Accelerated gravitational-wave parameter estimation with reduced order modeling

Priscilla Canizares,1,2 Scott E. Field,3,4 Jonathan Gair,1 Vivien Raymond,5 Rory Smith,5 and Manuel Tiglio6

1Institute of Astronomy, Madingley Road, Cambridge, CB30HA, United Kingdom
2Department of Applied Mathematics and Theoretical Physics, Wilberforce Road, Cambridge CB3 0WA, United Kingdom
3Department of Physics, Joint Space Sciences Institute, Maryland Center for Fundamental Physics, University of Maryland, College Park, MD 20742, USA
4Center for Radiophysics and Space Research, Cornell University, Ithaca, New York 14853, USA
5LIGO, California Institute of Technology, Pasadena, CA 91125, USA
6Center for Astrophysics and Space Sciences, Center for Computational Mathematics, San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92037-0424

(Dated: January 29, 2015)

Inferring the astrophysical parameters of coalescing compact binaries is a key science goal of the upcoming advanced LIGO-Virgo gravitational-wave detector network and, more generally, gravitational-wave astronomy. However, current approaches to parameter estimation for these detectors require computationally expensive algorithms. Therefore there is a pressing need for new, fast and accurate Bayesian inference techniques. In this Letter we demonstrate that a reduced order modeling approach enables rapid parameter estimation to be performed. By implementing a reduced order quadrature scheme within the LIGO Algorithm Library, we show that Bayesian inference on the 9-dimensional parameter space of non-spinning binary neutron star inspirals can be sped up by a factor of ~30 for the early advanced detectors’ configurations (with sensitivities down to around 40 Hz) and ~70 for sensitivities down to around 20 Hz. This speed-up will increase to about 150 as the detectors improve their low-frequency limit to 10 Hz, reducing to hours analyses which could otherwise take months to complete. Although these results focus on interferometric gravitational wave detectors, the techniques are broadly applicable to any experiment where fast Bayesian analysis is desirable.

Introduction—Advanced LIGO (aLIGO)1 and advanced Virgo (AdV)2 are expected to yield the first direct detections of gravitational waves (GWs) from astrophysical sources in the next few years. Compact binary coalescences (CBCs) are the most promising GW sources, with expected detection rates between a few and tens per year3. Effective parameter estimation for CBCs has been demonstrated4–6, but approaches to date carry high computational costs for the cases of interest, even when using efficient algorithms such as Markov chain Monte Carlo (MCMC) or nested sampling7. For the advanced detectors, which will start taking data within a year or two, current approaches will lead to months or years of computational wall (clock) time for the analysis of each detected signal. Therefore, given the expected detection rates, new techniques which can estimate the astrophysical source parameters in short timescales are highly desirable. Such techniques are also important for large scale mock data challenges.

In parameter estimation studies, the posterior probability density function (PDF) of a set of parameters, \( \theta \), is computed from a GW model, \( h(\theta) \), assumed to describe the detector’s signal \( d \). The PDF is related to the likelihood function, \( L(d|\theta) \), and the prior probability on the model parameters, \( P(\theta) \), via Bayes’ theorem:

\[
p(\theta|d) \propto P(\theta) L(d|\theta).
\]

Assuming that the detector data \( d \) contains the source’s signal \( h(\theta_{\text{true}}) \) and stationary Gaussian noise \( n \), the likelihood function is given by

\[
\log L(d|\bar{\theta}) = (d|h(\bar{\theta})) - \frac{1}{2} \left[ (h(\bar{\theta})|h(\bar{\theta})) + (d|d) \right],
\]

where \( d = h(\theta_{\text{true}}) + n \) and \( (a|b) \) is a weighted inner product for discretely sampled noisy data

\[
(d|h(\bar{\theta})) = 4R \Delta f \sum_{k=1}^{L} \frac{\hat{d}^*(f_k)\hat{h}(\bar{\theta}; f_k)}{S_n(f_k)},
\]

where \( \hat{d}(f_k) \) and \( \hat{h}(\bar{\theta}; f_k) \) are the discrete Fourier transforms at frequencies \( f_k \) (with units of Hz), \( * \) denotes complex conjugation, and the power spectral density (PSD) \( S_n(f_k) \) characterizes the detector’s noise.

