ELECTROMAGNETIC FOLLOW-UP OF GRAVITATIONAL WAVE TRIGGERS AND EFFICIENT PARALLEL-TEMPERED MARKOV CHAIN MONTE CARLO INFERENCE

by William Dominic Vousden

A thesis submitted to the University of Birmingham
for the degree of
Doctor of Philosophy

Astrophysics and Space Research Group
School of Physics and Astronomy
University of Birmingham
September 2015

Abstract

A new generation of ground-based interferometric gravitational wave (GW) detectors is due to begin operation this year, with routine detections anticipated within the next decade. Compact binary coalescences (CBCs), comprising pairs of neutron stars and/or black holes, are among the most promising sources for these detectors. In this work, we focus on two aspects of the science effort in GW astronomy with CBCs.

Firstly, an attractive prospect for GW astronomers, in the wake of a CBC detection, is to observe its electromagnetic counterpart using a conventional telescope. In the first part of this thesis, I investigate our prospects for timely electromagnetic follow-up of such events and the degree to which a galaxy catalogue might aid such observation campaigns.

Secondly, an important aspect of the science effort for GW detections is to efficiently estimate the parameters of the system from which a detected signal originated. In the latter part of this thesis I describe a refinement on existing Bayesian inference techniques used for this purpose. I follow this description with a reference implementation and an application to parameter estimation for CBCs.

Acknowledgements

I am indebted to the following people for their support throughout my PhD: to my supervisors, Ilya Mandel and Will Farr, for their mentorship and patience; to David Stops, for his friendship and indispensable technical wisdom; to the patrons of 11 o'clock coffee, for laughter, biscuits, and respite from real-world problems; and most of all to Miranda Bradshaw for her unwavering support and encouragement.

Some of the work presented here is the result of collaborations and has benefited from discussions with several people. I am grateful to Chad Hanna and Ilya Mandel for their collaboration on the work presented in Chapter 2, and for useful discussions with Darren White, Walter Del Pozzo, Will Farr, Jonah Kanner, Luke Kelley, Trevor Sidery, and Alberto Vecchio. Meanwhile, Chapters 3 and 4 are the fruit of a collaboration with Will Farr and Ilya Mandel, benefiting from discussions with Ewan Cameron and technical assistance from Christopher Berry, John Veitch, Carl-Johan Haster, David Stops, and Paul Hopkins. In particular, I am grateful to Daniel Foreman-Mackey for his work on the *emcee* sampling package, which enabled much of the work in this thesis, and for his help in repackaging it with the modifications described herein.

Finally, I am grateful to Patrick Sutton and Sean McGee for their detailed critique of this thesis, to the Science and Technology Facilities Council for supporting my work, and for computational resources provided by Cardiff University.

Contents

List of figures

List of tables

Chapter 1

Introduction

The first detections of gravitational waves (GWs) by ground-based interferometric GW detectors are anticipated within a decade of this writing as a new generation of detectors begin operation.

While gravitational waves were first predicted almost a century ago, and arise naturally from the Einstein field equations, observational evidence for their existence remained elusive for many decades. In 1993, Russell Hulse and Joseph Taylor were awarded the Nobel Prize in Physics for presenting the first such evidence following their discovery of the binary pulsar PSR B1913+16, a system of two neutron stars whose orbital period is measured from precise timing of the radio emission of one of the neutron stars. Measurements of the decreasing orbital period of this system, spanning some 30 years, have demonstrated striking agreement with the predicted decay from gravitational radiation (Weisberg & Taylor, 2005). Nonetheless, while the orbital decay of the Hulse–Taylor binary is especially convincing evidence for the existence of GWs, we still lack direct measurements.

However, recent technological progress has made the detection of GWs by ground-based interferometers a realistic prospect. This has led to the construction of several kilometre-scale interferometric GW detectors, most notably the LIGO detectors in Hanford, Washington and Livingston, Louisiana (Abbott et al., 2009a) and the Virgo detector in Cascina, Italy (Accadia et al., 2012).

The most promising detection candidates (Aasi et al., 2013b; Abadie et al., 2010b, 2011, 2012a) for these instruments are the signals generated by compact binary coalescences (CBCs): the energetic inspiral and merger of a binary star system comprising either a pair of neutron stars (NSs), a pair of black holes (BHs), or one of each. While the initial designs of LIGO and Virgo were insufficiently sensitive to detect GWs from these sources, upper limits were set on the rate of CBCs in the nearby universe that are consistent with predictions (Abadie et al., 2010a).

However, following recent upgrades, these interferometers – now known as Advanced LIGO and Advanced Virgo (Aasi et al., 2015; Acernese et al., 2015) – are due to start taking data later this year, and are expected to reach design sensitivity in the coming years – an order of magnitude greater than for the initial designs. This will mark the beginning of the "advanced detector era", during which GW detections from CBCs and possibly other sources (Cutler & Thorne, 2002) are expected to become routine, with recent estimates suggesting between one detection per few years to one per few days (Abadie et al., 2010a).

GW detections will offer a new channel through which to observe the universe, along with new insights into a great variety of astrophysical events: from the merging compact binary systems whose transient signals we expect LIGO and Virgo to detect, through to continuous and stochastic sources of gravitational waves, accessible to space-based detectors such as eLISA/NGO (Amaro-Seoane et al., 2012) and galactic-scale pulsar timing arrays (e.g., Hobbs et al., 2010).

The focus of this thesis will be on the compact binary coalescences that LIGO and Virgo – along with similar second-generation detectors planned for Japan and India (Iyer et al., 2012; Somiya, 2012) – hope to detect. These exotic events take place in highly curved space-time, well beyond the Newtonian approximation of gravity, and are therefore ideal probes of strong-field gravitation. Since second-generation detectors are expected to be sensitive to cosmological distances, CBC detections might expose new cosmology and further constrain existing estimates of cosmological parameters (Schutz, 1986). Meanwhile, potential joint electromagnetic observations offer a rich source of information about their astrophysical provenance.

1.1 Measuring gravitational waves

Gravitational waves are perturbations in the metric tensor describing spacetime that propagate outward from their source. This perturbation manifests as a fractional change in length – or strain – at a remote observer, denoted h. This strain presents itself as a transverse wave with two polarisation states, denoted $+$ and $\times$, oriented at 45° to one another. The effect of a GW on a ring of test particles is illustrated in Fig. 1.1. The total strain h induced by a GW can be represented as a linear combination of the strain due to each of these polarisations, so that $h = A_+ h_+ + A_\times h_\times$.

It is this dimensionless strain that is measured by detectors such as LIGO and Virgo. Michelson interferometric GW detectors are laid out so that a passing GW induces a fractional change in the length of each arm. Since the projection of the

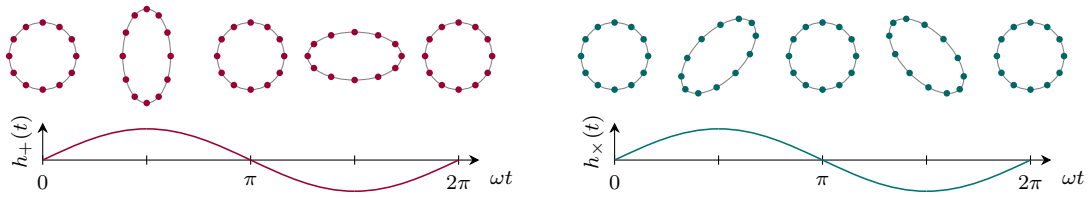

Figure 1.1: The effect on a ring of test particles of a gravitational wave of angular frequency ω propagating along the axis of the ring. On the left is the + polarisation and on the right is the $\times$ polarisation.

transverse strain onto each arm of the interferometer differs, there is a difference in optical path length between them that allows a differential strain h to be measured from the interference pattern that is generated at the output photodiode.

The magnitude of the strain from a GW that can be measured by a detector is limited by the detector's sensitivity. This is characterised by the noise power spectral density (PSD) of the detector, which describes the contribution to noise variations in the strain output per unit frequency. Figure 1.2 shows the square root of the anticipated noise PSDs of Advanced LIGO and its decomposition into some of the expected sources.

By integrating the noise PSD over a frequency band, we can estimate to zeroth order the noise amplitude of the detector in that band. For example, between 10 Hz and 1000 Hz, the noise variations in the output signal of Advanced LIGO, in its zero-detuning high-power configuration (Aasi et al., 2015), will be of order 10^{-22}. Given that the effective arm length[1] of Advanced LIGO is of order 10^5 m, this corresponds to fluctuations in arm length of order 10^{-17} m.

The criteria for a detectable GW signal are somewhat more complicated, however, depending on both the observation time and the spectral composition of the signal in this window. The specifics are discussed later, but this rough calculation indicates the regime in which GWs will be detected.

As illustrated in Fig. 1.2, the total noise in the detector in fact arises from the many separate noise sources that are inevitable in ground-based interferometric detectors. These sources include, among others (Aasi et al., 2015),

(i) seismic motion at the observatory,

(ii) quantum noise arising from Poisson fluctuations in the discrete photon arrival times (also called shot noise),

(iii) thermal noise in the test masses at the ends of each arm caused by absorption

[1]The Advanced LIGO and Virgo detectors use Fabry–Pérot cavities in their arms to increase the effective optical path length.

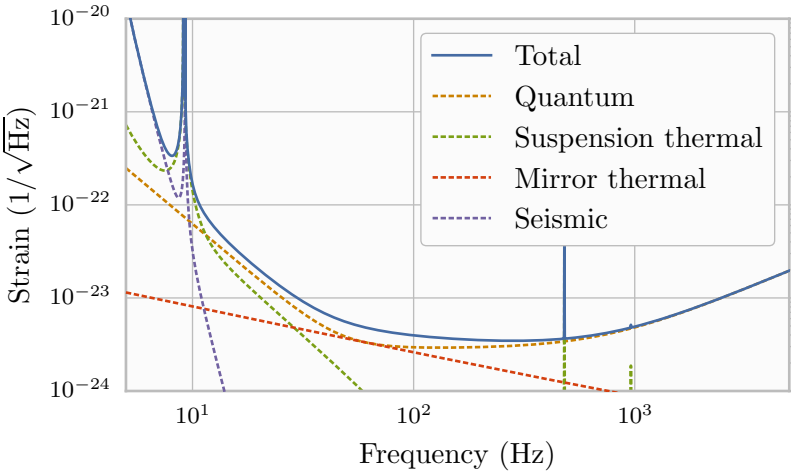

Figure 1.2: The anticipated noise amplitude spectral density for Advanced LIGO in its standard configuration and at design sensitivity (i.e., zero-detuning, high-power; see Aasi et al., 2015; The LIGO Scientific Collaboration, 2010). Also shown are some of the individual noise sources that contribute to the total noise curve (dashed lines). This plot was generated using version 3 of the GWINC tool (Finn et al., 2015).

and dissipation of laser power, and

(iv) thermal noise in the suspension used to isolate test masses from external mechanical noise sources.

Each noise source in the detector is associated with a frequency band over which it is significant. For example, the quantum noise dominates at high frequencies, setting the high-frequency limit of sensitivity. Meanwhile, seismic noise dominates at lower frequencies, setting the low-frequency cut-off for the detector. Note the resonances visible in Fig. 1.2 at 9 Hz and at multiples of $\sim$500 Hz, caused by the vertical stretching mode and violin mode harmonics[2] of the suspension fibre, respectively (Aasi et al., 2015).

1.2 Compact binaries as gravitational wave sources

In the linearised approximation to general relativity, the metric perturbation responsible for GWs is determined by the second time derivative of the mass quadrupole moment of the source. Gravitational radiation is therefore emitted by any system whose mass quadrupole moment has a non-zero *third* time derivative; i.e., whose

[2]The violin mode fundamental frequency is in fact 510 Hz, but limitations in GWINC – the tool used to generate Fig. 1.2 – mean that these resonances appear incorrectly at slightly lower frequencies (Finn et al., 2015).

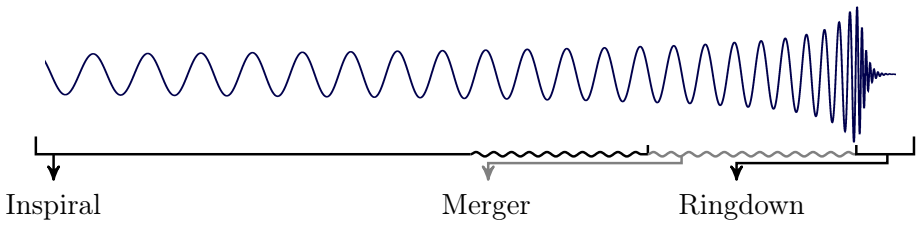

Figure 1.3: The inspiral-merger-ringdown gravitational waveform characteristic of a CBC. This figure is adapted from Ohme (2012).

acceleration is neither spherically nor cylindrically symmetric (Maggiore, 2007).

Since GWs interact with matter to induce motion, they must carry energy away from their source as they are emitted. In the case of a binary system, this energy is extracted from the orbital motion of the component masses, causing the orbital period and separation to decrease, as observed in the Hulse–Taylor binary. Since the frequency of the GWs emitted by an inspiralling binary is proportional to its orbital frequency, the GW frequency consequently increases with the orbital decay of the binary during the inspiral stage of its evolution.

The evolution of a compact binary surrounding its coalescence can be divided approximately into three stages: (i) the "inspiral", during which the binary orbit hardens due to gravitational radiation, (ii) the "merger", when strong-field effects begin to dominate and the component masses combine to form a single object, and (iii) the "ringdown", wherein the resulting object undergoes damped, quasi-normal ringing (Buonanno et al., 2009). Together, these three stages comprise the "inspiral-merger-ringdown" (IMR) gravitational waveform that characterises a CBC. While the exact form of the strain signal $h(t)$ depends strongly on the parameters of the source, and is not known exactly, Fig. 1.3 illustrates its general shape.

During the inspiral stage of a binary system's evolution, the frequency and amplitude evolution of the GW signal measured by a distant observer are determined by the increasing frequency and shrinking separation of the binary's orbit. As the orbit loses energy, it evolves along a path of shrinking quasi-circular orbits[3], modelled by post-Newtonian (PN) expansion of the system's motion in v/c (Buonanno et al., 2009).

When the orbital separation of the objects is comparable to their combined Schwarzschild radius, the initial quasi-circular orbit of the binary becomes unstable due to strong-field effects in the local spacetime geometry. In the case of a test particle orbiting a massive body, this is known as the innermost stable circular orbit

[3]While elliptically inspiralling binaries are also possible, their orbits are likely to circularise by the time they become detectable by current detectors (Peters & Mathews, 1963).

(ISCO) and occurs at an orbital radius of

$$r_{\mathrm{ISCO}} = \frac{6Gm}{c^2}, \tag{1.1}$$

where m is the total mass of the system (Allen et al., 2012; Maggiore, 2007).

In the case of a binary comprising two massive objects, the masses will plunge inward and coalesce at this point in their orbital decay[4]. From Kepler's third law, the orbital frequency f_{s} of the binary at this radius scales as the total mass. Inserting numerical values, the GW frequency at ISCO is, to leading order,

$$(f_{\mathrm{gw}})_{\mathrm{ISCO}} = 2(f_{\mathrm{s}})_{\mathrm{ISCO}} \approx 4.4\,\mathrm{kHz}\left(\frac{\mathrm{M}_\odot}{m}\right). \tag{1.2}$$

For a fiducial GW frequency f_{gw}, the time τ until coalescence is given, to leading order, by

$$\tau \approx 6.47 \times 10^5\,\mathrm{s}\left(\frac{\mathrm{M}_\odot}{\mathcal{M}}\right)^{\frac{5}{3}}\left(\frac{\mathrm{Hz}}{f_{\mathrm{gw}}}\right)^{\frac{8}{3}}, \tag{1.3}$$

where $\mathcal{M} \equiv (m_1 m_2)^{3/5}/(m_1 + m_2)^{1/5}$ is the "chirp mass" of the system (Allen et al., 2012; Maggiore, 2007).

For example, the low-frequency cut-off in the Advanced LIGO sensitivity curves is at around 10 Hz. Therefore, for a system of two $1.4\,\mathrm{M}_\odot$ neutron stars, Advanced LIGO will be sensitive to the final 17 minutes or so of the inspiral signal before the merger occurs. From Eq. (1.2), the GW frequency at this point will be approximately 1.6 kHz: close to the high-frequency cut-off.

For the merger of non-spinning component masses, the quasi-normal ringing frequency of the remnant Schwarzschild black hole also scales as the total mass of the system, such that

$$f_0 \approx 12\,\mathrm{kHz}\left(\frac{\mathrm{M}_\odot}{m}\right), \tag{1.4}$$

while the exponential decay time-scale is $2/\pi f_0$ (Abbott et al., 2009b; Berti & Cardoso, 2006). The binary neutron star (BNS) system considered above therefore rings at approximately 4.3 kHz with a decay time-scale of 150 µs. Corrections are required for the Kerr black hole that is produced by the merger of spinning component masses (Berti & Cardoso, 2006).

Between inspiral and ringdown, the merger itself occurs in the strong-field regime of gravity, where the system's dynamics are analytically intractable. This part of the waveform is therefore modelled with numerical relativity (NR) (e.g., Ohme, 2012).

[4]While the ISCO is only well-defined for a test particle, it nonetheless adequately approximates the point of merger for massive systems.

Which parts of the complete IMR waveform are relevant for detection and analysis depends largely on the total mass of the system. Since both the ISCO and ringdown frequencies are inversely proportional to m, higher mass systems will merge at lower frequencies and spend less of their inspiral in the detector's sensitivity band. For these systems, therefore, the merger and ringdown components of their waveforms will be more important than for lower mass systems.

1.3 Gravitational wave data analysis

Detecting a gravitational wave

Before analysis of a GW event can begin, it must first be identified in the stream of data that is produced by a GW detector. We denote the detector output by the time series $s(t)$, which – in the presence of a GW signal – we can decompose into a sum of the signal $h(t)$ and the noise $n(t)$, so that

$$s(t) = h(t) + n(t), \tag{1.5}$$

where s, h, and n have Fourier transforms $\tilde{s}$, $\tilde{h}$, and $\tilde{n}$ respectively.

To identify putative events in the data s, we filter it in a way that maximises a given detection statistic. If the detection statistic generated by the filter exceeds a threshold that is chosen beforehand, we claim a detection.

