AEI-2009-115

Random template placement and prior information

Christian Röver

Max-Plack-Institut für Gravitationsphysik (Albert-Einstein-Institut), Callinstraße 38, 30167 Hannover, Germany.

Abstract. In signal detection problems, one is usually faced with the task of searching a parameter space for peaks in the likelihood function which indicate the presence of a signal. Random searches have proven to be very efficient as well as easy to implement, compared e.g. to searches along regular grids in parameter space. Knowledge of the parameterised shape of the signal searched for adds structure to the parameter space, i.e., there are usually regions requiring to be densely searched while in other regions a coarser search is sufficient. On the other hand, prior information identifies the regions in which a search will actually be promising or may likely be in vain. Defining specific figures of merit allows one to combine both template metric and prior distribution and devise optimal sampling schemes over the parameter space. We show an example related to the gravitational wave signal from a binary inspiral event. Here the template metric and prior information are particularly contradictory, since signals from low-mass systems tolerate the least mismatch in parameter space while high-mass systems are far more likely, as they imply a greater signal-to-noise ratio (SNR) and hence are detectable to greater distances. The derived sampling strategy is implemented in a Markov chain Monte Carlo (MCMC) algorithm where it improves convergence.

1. Introduction

Signal detection, in gravitational wave detection in particular, frequently entails the problem of performing a computationally expensive numerical search over a large parameter space. The *search* here means a search for a peak in the likelihood function, or another detection statistic, based on the data at hand and varying the unknown signal parameters. A peak or a threshold excess then indicates the presence of a signal [1, 2]. Such "brute-force" searches may be implemented as *grid searches*, evaluating the detection statistic at regularly placed points in parameter space. Computing the detection statistic usually means evaluating the *match* between a *signal template* and the data; the spacing between evaluated points in parameter space is then usually based on a *template metric* which ensures that all possible signals (corresponding to points in parameter space) have at least a certain minimal match with one of the evaluated templates (corresponding to the grid points). Instead of using regularly spaced template banks, the use of *random template banks* has recently gained popularity, as these are often very easily implemented, and have also be shown to be very efficient, especially in higher dimensions [3]. Here the idea is to populate the parameter space randomly, but uniformly with respect to the template metric.

These template placement strategies have by now usually been based on "*minimax*" reasoning, by aiming at minimizing the maximal (worst-case) mismatch across the whole parameter space. Once one takes prior information on the unknown parameters into consideration, by accounting for a priori probabilities attached to different regions of parameter space, a decision-theoretic

approach allows us to devise other strategies, effectively concentrating efforts on the more promising regions of parameter space in pursuit of a certain optimality criterion [4, 5]. In fact, a minimax strategy may often only exist once one imposes hard bounds on the parameter space (and by that ensuring the existence of an absolute *worst case*).

Markov chain Monte Carlo (MCMC) methods are meanwhile widely used for (Bayesian) parameter estimation in the signal processing stage for gravitational-wave signals [6, 7]. MCMC algorithms are, first of all, methods for *stochastic integration* [8, 9], although by the way they work they often behave similarly to *stochastic search* algorithms as well. This is in fact a most welcome property, as part of the parameter estimation problem is usually also a search/optimization problem, as, besides integration over the parameters' posterior distribution, it requires finding the global mode or secondary modes. *Parallel tempering* [10, 11] is a variety of the Metropolis-Hastings MCMC algorithm (and a special case of Metropolis-coupled MCMC algorithm [12, 9]) aimed at enhancing these stochastic search capabilities. This is done by basically running several MCMC chains in parallel, where *tempering* at increasing temperature values is applied to subsequent chains (as in simulated annealing methods [13]), and additional steps are introduced to allow for communication between chains [14]. Parallel tempering methods have been applied to gravitational-wave data analysis for binary inspiral signals in the context of ground-based [15] and space-based (LISA) measurements [14], where they have proven advantageous especially in cases of high SNR and of posterior distributions exhibiting multiple modes or degeneracies [16, 17]. They have meanwhile also been adopted for the analysis of burst signals [18].