For a given observation time \( T = 1/\Delta f \) and detection frequency window \( (f_{\text{high}} - f_{\text{low}}) \) there are

\[
L = \int_{[f_{\text{high}} - f_{\text{low}}]} T
\]

sampling points in the sum8. When \( L \) is large, as in the cases of interest for this Letter, there are two major bottlenecks: (i) evaluation of the model at each \( f_k \) and, (ii) assembly of the likelihood1. In general, smoothly parameterized models are amenable to dimensional reduction which, in turn, provides computationally efficient representations. The specific application of dimensional reduction we consider in this Letter tackles the two aforementioned bottlenecks by permitting the inner product2 to be computed with significantly fewer terms. In summary: if a reduced set
of \( N < L \) basis can be found which accurately spans the model space, it is possible to replace the inner product \( \langle \rangle \) with a reduced order quadrature (ROQ) rule \( \{ \cdot \} \) containing only \( N \) terms, reducing the overall parameter estimation analysis cost by a factor of \( L/N \), provided the waveforms can be directly evaluated. For other models, in particular those described by partial or ordinary differential equations, direct evaluation may be accomplished using surrogates \( \{ \cdot \} \).

In this Letter, we demonstrate a ROQ accelerated GW parameter estimation study. While the approach is applicable to any GW model, here we focus on binary neutron star (BNS) inspirals, as these are expected to have the highest detection rates with the lowest uncertainty \( \{ \cdot \} \). We show, both through operation counts and an implementation in the LIGO Algorithm Library (LAL) pipeline \( \{ \cdot \} \), that ROQs provide a factor of \( \sim 30 \) speedup for the early advanced detectors’ configuration \( \{ \cdot \} \) and \( \sim 70 \) as the detectors’ low-frequency sensitivity reaches 20 Hz. This speedup will rise to \( \sim 150 \) as the sensitivity band is lowered to a target of 10 Hz, allowing for significant reduction of the computational cost of Bayesian parameter estimation analyses on BNS sources.

**Compressed likelihood evaluations**—Compared to previous work on which this Letter is based \( \{ \cdot \} \), parameter estimation of gravitational waves from binary neutron stars carries a number of challenges unique to large datasets which contain long gravitational waveforms with many in-band wave cycles. We briefly summarize the construction of ROQs while focusing on technical but essential solutions to these challenges.

Reduced order quadratures can be used for fast parameter estimation whenever the waveform model is amenable to dimensional reduction, through three steps. The first two are carried out offline, while the final, data-dependent step is performed at the beginning of the parameter estimation analysis. (1) Construct a reduced basis, i.e. a set of \( N \) elements whose span reproduces the GW model within a specified precision. (2) Construct an empirical interpolant by requiring it to exactly match any template at \( N \) carefully chosen frequency subsamples \( \{ F_k \}_{1 \leq k \leq N} \). (3) The empirical interpolant is used to replace, without loss of accuracy, inner product evaluations \( \{ \cdot \} \) by ROQ compressed ones \( \{ \cdot \} \).

**Step 1.** The reduced basis set only needs to be built over the space of intrinsic parameters for the waveform family. Furthermore, if the basis is generated using a PSD of unity the representation of the waveform family can be used with any PSD whenever the weights are built as in Eq. (6).

Basis generation in this Letter proceeds in two stages. A greedy algorithm first identifies a preliminary basis suitable for any value of \( \Delta f \). Next, this preliminary basis is evaluated at \( L \) equally spaced frequency samples appropriate for the detector. The elements of this “resampled basis” are neither orthogonal nor linearly independent, and so a second similar dimensional reduction is necessary. During these steps appropriately conditioned numerical algorithms \( \{ \cdot \} \) are used to avoid poor conditioning, which, for large values of \( L \), would otherwise lead to bases with no accuracy whatsoever.