Since the form of the GW signal from a CBC is well-modelled (Buonanno et al., 2009), the conventional approach is to use "matched filtering", wherein a known template signal u is correlated with the data s to detect the presence of the template.

The detection statistic in this case is the signal-to-noise ratio (SNR), defined as the ratio of the power in s due to h to the power in s due to n, to wit

$$\mathrm{SNR} = \frac{\langle u|s\rangle}{\sqrt{\langle u|u\rangle}}, \tag{1.6}$$

where u is a template GW signal, $\langle \cdot | \cdot \rangle$ is the noise-weighted inner product defined by

$$\langle a|b\rangle = 4\,\mathrm{Re}\int_{-\infty}^{\infty} \frac{\tilde{a}(f)^*\,\tilde{b}(f)}{S_n(f)}\,\mathrm{d}f, \tag{1.7}$$

and S_n is the two-sided noise PSD that describes n. If the true signal present in s is h and the noise in the system is both stationary and Gaussian, then the expectation of the SNR over all noise realisations is maximised by setting $u \propto h$; this is the "matched" filter.

In this simple model, a threshold is set on the SNR required to claim a detection. If this threshold is too low, the detection pipeline will report many false positives that are in fact noise artefacts, while if it is too high, genuine signals in the noise will be missed. While the choice of this threshold depends, among other things, on the detector network configuration, a reasonable threshold SNR is of order 10 for a false alarm rate of order $1\,\mathrm{yr}^{-1}$ (Abadie et al., 2012a; Maggiore, 2007).

In reality, of course, the detector noise is generally non-stationary and non-Gaussian, and significant effort is spent on optimising searches for signals obscured by more realistic noise. These methods include, for example, signal quality checks and chi-squared discriminants that reject spurious responses of the matched filter to instrumental artefacts in the data stream (e.g., Allen et al., 2012; Babak et al., 2013; Cannon et al., 2012).

The template GW signal $u(t;\vec{\theta})$ is a function of the many physical parameters contained in $\vec{\theta}$ (described on page 10), so we require a set of templates evaluated at different points in the parameter space – called a template bank – to match against the detector data s. Since the optimal SNR is achieved with h itself, it is important that the template bank can accurately represent the GW signals we expect to encounter in the detector output. A maximum fractional loss in SNR that is acceptable for the search (typically $\sim 3\,\%$) is used to decide the points on which the template bank is constructed (Balasubramanian et al., 1996; Owen, 1996).

Estimating the parameters of a compact binary coalescence

The main scientific value of a GW detection lies in the information about its source parameters that is encoded in the signal.

We can extract some of this information by examining the template waveform that yielded the highest SNR by matched filtering when the signal was detected. This corresponds to the maximum likelihood estimator (MLE) for the signal parameters under the model that a signal is present in the detector strain data. Since the template waveforms $u(t;\vec{\theta})$ are parameterised by the GW source parameters $\vec{\theta}$, we can simply read off the parameter values that maximise the SNR.

While this method provides a point estimate of $\vec{\theta}$ at no computational cost beyond that of detecting the event in the first place, it conveys no information about the uncertainty on these values, which is needed to draw useful conclusions about the physics of the events; neither does it allow us to express any prior belief about the values of the source parameters. For example, we might expect sources to be distributed uniformly in volume – and therefore with density proportional to r^2 – but we lack a mechanism to express this bias.

Instead, we are motivated to construct the full probability density function (PDF) of $\vec{\theta}$, given the strain data s, that accounts for prior beliefs on $\vec{\theta}$. To this end we turn to Bayes's theorem, through which we can express the "posterior" PDF $p(\vec{\theta}|s, H)$ in terms of the "likelihood" $p(s|\vec{\theta}, H)$, the "prior" $p(\vec{\theta})$, and the "evidence" $p(s|H)$, such that

$$p(\vec{\theta}|s, H) = \frac{p(s|\vec{\theta}, H)\, p(\vec{\theta})}{p(s|H)}, \tag{1.8}$$

where H is a model, parameterised by $\vec{\theta}$, describing the presence of a signal in s.

The most important element in the Bayesian formalism is the likelihood function, usually denoted $L(\vec{\theta})$ when we are not concerned with other models than H. Assuming a noise model for the detector, the noise realisation n is then a random variable with a known distribution, from which we can define $L(\vec{\theta})$. If we approximate the noise in the detector as stationary and Gaussian, then the likelihood function can be written in terms of $n = s - h$, so that

$$\log L(\vec{\theta}; s) = -\frac{1}{2} \left\langle s - h(\vec{\theta}) | s - h(\vec{\theta}) \right\rangle , \tag{1.9}$$

where the inner product is defined as in Eq. (1.7).

This likelihood function therefore relies on the evaluation of a waveform approximant $h(\vec{\theta})$. Most approximants do not have analytical forms, but are instead costly solutions to differential equations and, since the inner product is defined in the frequency domain, many must also be Fourier transformed in order to compute L (e.g., Buonanno et al., 2009). Waveform computations therefore tend to dominate the computational expense of parameter estimation.

The CBC waveforms that define the posterior are described by between 9 and 15 physical parameters. With the high dimension of this parameter space and the cost of computing waveforms, the curse of dimensionality ensures that we cannot reconstruct the posterior PDF simply by evaluating it over a regular grid. We must instead turn to stochastic techniques that concentrate sample points in regions of the parameter space containing the bulk of the probability mass. Two such methods are nested sampling (Skilling, 2006) and Markov chain Monte Carlo (MCMC), both of which have been applied to the problem of CBC parameter estimation (Veitch et al., 2015).

An inspiralling binary with arbitrarily spinning component masses generates a gravitational waveform that is described by 15 physical parameters, which can be partitioned into two disjoint groups. One possible parameterisation, which is favoured for parameter estimation purposes (Veitch et al., 2015), is the following.

Firstly, the dynamics of the system are described by up to 8 "intrinsic" parameters, comprising

(i) q and $\mathcal{M}$: the mass ratio and chirp mass of the binary, system,

(ii) a_1 and a_2: the spin magnitudes of the binary components, and

(iii) t_1, t_2, ϕ_{JL}, and ϕ_{12}: the four angles describing the spin orientations of the binary components.

Secondly, the location and orientation of the binary with respect to the detector network are described by a further 7 "extrinsic" parameters, viz

(iv) d_L: the luminosity distance between the binary system and the detector,

(v) α and δ: the right ascension and declination of the event,

(vi) ψ and θ_{JN}: the two angles describing the orbital orientation of the binary system, and

(vii) t_{c} and ϕ_{c}: a reference time and phase for the waveform.

If the spins of the binary's component masses are aligned with the orbital axis, then the spin angles t_1, t_2, ϕ_{JL}, and ϕ_{12} may be discarded, leaving 11 free parameters. Likewise, if the system is non-spinning, the spin magnitudes a_1 and a_2 may also be neglected, and the parameter space is reduced to 9 dimensions.

Despite methods such as nested sampling and MCMC that are well-suited to high-dimensional problems, the posterior distribution generated by a CBC gravitational waveform remains difficult to sample, with many correlations and well-separated modes.

For example, under MCMC, samples are generated from a random walk in the parameter space of the target distribution such that the density of samples is proportional to the target probability density. Such a random walk is prone to getting "stuck" on the first mode it finds, being unable to traverse regions of very low probability density in order to find other modes.

In the case of MCMC, one solution to this problem is parallel tempering (Earl & Deem, 2005; Geyer, 1991; Swendsen & Wang, 1986). In this formalism, several chains sample independently from "tempered" versions of the posterior distribution, with "cold" chains being efficient at sampling individual modes, and "hot" chains being able to migrate between them (in close analogy to a particle moving between potential wells). Swaps are proposed periodically between temperatures to enable samples on cold chains to efficiently explore multi-modal parameter spaces.

The details of parallel tempering and a new strategy for selecting temperatures are the subject of Chapters 3 and 4. This strategy improves the performance of

parallel tempered samplers and their robustness against multi-modal distributions, for example those generated by GW signals from CBC events.

1.4 Electromagnetic follow-up of compact binary coalescences

While the anticipated detection of CBCs and the extraction of their parameters is an attractive prospect in its own right, a major science objective for GW astronomers is to observe the coincident electromagnetic (EM) counterparts that are expected to accompany of them. These transients, which are expected to span most of the EM spectrum on a broad range of time-scales (Metzger et al., 2013), encode a wealth of astrophysical information about the merger process to complement that carried by the GW strain signal.

For example, the GWs emitted by a CBC encode the luminosity distance to the event (Schutz, 1986). While the redshift measurement from the GW signal is degenerate with the mass of the binary system, an EM counterpart observation might provide an independent spectroscopic redshift measurement – for example, by association with a host galaxy (Metzger & Berger, 2012). CBCs could therefore provide a new class of "standard candle" and, with it, independent measurements of the Hubble constant and other cosmological parameters (Chernoff & Finn, 1993; Holz & Hughes, 2005; Nissanke et al., 2010; Schutz, 1986).

An important class of EM transient that is expected to accompany some CBCs is a short γ-ray burst (SGRB) (Eichler et al., 1989; Nakar et al., 2006; Narayan et al., 1992). Spanning the seconds following the merger of a BNS or neutron star/black hole (NSBH) system, an SGRB is a collimated jet of γ radiation predicted to be emitted along the orbital axis of the binary. Given their temporal coincidence with the GW signal expected from such an event, a simultaneous SGRB observation and GW detection would confirm current suspicions (Fong & Berger, 2013) of compact object binaries as progenitors of SGRBs.

There are, broadly, two categories into which joint EM and GW detections can be grouped: (i) GW searches triggered by EM observations and (ii) GW-triggered EM follow-up searches.

In the first instance, an advance observation of an EM transient can be used to inform a search for the accompanying GW signal. If the EM transient can sufficiently constrain the coalescence time of the binary, a targeted GW search can be performed on the appropriate stretch of detector data (Abadie et al., 2010c; Finn et al., 1999; Mohanty et al., 2004), allowing a lower threshold on the SNR required to claim a

detection (Kochanek & Piran, 1993, but see Kelley et al., 2013). The viability of this strategy depends both on the precision to which the coalescence can be located in time from the EM transient and on the fraction of transients that are accompanied by a detectable GW signal. While the $\mathcal{O}$(seconds) duration of SGRBs allows a precisely targeted GW search, most will occur outside the sensitivity volume of Advanced LIGO/Virgo (Kelley et al., 2013; Metzger & Berger, 2012). Kelley et al. (2013) instead argue that a more promising route to such a GW detection might be the optical and near-infrared emission of r-process heating in the merger ejecta – commonly known as a "kilonova" – discussed shortly.

Conversely, since the GW signal from a CBC encodes its location on the sky (typically to within 10–100 deg^2, e.g., Aasi et al., 2013a; Nuttall & Sutton, 2010; Sidery et al., 2014; Singer et al., 2014), such a detection could allow for a direct observation of accompanying EM counterparts with a co-ordinated telescope follow-up campaign (e.g., Aasi et al., 2014). This strategy allows the joint detection of less energetic EM counterparts, predicted to have longer post-merger delays, more extended light curves, and less beamed emission than SGRBs.

Potential electromagnetic counterparts

While SGRBs are exceptionally bright, their collimation means that they will accompany only a small fraction of detected CBC events as a visible counterpart. Specifically, Metzger & Berger (2012) estimate, from previous observations (e.g., Burrows et al., 2006; Soderberg et al., 2006) and for consistency with predicted GW merger detection rates (Abadie et al., 2010a), that the jet half-opening angle θ_j illustrated in Fig. 1.4 is $\sim 7°$. Including a GW Malmquist bias toward face-on binaries of a factor of ~ 3.4 (Kelley et al., 2013), approximately 2 % of detected CBC events would then be accompanied by an observable SGRB (assuming all mergers generate such bursts).

Following the SGRB an "afterglow", generated by the interaction of the axial jet with the surrounding interstellar medium (ISM), is visible between optical and radio wavelengths on time-scales of days (optical) to months (radio) (e.g., Berger, 2010; Nakar & Piran, 2011; Perley et al., 2009; van Eerten & MacFadyen, 2011). This afterglow is less beamed than the SGRB that precedes it, with the beam opening angle increasing with the wavelength and post-merger delay. Nonetheless, the accessibility of afterglows will be limited by their beaming angle in the optical band, with ~ 10 % being detectable, and by magnitude in the case of almost-isotropic radio afterglows (Metzger & Berger, 2012).

Perhaps a more promising EM counterpart, however, is the (predicted) optical

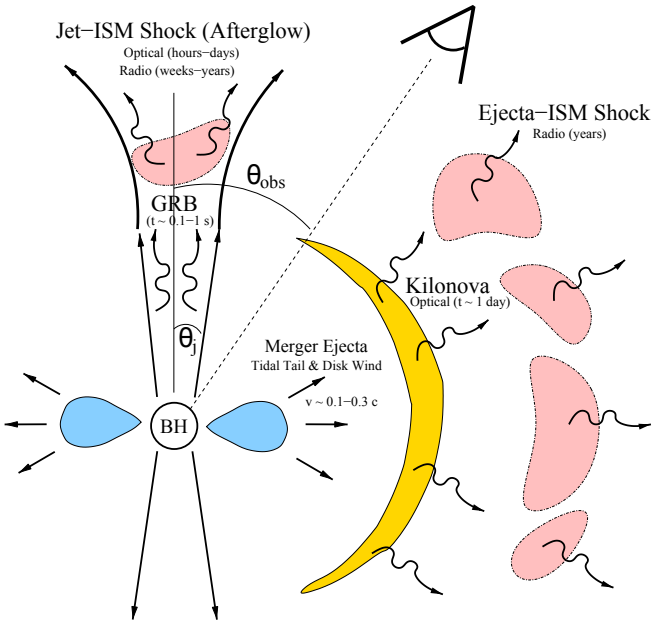

Figure 1.4: A schematic of the four most promising EM counterparts expected to accompany a BNS or NSBH merger. This figure is reproduced from Metzger & Berger (2012).

and near-infrared "kilonova" emitted by the radioactive decay of heavy nuclei following r-process nucleosynthesis in the merger ejecta (Kasen et al., 2013). This emission occurs on the time-scale of $\mathcal{O}$(days): short enough for confident association with the GW, but long enough to relax the latency requirements between detecting the GW signal and pointing telescopes. Unlike SGRBs and afterglows, kilonovae are both isotropic and independent of the density of the ISM in the neighbourhood of the merger. However, hints from recent observations (Berger et al., 2013; Tanvir et al., 2013) suggest that kilonovae may be predominantly infrared due to optical extinction in the merger ejecta. In this case, they may prove very difficult to observe serendipitously without direction from an associated EM observation, given the narrow field-of-view (FOV) of most infrared telescopes with respect to the sky localisation area from a GW CBC detection.

Each of the EM emission mechanisms mentioned above presents its own merits and difficulties for this purpose, a detailed analysis of which is presented by Metzger & Berger (2012). However, for the purposes of EM follow-up searches for GW trigger events, Metzger & Berger suggest that the optical and near-infrared bands will offer the most confident and frequent joint GW–EM observations.

1.5 Overview of the thesis

Chapter 2

In Chapter 2 I present a paper (Hanna et al., 2014) written in collaboration with Chad Hanna and Ilya Mandel that addresses the utility of galaxy catalogues in directing EM follow-up searches.

This chapter considers the prospect of following up a GW detection of a BNS merger with an observation of one of its EM counterparts, and the challenges associated with such an effort. In particular, a major obstacle for the GW astronomy community in obtaining a joint observation of this format is the allocation of telescope resources that are severely limited with respect to the large area of the sky that must be imaged.

Typical uncertainties on the sky location recovered from a CBC detection from advanced detectors will be of order 10–100 deg^2 (Sidery et al., 2014). In contrast, current and upcoming wide-field optical telescopes have FOVs of 0.1–1 deg^2 (e.g., Singer et al., 2012, and references therein), so complete and timely coverage of the GW error region on the sky is impractical.

Fairhurst (2009) and Kanner et al. (2008) suggest that a galaxy catalogue could be used to identify potential host galaxies for an event and therefore to reduce the observational resources required to confidently image an associated EM counterpart. However, while there are few enough galaxies within the sensitivity volume of the initial LIGO/Virgo network that they can be imaged individually (Abadie, J. et al., 2012; Abadie et al., 2012b; Kopparapu et al., 2008), the catchment volume of the Advanced LIGO/Virgo network will contain $\sim 10^3$ times as many galaxies.

This chapter examines the usefulness of a galaxy catalogue in the regime where there are many galaxies per telescope FOV – as we anticipate for advanced detector networks – and the speed and depth required of the survey make complete coverage of the error region impractical. In this case, statistical fluctuations in the distribution of galaxies across the sky can still improve prospects for imaging an EM counterpart, even in the case of an incomplete catalogue.

Chapter 3

In this chapter I present a second paper (Vousden et al., 2015), written in collaboration with Will Farr and Ilya Mandel, that describes a new method for improving the efficiency of parallel tempered MCMC sampling, described in Section 1.3.

The "golden question" in the application of parallel tempering is how to choose

an appropriate set of temperatures at which to sample the posterior distribution. Since the optimal temperature allocation depends strongly on the likelihood that is being sampled, there is no universal prescription for an effective set of temperatures.

In Chapter 3 we examine this problem in detail and develop a method for selecting temperatures dynamically as a sampler explores the likelihood distribution. We present an implementation of the method in the ensemble-based MCMC sampler *emcee* (Foreman-Mackey et al., 2013) and show that the method increases the efficiency of the sampler, as measured by its autocorrelation time (ACT).

Chapter 4

The parameter estimation for GW signals produced by CBC events is an important problem and the subject of significant effort in the field of GW astronomy. This Bayesian inference problem – described earlier in Section 1.3 – requires the characterisation of complicated, multimodal probability distributions in high-dimensional parameter spaces. Parallel tempering is therefore well-suited to this problem, but current implementations for CBC parameter estimation lack a robust method of temperature selection, which is critical to their efficacy.

In Chapter 4 I extend the work presented in Chapter 3 by applying the method to the problem of CBC parameter estimation, under the *LALInference* library that was developed for this purpose (Veitch et al., 2015).

The first step in this application was to implement the temperature selection algorithm of Chapter 3 under the *LALInference* MCMC sampler. I describe this implementation and the adjustments that were made to *LALInference* in its service.