Among the parallel Markov chains being run at different 'temperatures' within the parallel tempering implementation, the 'cool' ones with no tempering applied produce samples from the posterior distribution for the stochastic integration part, while the high-temperature chains are producing samples for the stochastic search. The question now is how to set up the algorithm so that the search is most efficient, given our knowledge of prior and template metric, i.e., our knowledge of "where the true parameters are (un-) likely to be", and "how hard one needs to look" across the parameter space. The problem is of special interest in the context of binary inspiral signals, as prior and template metric are particularly contradictory: a priori one is most likely do detect an inspiral involving high masses, as these result in a high-SNR signal that is detectable to a greater distance. On the other hand, considering the template metric only, one might want to mostly try low-mass templates, since at low masses the template's and true signal parameters need to be in very close agreement in order for them to match, while at high masses greater discrepancies still yield a good match. What needs to be defined is the distribution to sample from in order to find the mode(s) fastest, which is very similar to setting up a random template bank, the difference being that one does not settle on some fixed number of templates, as the MCMC sampler in principle is thought to sample indefinitely.

In the following Sec. 2, we will introduce the problem for the case of binary inspiral signals, and Sec. 3 briefly introduces the parallel tempering context. In Sec. 4 the general problem is formulated in decision-theoretic terms and solved for a particular optimality criterion. Sec. 5 shows some illustrative examples, and Sec. 6 eventually closes with conclusions and perspectives.

2. Binary inspiral parameters

In the simplest description, a binary inspiral signal as measured by ground-based interferometers is determined by 9 parameters: sky location (declination δ, right ascension α), polaristion (ψ), companion masses (m_1, m_2), luminosity distance (d_L), time of arrival (t_c), phase (ϕ) and inclination angle (ι). Assuming some prior distribution for the masses (in the following simply defined to be uniform, $m_1, m_2 \in [1\,\mathrm{M}_\odot, 10\,\mathrm{M}_\odot]$), and an isotropic distribution of events across space while folding in the detectability as a function of signal-to-noise ratio (SNR), one can derive a joint prior distribution whose marginal distribution of masses is shown in Fig. 1

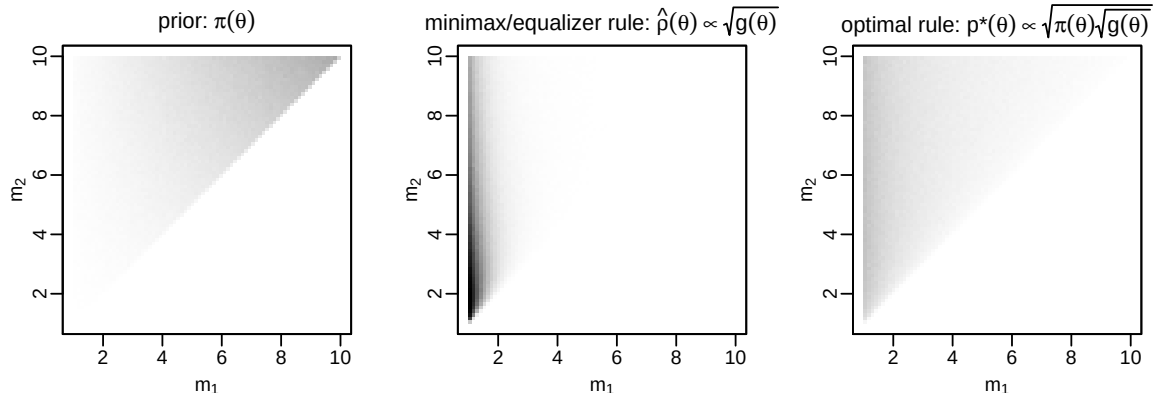

Figure 1. (Marginal) densities of the distributions π, $\hat{\rho}$ and $p^\star$ for the two mass parameters (m_1, m_2) of a binary inspiral signal. The prior (left plot) indicates that high masses are most likely, which is because they result in stronger signals that are detectable to greater distances. The template metric on the other hand implies that low masses require a dense template spacing (middle plot).

[19, 14]. A template metric may be defined following [20, 21], assuming the metric to be constant in the space of the *Newtonian* and *1.5 PN chirp times* λ_1 and λ_2, which are functions of the mass parameters. For the remaining parameters, for now, we again assume the metric to be uniform (t_c, $\log(d_L)$) and isotropic (δ, α, ψ, ι, ϕ). The implied distribution in terms of (m_1, m_2) following from a uniform spacing in (λ_1, λ_2) may be derived using the reparametrisation explicated in [22]. This distribution is shown in Fig. 1.