**Step 2.** Given an \( N \)-size basis it is possible to uniquely and accurately reconstruct any waveform from only \( N \) evaluations \( \{ \tilde{h}(\tilde{\theta}; F_k) \}_{k=1}^N \). The special frequencies \( \{ F_k \}_{k=1}^N \), selected from the full set \( \{ f_i \}_{i=1}^L \), can be found from Algorithm 5 of Ref. \( \{ \cdot \} \) without modification. This step provides a near-optimal compression strategy in frequency which is complimentary to the parameter one of Step (1). The model’s empirical interpolant, valid for all parameters, can be written as (cf. Eq. (19) of Ref. \( \{ \cdot \} \))

\[
\tilde{h}(\bar{\theta}; f_i) \approx e^{-2\pi i t_c f_i} \sum_{j=1}^{N} B_j(f_i) \tilde{h}(\bar{\theta}, t_c = 0; F_j),
\]

a sum over the basis set \( \{ B_j \}_{j=1}^N \) and where, for the sake of the discussion below, we have temporarily isolated the coalescence time \( t_c \) from the other parameters.

**Step 3.** All extrinsic parameters, except the coalescence time \( t_c \), do not affect the frequency evolution of the binary and simply scale the inner product \( \{ \cdot \} \), thereby sharing the same ROQs. The coalescence time, however, requires special treatment. Substituting Eq. (4) into Eq. (2),

\[
(d|h(\bar{\theta})|h(\bar{\theta})) = \sum_{k=1}^{N} \omega_k(t_c) \tilde{h}(\bar{\theta}, t_c = 0; F_k),
\]

with the ROQ weights given by

\[
\omega_k(t_c) = 4 \mathcal{R} \sum_{i=1}^{L} \frac{\tilde{d}^i(f_i) B_k(f_i)}{S_n(f_i)} e^{-2\pi i t_c f_i}.
\]

Our approach for the dependence of (6) on \( t_c \) is through domain decomposition: an estimate for the time window \( W \) centered around the coalescence time \( t_{\text{trigger}} \) is given by the GW search pipeline. This suggests a prior interval \( [t_{\text{trigger}} - W, t_{\text{trigger}} + W] \) be used for \( t_c \). The prior interval is then split into \( n_c \) equal subintervals of size \( \Delta t_c \). The number of subintervals is chosen so that the discretization error is below the measurement uncertainty on the coalescence time. Finally, on each subinterval a unique set of ROQ weights is constructed.

Since Step (3) is currently implemented in the LAL pipeline, we summarize it in Algorithm \( \{ \cdot \} \). The offline steps (1) and (2) are carried out independently. Our approach guarantees, though, that those steps need be to carried out only once for each waveform model.

To quickly compute the likelihood we also need an inexpensive rule for \( (h(\bar{\theta})|h(\bar{\theta})) \), whose evaluation no longer depends on the data stream or coalescence time. Con-
sequently, such expressions are typically simple. For example, the norm of the restricted TaylorF2 gravitational waveform model considered below is exactly computable \[19\]. In the general case, building a basis for \(\{h^2\} \) and an associated ROQ provides fast norm evaluations. The TaylorF2 gravitational waveform model’s norm is computable as a 1-term ROQ rule, for example.

By design, weight generation is computed in the startup stage for each detection-triggered data set, requiring \(N\) full inner product \[2\] evaluations for each \(t_c\) interval. This cost is negligible, while each likelihood is subsequently calculated millions of times, leading to significant speedups in parameter estimation studies. The latter scales as the fractional reduction \(L/N\) of the number of terms in the quadrature rules \[2\] and \[5\].

**Parameter estimation acceleration for binary neutron star signals** — The majority of a binary neutron star’s GW signal will be in the inspiral regime \[20\], which can be described by the closed-form TaylorF2 approximation \[21\]. While TaylorF2 does not incorporate spins or the merger-ringdown phases of the binary’s evolution, these should not be important for BNS parameter estimation and can therefore be neglected \[22\]. Even for this simple to evaluate waveform family, inference on a single data set requires significant computational wall-time with standard parameter estimation methods \[6\]. We now report on the anticipated speedup \(L/N\) achieved by ROQ compressed likelihood evaluations. First, we compute the observation time \(T\) required to contain a typical BNS signal. Next, we find the number of reduced basis elements \(N\) needed to represent this model for any pair of BNS masses. In our studies we fix \(f_{\text{high}}\) to 1024Hz while \(f_{\text{low}}\) varies between 10Hz and 40Hz.