Secondly, I describe and present tests of this implementation on synthetic but astrophysically relevant CBC events. These tests consistently demonstrate improvements in the efficiency of the sampler as measured by its ACT. However, the tests also expose problems with the assumption made in Chapter 3 that uniform acceptance ratios between all adjacent pairs of temperatures will always lead to efficient – if not optimal – sampling. I present the results of diagnostic tests, a plausible explanation of this behaviour, and suggestions for how one might address the problem.

Chapter 2

Utility of galaxy catalogues for electromagnetic follow-up of gravitational waves from binary neutron star mergers

The text and figures that follow are reproduced from Hanna et al. (2014), a paper written in collaboration with Chad Hanna and Ilya Mandel that addresses the utility of galaxy catalogues in directing electromagnetic follow-up searches.

In this project, I contributed in roughly equal measure with my co-authors to writing the text of the paper, and was responsible for generating the plots and writing the supporting code. The idea for the project was IM's, while the discussion, interpretation, and editing were a joint effort between myself, IM, and CH. The author list is arranged alphabetically to reflect equal contributions from all three authors.

2.1 Introduction

Abadie, J. et al. (2012); Abadie et al. (2012b) present the first *low-latency* searches for gravitational waves (GWs) that triggered electromagnetic (EM) follow-up observations with a $\sim 30\,\mathrm{min}$ response time[1]. No GWs were detected, but several GW candidate events consistent with noise were followed up with telescopes at a variety of wavelengths (Aasi et al., 2014; Evans et al., 2012). Later this decade, a network of advanced GW detectors including LIGO and Virgo (Aasi et al., 2015; Acernese et al., 2015) may detect tens of binary neutron star (BNS) mergers per year once

[1] $\lesssim 5\,\mathrm{min}$ not including a human intervention time scale that could be removed if automated.

at full sensitivity (with a plausible range of one detection in a few years to a few hundred detections per year) (Abadie et al., 2010a). Some of these detections may be accompanied by EM counterparts (e.g., Bloom et al., 2009; Kelley et al., 2013; Metzger & Berger, 2012), summarised below[2]. Several nearly "instantaneous" search methods for GWs from BNS mergers have been proposed, introducing the possibility of transmitting information about the candidate to EM telescope partners within tens of seconds of a binary merger (Cannon et al., 2012; Luan et al., 2012).

BNS mergers are thought to generate several distinct EM counterparts spanning most of the EM spectrum; Fig. 1.4 illustrates the counterpart emission mechanisms. Short, hard gamma ray bursts occur on timescales of $\lesssim 2$ seconds (Nakar et al., 2006) and are strongly beamed. Afterglows from shock waves produced when the emitted jet encounters the interstellar medium span the spectrum from X-rays to radio waves (e.g., Berger, 2010; Nakar & Piran, 2011; Perley et al., 2009; van Eerten & MacFadyen, 2011). Thermal emission from r-process nucleosynthesis in the merger ejecta has been predicted to peak in the infrared (Kasen et al., 2013); the first hint of such a kilonova signal has been recently observed (Tanvir et al., 2013).

Several transient telescope networks exist with wide-field coverage and it is important to understand what is the best way to tile pointings within the GW localisation region using wide-field instruments. This question has been addressed partly by Singer et al. (2012), who present a framework for allocating telescope resources to optimally cover the available sky localisation region. In this work we consider the situation in which only a fraction of this area can be surveyed in a timely fashion, where it is important to choose the tiles that represent the most likely source location first. Both Singer et al. (2012) and Fairhurst (2009) focus on the assumption of a uniform-on-the-celestial-sphere prior on the GW source location. However, given the broad GW localisation region, pointing might be strongly influenced by a sharply peaked prior expectation for the signal location. Kanner et al. (2008) and Fairhurst (2009) mention that a galaxy catalogue could serve as a better prior and, indeed, Abadie, J. et al. (2012); Abadie et al. (2012b) demonstrate that the use of a galaxy catalogue (Kopparapu et al., 2008) greatly increases the chance of imaging an EM counterpart in simulations for the initial GW detectors with $\lesssim 20$ Mpc range for GWs from merging BNSs. At this range, nature provides few galaxies as potential hosts for the merger, corresponding to sharp peaks in the prior probability.

[2]While we focus on BNS mergers in this paper, mergers of neutron star/black hole (NSBH) binaries may also produce electromagnetic counterparts, provided tidal forces remove a significant amount of material from the neutron star before it plunges into the black hole. The BNS analysis presented here also applies to these mixed binaries, although typical mixed-binary detections will happen at greater ranges, making a catalogue less useful.

The same angular scale will encompass many more galaxies in the advanced GW detector era and the usefulness of a galaxy catalogue prior comes into question. Metzger & Berger (2012) suggest that the number of bright galaxies in the localisation region will be too large to improve the prospects of imaging the EM counterpart. For example, GW detections in the advanced detector era will occur at a median distance of $\sim 200\,\mathrm{Mpc}$. A source at this distance may be optimistically localised to a sky area of $20\,\mathrm{deg}^2$ and a fractional distance error of $\sim 30\,\%$ by GW measurements alone (Aasi et al., 2013a; Fairhurst, 2009; Nissanke et al., 2011; Rodriguez et al., 2014; Veitch et al., 2012) with a network of three or more GWs detectors[3]. The volume defined by this solid angle and distance range will contain more than 500 galaxies brighter than $0.1L^*$ (see Section 2.3) – more than can realistically be imaged individually on short timescales.

If wide-field instruments are used to tile the GW localisation region and the requirements on the speed and depth of the search make it impossible to follow up the entire localisation region, the question arises of how to prioritise which tiles should be observed. Nuttall & Sutton (2010) partly address this problem by simulating follow-up searches within 100 Mpc in the advanced detector era using the Gravitational Wave Galaxy Catalog (GWGC) of White et al. (2011). Individual galaxies are targeted on the basis of a ranking algorithm that accounts for luminosity and distance to putative host galaxies. Meanwhile, Nissanke et al. (2013) provide case studies for the process of detecting a GW event and locating and identifying its EM counterpart, using galaxy catalogues to eliminate false-positive EM signals. Both studies find that catalogues can be useful both for locating and identifying an EM counterpart when there are insufficient resources to point individually at each galaxy.

In this work, we revisit the utility of a galaxy catalogue in the regime where there are too many galaxies in the GW localisation region to be followed up individually, and observational constraints on the speed and depth of the search prevent complete coverage with wide-field instrument pointings. We quantify the utility of a galaxy catalogue as a function of the three-dimensional volume within the field-of-view (FOV) of the follow-up telescope (after accounting for the distance measurement uncertainty from GW measurements) and of the fraction of the GW localisation region that can be covered. We consider realistic catalogues, which are likely to be significantly incomplete within the large volumes in which advanced detectors are sensitive.

We find that even in the advanced-detector era, galaxy catalogues can still confer

[3]With significantly larger uncertainties expected for a two-detector network for the early runs of Advanced LIGO alone (Aasi et al., 2013a; Kasliwal & Nissanke, 2014).

benefits through the inherent fluctuations in luminosity density on the sky. Galaxy luminosity and count fluctuations will help to prioritise tiles and increase the relative probability of imaging a GW EM counterpart. In particular, we will show that catalogues are most relevant for narrow and shallow follow-up searches (that is, smaller FOVs and shorter ranges) and that improving the completeness and range of existing catalogues is important for EM follow-up efforts.

This work is organised as follows. In Section 2.2 we define our condition for a "successful" follow-up and describe the algorithm we use to select tiles for pointing, given a galaxy catalogue. In Section 2.3 we discuss the characteristics of the galaxy luminosity distribution and show that there can be significant variations in luminosity between tiles. Sections 2.4 and 2.5 present the results of simulated follow-up searches for several detection and observation scenarios and discuss the effects of incompleteness of the galaxy catalogue on the results. Section 2.6 concludes with a brief discussion of our results and additional suggestions for future work.

2.2 Problem statement

In this work we are not concerned specifically with identifying host galaxies, but rather with choosing the most probable sky regions commensurate with a given FOV by using galaxy catalogue information. We neglect many of the practicalities considered by Nissanke et al. (2013) and Singer et al. (2012) (e.g., telescope slew time, limiting depth, day/night observation time, etc.) to isolate the utility of galaxy catalogues on their own merits. We do, however, assess the effect of incompleteness of galaxy catalogues in our method.

Throughout this work we will use a blue-band galaxy catalogue as a proxy for merger rate density. This assumes that the rate of BNS mergers is proportional to the instantaneous massive star formation rate (with negligible time delays between formation and merger) and is therefore tracked by blue-light luminosity (Phinney, 1991). On the contrary, observational evidence indicates that a quarter of short gamma ray bursts occur in elliptical galaxies with no signs of ongoing star formation (Fong & Berger, 2013). However, the choice of colour is not critical for the modelling below; it is sufficient to assume that we have a catalogue that is an accurate tracer of merger rate density. We discuss the validity of this assumption in Section 2.6.

To model the effect of using galaxy catalogues to assist in EM follow-up we begin by dividing the GW localisation area, A, into $\mathcal{N}$ tiles (assumed to be non-overlapping for simplicity), each representing a telescope FOV P, where $\mathcal{N} = \lceil \frac{A}{P} \rceil$.

We define a **successful follow-up** as a GW-triggered EM transient search in

which one of the tiles selected for imaging contains the GW source. For simplicity, we require only that the source *reside* in one of the tiles, and not that the expected EM counterpart is actually *detectable* by a given follow-up instrument[4] or distinguishable from background events. We therefore assume the transient search to be limited in range only by the capabilities of the GW detector network and not by the depth of the follow-up instrument. Considering the above assumptions, the probability of success is 1 if all tiles in the sky are searched, regardless of whether the correct transient is identified.

In practice, sky localisation from GW data will yield regions of non-uniform probability on the sky; in the high signal-to-noise ratio (SNR) limit, the probability distribution on the sky will have a Gaussian shape. The probability density function on the sky will be computed through coherent parameter estimation on GW detector data (Aasi et al., 2013c). Here, we treat the event localisation area A as a suitable "effective" area, and consider the GW localisation probability to be uniform over A.

We define the success fraction $\mathcal{F}$ as the fraction of GW events that are expected to be successfully followed up for a given follow-up strategy according to the definition above. If one ignores the galaxy distribution, the relative probability that a GW is in a given tile is uniform amongst the tiles. The success fraction is

$$\mathcal{F} = \sum_{i}^{N} \frac{1}{\mathcal{N}} = \frac{N}{\mathcal{N}} \equiv f, \tag{2.1}$$

where N is the number of telescope pointings compatible with search speed and depth requirements, and f is simply the fraction of the GW localisation area[5] that is followed up, $f = N\frac{P}{A}$.

This should be compared to the case where each tile has a relative probability of containing the GW event proportional to its blue light luminosity L_i, and a greedy pointing algorithm is used whereby the brightest tiles are pointed at first, $L_i \geq L_{i+1}$

[4]In fact, the depth to which the available telescopes can detect an EM transient may influence the optimal choice of follow-up target. For example, there is little point in targeting galaxies that are so distant that the transients they might contain would not be detectable by a given telescope. Moreover, for consistency with recent observations (Tanvir et al., 2013), we might expect the kilonova generated by a BNS merger at a luminosity distance of 200 Mpc to peak in the optical band at $\sim$23 mag. With currently available wide-field telescopes, the $\sim$50 % of events occurring *beyond* this distance will be difficult to observe (Metzger et al., 2013). Given a high distance measurement from the GW signal, we might therefore choose not to follow up events at extreme distances at all. As we shall see in Section 2.4, it is in this regime that a galaxy catalogue is least helpful anyway.

[5]Realistically, this area will in fact be a Bayesian credible region on the sky containing a set fraction C of the marginal sky location probability mass. Therefore, the success fractions in Eqs. (2.1) and (2.2) should, strictly, be multiplied by C.

for all i:

$$\mathcal{F} = \frac{1}{L} \sum_{i}^{N} L_i \,, \tag{2.2}$$

where $L \equiv \sum_i^{\mathcal{N}} L_i$. With the greedy strategy, $\mathcal{F} \geq f$; in other words, if GW sources are distributed according to blue luminosity then using that information never hurts the success fraction.

The GW amplitude depends on the inclination and orientation of the source relative to the line of sight, with the highest detector response for face-on sources (e.g., Finn & Chernoff, 1993; Kelley et al., 2013). This allows us to compute the probability that a source in a given galaxy at a known distance and sky location would pass a signal-to-noise-ratio detection threshold under the assumption that the binary's inclination and orientation are isotropically distributed. This probability decreases from ~ 1 for a very nearby galaxy to 0 for a galaxy at the maximum distance for a given sky location and detector network configuration (where only face-on sources would be detectable); the decrease is roughly linear in the distance to the source (cf. the ad hoc weighting by Nuttall & Sutton (2010) of galaxies by one over distance or one over distance squared). In principle, this detection probability should be included in the prior weighting of galaxies in the catalogue, giving each galaxy an effective luminosity that is the product of its actual luminosity and the probability that a source in this galaxy would be detectable in a GW search with the given detector network.

However, in practice, the analysis of GW data will yield (strongly correlated) constraints on distance and inclination, so this prior probability should not be assigned independently of the detector data. As discussed in Section 2.6, the correct approach would be to include the galaxy catalogue directly in coherent Bayesian parameter estimation as a prior, which would allow for a self-consistent application of all information, rather than attempting an a posteriori correction as we are doing here. However, for the purposes of estimating the utility of a galaxy catalogue, we take the simplified approach of considering only galaxies in a range of distances consistent with the distance measurement accuracy expected for multi-detector networks: $\sim 30\,\%$ in fractional distance uncertainty for an event at the detection threshold (Veitch et al., 2012). Within this range, we will neglect the detection probability in the galaxy prior, and consider only priors proportional to blue-light luminosity. We expect this to be conservative, since the effective galaxy luminosity with the detection probability included would have had greater fluctuations than the absolute luminosity, and, as we will see below, luminosity fluctuations increase the utility of galaxy catalogues.

We will thus assume that the detector network is able to localise a source at distance D to within a range $[D_{\min}, D_{\max}]$, with $D_{\min} = 0.7D$ and $D_{\max} = 1.3D$. Combining the solid angle P of a telescope pointing with this range, we can define the *pointing volume*, i.e., the volume of each pointing within the measured distance range, as

$$V \equiv \frac{4\pi P}{3\Omega}(D_{\max}^3 - D_{\min}^3),$$

where $\Omega \approx 4.1 \times 10^4\,\mathrm{deg}^2$ is the solid angle of the whole sky. The average luminosity per pointing volume is then given by $\langle L_i \rangle = V\rho_L$, where ρ_L is the average spatial density of luminosity. We will use a luminosity density $\rho_L = 0.02\,\mathrm{L}_{10}\,\mathrm{Mpc}^{-3}$, where L_{10} is defined as 10^{10} times the solar blue-light luminosity $\mathrm{L}_{\mathrm{B},\odot}$ (Abadie et al., 2010a; Kopparapu et al., 2008).

2.3 Luminosity fluctuations

In this section, we incorporate the distribution of intrinsic galaxy luminosity and the counting fluctuations in the number of galaxies in different pointings into the expected distribution of L_i. We neglect spatial correlations of galaxies (e.g., due to the presence of galaxy clusters – a conservative assumption since greater clustering improves the utility of galaxy catalogues, as we shall see shortly) and assume they are homogeneously distributed in volume. We model the distribution of galaxies in blue luminosity and volume as a Schechter function (Schechter, 1976)

$$n(x)\,\mathrm{d}x \propto x^{\alpha} e^{-x}\,\mathrm{d}x, \tag{2.3}$$

where $x \equiv L/L^*$ and $n(x)\,\mathrm{d}x$ is the expected number of galaxies per Mpc^3 in the interval $[x, x + \mathrm{d}x]$. We use the GWGC (White et al., 2011) within 20 Mpc, where it is complete, to estimate $\alpha = -1.1$ and $L^* = 2.2\,\mathrm{L}_{10}$, slightly brighter than the Milky Way's blue-band luminosity of $\sim 1.7\,\mathrm{L}_{10}$.[6] We normalise the luminosity function to yield $\rho_L = 0.02\,\mathrm{L}_{10}\,\mathrm{Mpc}^{-3}$ on the interval $L \in [0.001, 20]\,\mathrm{L}_{10}$ (Kopparapu et al., 2008). All following results are based on this Schechter luminosity distribution. Figure 2.1 shows the luminosity function of the GWGC within 20 Mpc as well as the Schechter model.

If we assume that a pointing tile has volume V, containing a random integer sample of galaxies taken from the distribution in Eq. (2.3), then the resulting luminosity L in that volume can be described by a random variable of mean $V\rho_L$. The

[6] Similar values of these parameters are quoted in the literature, e.g., $\alpha = -1.07$, $L^* = 2.4\,\mathrm{L}_{10}$ (Schneider, 2006); our conclusions are insensitive to small changes in these parameters.

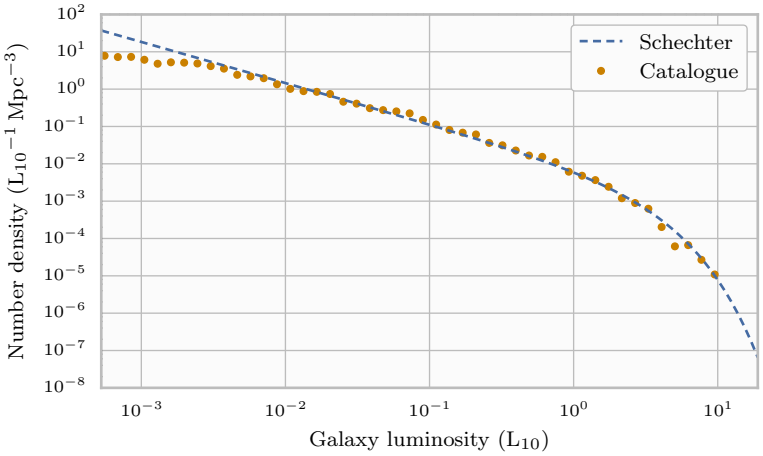

Figure 2.1: The GWGC luminosity function within 20 Mpc compared to a fit of Eq. (2.3) with $L^* = 2.2\,\mathrm{L}_{10}$ and $\alpha = -1.1$. GWGC luminosities are divided into 50 logarithmically spaced bins covering the interval $[0.001, 20]\,\mathrm{L}_{10}$; the fit is normalised to match the catalogue luminosity density of $\sim 0.027\,\mathrm{L}_{10}\,\mathrm{Mpc}^{-3}$ within 20 Mpc.

results of a direct Monte Carlo simulation of Eq. (2.3) for 100 Mpc^3 and 1000 Mpc^3 are shown in Fig. 2.2. To understand these results, one can crudely approximate the Schechter galaxy population as a Poisson scattering of identical galaxies of "typical" luminosity L^*. In this case, the luminosity in a volume V is simply $L = nL^*$, where n is drawn from a Poisson distribution of mean $V\rho_L/L^*$. For 100 Mpc^3 and 1000 Mpc^3 pointing volumes, for example, we should expect 0.9 ± 0.95 and 9 ± 3 galaxies per pointing volume, with corresponding luminosities of $(2.0 \pm 2.1)\,\mathrm{L}_{10}$ and $(20 \pm 6.6)\,\mathrm{L}_{10}$, respectively. This closely matches the fluctuations of $(2.0 \pm 2.0)\,\mathrm{L}_{10}$ and $(20 \pm 6.3)\,\mathrm{L}_{10}$, respectively, measured via a Monte Carlo simulation of the actual Schechter distribution.