3. Parallel tempering

In the context of Monte Carlo integration, tempering is utilised to prevent the integration algorithm from getting stuck in local modes of the distribution from which it is sampling. A temperature parameter $T \geq 1$ is introduced, and instead of sampling from the distribution of actual interest, with density function $f(\theta)$, the modified distribution

$$f_{(T)}(\theta) \;\propto\; f(\theta)^{\frac{1}{T}} \tag{1}$$

is used. The introduced exponent is supposed to make the distribution more tractable, as it has a "flattening" effect on the density; the same effect is also taken advantage of in simulated annealing methods [13]. In the limit of $T \to \infty$, the density $f_{(T)}(\theta)$ then approaches a uniform distribution [9]. In the context of posterior inference, when the target distribution $f(\theta)$ is the product of prior $\pi(\theta)$ and likelihood $\mathcal{L}(\theta)$, it may be more sensible to use a scheme only tempering the likelihood part:

$$f_{(T)}(\theta) \;\propto\; \pi(\theta)\,\mathcal{L}(\theta)^{\frac{1}{T}}, \tag{2}$$

in which case $f_{(T)}(\theta)$ goes towards the prior $\pi(\theta)$ for $T \to \infty$ [14]. Both uniform distribution and prior distribution may in general not be the most sensible choice, as was pointed out above, since the tempering is also supposed to enhance the algorithm's *stochastic search* properties. Assume that one had a distribution $p^\star(\theta)$ available, which leads to an optimal sampling (w.r.t. to some pre-specified criterion), and which is then the desired density for $T \to \infty$. This suggests a generalized tempering parametrisation:

$$f_{(T)}(\theta) \;\propto\; p^\star(\theta)\left(\frac{f(\theta)}{p^\star(\theta)}\right)^{\frac{1}{T}} \;=\; p^\star(\theta)^{1-\frac{1}{T}}\, f(\theta)^{\frac{1}{T}} \tag{3}$$

which in the special cases of $p^\star(\theta) \propto 1$ and $p^\star(\theta) = \pi(\theta)$ again yields the tempering schemes from (1) and (2) above. The question now is how to choose such a limiting distribution $p^\star(\theta)$ based on given prior information and template metric.

4. The decision theoretic approach

Let $g(\theta)$ be the determinant of the template metric as a function of the signal parameters. A large value of g means that that templates need to be densely spaced around θ, while a smaller g indicates that a coarser spacing is sufficient. The volume "covered" by a template placed at parameter θ is proportional to $g(\theta)^{-\frac{1}{2}}$, and hence the probability density to sample from for setting up a random template bank is given by $\hat{\rho}(\theta) \propto \sqrt{g(\theta)}$ [3].

Now consider the case of the true parameter value being $\theta_0 \in \Theta$. The actual value θ_0 is unknown, what is known is the prior probability density $\pi(\theta)$. Whenever a template θ^* is placed in parameter space, it is considered a *match* if it was sufficiently close to the true value θ_0. What exactly is "sufficiently close" is determined via mismatch considerations and is expressed through the template metric. Then the probability of a match is

$$\mathrm{P}(\mathrm{match}\,|\,\theta^*) \;=\; c\,\frac{1}{\sqrt{g(\theta^*)}}\,\pi(\theta^*), \tag{4}$$

where $c \in \mathbb{R}^+$ is a constant depending on how close a match actually is required to be. If one was to pick a *single* template θ^*, the chances for success would obviously be maximal where the above product reaches its maximum. Analogously, consider the case of a *given* true value θ_0 and repeated, independent "guesses" drawn from $p^\star(\theta)$. Then for each single guess the probability of success is

$$\mathrm{P}(\mathrm{match}\,|\,\theta_0) \;=\; c\,\frac{1}{\sqrt{g(\theta_0)}}\,p^\star(\theta_0). \tag{5}$$

What is desired is a distribution $p^\star$ from which to generate independent draws so that the chances of getting a match are "optimal". Whether or when one will get a match is a matter of chance, depending on both the true value $\theta_0 \in \Theta$ and the choice of $p^\star \in \mathcal{P}^\star$, where $\mathcal{P}^\star$ is the space of probability distributions over Θ. Suppose we are interested in minimizing the expected number of trials T (or *waiting time*) until the first match. Any choice of $p^\star$ implies a probability distribution for T; for a *given* true value θ_0 and a sampling distribution $p^\star$, T follows a geometric distribution with density and expectation:

$$\mathrm{P}(T{=}t\,|\,\theta_0) \;=\; \Big(1{-}c\,\frac{1}{\sqrt{g(\theta_0)}}\,p^\star(\theta_0)\Big)^{t-1}\,\Big(c\,\frac{1}{\sqrt{g(\theta_0)}}\,p^\star(\theta_0)\Big), \quad \mathrm{E}[T\,|\,\theta_0] \;=\; \frac{1}{c\,\frac{1}{\sqrt{g(\theta_0)}}\,p^\star(\theta_0)}. \tag{6}$$