The time taken for a BNS system with an initial GW frequency of \(f_{\text{low}}\) to inspiral to 1024Hz,

\[
T_{\text{BNS}} = \left[ 6.32 + 2.07 \times \frac{10^6}{(f_{\text{low}}/\text{Hz})^3 + 5.86 (f_{\text{low}}/\text{Hz})^2} \right] \text{s},
\]

is empirically found by generating a \((1 + 1) M_\odot\) waveform (directly given in the frequency domain) and Fourier transforming to the time domain where we measure the duration up to when the waveform’s evolution terminates. Equation \[7\] and subsequent fits were found using a genetic algorithm-based symbolic regression software, Eureqa \[23, 24\]. The length \(L\), as implied by Eq. \[3\], is plotted in the top panel of Fig. 1.

As discussed, each basis only needs to be constructed over the space of intrinsic parameters — in this case the two-dimensional space of component masses in the range \([1, 4]\) \(M_\odot\). This range is wider than expected for neutron stars, but ensures that the resulting PDFs do not have sharp cut-offs \[25\]. The number of reduced basis required to represent the TaylorF2 model within this range with a representation error around double precision \((\sim 10^{-14})\) can be fit by

\[
N_{\text{BNS}} = 3.12 \times 10^5 (f_{\text{low}}/\text{Hz})^{-1.543},
\]

and is depicted in the middle panel of Fig. 1. We have found that increasing the high-frequency cutoff to 4096 Hz only adds a handful of basis elements, while \(L\) changes by a factor of 4, thus indicating that the speedup for an inspiral-merger-ringdown model might be higher, especially given that not many empirical interpolation nodes are needed for the merger and ringdown regimes \[8\].

Recalling equation \[3\], the speedup from standard to ROQ-compressed likelihood evaluations is given by

\[
\frac{L}{N} \approx (1024\text{Hz} - f_{\text{low}}) \frac{T_{\text{BNS}}}{N_{\text{BNS}}},
\]

with \(T_{\text{BNS}}\) and \(N_{\text{BNS}}\) given by Eqs. \[7\] and \[8\]. This speedup is shown in Fig. 1 (bottom), with a reduction in computational cost and time of \(\sim 30\) for the initial detectors (with a cutoff of \(f_{\text{low}} = 40\text{Hz}\)) and \(\sim 150\) once the advanced detectors reach \(f_{\text{low}} \sim 10\text{Hz}\).
Algorithm 1 Computing the ROQ integration weights

1: **Input:** $d, S_n, \{B_j\}_{j=1}^N, \Delta f, t_{\text{trigger}}, W, \Delta t_c.$
2: Set $n_c = \text{int} \left( \left(2W\right)/\Delta t_c \right) + 1$
3: for $j = 1 \rightarrow n_c$ do
4: $T_j = t_{\text{trigger}} - W + (j - 1) \Delta t_c$
5: for $k = 1 \rightarrow N$ do
6: Compute $\omega_k(T_j)$ via Eq. (6)
7: end for
8: end for
9: **Output:** $\{T_j\}_{j=1}^{n_c}, \{\omega_k(T_j)\}_{k=1}^N \}_{j=1}^{n_c}.$

**Implementation and numerical studies—** We have implemented compressed likelihood evaluations and Algorithm (1) in the LAL parameter estimation pipeline, known as LALInference, naming the resulting variation LALInferenceROQ.

