\min})^{N-1}$, since we could pick any of the N as the minimum, and then its probability of being in $[x_{\min}, x_{\min} + dx]$ is just dx times the probability that the other $N - 1$ are in $[x_{\min}, 1]$. The minimum $y_{\min}$ among N variables with distribution $p(y_{\min})$ follows from the transformation $x = \mathrm{cdf}(y)$, from which $p(y_{\min}) = p(x_{\min})\frac{dx}{dy}|_{y_{\min}}$.

VI. DISCUSSION

In this paper, I have shown that, under the assumptions of strong signals and Gaussian detector noise, the prospects for detecting alternative-gravity corrections to general relativity can be characterized very simply. A single number, the fitting factor FF between the GR and AG waveform families, determines the source SNR required for the alternative-gravity hypothesis to be favored in a decision scheme based on the Bayesian odds ratio (see Fig. 1).

This happens because the FF is an SNR-independent measure of the strength of the AG corrections $\Delta h_{\perp}$ that cannot be reabsorbed by changing the GR source parameters from their true values. The GR parameters are not known *a priori*, but must be determined from the same observation, so such "reabsorbable" AG effects cannot be detected positively, and they would result in a "fundamental bias" [23] on the GR parameters if AG is true, but post-detection parameter estimation is performed with GR model templates. In Ref. [25], Cornish and colleagues call such errors "stealth bias" if they are comparable to or larger than the noise-induced statistical errors in the GR parameters, and yet AG cannot be detected positively. In the terms of this paper, stealth bias corresponds to FF very close to one and AG-induced errors $(F^{-1})^{ij}(\Delta h, h_j)$ that are large compared to the Fisher-matrix statistical errors $\sqrt{(F^{-1})^{ii}}$.

My formalism can also be applied to other contexts[7] where we need to decide between a simpler model and one with additional parameters, such as binary inspirals of nonspinning vs spinning compact objects, orbit-aligned vs precessing spins, or pointlike vs extended-object dynamics.

My formulas cannot predict what happens when the high-SNR, linearized-parameter approximation is not warranted; whether that is the case can be determined using the test described in Sec. VI of Ref. [39]. At low SNRs, full-fledged Monte Carlo integration [24,25,29] would be required for accurate predictions, although the FF formula could be used as a preliminary step, and its comparison with the full result would be very instructive. I note however that it is for the strongest signals that GR-by-GW tests become most interesting, and that the results discussed above would persist as the leading-order contributions to the evidence at any SNR (see again [39], Sec. VII).

Beyond the statistical characterization of the tests, we should always ask ourselves *what it is* that we could really detect, and whether *we should really believe* a positive AG detection if we get it. These are very hard questions, and here I offer only some qualitative considerations that should be kept in mind whenever we discuss the sensitivity of GR-by-GW tests.

First, it seems evident that a test based on matching AG corrections of a certain functional form Δh would only be sensitive to non-GR effects that have nonzero projection

---

[7]I thank Ilya Mandel for pointing this out and providing these examples.

along Δh. (For instance, AG waveforms with additional phasing parameters would not be sensitive to amplitude corrections.) Now, both the consistency checks based on altering PN coefficients and the ppE framework consider rather general corrections, so it may be hard to imagine that the waveform imprint from any reasonable AG theory would be fully orthogonal to them. Indeed, Ref. [26] argues that for quasicircular binary inspirals, the well-posedness of the initial-value problem restricts possible phasing terms to frequency powers $f^{n/3}$ (where n can be negative), which could be covered in the ppE scheme. However, if the projection is small, the resulting $1 -$ FF would be strongly reduced, and the test would be sensitive only to much larger effects.

Second, any positive detection of an AG correction Δh could also be explained as one of many *systematic* waveform corrections [45] that have nonzero projection along Δh, such as the effects of detector calibration and non-Gaussian detector noise, of standard-GR physics not included in the waveforms (spins, eccentricity, higher-PN terms), and of astrophysical perturbations (accretion disks, three-body systems). All of these effects should be considered *a priori* more likely than a modification of the extensively well tested GR, so they must be controlled by including them explicitly in the GR model, or at least by establishing that they are sufficiently orthogonal to AG corrections. On the plus side, instrumental systematics

would be different for the same signal as observed in multiple detectors, and GR-theoretical and astrophysical systematics would be different for multiple signals from similar sources, which would help discriminate AG corrections [46]. Nevertheless, preliminary claims of sensitivity to specific AG corrections may be overoptimistic, because Δh could be largely reabsorbed by systematic effects that are initially neglected.

Testing GR with GWs remains one of the exciting frontiers of GW astronomy, but appropriate caution is needed to provide the proper context for current and future investigations, and to allocate research efforts wisely as we move toward the GW detection era. Computing some FFs will help.

ACKNOWLEDGMENTS

[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://www.livingreviews.org/lrr-2009-2].

[3] C. M. Will, Phys. Rev. D **50**, 6058 (1994).

[4] A. Królak, K. D. Kokkotas, and G. Schäfer, Phys. Rev. D **52**, 2089 (1995).