In decision theoretic terms, we are given a *state-of-nature space* Θ, an *action space* $\mathcal{P}^\star$, and a *loss function* $L : \Theta \times \mathcal{P}^\star \to \mathbb{R}$ with $L(\theta_0, p^\star) = \mathrm{E}_{p^\star}[T|\theta_0]$ [4, 5]. An optimal choice of $p^\star$ may now be determined by minimizing the expected loss; integrating over the possible values that θ_0 could take, that (prior) expectation is

$$\mathrm{E}[T] \;=\; \frac{1}{c}\int_\Theta \frac{1}{\frac{1}{\sqrt{g(\theta_0)}}\,p^\star(\theta_0)}\,\pi(\theta)\,\mathrm{d}\theta, \tag{7}$$

which is minimized by choosing

$$p^\star(\theta) \;\propto\; \sqrt{\pi(\theta)\,\sqrt{g(\theta)}} \;=\; \sqrt{\pi(\theta)\,\hat{\rho}(\theta)}, \tag{8}$$

i.e., the optimal $p^\star$ here is proportional to the geometric mean of π and $\hat{\rho}$, and independent of c.

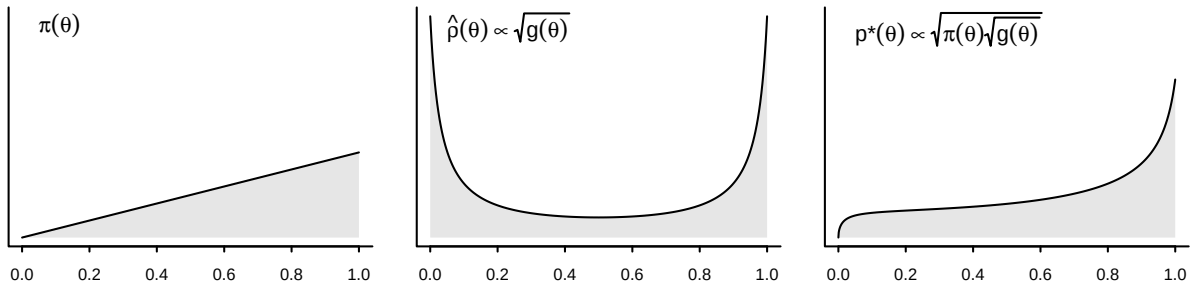

Figure 2. Densities of the distributions π, $\hat{\rho}$ and $p^\star$ for the toy example discussed in Sec. 5.2.

The distribution defined through the density $\hat{\rho}$ that is usually utilized for random template banks [3] plays a particular role in this context. From equation (5) one can see that by setting $p^\star := \hat{\rho}$ the probability of a match (and with that also the waiting time) becomes independent of the actual parameter value θ_0, so that $\hat{\rho}$ constitutes an *equalizer rule*. From (8) it follows that $\hat{\rho}$ will be optimal in the case that the prior happens to be $\pi = \hat{\rho}$. This implies that $\pi = \hat{\rho}$ defines the "*least favourable prior distribution*" for this case, and that $p^\star = \hat{\rho}$ also constitutes the *minimax* strategy (independent from the particular prior π), as it minimizes the maximum of $\mathrm{E}[T \,|\, \theta_0]$ across all possible true values θ_0 [4]. Since $p^\star = \hat{\rho}$ leads to a uniform match probability in (5), it actually constitutes the equalizer rule for the wider family of optimality criteria that are functions of $\mathrm{P}(\mathrm{match} \,|\, \theta_0)$.

5. Examples

5.1. Toy example 1: Gaussian prior

Consider a parameter space $\Theta = \mathbb{R}$ where the prior is Gaussian with mean μ and variance σ^2: $\pi = \mathrm{N}(\mu, \sigma^2)$, and the template metric is *flat*, i.e., $g(\theta) = \gamma$ is independent of θ. Then the equalizer rule $\hat{\rho}$ does not exist, and the optimal rule would be $p^\star = \mathrm{N}(\mu, 2\sigma^2)$.