The large variations in tile luminosity – $(2 \pm 2)\,\mathrm{L}_{10}$ for the 100 Mpc^3 volume – suggest that there can be substantial advantage to following up the brightest tiles first in a survey with limited pointings. For lower pointing volumes, the distribution becomes increasingly non-Gaussian, and its skewness amplifies the advantage of luminosity-directed surveys.

2.4 Results: complete galaxy catalogue

We show the success fraction $\mathcal{F}$ when using a complete, ideal galaxy catalogue as a function of pointing volume in Fig. 2.3, for four choices of the fraction $f \in \{0.01, 0.05, 0.10, 0.50\}$ of the GW localisation region being followed up. Recall from

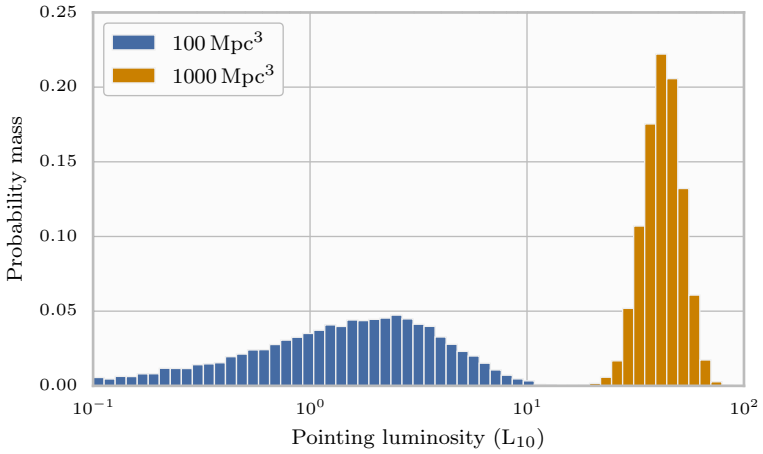

Figure 2.2: Distribution of luminosity drawn from Eq. (2.3) for fixed volumes of 100 Mpc^3 and 1000 Mpc^3. The means are 2.0 L_{10} and 20.0 L_{10} and the standard deviations are 2.0 L_{10} and 6.3 L_{10} respectively. These values correspond approximately to a Poisson scattering of galaxies of "typical" luminosity L^*.

Section 2.3 that in the case where no galaxy catalogue is used we would expect on average that $\mathcal{F} = f$. We find in Fig. 2.3 that in all cases when using the galaxy catalogue $\mathcal{F} > f$ as we would expect, with the advantage of the catalogue being more pronounced for smaller pointing volumes where the variation of luminosity per pointing is larger.

It is useful to apply the results of Fig. 2.3 to a few potential scenarios in order to understand the impact that an ideal galaxy catalogue would have. The distance at which a single Advanced LIGO detector is capable of detecting an optimally oriented and located BNS merger at an SNR of 8 – known as the *horizon distance* – is $\sim$ 450 Mpc (Abadie et al., 2010a). However, averaging over sky locations and orientations, we expect 75 % of detections to come from within $\sim$ 250 Mpc, or 50 % from within $\sim$ 200 Mpc. For early versions of the Advanced LIGO/Virgo network, the median distance could be reduced to as little as $\sim$ 100 Mpc (Aasi et al., 2013a). We therefore consider the median distance of 200 Mpc as a typical distance to a detection with the advanced-detector network operating at design sensitivity in cases (i) and (ii) below. Meanwhile, case (iii) represents the rarer scenario of a closer source at an estimated distance of 100 Mpc.

(i) Consider a 10 deg^2 FOV telescope following up with a single pointing a GW source estimated to be at a distance of 200 Mpc, with a 100 deg^2 localisation region. The pointing volume to this source, assuming GW observations constrain the distance to be $\in [140, 260]$ Mpc, is 15 000 Mpc^3. (For comparison,

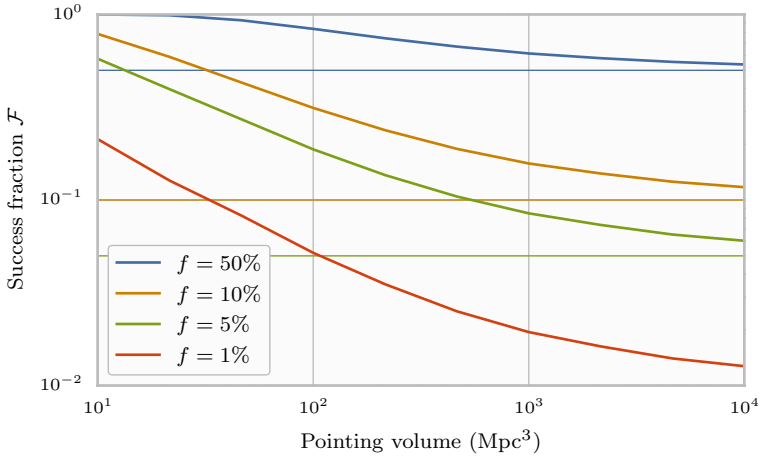

Figure 2.3: A comparison of the success fraction $\mathcal{F}$ relative to the follow-up fraction as a function of pointing volume. Cases (i), (ii), and (iii) in the text correspond to pointing volumes of 15 000 Mpc^3, 1500 Mpc^3 and 60 Mpc^3, respectively. The horizontal grid lines represent follow-up searches that do not use galaxy catalogues, wherein $\mathcal{F} = f$. The large variation in luminosity per tile can cause certain FOVs within the GW localisation region to be more likely to contain the source, suggesting an obvious pointing priority in the case where the entire localisation region cannot be followed up in a timely fashion.

a 1 deg^2 conical FOV contains a volume of ~ 100 Mpc^3 out to a distance of 100 Mpc.) Without a galaxy catalogue we would expect that the success fraction $\mathcal{F} = f = 10\,\%$. However, Fig. 2.3 shows that using a galaxy catalogue we might expect to have a $\sim 11\,\%$ success fraction.

(ii) Consider the same GW source as case (i) but with a 1 deg^2 FOV follow-up instrument having 10 pointings. While the overall coverage is still $f = 10\,\%$, the pointing volume is reduced to 1500 Mpc^3, so the fluctuations in luminosity between tiles are more significant. As a result, the success fraction improves to $\sim 16\,\%$.

(iii) Finally, consider a loud GW signal with a 100 Mpc distance estimate being followed up by a single pointing of a 1 deg^2 FOV instrument. The sky-localisation and distance measurement accuracy improve for high signal-to-noise ratio GW detections. We therefore consider a 10 deg^2 localisation region and a reduced distance uncertainty range $\in [90, 110]$ Mpc. In this case, the pointing volume is only 60 Mpc^3, and the success fraction is $\sim 40\,\%$, a 4-fold improvement over the nominal $\mathcal{F} = f = 10\,\%$ success fraction in the absence of a galaxy catalogue.

These cases are meant as illustrations only. Distances to optimally located and oriented sources may range to 450 Mpc for advanced detectors at design sensitivity.

Meanwhile, sources in the early phases of advanced detector commissioning, when detectors are sensitive within a smaller range, may resemble case (iii) in typical distance estimates, but with poorer sky localisation and distance measurements.

While the overall fraction f of the GW localisation region is 10 % for each of the above cases, the pointing volumes are respectively $\sim 15\,000\,\mathrm{Mpc}^3$, $\sim 1500\,\mathrm{Mpc}^3$, and $\sim 60\,\mathrm{Mpc}^3$. The progressively larger success fractions for each case illustrate how the utility of the catalogue depends on pointing volume.

The effectiveness of a given follow-up telescope, as characterised by its FOV P and the number of pointings N that can be taken within the allotted time while observing to a sufficient depth, may be influenced by the sensitivity of the GW search and the distance estimate it yields. For case (iii), an instrument with a larger FOV might be chosen at the expense of depth, since the EM signal is expected to be louder. Similarly, the greater imaging depth required to detect transients at 200 Mpc might mean that fewer pointings are available for the more distant sources in cases (i) and (ii).

2.5 The effect of galaxy catalogue completeness

The previous discussion assumed that an ideal, complete galaxy catalogue is available; however, the GWGC (White et al., 2011) is incomplete beyond ~ 30 Mpc, and there are limitations to how complete catalogues can be at ~ 200 Mpc distances (Metzger et al., 2013). In practice, catalogues may comprise many different surveys with different characteristics and selection criteria, and will be influenced by spatially dependent factors such as extinction in the Galactic plane. However, the simplest model, and the one we consider here, is incompleteness from a flux-limited survey. As a simple example, we considered an extremely flux-limited survey that does not resolve galaxies fainter than apparent magnitude ~ 15.5 mag in the blue band, which is a rough approximation to the GWGC. The resulting luminosity function of this hypothetical survey, for galaxies within 200 Mpc, is shown in Fig. 2.4. The catalogue is only 33 % complete within this volume (i.e., contains 33 % of the total absolute luminosity) compared to the model presented in Fig. 2.1. However, it contains most of the rare bright galaxies, whose distribution on the sky shows significant fluctuations, making them most useful for informing pointing strategy, while missing common dim galaxies which are nearly homogeneous on the sky and are therefore less useful for pointing.

The shape of the luminosity function for a flux-limited catalogue is sensitive only to the overall completeness of the catalogue, not the specific range and cut-off

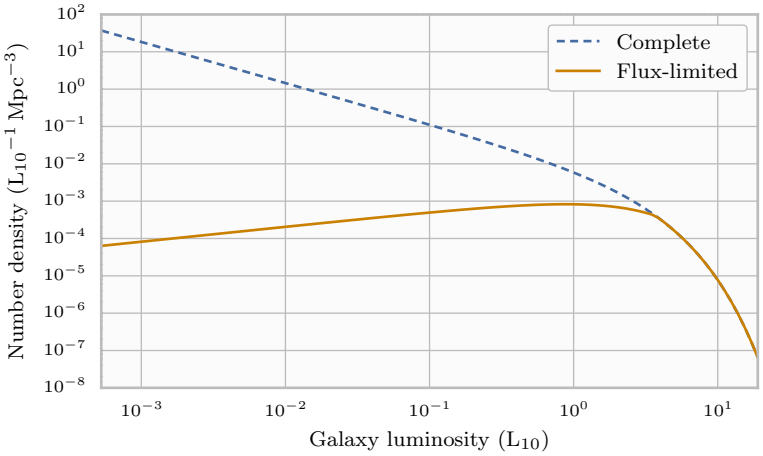

Figure 2.4: Comparison of an ideal complete luminosity function (dashed) to a hypothetical flux-limited survey (solid) at apparent magnitude $m_B = 15.5$ mag out to 200 Mpc. This example is only 33 % complete within 200 Mpc relative to the expected $0.02\,L_{10}\,\mathrm{Mpc}^{-3}$.

magnitude. Therefore, in order to express the follow-up success probability for a flux-limited catalogue in terms of pointing volume, which incorporates FOV and depth in a single variable, we fix the *completeness* of the catalogue, rather than the cut-off magnitude. This success probability is plotted in Fig. 2.5 for three choices of completeness: 33 %, 75 %, and 100 %.

In our flux-limited survey model for incompleteness, the catalogue luminosity function agrees with a hypothetical complete luminosity function for the most luminous galaxies. It is therefore not surprising that incompleteness in a catalogue has little effect on the scenarios where only a small fraction of the sky uncertainty region will be followed up, since both complete and incomplete catalogues will tend to agree on the most luminous tiles, which are the only ones that will be pointed at for small follow-up fractions. For example, in Fig. 2.5, the line corresponding to a follow-up fraction of $f = 0.01$ is virtually unchanged from the corresponding line for a complete catalogue.

When the follow-up fraction is large, incomplete catalogues still yield similar success fractions to complete catalogues as long as the pointing volume is also sufficiently large. Of course, at very large pointing volumes $\mathcal{F}$ asymptotes to f, as the FOVs become increasingly uniform due to the very large numbers of galaxies they contain, and a galaxy catalogue ceases to be useful even when complete. Even at moderate pointing volumes and moderate follow-up fractions, catalogue incompleteness is not necessarily a concern if it only leads to missing the many dim galaxies which are nearly homogeneously distributed on the sky. A few bright

galaxies can still dominate the prior and since these are included even in incomplete flux-limited catalogues, the success fraction is still relatively insensitive to completeness. When pointing volumes are small, even the dimmest galaxies, which are missed out in incomplete catalogues, contribute to the variability between different FOVs. When follow-up fractions are large at small pointing volumes, the success probability asymptotes to the maximum possible success fraction $\mathcal{F} \to \lambda + f(1-\lambda)$, where λ is the catalogue completeness[7], and incompleteness limits catalogue utility. This happens for a 33 %-complete catalogue when the pointing volume is $V \lesssim 100\,\mathrm{Mpc}^3$ and the follow-up fraction is $f \gtrsim 10\,\%$. Nevertheless, even for larger follow-up fractions, $\mathcal{F}$ is significantly larger than f for a large range of pointing volumes, suggesting that even a moderately complete (33 % complete) catalogue is still useful for pointing at the sky region hosting the source of a GW transient.

As discussed above, an incomplete catalogue is most useful when it contains a high fraction of intrinsically luminous galaxies at the expense of missing galaxies with low absolute magnitudes. Therefore, a simple flux limit is an optimistic model of a catalogue's incompleteness. If a catalogue instead has a more gradual cut-off with apparent magnitude, its utility for a given completeness fraction λ could be lower than estimated here. For example, for $f = 10\,\%$, the largest discrepancy between a 33 %-complete flux-limited catalogue and the GWGC is at $V \approx 20\,\mathrm{Mpc}^3$, where they yield success fractions of $\mathcal{F} \approx 63\,\%$ and $\mathcal{F} \approx 60\,\%$, respectively. The difference drops to 1 % for a pointing volume of $100\,\mathrm{Mpc}^3$ and is below the statistical error of the simulation for $V = 1000\,\mathrm{Mpc}^3$.

2.6 Conclusion and future work

Electromagnetic follow-up prospects in the advanced GW detector era can be aided by the use of galaxy catalogues to direct follow-up surveys. The relevance of catalogue-directed wide-field follow-ups is limited mostly by the modest spatial fluctuations of luminosity on the sky for the large three-dimensional localisation uncertainty volumes of the advanced-detector network.

We have shown in Figs. 2.3 and 2.5 that the utility of a catalogue depends on the volume of individual telescope pointings and on the fractional coverage of the GW localisation area. Catalogues are therefore most relevant for shallow and narrow follow-up searches, although narrow-field instruments are unlikely to follow up a

[7]Consider a situation in which the number of bright galaxies that enter the catalogue is sufficiently small that all of them can be followed up with a negligible number of pointings; the remaining allowed pointings will capture a fraction f of the other FOVs, which have no galaxies in the incomplete catalogue and have a uniform prior density $1-\lambda$ painted across them.

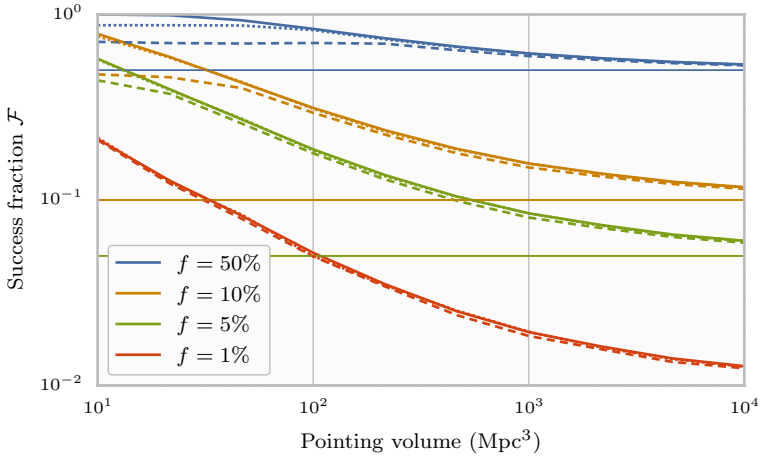

Figure 2.5: The success fraction $\mathcal{F}$ as a function of pointing volume for hypothetical flux-limited catalogues of 33 % (dashed), 75 % (dotted) and 100 % (solid) completeness within the pointing volume being considered. Incomplete catalogues yield similar success fractions to complete catalogues except at small pointing volumes and large follow-up fractions.

sufficient fraction f of the GW localisation region for a successful follow-up to be realistic (with the possible exception of short-range observations, where individual galaxies could be followed up). Loud, nearby GW triggers are an obvious scenario where catalogues will be particularly useful. It is possible, for example, that they confer as much as a four-fold increase in success fraction over a follow-up that does not use a catalogue; e.g., case (iii) in Section 2.3 (expected to account for ~ 10 % of detections in the advanced detector era). Similarly, follow-ups from shallower GW searches – during the early commissioning phases of advanced detectors, for example – will also benefit from the use of catalogues.

However, even for sources located at the median 200 Mpc distance expected for detections with advanced-detector networks, we have shown that catalogues are still relevant for sufficiently small telescope FOVs. For example, a catalogue might confer as much as a 70 % increase in the probability of imaging the EM counterpart relative to a follow-up without the benefit of a catalogue, as in case (ii) of Section 2.5.