Next we compare MCMC parameter estimation results using the standard version of LALInference to ROQ accelerated studies using LALInferenceROQ. We consider gravitational waveforms emitted from binary neutron star systems with TaylorF2 as the waveform model. We inject synthetic signals embedded in simulated Gaussian noise into the LAL pipeline, for settings anticipating the initial configuration of aLIGO, which should be online within the next two years, using the zero detuned high power PSD, and the initial configuration of AdV, using Eqn. (6) of [28], with in both cases $f_{\text{low}} = 40 \text{ Hz}.$

We take $W = 0.1s$ as the typical time window for the coalescence time $t_c$ of a binary neutron star signal centered around the trigger time $\sim 10^{-5}$. Following the procedure discussed above, LALInferenceROQ discretizes this prior into $n_c = 2,000$ sub-intervals, each of size $\Delta t_c = 10^{-5}s$, for which it constructs a unique set of ROQ weights on each sub-interval. A width of $10^{-5}s$ ensures that this discretization error is below the measurement uncertainty on the coalescence time, which is typically $10^{-3}s$.

We found that, as expected, the ROQ and standard likelihood approaches produce statistically indistinguishable results for posterior probability density functions over the full 9-dimensional parameter space. Figure 2 and Table I describe results for the intrinsic mass parameters obtained in one particular MCMC simulation; other simulations were qualitatively similar. It is also useful to quantify the fractional difference in the 9D likelihood function computed using ROQs and the standard approach. We have observed this fractional error to be

$$\Delta \log \mathcal{L} = 1 - \left( \frac{\log \mathcal{L}}{\log \mathcal{L}_{\text{ROQ}}} \right) \lesssim 10^{-6}$$

in all cases. That is, both approaches are indistinguishable for all practical purposes.

In addition to providing indistinguishable results, ROQ accelerated inference is significantly faster: The

<table>
<thead>
<tr>
<th>$\mathcal{M}<em>c(M</em>\odot)$</th>
<th>$\eta$</th>
<th>$m_1 (M_\odot)$</th>
<th>$m_2 (M_\odot)$</th>
<th>SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>injection</td>
<td>1.2188</td>
<td>0.25</td>
<td>1.4</td>
<td>14.4</td>
</tr>
<tr>
<td>standard</td>
<td>1.2188</td>
<td>0.25</td>
<td>1.4</td>
<td>14.4</td>
</tr>
<tr>
<td>ROQ</td>
<td>1.2188</td>
<td>0.25</td>
<td>1.4</td>
<td>14.4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\mathcal{M}<em>c(M</em>\odot)$</th>
<th>$\eta$</th>
<th>$m_1 (M_\odot)$</th>
<th>$m_2 (M_\odot)$</th>
<th>SNR</th>
</tr>
</thead>
<tbody>
<tr>
<td>injection</td>
<td>0.236</td>
<td>0.240</td>
<td>0.244</td>
<td>0.248</td>
</tr>
<tr>
<td>standard</td>
<td>0.236</td>
<td>0.240</td>
<td>0.244</td>
<td>0.248</td>
</tr>
<tr>
<td>ROQ</td>
<td>0.236</td>
<td>0.240</td>
<td>0.244</td>
<td>0.248</td>
</tr>
</tbody>
</table>

TABLE I. Chirp mass $\mathcal{M}_c$, symmetric mass ratio $\eta$, component masses $m_1$ and $m_2$, and Signal-to-Noise Ratio (SNR) of the analysis from Figure 2. Median value and 90% credible intervals are provided for both the standard likelihood (second line) and the ROQ compressed likelihood (third line). The SNR is empirically measured from Likelihood$_{\text{max}} \approx \text{SNR}^2/2$. The differences between the two methods are dominated by statistics from computing intervals with a finite number of samples. In our analysis, the masses are subject to the constraint $m_1 < m_2$, leading to the true values (where $m_1 = m_2$) being at the edge of the confidence interval.

FIG. 2. Probability density function for the chirp mass $\mathcal{M}_c$ and symmetric mass ratio $\eta$ of a simulated event in LIGO/Virgo data. In green as obtained in $\sim 30$ hours by the standard likelihood, and in blue as obtained in 1 hour with the ROQ. The injection values are in red, and are listed in Table I. The overlap region of the sets of PDFs is the hatched region. Plotting based on [29].