5.2. Toy example 2: Numerical simulation

Consider a parameter space $\Theta = [0, 1]$, where the prior and template metric behave as shown in Fig. 2. For this simple case the behaviour of different sampling strategies can be simulated numerically, by drawing "true" parameter values θ_0 from the prior distribution and then drawing "guesses" θ^* from either $\hat{\rho}$ or $p^\star$ in order to see how the strategies differ.

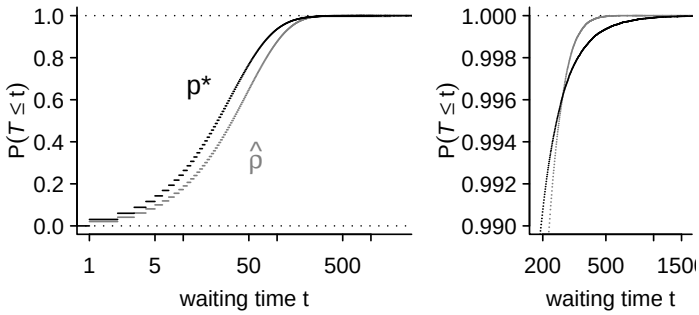

Figure 3. Cumulative distributions of the resulting waiting times T when using sampling strategies $p^\star$ and $\hat{\rho}$ in the toy example of Sec. 5.2. The right panel shows a zoom-in on the differing tail behaviour.

Fig. 3 illustrates the distribution of the resulting times T, for both the minimax and optimal strategies $\hat{\rho}$ and $p^\star$. As expected, the average waiting time is lower for $p^\star$, and one can see that the minimax strategy performs better in the unlikely "worst cases".

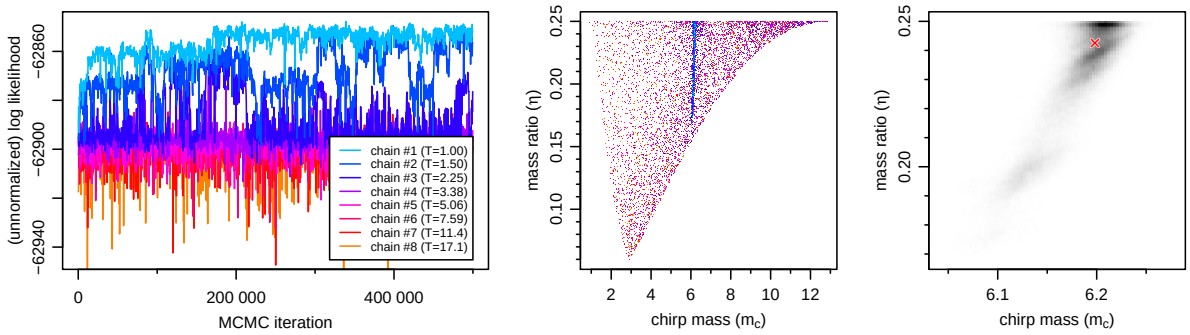

Figure 4. This plot illustrates the behaviour of a Parallel Tempering algorithm utilizing the distribution $p^\star$ when running on simulated data. The left panel shows how the algorithm's 'cool' chains manage to ascend to greater likelihood values while the tempered chains keep sampling at lower likelihood values. The 2nd panel is a scatter plot of mass parameter samples from all the different chains (after the algorithm's burn-in phase). The right panel eventually shows the resulting mass parameters' marginal posterior density derived from the 'cool' chain #1 alone; the cross indicates the true parameter value.

5.3. Binary inspiral example

The prior π and minimax sampling rule $\hat{\rho}$ for the mass parameters of a binary inspiral event were shown in Fig. 1. The right panel of the same figure also shows the resulting optimized sampling distribution $p^\star$. The obvious discrepancy between least favourable ($\hat{\rho}$) and actual prior (π) suggests that there actually is a gain in doing the optimization. Fig. 4 shows how a parallel tempering algorithm for parameter estimation behaves when utilizing the distribution $p^\star$ for high-temperature chains as described in Sec. 3 (3). The MCMC chains quickly converge to the true parameter values, while the higher-temperature chains keep scanning the parameter space efficiently.