Realistic, incomplete galaxy catalogues are likely adequate for most follow-up campaigns. Metzger et al. (2013) propose that a catalogue complete to ~ 75 % with respect to B-band luminosity should be achievable. At $f = 10$ %, a hypothetical flux-limited catalogue of this completeness concedes a fraction < 1 % of the success fraction from a complete catalogue for both 100 Mpc^3 and 1000 Mpc^3 pointing volumes. Metzger et al. (2013) suggest that it will be difficult to construct a galaxy catalogue of more than ~ 33 % completeness with respect to K-band luminosity, a

tracer of total mass. Even in this case, the fractional loss of success fraction relative to a search with a complete catalogue is small: 7.5 % and 5 % respectively for 100 Mpc^3 and 1000 Mpc^3 pointings.

2.6.1 Imaging vs. identifying of the counterpart

Our study focuses on the probability of *imaging* the EM counterpart to a detected GW signal – i.e., pointing a telescope so that the EM counterpart is within the FOV – but not on the probability of detecting and *identifying* it among background sources. In reality, some telescopes may have trouble observing weak, distant EM counterparts (Aasi et al., 2014).

For example, Metzger & Berger (2012) suggest that the orphan optical afterglow expected to accompany a BNS merger at 200 Mpc will have a peak optical brightness as faint as ~ 23 mag when viewed slightly off-axis: beyond the limiting flux of many telescopes. Even if they are detected, contamination from background events may make it difficult to pick out the correct transient. Identification of GW EM counterparts among false positives is addressed by Nissanke et al. (2013). The detectability of EM counterparts could be further investigated by considering the capabilities of specific telescopes given the observing requirements of particular sources (for example, their peak luminosities, light-curve evolution, etc.).

2.6.2 Astrophysical assumptions

We have made a number of assumptions about the astrophysics underlying BNS merger signals:

(i) *B-band luminosity of the host galaxy – which traces its star formation rate – is a proxy for the merger rate.* In fact, if there are long time delays between star formation and binary merger, the total mass of the host galaxy, traced by K-band luminosity, might be the more relevant indicator of merger rate. For example, population synthesis modelling suggests that half of all BNS mergers may take place in elliptical galaxies with little ongoing star formation (O'Shaughnessy et al., 2010; de Freitas Pacheco et al., 2006). Meanwhile, observational evidence on short gamma ray bursts indicates that about a quarter of them occur in elliptical galaxies (Fong & Berger, 2013), though selection effects associated with the detection of afterglows that allow the host to be identified could influence this fraction.

(ii) *The completeness of the galaxy catalogue is known precisely.* In practice, the completeness of the catalogue is estimated from the expected spatial luminosity

density in the local Universe ($\sim 0.02\,\mathrm{L}_{10}\,\mathrm{Mpc}^{-3}$ for blue luminosity). Inaccuracy in the estimated completeness may lead to a less-than-optimal ranking of tiles on the sky. We can account for the incompleteness of a catalogue by changing the weighting we give to individual galaxies; if the catalogue completeness fraction is λ, then the catalogued luminosity of a given pixel, L_i, is multiplied by λ when computing the prior, with a prior fraction $1-\lambda$ painted uniformly over the entire GW sky uncertainty region to account for the galaxies missed in the catalogue.

(iii) *Mergers are spatially coincident with host galaxies on the celestial sphere.* Natal kicks accompanying supernovae that give birth to the neutron-star components of a binary (up to hundreds of $\mathrm{km\,s}^{-1}$; Fryer & Kalogera, 1997) can combine to give a significant velocity to the binary as a whole. As a result, mergers are distributed at larger distances from the galactic centre than typical stellar concentrations (Fong & Berger, 2013), and galaxies should properly be treated as extended objects rather than point sources. However, for telescope FOVs of order a square degree or more and typical source distances of 100 to 200 Mpc, treating galaxy sizes $\lesssim 100\,\mathrm{kpc}$ as point sources will not affect our results. On the other hand, binaries may be completely ejected from their host galaxies (e.g., Kelley et al., 2010), and some fraction of the "no-host" short gamma ray bursts (Berger, 2010; Tunnicliffe et al., 2013) may provide evidence for this population of merging ejected binaries, which may be separated by more than $\sim 1\,\mathrm{Mpc}$ from their host galaxy (but see discussion by Kanner et al., 2013 and references therein).

We suggest a future study of the importance of these effects – given our ignorance – as parameterised priors. One would allow nature to choose a *true* value of a given parameter (e.g., the relative contribution of blue and red luminosity tracers to merger rates) and attempt to image counterparts from the resulting GW events by ranking tiles according to an *assumed* parameter value representing our own knowledge. The effects of our ignorance of the true values of each parameter could thus be described by a matrix in which one dimension represents nature's choice of prior, and the other our assumed knowledge.

2.6.3 Coherent use of galaxy catalogues

Finally, we have investigated the utility of a galaxy catalogue when applied to the sky location posterior obtained from a parameter estimation pipeline. In practice, if a galaxy catalogue were to be used for follow-up, it should be applied as a prior during

coherent Bayesian parameter estimation (Aasi et al., 2013c). Doing so would make it possible to consistently account for the probability that a given galaxy hosts the GW source, which depends not only on the galaxy luminosity but also on the distance to the galaxy and the inclination and orientation of the binary, which must yield a GW signal amplitude consistent with observations. This is particularly important when considering the correlations between the recovered GW signal parameters such as inclination and distance. Moreover, using coherent Bayesian parameter estimation would allow complex sky location posteriors to be accurately accounted for.

Chapter 3

Adaptive parallel tempering

The following text and figures are reproduced from a paper (Vousden et al., 2015) written in collaboration with Will Farr and Ilya Mandel that describes a new method for improving the efficiency of parallel-tempered Markov chain Monte Carlo samplers.

During this project I was responsible for writing the paper, developing the code, and executing the tests that yielded the results presented here. The idea for the project grew from discussions with WF, while the subsequent analysis, development of the algorithm, and manuscript editing were a joint effort with WF and IM.

3.1 Introduction

Many problems in astronomical data analysis and Bayesian statistical inference demand the characterisation of high-dimensional probability distributions with complicated structures. Lacking analytic forms, these distributions must be explored numerically, usually via Monte Carlo methods.

Parallel-tempered Markov chain Monte Carlo (MCMC), a development on standard MCMC, uses several Markov chains in parallel to explore a target distribution at different "temperatures" (Earl & Deem, 2005; Geyer, 1991; Swendsen & Wang, 1986). As the temperature increases, the posterior distribution asymptotes to the prior, allowing a chain to efficiently explore the whole prior volume without becoming stuck in regions of the parameter space with high probability density. At lower temperatures, a chain can more efficiently sample from such a high-probability region. Meanwhile, exchange of positions between chains allows colder chains to migrate between widely separated modes in the parameter space (Geyer, 1991). Parallel-tempered MCMC samplers are thus particularly well-suited to sampling posterior distributions with well-separated modes, where a regular MCMC sampler would take

many iterations to find its way between modes.

An open problem in the application of parallel tempering is selecting a specification, or ladder, of temperatures that minimises the autocorrelation time (ACT) of the chain sampling the posterior distribution of interest. The efficiency of a given ladder hinges critically on the rate at which it can transfer the positions in parameter space of samples between high and low temperatures.

In this paper we present a simple algorithm that adapts the temperature ladder of an ensemble-based parallel-tempered MCMC sampler (Goodman & Weare, 2010) such that the rate of exchange between chains is uniform over the entire ladder. The algorithm is easy to implement in existing code, and we provide an example implementation for the *emcee* sampler of Foreman-Mackey et al. (2013).

In Section 3.2 we describe the parallel tempering formalism and lay out the requirements for a good temperature ladder. We discuss previous work on temperature selection and suggest a definition of ladder optimality that, for simple cases, proposes a geometric spacing of temperatures. For illustration, we apply these ideas in Section 3.2.2 to the simple example of an unbounded Gaussian posterior distribution.

In Section 3.3 we describe the algorithm mentioned above and then apply it in Section 3.4 to a variety of test distributions. We show that, while our temperature selection strategy is not necessarily optimal in the ACT of the sampler, it nonetheless improves the ACT compared to the simple geometric spacing that is conventional in the literature (Earl & Deem, 2005; Kofke, 2002, 2004; Sugita & Okamoto, 1999) by factors of > 1.2 for our test cases.

We conclude in Section 3.5 with a discussion of our results and suggestions for further research.

3.2 Parallel tempering

Parallel tempering (Earl & Deem, 2005; Geyer, 1991; Swendsen & Wang, 1986) is a development on the standard MCMC formalism that uses several Markov chains in parallel to sample from tempered versions of the posterior distribution π,

$$\pi_T(\vec{\theta}) \propto L(\vec{\theta})^{1/T} p(\vec{\theta}), \tag{3.1}$$

where L and p are respectively the likelihood and prior distributions.

For high T, individual peaks in L become flatter and broader, making the distribution easier to sample via MCMC. A set of N chains is assigned temperatures in a ladder $T_1 < T_2 < \ldots < T_N$, with $T_1 = 1$ (the target temperature). The tempera-

tures are typically geometrically spaced from 1 up to some $T_{\max}$, decided in advance (a convention that we shall discuss in more detail in Section 3.2.2).

Each chain is allowed to explore its tempered distribution π_T under an MCMC algorithm, while at pre-determined intervals "swaps" are proposed between (usually adjacent[1]) pairs of chains and accepted with probability

$$A_{i,j} = \min\left\{\left(\frac{L(\vec{\theta}_i)}{L(\vec{\theta}_j)}\right)^{\beta_j-\beta_i}, 1\right\}, \tag{3.2}$$

where $\vec{\theta}_i$ is the current position in the parameter space of the i^{th} chain and $\beta_i \equiv 1/T_i$ is the inverse temperature of this chain. This acceptance probability is chosen to maintain detailed balance for the *joint* Markov chain whose state space is the product of those of the individual chains at each temperature.

When a swap is accepted, the chains exchange their positions in the parameter space, so that chain (i) is at $\vec{\theta}_j$ and chain (j) is at $\vec{\theta}_i$. Since the hottest chains can access all of the modes of π (as long as $T_{\max}$ is chosen appropriately), their locations propagate to colder chains, ultimately allowing the $T = 1$ (cold) chain to efficiently explore the entire target distribution. At the same time, the positions of the colder chains propagate upward to higher temperature chains, where they are free to explore the entire prior volume.

The goal in choosing an effective ladder of temperatures is to minimise the ACT of the cold chain (our measure of the efficiency of the sampler). The requirements to this end are two-fold:

(i) $T_{\max}$ must be large enough that isolated modes of L broaden sufficiently that an individual MCMC chain can efficiently access all of these modes when sampling under the tempered posterior π_T in Eq. (3.1) at $T = T_{\max}$. We denote this temperature T_{prior}.

(ii) Since $A_{i,j}$ depends on $\beta_i - \beta_j$, the differences between temperatures must be small enough that neighbouring chains can communicate their positions efficiently with one another.

Both requirements depend sensitively on the (unknown) shape of the target distribution, so it is difficult to select temperatures appropriately in advance.

In choosing $T_{\max}$, one must know roughly the relative size and separation of the modes to be explored. As an example, consider a one-dimensional likelihood with

[1]In principle, swaps can be proposed between any pair of chains. However, since the swap acceptance ratio (3.2) decays exponentially with the separation of inverse temperatures, $\Delta\beta$, it is generally sufficient only to propose swaps between adjacent chains.

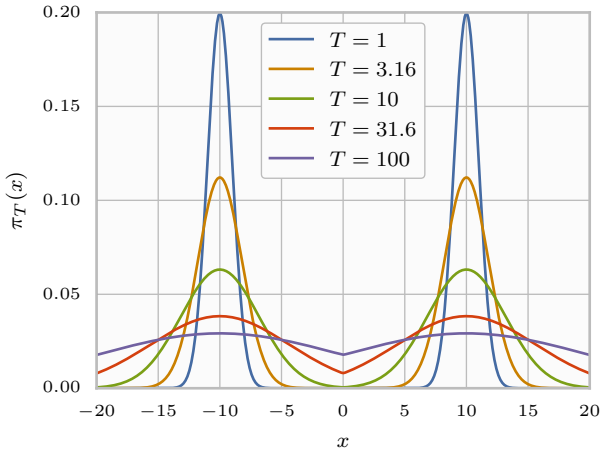

Figure 3.1: A one-dimensional target distribution with two Gaussian peaks of width $\sigma = 1$ at $\mu = \pm 10$ normalised for a uniform prior over $[-20, 20]$. At $T = 100$, the peaks broaden to $\sigma = 10$, allowing an MCMC chain sampling at this temperature to find both modes quickly, starting from anywhere within the prior volume.

two Gaussian modes of width $\sigma = 1$ and centres $\mu = \pm 10$. In order to prevent a sampler from getting stuck on one of the modes, they must be widened to roughly the separation between them[2], giving $\sigma = \mathcal{O}(10)$. The width of a Gaussian peak scales with the temperature as $\sqrt{T}$, so we might choose $T_{\max} = 100$; Fig. 3.1 illustrates the resulting coalescence of the modes. A different configuration of modes will, of course, require a different $T_{\max}$.

On the other hand, the swap acceptance probability $A_{i,j}$ depends on the distribution of likelihood values at temperatures T_i and T_j. In the case of a likelihood distribution comprising a single Gaussian mode, the time-averaged acceptance ratios between chains, $\mathrm{E}[A_{i,j}]$, can be computed analytically (see Section 3.2.2).

In general, we don't know in advance what the target distribution looks like, and so choosing an effective ladder becomes a heuristic exercise, relying largely on educated guesswork. We are therefore motivated to find some method of empirically determining an effective ladder.

3.2.1 Ladder selection

For an n-dimensional problem, the conventional choice of temperatures is a geometrically spaced ladder constructed so that approximately 23 % of swaps proposed between chains will be accepted when sampling from an n-dimensional, unbounded

[2]Ideally, the modes must also be widened enough that they extend to the edges of the prior volume. A likelihood distribution with a single mode that occupies only a small fraction of the prior volume will take a long time to burn in.

Gaussian distribution (Earl & Deem, 2005; Roberts & Rosenthal, 1998; ter Braak & Vrugt, 2008). We shall discuss this convention in more detail in Section 3.2.2.

A consequence of this strategy is that increasing the number of chains N does not improve communication between existing chains, which is determined by $\mathrm{E}[A_{i,j}] = 0.23$. Instead, adding new chains extends the ladder to higher temperatures. This may be appropriate for an unbounded posterior, but for a realistic problem with a finite prior volume, the acceptance ratio between adjacent chains saturates to $\sim 100\,\%$ at some temperature T_{prior}, at which the posterior π_T begins to look like the prior p.

For this geometric spacing scheme – where T_{prior} is unknown – there is therefore an optimal number of chains, N_{opt}, such that $T_{\mathrm{prior}} \approx T_{N_{\mathrm{opt}}} \equiv T_{\mathrm{max}}$. For $N < N_{\mathrm{opt}}$ none of the chains will be sampling from the prior (so the sampler may not find all of the modes), while for $N > N_{\mathrm{opt}}$ we end up with several chains sampling redundantly from the prior.

Since we are generally ignorant of T_{prior} for the problem at hand, we are motivated to find an alternative temperature selection strategy.

It has been suggested in the literature (Earl & Deem, 2005; Kofke, 2002, 2004; Sugita & Okamoto, 1999) that one could select temperatures such that the acceptance ratios $A_{i,j}$ are uniform for all pairs (i, j) of adjacent chains, in an attempt to ensure that each sample sequence $\vec{\theta}(t)$ for $t = 1, 2, \ldots$, as it moves between chains, spends an equal amount of time at every temperature. Sugita & Okamoto (1999) justify this notion experimentally – in the context of molecular dynamics – with test cases in which such a ladder indeed performs well. They use an algorithm derived from that of Hukushima & Nemoto (1996), which selects temperatures according to an iterative process for which a uniform-A ladder is a fixed point. Earl & Deem (2005) provide further references for similar methods of determining temperature ladders that yield a given a target acceptance ratio (Rathore et al., 2005; Sanbonmatsu & García, 2002; Schug et al., 2004). However, these methods do not address requirement (i), discussed above, that the temperature ladder should reach a T_{max} sufficient for all of the modes of L to mix (specified by T_{prior}).

Kofke (2002) discusses the selection of temperature ladders in the context of molecular simulations. He shows that, in simulations of such thermodynamic systems, there is a close relation between the specific heat of the system, C_V, and the acceptance ratios between adjacent temperatures. In particular, when C_V is constant with respect to T over a given temperature interval, then a geometric spacing of temperatures on that interval yields uniform acceptance ratios between adjacent temperatures.

In the language of thermodynamics, the energy of the system, U, is analogous to $-\log L$, and an analogue to the specific heat can therefore be defined as

$$C_V(T) = -\frac{\mathrm{d}}{\mathrm{d}T}\,\mathrm{E}[\log L]_T, \tag{3.3}$$

where $\mathrm{E}[\,\cdot\,]_T$ denotes the expectation operator over $\vec{\theta}$ under the distribution $\pi_T(\vec{\theta})$. $\mathrm{E}[\log L]_T$ is therefore the expectation of the *untempered* log likelihood collected when sampling from the posterior at temperature T.

In the context of Bayesian inference, Kofke's result therefore tells us that if the mean log likelihood collected by a sampler responds linearly to changes in temperature, then a geometrically spaced temperature ladder will achieve uniform acceptance ratios between adjacent chains. Conversely, temperature intervals on which $\mathrm{E}[\log L]_T$ is strongly non-linear in T represent a phase transition that will require more careful placement of temperatures, as we shall show in Section 3.4.

3.2.2 The ideal Gaussian distribution: a simple example

In the simple case of a unimodal Gaussian likelihood under a flat prior[3], the optimal temperature spacing at low temperatures – where very little likelihood mass is truncated by the prior – can be analysed by approximating the prior to be unbounded[4]. We show that, for this tractable example, a geometric temperature spacing is consistent with both the uniform-A criterion and also with the alternative criterion that the Kullback–Leibler (KL) divergence is uniform between all pairs of adjacent chains (see Section 3.2.3). We use the example to illustrate the relationship between the analytical distribution of $\log L$, the acceptance ratio $A_{i,j}$, and the temperature T.

We shall work with an n-dimensional unit Gaussian centred on the origin (the same result can be achieved for a general Gaussian through a simple change of coordinates). Since the prior is uniform and unbounded, we can restrict attention to the likelihood distribution L. In this case, the probability density $\tilde{p}$ for the values of $\log L(\vec{\theta})$ collected by the sampler is

$$\tilde{p}(\log L) = \frac{e^{\log L}(-\log L)^{\frac{n}{2}-1}}{\Gamma(\frac{n}{2})}, \tag{3.4}$$

where L is normalised so that $\log L(\vec{0}) = 0$ and n is the number of parameters.