[5] C. M. Will, Phys. Rev. D **57**, 2061 (1998).

[6] P. D. Scharre and C. M. Will, Phys. Rev. D **65**, 042002 (2002).

[7] C. M. Will, Classical Quantum Gravity **20**, S219 (2003).

[8] C. M. Will and N. Yunes, Classical Quantum Gravity **21**, 4367 (2004).

[9] E. Berti, A. Buonanno, and C. M. Will, Classical Quantum Gravity **22**, S943 (2005).

[10] D. I. Jones, Astrophys. J. Lett. **618**, L115 (2005).

[11] E. Berti, A. Buonanno, and C. M. Will, Phys. Rev. D **71**, 084025 (2005).

[12] K. G. Arun and C. M. Will, Classical Quantum Gravity **26**, 155002 (2009).

[13] A. Stavridis and C. M. Will, Phys. Rev. D **80**, 044002 (2009).

[14] D. Keppel and P. Ajith, Phys. Rev. D **82**, 122001 (2010).

[15] K. Yagi and T. Tanaka, Phys. Rev. D **81**, 064008 (2010).

[16] E. Berti, J. Gair, and A. Sesana, Phys. Rev. D **84**, 101501 (R) (2011).

[17] S. Mirshekari, N. Yunes, and C. M. Will, Phys. Rev. D **85**, 024041 (2012).

[18] E. Berti, L. Gualtieri, M. Horbatsch, and J. Alsing, Phys. Rev. D **85**, 122005 (2012).

[19] K. G. Arun, B. R. Iyer, M. S. S. Qusailah, and B. S. Sathyaprakash, Classical Quantum Gravity **23**, L37 (2006).

[20] K. G. Arun, B. R. Iyer, M. S. S. Qusailah, and B. S. Sathyaprakash, Phys. Rev. D **74**, 024006 (2006).

[21] C. K. Mishra, K. G. Arun, B. R. Iyer, and B. S. Sathyaprakash, Phys. Rev. D **82**, 064010 (2010).

[22] C. Huwyler, A. Klein, and P. Jetzer, arXiv:1108.1826 [Phys. Rev. D (to be published)].

[23] N. Yunes and F. Pretorius, Phys. Rev. D **80**, 122003 (2009).

[24] W. Del Pozzo, J. Veitch, and A. Vecchio, Phys. Rev. D **83**, 082002 (2011).

[25] N. Cornish, L. Sampson, N. Yunes, and F. Pretorius, Phys. Rev. D **84**, 062003 (2011).

[26] K. Chatziioannou, N. Yunes, and N. Cornish, Phys. Rev. D **86**, 022004 (2012).

[27] E. T. Jaynes, *Probability Theory: The Logic of Science*, edited by G. L. Bretthorst (Cambridge University Press, Cambridge, UK, 2003).

[28] P. C. Gregory, *Bayesian Logical Data Analysis for The Physical Sciences: A Comparative Approach with Mathematica Support* (Cambridge University Press, Cambridge, UK, 2005).

[29] T. G. F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, and A. Vecchio, Phys. Rev. D **85**, 082003 (2012).

[30] T. G. F. Li, W. Del Pozzo, S. Vitale, C. Van Den Broeck, M. Agathos, J. Veitch, K. Grover, T. Sidery, R. Sturani, and A. Vecchio, J. Phys. Conf. Ser. **363**, 012028 (2012).

[31] J. Veitch and A. Vecchio, Classical Quantum Gravity **25**, 184010 (2008).

[32] J. Veitch and A. Vecchio, Phys. Rev. D **78**, 022001 (2008).

[33] R. Umstätter and M. Tinto, Phys. Rev. D **77**, 082002 (2008).

[34] T. B. Littenberg and N. J. Cornish, Phys. Rev. D **80**, 063007 (2009).

[35] A. C. Searle, arXiv:0804.1161.

[36] T. A. Apostolatos, Phys. Rev. D **52**, 605 (1995).

[37] C. Cutler and É. E. Flanagan, Phys. Rev. D **49**, 2658 (1994).

[38] C. Cutler and M. Vallisneri, Phys. Rev. D **76**, 104018 (2007).

[39] M. Vallisneri, Phys. Rev. D **77**, 042001 (2008).

[40] J. Abadie *et al.*, Classical Quantum Gravity **27**, 173001 (2010).

[41] B. Sathyaprakash *et al.*, Classical Quantum Gravity **29**, 124013 (2012).

[42] I. Mandel, Phys. Rev. D **81**, 084029 (2010).

[43] T. A. Prince *et al.*, LISA: Probing the Universe with Gravitational Waves [http://list.caltech.edu/mission_documents], 2009.

[44] P. Amaro-Seoane *et al.*, Classical Quantum Gravity **29**, 124016 (2012).

[45] B. Kocsis, N. Yunes, and A. Loeb, Phys. Rev. D **84**, 024032 (2011).

[46] N. Yunes, B. Kocsis, A. Loeb, and Z. Haiman, Phys. Rev. Lett. **107**, 171103 (2011).