6. Conclusions and outlook

We have applied a decision-theoretic approach in order to derive an optimized sampling distribution to be used within a parallel tempering MCMC implementation. The optimization step here provides a natural link between the parameter space metric and the prior information about the parameter values. The particular optimality criterion chosen here (the expected time until a matching template is found, $\mathrm{E}[T|\theta_0]$) turns out to be computationally convenient, as the resulting sampling distribution $p^\star$ is independent of the particular mismatch threshold c, and is almost trivial to implement within an MCMC application. Other criteria are conceivable though, like the probability of a missed detection within N samples $\mathrm{P}(T > N|\theta_0)$ for example, which may then lead to more complicated results.

The general approach used here should also be useful in other contexts; it turns out that the distribution usually used for setting up random template banks here constitutes the special case of a *minimax* strategy, which implies that the explicit specification of particular figures-of-merit and the consideration of prior information may yield great efficiency improvements, especially in cases where the implicitly assumed *least favourable prior* greatly deviates from the actual prior information as in the binary inspiral case. In the framework discussed above, the resulting optimized sampling distribution $p^\star$ even exists for cases where the minimax rule does not (as in the example of Sec. 5.1 above). This suggests that a similar approach may also make other ad-hoc fixes like the mass parameter bounds in the binary inspiral example dispensable, as it would naturally focus in on the promising parameter range while ruling out too unlikely and

too costly regions of parameter space.

Acknowledgments
The author would like to thank Chris Messenger, Reinhard Prix and Graham Woan for helpful discussions. This work was supported by the Max-Planck-Society.

References

[1] McDonough R N and Whalen A D 1995 *Detection of signals in noise* 2nd ed (New York: Academic Press)
[2] Wainstein L A and Zubakov V D 1962 *Extraction of signals from noise* (Englewood Cliffs, NJ: Prentice-Hall)
[3] Messenger C, Prix R and Papa M A 2009 *Physical Review D* **79** 104017
[4] Berger J O 1985 *Statistical decision theory and Bayesian analysis* 2nd ed (Springer-Verlag)
[5] Ferguson T S 1967 *Mathematical Statistics: A Decision Theoretic Approach* (New York: Academic Press)
[6] Christensen N and Meyer R 1998 *Physical Review D* **58** 082001
[7] Umstätter R, Meyer R, Dupuis R, Veitch J, Woan G and Christensen N 2004 *Classical and Quantum Gravity* **21** S1655–S1665
[8] Metropolis N and Ulam S 1949 *Journal of the American Statistical Association* **44** 335–341
[9] Gilks W R, Richardson S and Spiegelhalter D J 1996 *Markov chain Monte Carlo in practice* (Boca Raton: Chapman & Hall / CRC)
[10] Hukushima K and Nemoto K 1996 *Journal of the Physical Society of Japan* **65** 1604–1608
[11] Hansmann U H E 1997 *Chemical Physics Letters* **281** 140–150
[12] Geyer C J 1991 *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* ed Keramidas E M (Fairfax Station: Interface Foundation) pp 156–163
[13] Press W H, Teukolsky S A, Vetterling W T and Flannery B P 1992 *Numerical recipes in C: The art of scientific computing* (Cambridge: Cambridge University Press)
[14] Röver C 2007 *Bayesian inference on astrophysical binary inspirals based on gravitational-wave measurements* Ph.D. thesis The University of Auckland URL `http://hdl.handle.net/2292/2356`
[15] Röver C, Meyer R and Christensen N 2007 *Physical Review D* **75** 062004
[16] van der Sluys M V, Röver C, Stroeer A, Christensen N, Kalogera V, Meyer R and Vecchio A 2008 *The Astrophysical Journal Letters* **688** L61–L64
[17] Raymond V, van der Sluys M V, Mandel I, Kalogera V, Röver C and Christensen N 2009 *Classical and Quantum Gravity* **26** 114007
[18] Key J S and Cornish N J 2009 *Physical Review D* **79** 043014
[19] Röver C, Meyer R, Guidi G M, Viceré A and Christensen N 2007 *Classical and Quantum Gravity* **24** S607–S615
[20] Owen B J and Sathyaprakash B S 1999 *Physical Review D* **60** 022002
[21] Chronopoulos A E and Apostolatos T A 2001 *Physical Review D* **64** 042003
[22] Umstätter R and Tinto M 2008 *Physical Review D* **77** 082002