[3] See Freeman (2006) for a comparison with a less conventional example.

[4] The approximation breaks down at higher temperatures, where boundary effects become significant. Indeed, with no prior boundaries, there is no T_{prior} at which the mode is spread over the entire prior volume.

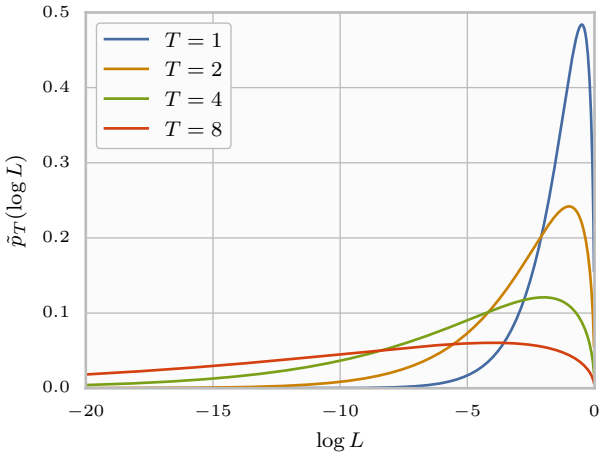

Figure 3.2: The distribution of $\log L$ under a three-dimensional, unimodal Gaussian at various temperatures, where L is normalised so that $\log L(\vec{0}) = 0$. As $T \to \infty$, the variance of $\log L$ diverges. The legend is ordered to match the vertical order of the lines' peaks.

At a temperature T, $-\log L$ simply follows a gamma distribution $\Gamma(\alpha, \beta)$ with shape parameter $\alpha = n/2$ and rate parameter $\beta = 1/T$. Thus, for a chain sampling at temperature T, the log likelihood distribution is $\tilde{p}_T(\log L) = T\,\tilde{p}(\log L/T)$.

Over long time-scales, the average acceptance ratio between chains i and j is

$$\begin{aligned}
\mathrm{E}[A_{i,j}] &= \iint_{(-\infty,0]^2} A_{i,j}\,\tilde{p}_{T_i}(\log L_i)\,\tilde{p}_{T_j}(\log L_j)\;\mathrm{d}\log L_i\;\mathrm{d}\log L_j \\
&= \left(\frac{1}{\sqrt{\pi}}\,2^{n-1}\gamma_{i,j}^{-n/2}\,\Gamma\left(\frac{n+1}{2}\right)\right) \\
&\quad\cdot\left({}_2\tilde{F}_1\left(\frac{n}{2}, n; \frac{n}{2}+1; -\frac{1}{\gamma_{i,j}}\right) - \right. \\
&\qquad \left.\gamma_{i,j}^{n}\,{}_2\tilde{F}_1\left(\frac{n}{2}, n; \frac{n}{2}+1; -\gamma_{i,j}\right)\right) + 1,
\end{aligned} \tag{3.5}$$

where ${}_2\tilde{F}_1$ is the regularised Gauss hypergeometric function and $\gamma_{i,j} = T_j/T_i$ is the ratio between the temperatures of two chains. Since $\mathrm{E}[A_{i,j}]$ depends on T_i and T_j only through the ratio $\gamma_{i,j}$, uniform acceptance ratios between all adjacent pairs of chains can be achieved with a geometric spacing of temperatures – where $\gamma_{i,i+1}$ is constant – for a unimodal Gaussian likelihood.

The log spacing required for a particular acceptance ratio also depends on the dimension of the parameter space, with more parameters requiring a closer spacing

of temperatures, illustrated by Fig. 3.3. This can be understood by looking at the expectation and variance of $\log L$ at a particular temperature (see Fig. 3.2),

$$\mathrm{E}[\log L]_T = -\frac{nT}{2} \quad \text{and} \quad \mathrm{Var}[\log L]_T = \frac{nT^2}{2}. \tag{3.6}$$

Note that the specific heat from Eq. (3.3) is a constant $n/2$, as expected.

Since the acceptance ratio $A_{i,j}$ depends on $\log L_i - \log L_j$, the more separate the distributions of $\log L_i$ and $\log L_j$ at their respective temperatures, T_i and T_j, the lower the acceptance ratio between such chains will be. For two chains at temperatures T and γT, the separation of the means of $\tilde{p}_T$ and $\tilde{p}_{\gamma T}$, in units of the standard deviation at T, will be

$$\frac{\mathrm{E}[\log L]_T - \mathrm{E}[\log L]_{\gamma T}}{\sqrt{\mathrm{Var}[\log L]_T}} = (\gamma - 1)\sqrt{\frac{n}{2}}. \tag{3.7}$$

It follows that – for constant γ – as the dimension n increases, so the acceptance ratio between chains at temperatures T and γT falls. For a higher dimensional target distribution, therefore, a closer spacing of temperatures is required for a given acceptance ratio.

For more general distributions, by considering the overlap of $\tilde{p}_T(\log L)$ at different temperatures, Falcioni & Deem (1999) argue that the number of temperatures N required to efficiently sample the posterior distribution should scale with $\Delta \log L / \sqrt{n}$, where $\Delta \log L$ is the range of $\mathrm{E}[\log L]_T$ between $T = 1$ and $T = T_{\text{prior}}$. That is:

$$N \propto \frac{\mathrm{E}[\log L]_1 - \mathrm{E}[\log L]_{T_{\text{prior}}}}{\sqrt{n}}. \tag{3.8}$$

Since the log likelihood range $\Delta \log L$ itself depends on the dimension of the system n, it is difficult to apply this relation in practice. However, for the ideal Gaussian, we can see from Eq. (3.6) that $\Delta \log L$ scales with n, and so N scales with $\sqrt{n}$, as we might expect.

3.2.3 The Kullback–Leibler divergence

Another measure of the optimal spacing of temperatures is the Kullback–Leibler (KL) divergence between adjacent chains. The KL divergence from a hot distribution π_{T_j} to a cold distribution π_{T_i},

$$D_{\mathrm{KL}}(\pi_{T_i} \| \pi_{T_j}) = \int \pi_{T_i}(\vec{\theta}) \log \frac{\pi_{T_i}(\vec{\theta})}{\pi_{T_j}(\vec{\theta})} \, \mathrm{d}\vec{\theta}, \tag{3.9}$$

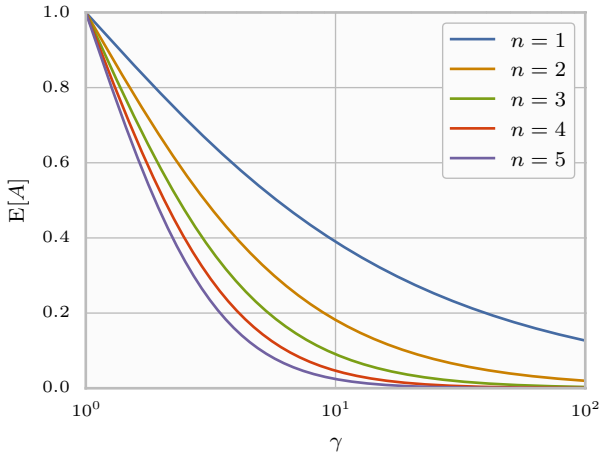

Figure 3.3: The time-averaged acceptance ratio, $\mathrm{E}[A]$, between two chains of a parallel-tempered MCMC sampler on a unimodal, n-dimensional Gaussian likelihood distribution. The chains have temperatures T and γT. The lines are ordered vertically to match the legend.

quantifies the information gained about the posterior with each step down the temperature ladder, from the prior $p = \pi_{T=\infty}$ to the posterior $\pi = \pi_{T=1}$. It is reasonable to expect that for an optimally-spaced ladder – that is, one with a minimal ACT on the cold chain for a given number of chains – the information gain should be uniform for every step down the ladder.

For the example of the ideal Gaussian of Section 3.2.2, the KL divergence is, straightforwardly,

$$D_{\mathrm{KL}}(\pi_{T_i} \| \pi_{T_j}) = \frac{n}{2} \left(\frac{1}{\gamma_{i,j}} + \log \gamma_{i,j} - 1 \right). \tag{3.10}$$

Like the swap acceptance ratio, therefore, uniform KL divergence over the entire ladder is also achieved by a geometric spacing of temperatures for the ideal Gaussian.

Unfortunately, unlike the acceptance ratio, the KL divergence is difficult to compute numerically while sampling, owing to the unknown – and temperature-dependent – evidence (normalisation) values on π_{T_i} and π_{T_j}.

We henceforth assume that spacing temperatures for uniform acceptance ratios is a reasonable approximation of a ladder that is optimal in the ACT of the cold chain. We make this assumption on faith and, while we briefly examine its validity in Section 3.4.1, it invites a more careful study.

3.3 Adaptive temperature ladders

From the arguments in Section 3.2 and the references therein, we shall assume that uniformity of acceptance ratios provides a good approximation to the optimal temperature ladder for parallel tempering problems. In this section, we describe an algorithm for dynamically adapting chain temperatures to achieve uniform acceptance ratios for inter-chain swaps.

From Eq. (3.2), as $1/T_j - 1/T_i \to 0$, $A_{i,j} \to 1$, so in order to increase the expected acceptance ratio between chains, it suffices to move them closer together in temperature space; conversely, to reduce $\mathrm{E}[A_{i,j}]$, we can push the chains apart. We will henceforth adopt the notation that $A_i \equiv A_{i,i-1}$ and that $T_i < T_{i+1}$, with $T_1 = 1$ being the untempered or cold chain (which samples from the target distribution, π). Here, $A_i(t)$ are the instantaneous acceptance ratios between chains, but we shall shortly describe the discrete case where empirical measurements of A_i are collected with each iteration of the sampler.

3.3.1 Dynamics

Our goal is to dynamically adjust the temperatures of the chains to achieve uniform acceptance ratios as we sample the target distribution. We define our temperature dynamics in terms of the log of the temperature difference between chains,

$$S_i \equiv \log(T_i - T_{i-1}). \tag{3.11}$$

Under this scheme, finite changes to S_i will always preserve the correct ordering of temperatures ($T_1 < ... < T_N$).

To achieve the same A_i for all chains, we can drive the gap S_i according to the acceptance ratios between chain (i) and those immediately above and below, to wit

$$\frac{\mathrm{d}S_i}{\mathrm{d}t} = \kappa(t)\left[A_i(t) - A_{i+1}(t)\right], \tag{3.12}$$

for $1 < i < N$, where κ is a positive constant controlling the time-scale of the evolution of T_i. κ can be interpreted as the instantaneous exponential time-constant for temperature adjustments. The two extremal temperatures, T_1 and T_N, are fixed (see below).

Under this scheme, chain (i) will attempt to increase the gap in temperature space between itself and chain $(i+1)$ if swaps are accepted too often and close it when they are accepted too seldom — and similarly for chain $(i-1)$ — equilibrating at A_i that are uniform over i. Therefore, for an appropriate choice of κ – discussed

momentarily – these rules drive the chains $i = \{2, \ldots, N-1\}$ toward even acceptance spacing.

However, in order to efficiently sample a target distribution with strongly separated modes (such that a traditional MCMC sampler would be unable to traverse the "valleys" between them), T_N must be high enough that the modes are flattened out and the chain can explore the entire parameter space unhindered. This amounts to the topmost chain sampling from the prior distribution[5], which we achieve trivially by setting the inverse temperature of this chain as $\beta_N = 0$.

This continuous system is discretised as

$$S_i(t+1) - S_i(t) = \kappa(t) \left[A_i(t) - A_{i+1}(t) \right], \tag{3.13}$$

where $A_i(t)$ are the acceptance ratios accumulated by the sampler at the current iteration.

The values of A_i are measured instantaneously as the fraction of swap proposals between chains that were accepted *for that iteration alone*. For a traditional sampler comprising one sample per chain, these will be either 0 or 1. For ensemble samplers, however, comprising n_w distinct walkers per temperature, the measurements of A_i are less granular, such that $A_i \in \{x \in [0,1] | n_w x \in \mathbb{Z}\}$. In general, fewer walkers require a longer averaging time-scale – discussed below – in order to smooth out this granularity.

Importantly, the temperature adjustment scheme we have proposed – and, more generally, any adaptive sampling scheme – in fact violates the condition for detailed balance that ensures that an MCMC sampler will converge to the target distribution. Roberts & Rosenthal (2007) investigate the conditions required of such an adaptive sampler for it to be ergodic in the target distribution – that is, that it will converge on long time-scales. They determine (from their Theorem 1 and Corollary 4) that diminishing the amplitude of adaptations in the transition kernel with each iteration is sufficient for the sampler to be ergodic in the target distribution. We therefore suppress temperature adjustments to ensure that the sampler is Markovian on sufficiently long time-scales[6].

The rate of diminution of temperature adjustments is a trade-off between the rate of convergence of the temperature ladder and that of the sampler itself toward its stationary distribution. We modulate the dynamics with hyperbolic decay to

[5] For analytic priors, this special case, where the likelihood is ignored, can be treated separately by having the sampler draw independent samples directly from the prior.

[6] In principle, of course, we could stop temperature adjustments altogether once the temperatures have reached an equilibrium, discarding the previous samples as part of the burn-in.

suppress the dynamics on long time-scales,

$$\kappa(t) = \frac{1}{\nu}\frac{t_0}{t + t_0}, \tag{3.14}$$

where t_0 is the time at which the temperature adjustments have been reduced to half their initial amplitude. The initial amplitude of adjustments is in turn set by ν, the time-scale on which the temperatures evolve at early time.

3.3.2 Parameter choice

In the scheme of Eqs. (3.13) and (3.14), there are two parameters to choose: t_0 and ν. The dynamical time parameter t in Eq. (3.13) is measured in units of intra-chain jumps of the sampler, with temperature adjustments being made at every iteration.

The lag parameter t_0 sets the time-scale for the attenuation of temperature adjustments. This decay factor in κ is included as a fail-safe mechanism to ensure that, even for target distributions on which the temperature dynamics fail to find an equilibrium set of temperatures, the ladder will always converge over long time-scales. This condition guarantees that the sampler correctly explores the target distribution.

From Eq. (3.14), the time-scale of the dynamics at late time – when $t \gg t_0$ – is $\nu t/t_0$. To ensure that temperatures have time to find an equilibrium over the course of a run, we therefore require that $t_0 \gg \nu$, so that the dynamics will always be on a time-scale much shorter than the current run time. However, we should also ensure that, over the course of the run, the dynamical time-scale is *longer* than the ACT of the sampler, so that the temperatures respond to the correct posterior distribution. To this end, we require that $\nu N_\tau \gg t_0$, where N_τ is the number of independent samples gathered over the course of the run. For example, if $N_\tau = 100$, these two conditions are satisfied by $t_0 = 10\nu$, and for our test cases, we have indeed found this choice to work well.

Meanwhile, the time-scale of the dynamics at early time – when $t \ll t_0$ – is ν. A good choice of ν should therefore ensure that the sampler is not susceptible to large statistical errors on the measurements of the acceptance ratios A_i.

In general, for n_s swap proposals, the acceptance count $n_s A_i$ is a random variable that follows a binomial distribution $B(n_s, \mathrm{E}[A_i])$, so that A_i has variance

$$\mathrm{Var}[A_i] = \frac{\mathrm{E}[A_i](1 - \mathrm{E}[A_i])}{n_s}.$$

Since the dynamical equations (3.13) are linear in A_i, they will be driven by the

means, $\mathrm{E}[A_i]$, on long time-scales, assuming that the noise in the system from counting errors – proportional to $1/\sqrt{n_s}$ – does not cause short-term changes in $\mathrm{E}[A_i]$.

Given a sampler of n_w walkers, n_w swaps are proposed with each iteration, so that $n_s = n_w\nu$. To ensure stable dynamics at early time, we should therefore choose $n_w\nu \gg 1$.

A good choice of ν depends on the response of $\mathrm{E}[A_i]$ to changes in the relevant chains' temperatures, and therefore depends on the particular likelihood function that is being sampled. However, if $\mathrm{E}[A_i]$ will eventually be of order, say, 0.25, and we want the measurements of A_i to be between 0.2 and 0.3, then we should average A_i over at least 100 swap proposals, giving $\nu \gtrsim 100/n_w$.

Combining these criteria on ν and t_0, we therefore suggest default parameter values of $\nu = 10^2/n_w$ and $t_0 = 10^3/n_w$.

3.4 Examples

We have implemented the algorithm proposed above as a modification to the ensemble sampler *emcee* of Foreman-Mackey et al. (2013). Our implementation can be found at `https://github.com/willvousden/ptemcee`.

In this section we apply our implementation to specific examples in order to understand how and when the traditional geometric spacing fails and how much the uniform-A strategy might help us. We present the following test cases.

(i) In Section 3.4.1 we compare the uniform-A strategy used by the temperature dynamics of Section 3.3 with the alternative strategy of uniform KL divergence discussed in Section 3.2.2 on the example of a unimodal truncated Gaussian likelihood.

(ii) In Section 3.4.2 we test the dynamics on a more complex, bimodal distribution for various choices of the number of chains N. We compare the resulting ACTs of the sampler with those of another sampler using a geometric ladder whose maximum temperature is fixed such that $T_{\mathrm{max}} \approx T_{\mathrm{prior}}$.

(iii) In Section 3.4.3 we test the algorithm against the more difficult egg-box distribution with 243 modes. For comparison, we sample from the same distribution with a geometric ladder constructed to yield 25 % acceptance ratios when applied to the ideal Gaussian discussed in Section 3.2.2.

For all of these tests, $\nu = 10^2$ and $t_0 = 10^3$ are used to control the dynamics in Eq. (3.14), while the sampler uses 100 walkers. Note that these choices, while differ-

ent from the defaults proposed in Section 3.3.2, do satisfy the conditions described in that section[7].

3.4.1 Truncated Gaussian

Our first test case is an n-dimensional, unimodal, unit Gaussian similar to that of Section 3.2.2 but with finite prior volume. The simplicity of this case admits some exact analysis before recourse to numerics, allowing us to test the approximations made in Section 3.2.2.

At low temperatures, where the prior boundaries do not truncate much of the likelihood probability mass, the optimal temperature spacing should be similar to that of the ideal Gaussian. By imposing a step-like cut-off in the prior at a radius of R, there will be some temperature at which this approximation will fail and a geometric spacing becomes inappropriate.