ROQ-based MCMC study with the discussed settings takes $\sim 1$ hour, compared to $\sim 30$ hours using the standard likelihood approach, in remarkable agreement with the expected savings based on operation counts. The wall-time of the analysis is proportional to the total number of posterior samples of the MCMC simulation, which in this case was $\sim 10^7$. The startup stage required to build the ROQ weights has negligible cost and is completed in near real-time, $\sim 30$s, which is equivalent to $\sim 0.028\%$ of the total cost of a standard likelihood parameter estimation study.
For a lower cutoff frequency of $f_{\text{low}} \sim 20$Hz, the speedup reduction is from a couple of weeks to hours. Once the advanced detectors have achieved their target sensitivity, with $f_{\text{low}} \sim 10$Hz, the longest BNS signals will last around 2048s in duration. Assuming a fiducial high frequency cut-off of 1024 Hz, which is approaching the upper limit of where aLIGO/AdV will be sensitive, we estimate datasets as large as $L \sim 1024 \text{Hz}^{-1} \times 2048s \sim 10^6$. Assuming that the advanced detectors will require at least $\sim 10^7$ posterior samples, this implies runtimes upwards of $\sim 100$ days and one Petabyte worth of model evaluations using the standard approach. The results of this Letter indicate that an ROQ approach will reduce this to less than a day. Remarkably, this approach when applied to the advanced detectors operating at design sensitivity will be faster than even the standard likelihood one used for the initial detectors. Additionally, with parallelization of the sum in each likelihood evaluation essentially real-time full Bayesian analysis could be achieved. More details can be found at [30].

**Outlook**– For detectors operating at $f_{\text{low}} = 40$Hz, around three weeks of real (wall) time are needed to perform a precessing-spin parameter estimation study on a single data stream with standard likelihood evaluations [26]. With a cutoff of $f_{\text{low}} = 10$Hz, these analyses could take months to years, so techniques for accelerated inference on these models, such as the one presented in this Letter, are expected to play a central role in gravitational-wave astronomy for extracting the full science potential of the upcoming advanced gravitational-wave detectors.

In this Letter we have addressed the issue of fast likelihood evaluation for non-spinning binary neutron star inspirals. Results of previous work indicate that significant computational savings are to be expected for other waveform models. For example, in Ref. [17] it was found that the number of reduced basis waveforms barely changes as spins are included – at least in the non-precessing case – and, by construction, neither does the number of ROQ evaluations. While waveform evaluation is the dominant cost for models which incorporate precession, recent results [31] have shown that ultra-compact bases can also be constructed for fully precessing systems. This might provide a means for constructing fast to evaluate surrogate models of precessing binary inspirals [5] [6].

**Acknowledgments**– We thank Collin Capano, Peter Diener, Will Farr, Chad Galley, Larry Price, Jorge Pullin, Leo Singer, Alan Weinstein and the LSC CBC and parameter estimation groups for many useful discussions and encouragement throughout this project. PC’s work was supported by a Marie Curie Intra-European Fellowship within the 7th European Community Framework Programme (PIEF-GA-2011-299190). SEF thanks hospitality at and financial support from the Institute of Astronomy at Cambridge, UK, where part of this work was done. JG’s work was supported by the Royal Society and VR by the LIGO Laboratory and the California Institute of Technology (Caltech). This work was supported in part by NSF grant PHY-1208861 to the University of Maryland (UMD), NSF grant PHY-1500818 to the University of California at San Diego, NSF Grants PHY-1306125 and AST-1333129 to Cornell University, and the Sherman Fairchild Foundation. The authors also gratefully acknowledge the support of the United States National Science Foundation for the construction and operation of the LIGO Laboratory under cooperative agreement NSF-PHY-0757058. This Letter carries LIGO Document Number LIGO-P1400038. Some of the computations were carried out at the Center for Scientific Computation and Mathematical Modeling cluster at UMD and the LIGO Laboratory computer cluster at Caltech. Portions of this research were conducted with high performance computing resources provided by Louisiana State University (http://www.hpc.lsu.edu).