For the likelihood we use the same distribution as in Section 3.2.2, while for the prior we use a uniform distribution over the closed n-ball of radius $R = 30$, centred on the origin. The likelihood and prior are defined by

$$L(\vec{\theta}) \propto \exp\left(-\tfrac{1}{2}\left\|\vec{\theta}\right\|^2\right), \tag{3.15}$$

$$p(\vec{\theta}) \propto \begin{cases} 1 & \text{if } \left\|\vec{\theta}\right\| \leq R, \\ 0 & \text{otherwise,} \end{cases} \tag{3.16}$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^n$. Subsequently, the normalised posterior generated by Eqs. (3.15) and (3.16) at temperature T is

$$\pi_T(\vec{\theta}) = \begin{cases} \frac{(2\pi T)^{-\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}{\tilde{\gamma}\left(\frac{n}{2}, \frac{R^2}{2T}\right)} \exp\left(-\frac{\left\|\vec{\theta}\right\|^2}{2T}\right) & \text{if } \left\|\vec{\theta}\right\| \leq R, \\ 0 & \text{otherwise,} \end{cases} \tag{3.17}$$

where $\tilde{\gamma}(a, z)$ is the lower incomplete gamma function.

In the low-temperature limit, this distribution converges to the ideal Gaussian distribution. We should therefore expect the KL divergence for a step down the temperature ladder to asymptote to (3.10) as $T \to 0$, where the effects of the prior boundary are negligible[8]. Indeed, the KL divergence of (3.17) from T_2 to T_1 is

[7] These simulations were in fact carried out *before* deciding on the values for ν and t_0 suggested in Section 3.3.2.

[8] While we do not consider $T < 1$ in our simulations, the case of $T \to 0$ can equivalently be thought of as $R \to \infty$, since the width of the Gaussian scales with $\sqrt{T}$.

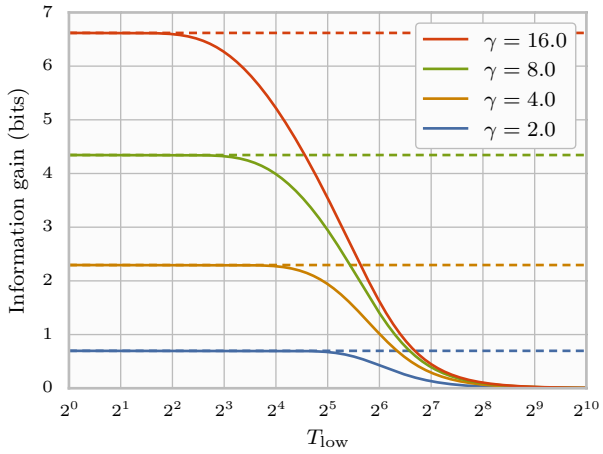

Figure 3.4: The KL divergence, or information gain, from a hot chain at temperature $\gamma\, T_{\text{low}}$ to a colder chain at temperature T_{low}, both sampling from (3.17) at $n = 5$ (solid lines). As $T_{\text{low}} \to 0$, the information gain tends to that of the ideal Gaussian of Section 3.2.2 (dashed lines). The lines are ordered vertically to match the legend.

available analytically as

$$D_{\text{KL}} = -\frac{(T_2 - T_1)\,\tilde{\gamma}\left(1 + \frac{n}{2}, \frac{R^2}{2T_2}\right)}{T_2\,\tilde{\gamma}\left(\frac{n}{2}, \frac{R^2}{2T_1}\right)} + \frac{n}{2}\log\left(\frac{T_2}{T_1}\right) + \log\left(\frac{\tilde{\gamma}\left(\frac{n}{2}, \frac{R^2}{2T_2}\right)}{\tilde{\gamma}\left(\frac{n}{2}, \frac{R^2}{2T_1}\right)}\right). \tag{3.18}$$

If we set $T_2 = \gamma T_1$ (with $\gamma T_1 \ll 1$), then $\tilde{\gamma}(a, z) \to \Gamma(a)$ as $T_1 \to 0$, and the expression reduces to (3.10), as expected.

Fig. 3.4 illustrates this convergence for $n = 5$. The point on this plot at which the solid line diverges from the dashed line, for each γ, predicts the temperature beyond which a geometric spacing of temperatures is no longer optimal (for optimality as defined by uniform KL divergence between chains). This is caused by truncation of the tempered likelihood by the prior boundaries.

For example, note how the KL divergence approaches zero at $T_{\text{low}} \approx 2^7$ for all values of γ. At this temperature, the Gaussian peak has broadened to $\sigma \approx 11$ – comparable to the cut-off radius of the prior ($R = 30$). Meanwhile, the KL divergence becomes *maximal* – i.e., approaches that of the ideal Gaussian – when $T_{\text{high}} \approx 2^7$.

Of course, since the KL divergence cannot easily be assessed empirically by an MCMC sampler, and we must instead resort to using acceptance ratios, we

would like to know how consistent these two schemes are outside the assumptions of Section 3.2.2.

Fig. 3.5 shows contours of constant D_{KL}, calculated from Eq. (3.18), and contours of constant A_i, illustrated by points representing temperature pairs (from ladders selected by the algorithm developed in Section 3.3). In the low temperature limit, as expected, both schemes select a geometric spacing of temperatures consistent with the ideal Gaussian of Section 3.2.2 (i.e., the contours remain constant in γ). At higher temperatures, both schemes depart from the geometric spacing, but they do so differently. The uniform acceptance scheme displays a more gradual departure from a geometric spacing than the contours of constant D_{KL}. The smaller γ selected by the uniform-A scheme outside the geometric regime, however, suggest that closer spacing is required in difficult temperature ranges (e.g., across a phase transition) in order to achieve uniform A than would be required for uniform D_{KL}. There is therefore less risk of a large gap in temperature across such a temperature range, at the cost of (potentially) slightly less efficient communication across the rest of the ladder. The uniform-A criterion for optimality is therefore conservative with respect to a uniform-D_{KL} criterion.

We can also visualise the ladder specification in terms of the density of chains over temperature. We define this density, in $\log T$, as

$$\eta(\log T) = \frac{\mathrm{d}N}{\mathrm{d}\log T} = \frac{1}{\Delta \log T} = \frac{1}{\log \gamma}, \tag{3.19}$$

with $\gamma = T_{i+1}/T_i$, where T_{i+1} and T_i are the chain temperatures to either side of T.

Fig. 3.6 shows this density for a temperature ladder of 20 chains that is in equilibrium under the temperature dynamics of Section 3.3 (the $N = 20$ contour of Fig. 3.5). The density exhibits the expected uniformity of γ for low temperatures but falls for $T \gtrsim 80$. The width σ of the unit Gaussian at temperature T is $\sqrt{T}$, so at this temperature the prior boundary is at $\sim 3\sigma$. At $T = 80$, $\sim 5\,\%$ of the likelihood mass is truncated – compared to $< 0.1\,\%$ for $T = 40$ and $\sim 35\,\%$ for $T = 160$ – indicating that the prior boundary becomes significant in this temperature regime.

This drop in density reflects the convergence of the tempered posterior distribution, π_T, toward the prior as $T \to \infty$. As π_T becomes flatter, fewer chains are needed per $\log T$ to maintain good communication.

Also shown on figure Fig. 3.6 is the square root of the estimated specific heat C_V of the system as discussed in Section 3.2.1, which can be seen to track closely the logarithmic chain density η when appropriately normalised.

From Eq. (3.19) the density η is constant on temperature intervals over which

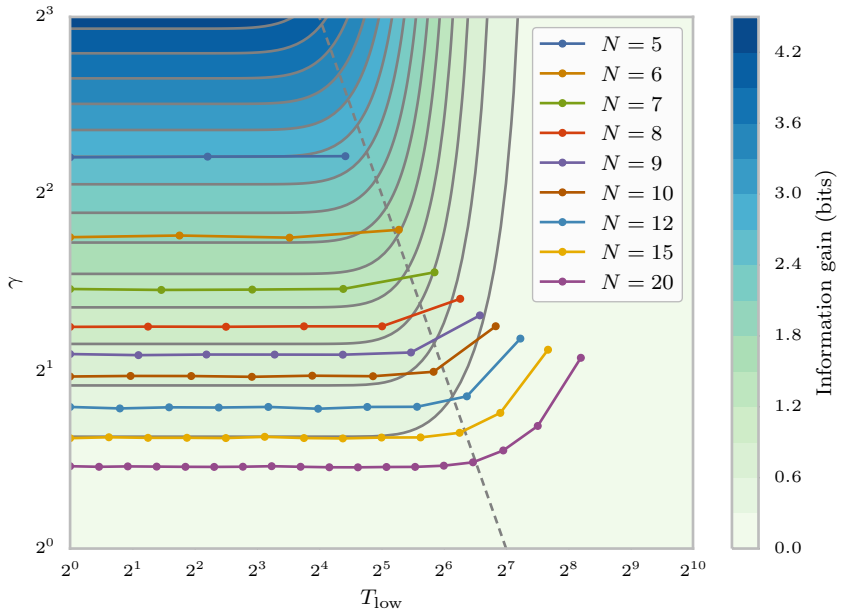

Figure 3.5: A contour plot of the KL divergence, or information gain, from a hot chain at temperature $T_{\text{high}} = \gamma T_{\text{low}}$ to a colder chain at temperature T_{low}, both sampling from the Gaussian likelihood in Eq. (3.17). The coloured lines show the equilibrium N-chain temperature ladders reached by the temperature dynamics algorithm of Section 3.3, where the acceptance ratio is the same between any pair of adjacent chains. The points on these lines represent pairs of adjacent temperatures $(T, \gamma T)$ (excluding the top-most, where $\gamma = \infty$). Overlaid is a contour of $T_{\text{high}} = 2^7$, showing the point at which the KL divergence begins to drop off from that of the ideal Gaussian.

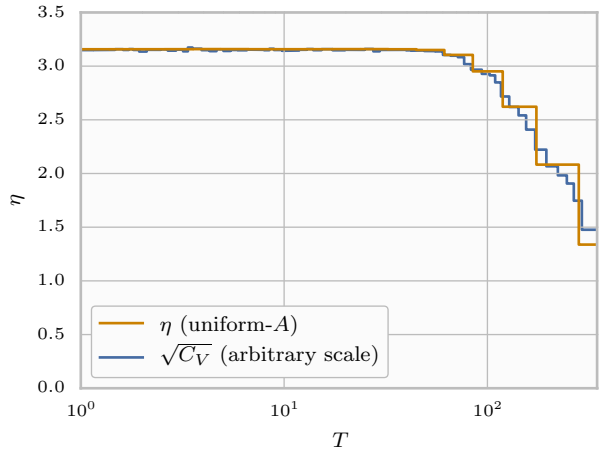

Figure 3.6: **Orange:** The density of chains per $\log T$ under the truncated Gaussian distribution in Eq. (3.17), where $N = 20$, $n = 25$, and temperatures are chosen for uniform acceptance ratios between chains. The chains have equilibrated to 77 % acceptance. **Blue:** The square root of the specific heat of the truncated Gaussian distribution, normalised to match the chain density η of the uniform-A ladder, between T_1 and T_{N-1}. The specific heat C_V, from Eq. (3.3), is estimated from the sample means of $\log L$ over many runs with different temperature ladders.

temperatures are spaced geometrically, consistent with the argument of Kofke (2002) that C_V should be constant over such an interval. Moreover, from Eq. (3.2), the acceptance ratio falls as the likelihood distributions of neighbouring chains become more distinct. This is consistent with the observed scaling of the density η (for a uniform-A ladder) with $\sqrt{C_V}$, since C_V describes the response of the likelihood distribution to changes in temperature.

While the exact provenance of this relationship is unclear, it demonstrates the relevance of the specific heat in determining an effective temperature ladder.

3.4.2 Double Rosenbrock function

The previous test demonstrated how a geometric ladder spaces temperatures too closely at higher temperatures, as the prior boundary becomes significant. While this may be an inefficient use of resources, it at least doesn't drastically inhibit communication between high temperatures and low temperatures. Instead, we now turn to a more complex, bimodal likelihood distribution for which a geometric spacing might cause bottlenecks in the communication between high and low temperatures.

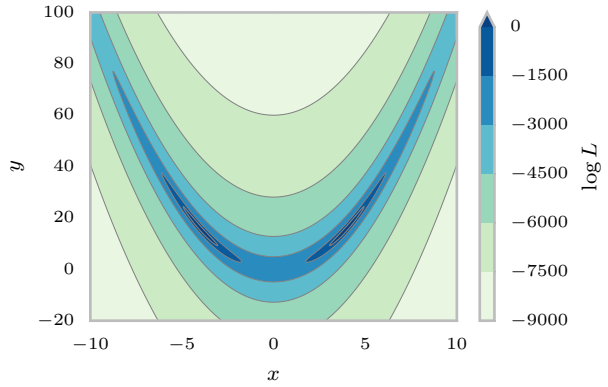

Figure 3.7: The Rosenbrock log likelihood, from Eq. (3.20).

We use a likelihood derived from the two-dimensional Rosenbrock function f:

$$L(x, y) \propto \left(\frac{1}{c + f(x, y)} + \frac{1}{c + f(-x, y)} \right)^{1/T_p}, \tag{3.20}$$

where

$$f(x, y) = (a - x)^2 + b(y - x^2)^2. \tag{3.21}$$

T_p is a pre-tempering factor chosen to increase the contrast of the distribution, making it harder to sample. When $T_p \ll 1$, each mode is locally Gaussian, making the results comparable to the Gaussian example considered in Section 3.2.2.

For the following tests, we use $a = 4$, $b = 1$, $c = 0.1$, and $T_p = 10^{-3}$. We use a flat prior on $[-10, 10] \times [-20, 100]$. Fig. 3.7 illustrates this likelihood over the prior volume.

Test: temperature evolution

As an illustrative example, we first tested the temperature dynamics of Section 3.3 with the double Rosenbrock posterior distribution in Eq. (3.20) using 13 chains. Fig. 3.8 shows the evolution of the temperature ladder according to these dynamics, while Fig. 3.9 shows the chain density $\eta(\log T)$ for the equilibrated temperature ladder.

While the equilibrated chains are distributed uniformly in $\log T$ for $T \lesssim 50$, there is a distinct peak in η at $T \approx 800$, where a simple geometric spacing of temperatures hinders communication between chains. This peak occurs at a phase transition where the two modes of the likelihood distribution begin to mix and $\mathrm{E}[\log L]$ changes rapidly with T, indicated by the sharp change in specific heat in

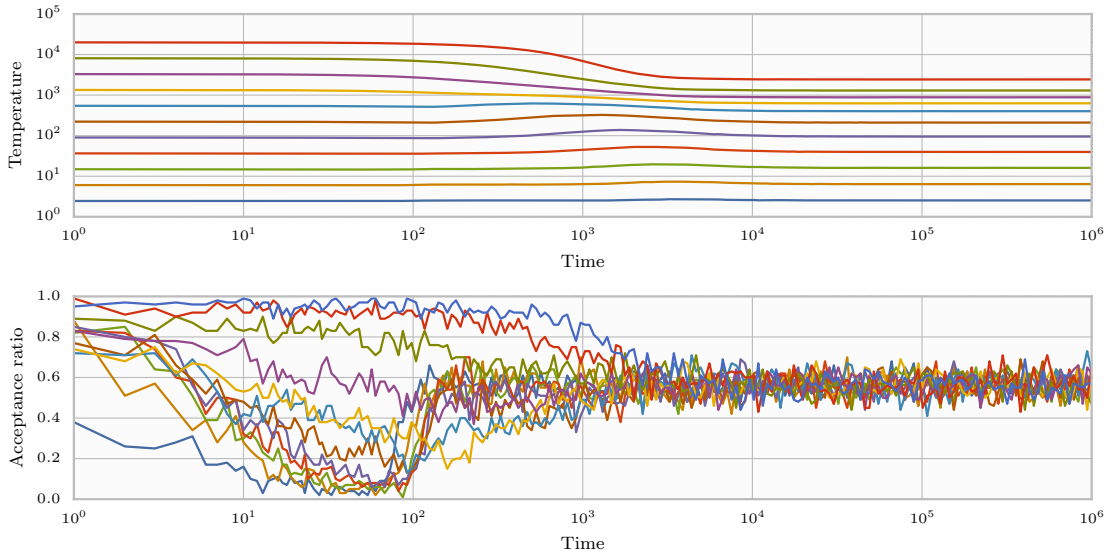

Figure 3.8: The evolution of a ladder of 13 temperatures T_i and acceptance ratios A_i over an *emcee* run of 10^6 iterations under the Rosenbrock likelihood in Eq. (3.20). The dynamical parameters described in Section 3.3.2 are chosen as $\nu = 10^2$ and $t_0 = 10^3$. Chains 1 and 13 are not shown, having fixed temperatures $T_1 = 1$ and $T_{13} = \infty$.

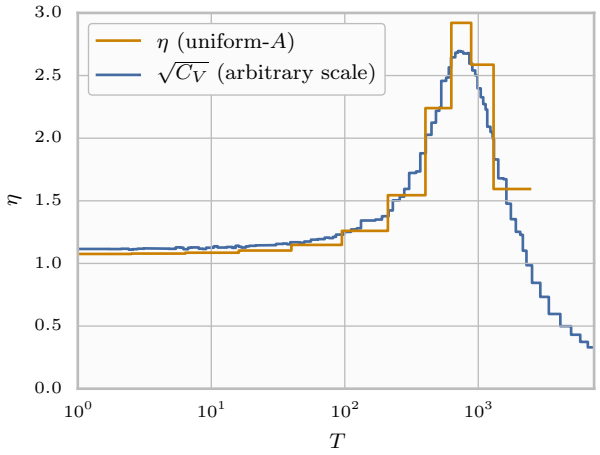

Figure 3.9: **Orange:** The equilibrium density of chains per $\log T$ for the double Rosenbrock run illustrated in Fig. 3.8, where the acceptance ratios have settled to $A_i \approx 0.57$. **Blue:** The square root of the specific heat for the double Rosenbrock distribution, as described in Fig. 3.6.

Fig. 3.9. Since the shape of the likelihood distribution in this regime becomes very sensitive to T, a higher density of chains is needed to maintain a given acceptance ratio. We also note that in the geometric regime (i.e., for low T) the specific heat is approximately $n/2 = 1$, with $\mathrm{E}[A] \approx 57\,\%$, consistent with the values derived for the ideal Gaussian from Eqs. (3.5) and (3.6) respectively.

Ultimately, however, the figure of merit for a temperature specification in a parallel-tempered MCMC simulation is the resulting ACT for the target temperature ($T = 1$) of the sampler. We must therefore test the performance of the sampler empirically.

We use the term ACT to refer to the *integrated autocorrelation time* discussed by Sokal (1997), which we estimate according to the algorithm used in the *acor* package (see Appendix A and Goodman, 2009). For the following tests, we use the ACT of the first parameter, x, as a measure of the efficiency of the sampler (since (3.20) is bimodal in x but unimodal in y).

Test: improvement over a geometric ladder

In Section 3.2 we claimed that aiming for uniform acceptance ratios between chains yields a good temperature ladder. Specifically, we expect that a ladder selected for uniform acceptance ratios should lead to a lower ACT for the $T = 1$ chain than that resulting from a plain geometric ladder.

The geometric ansatz that we use has a fixed maximum temperature such that $T_N = 2 \times 10^4$. As N increases, more chains are added between T_1 and T_N, maintaining the geometric spacing. Under this arrangement, the addition of new temperatures is not redundant even when T_N is already high enough to sample from the prior; the additional chains instead aid inter-chain communication at lower temperatures. Since T_N is close to the temperature at which the posterior becomes the prior, there is little CPU time wasted in sampling redundantly from the prior with several chains, while lower-temperature chains can still communicate with a chain sampling from the prior. Under this set-up, therefore, the ACT always decreases as N increases, per Fig. 3.10.

To test the improvement in ACT, τ, conferred by our temperature dynamics, we allowed *emcee* to explore the target distribution (3.20) with different numbers of chains, N, using both the uniform-A ladders and geometrically spaced ladders. The resulting ACTs, τ_{geo} and τ_{acc}, are plotted against N in Fig. 3.10.

In this example, an N-chain ladder dynamically adapted for uniform acceptance ratios clearly outperforms a geometrically spaced ladder of the same size for all N.

The benefit of a uniform-A ladder is most pronounced at low N – i.e., where there

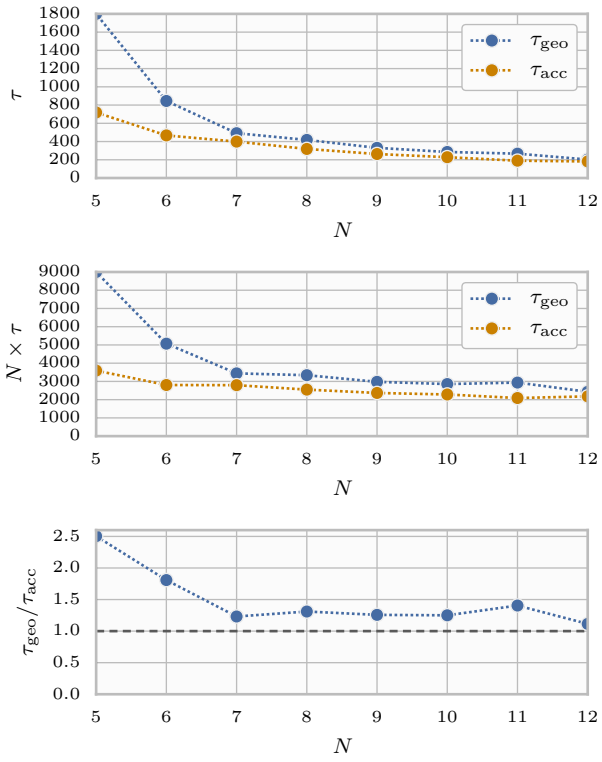

Figure 3.10: **Top:** the ACTs of x for the cold ($T = 1$) chain of a sampler exploring the double Rosenbrock distribution in Eq. (3.20), using both uniform-A and geometrically spaced temperature ladders as a function of the number of chains N. **Middle:** the total CPU time, $N \times \tau$, for the runs. **Bottom:** the relative improvement in the ACT for the uniform-A ladder over the geometric ladder. The joining lines are provided to guide the eye.

are few chains available. In this regime, the sampler will be more sensitive to phase transitions, since the bigger gaps in temperature could cause severe bottlenecks in communication across the temperature ladder.

When N is large, the differences in acceptance ratios between a geometric ladder and one chosen for uniform A becomes less significant. In this case, the difference between the limiting (minimum) acceptance ratio for a ladder and the ladder's average acceptance ratio is proportionally smaller.

In the case of the double Rosenbrock distribution defined by Eq. (3.20), we have found that, once the minimum acceptance ratio for a geometric ladder (terminating at $T_{\max} = 2 \times 10^4$) exceeds $\sim 10\,\%$, reallocating temperatures for uniform acceptance ratios does not reduce the measured ACT by more than $25\,\%$. This occurs when $N \approx 7$ in the current example. Nonetheless, there remains an overall improvement in ACT regardless of N.

Fig. 3.10 also shows, in the middle pane, the total number of iterations per independent sample across all chains. This quantity, given by $N \times \tau$, is proportional to the total CPU time of the simulation, while τ itself is proportional to the CPU time per chain, or *wall time*, of the simulation. In this instance, the CPU time of a run diminishes with N in much the same fashion as the wall time does. The fractional improvement in CPU time is of course the same as for wall time – $\tau_{\text{geo}}/\tau_{\text{acc}}$.

Test: chain removal

To determine whether a uniform-A temperature placement strategy is in fact close to optimal, we assess the contribution of each chain from such a temperature ladder to the efficiency of the sampler, as measured by its ACT. If this contribution is equal for all chains, then we can conclude that it is indeed optimal to have them all exchanging equally – that is, with uniform acceptance ratios.

To this end, we conducted the following test:

(i) Sample from (3.20) with $N = 7$ chains under the temperature dynamics of Section 3.3 until the temperatures have equilibrated to $(T_1, \ldots, T_7)$ to give uniform acceptance ratios.

(ii) Generate 5 new test ladders, each of 6 chains, formed by removing the i^{th} chain from that determined above – i.e., $(T_1, \ldots, T_{i-1}, T_{i+1}, \ldots, T_7)$ – for $i = 2, \ldots, 6$.

(iii) Sample from (3.20) with each of these 5 test ladders and calculate the ACTs on the cold chain, τ_{test}.

Fig. 3.11 shows the ACTs for the cold chain resulting from the test outlined above. While $\tau_{\text{acc}} < \tau_{\text{test}} < \tau_{\text{geo}}$ for ladders of the same N, τ_{test} increases with

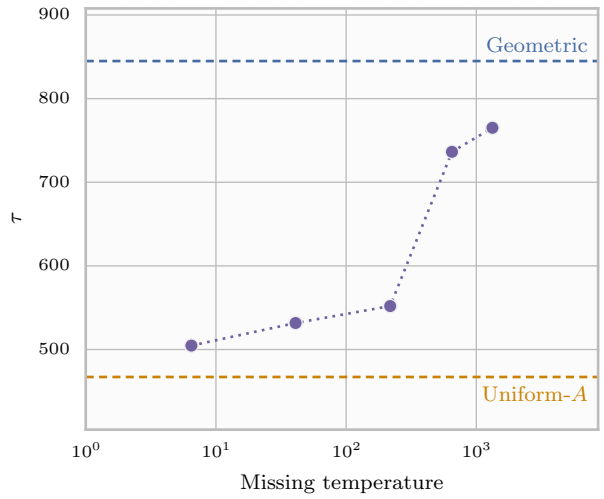

Figure 3.11: The cold-chain ACTs for samplers exploring the double Rosenbrock distribution in Eq. (3.20) per the test described in Section 3.4.2. The points denote the ACTs from ladders generated according to the scheme in Section 3.4.2. The dashed lines above and below identify the ACTs from geometric and uniform-A ladders, respectively, of $N = 6$.

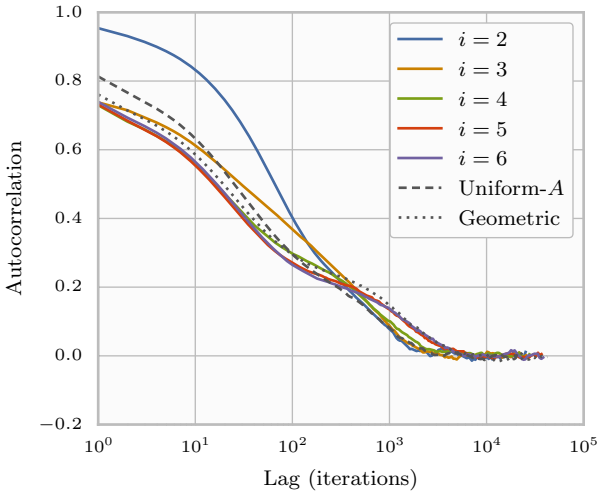

Figure 3.12: The cold-chain autocorrelation function for a sampler exploring the double Rosenbrock distribution in Eq. (3.20). The solid lines correspond to the ladders generated by the scheme outlined in Section 3.4.2, where i is the index of the removed chain. For comparison, the dashed and dotted lines represent respectively uniform-A and geometric ladders of the same size. The approximate ACTs are 504, 531, 551, 736, and 765, for $i = 2, \ldots, 6$; 467 for a uniform-A ladder; and 844 for a geometric ladder (see Fig. 3.11).

the temperature of the chain that is removed, suggesting that additional chains are more useful at higher temperatures. The sharp jump in τ_{test} when a chain above $T \approx 200$ is removed arises from the phase transition that occurs as T approaches T_{prior}, indicated by a peak in C_V (visible in Fig. 3.9).

We can understand this behaviour by examining the complete autocorrelation functions from which these ACTs are estimated. Illustrated in Fig. 3.12, these autocorrelation functions exhibit two distinct time-scales. Firstly, there is a large autocorrelation for lags $\lesssim 100$ for all i – particularly $i = 2$ – corresponding to the ACT of the sampler within one of the two modes: that is, the time taken for the sampler to generate an independent sample without changing mode. Secondly, there is a visible hump in the autocorrelation function for $100 \lesssim \text{lag} \lesssim 2000$, corresponding to the time taken for the sampler to migrate between modes. Removing the second chain from initial geometric ladder of 7 chains increases the intra-mode ACT in particular, but does not affect the inter-mode ACT. Meanwhile, while removing higher temperature chains pushes the secondary hump outward to larger lags, increasing the inter-mode ACT instead.

The overall autocorrelation time in which we are interested, discussed by Sokal (1997) and in Appendix A, represents the time between independent samples of the system. It is therefore set by the time-scale on which the sampler migrates to a new mode independently of the current mode. Removing a chain at higher temperatures increases the inter-modal ACT, and therefore damages the efficiency of the sampler.

Nonetheless, all of the tested temperature ladders yielded lower ACT than the default geometric ladder, despite the geometric ladder being chosen with prior knowledge of T_{prior}.

3.4.3 Egg-box in five dimensions

To test the algorithm's performance on a yet more strongly multi-modal distribution, we use an egg-box distribution defined by the likelihood

$$L(\vec{\theta}) \propto \left(\tfrac{1}{2} \prod_{i=1}^{n} \cos \theta_i + \tfrac{1}{2} \right)^{1/T_p} . \tag{3.22}$$

For a small value of the pre-tempering factor T_p the modes of this distribution become locally Gaussian, and in the low-T regime should therefore generate results similar to those of the Gaussian distributions examined in Section 3.2.2 and Section 3.4.1. For the following tests, we choose $T_p = 10^{-3}$.

We explore this likelihood distribution in 5 dimensions over a flat prior on

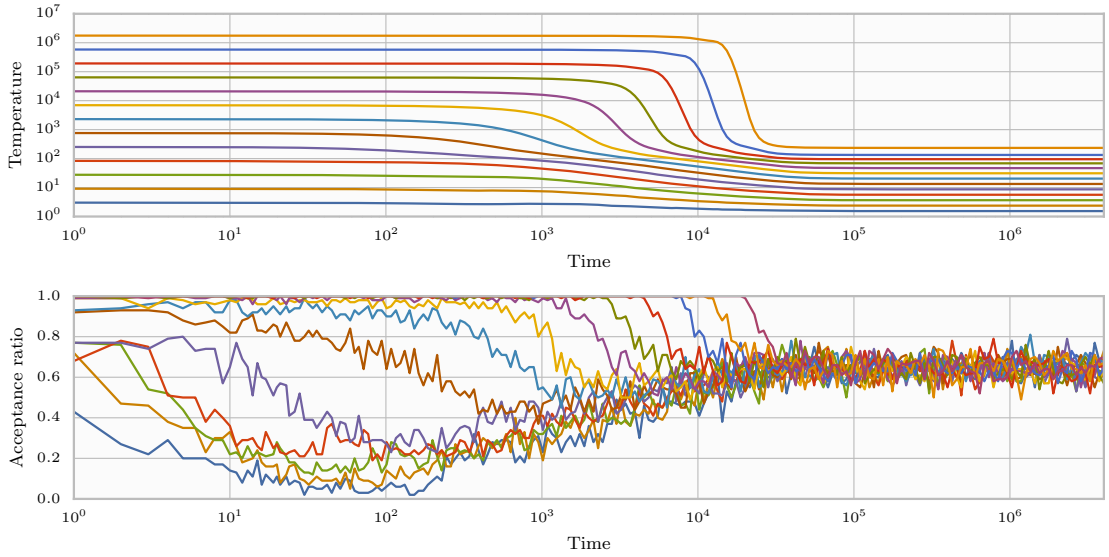

Figure 3.13: The evolution of temperatures T_i and acceptance ratios A_i while sampling with *emcee* from a 5-dimensional egg-box distribution, (3.22), with 15 chains. The dynamical parameters described in Section 3.3.2 are chosen as $\nu = 10^2$ and $t_0 = 10^3$. Chains 1 and 15 are not shown, having fixed temperatures $T_1 = 1$ and $T_{15} = \infty$.

$[-L/2, L/2]^n$, where we choose $L = 3\pi$, giving $3^n = 243$ modes.

Rather than compare our uniform-A temperature ladder against a geometric ladder with a fixed maximum temperature, as in Section 3.4.2, we instead use a geometric ladder constructed to give a fixed acceptance ratio of $\mathrm{E}[A] = 0.25$ when applied to the special case of an ideal Gaussian likelihood (per Section 3.2.2). Such a ladder will not, in general, give uniform acceptance ratios when applied to an arbitrary posterior distribution, but this choice reflects the more realistic scenario where we cannot guess at T_{prior}, and so we resort to assuming that the distribution indeed behaves like an ideal Gaussian.

Fig. 3.13 shows the evolution of the temperatures and acceptance ratios for an *emcee* sampler of 15 chains under the temperature dynamics of Section 3.3. Note the delayed response of the higher-T chains to the dynamical evolution of the lower-T chains. This is caused by a poor choice of T_{max}, such that the higher-T chain pairs start off with acceptance ratios of $\sim 100\,\%$, and are only disturbed when their colder neighbours begin to adjust in response to those below. Fig. 3.14 shows the equilibrium density $\eta(\log T)$ after the ladder has achieved uniform acceptance ratios.

Fig. 3.15 shows the ACTs of the cold chain ($T = 1$) under uniform-A and geometric ladders for the 5-dimensional egg-box problem as a function of the number of temperatures available. In this case, adding more temperatures to a geometric

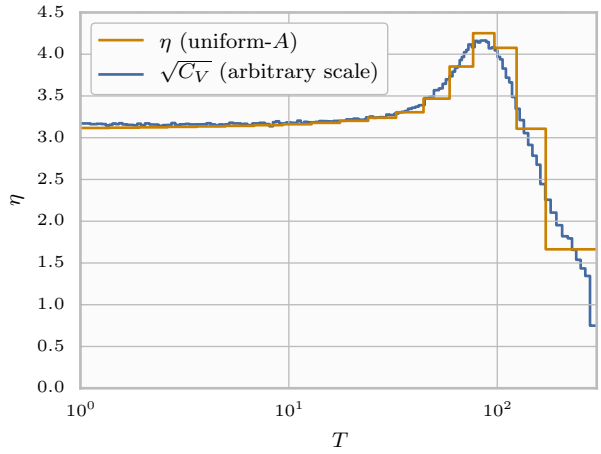

Figure 3.14: Orange: The equilibrium density of chains per $\log T$ for the egg-box run illustrated in Fig. 3.13, where the acceptance ratios have settled to $A_i \approx 0.65$. **Blue:** The square root of the specific heat for the egg-box distribution, as described in Fig. 3.6.

ladder does not reduce the measured ACT of the sampler for $N \geq 7$, since they are added to the high-T end of the ladder, above T_{prior}, and the ratios between lower temperatures do not change. Fig. 3.13 shows that from the initial geometric ladder only around 6 chains are within the range of temperatures spanned by the equilibrium ladder; the remaining 8 (excluding $T_1 = 1$ and $T_N = \infty$) are all above T_{prior} and effectively sample from the prior. In this case, therefore, the geometric spacing that would give uniform acceptance ratios of 25 % for an ideal Gaussian in fact spaces temperatures too widely for $\gtrsim 6$ chains.

Meanwhile, adding more chains to a dynamically adapted ladder clearly reduces the ACT of the sampler in this regime. Moreover, the ACT of a sampler using a uniform-A ladder in this instance is always lower – up to statistical error – than that of a sampler using the geometric ladder of the same N. In the egg-box example, which requires a relatively close spacing of temperatures, the improvement is dramatic when many chains are used: $\tau_{\text{geo}} > 2\tau_{\text{acc}}$ for $N \geq 12$.

The failure of the geometric ladders used in this example for $N \geq 7$ lies in the poor T_{max} chosen by assuming that the distribution behaves like an ideal Gaussian. A geometric spacing is in fact appropriate for a large portion of the temperature range, but its efficacy relies on the ladder terminating at the correct T_{prior}.

When $N < 7$, the geometric and uniform-A ladders show similar ACTs, and the geometric ladders in fact do slightly *better*. While unexpected, this is a consequence of the behaviour of the affine invariant ensemble sampler used in *emcee* (Foreman-Mackey et al., 2013; Goodman & Weare, 2010) as applied to the egg-box likelihood

defined in Eq. (3.22). When such a sampler is applied to a target distribution for which the number of modes n_m is greater than the number of walkers n_w used by the sampler, it behaves as though it is sampling from the prior (albeit inefficiently). There is therefore little benefit in having a chain sampling as high as T_{prior}, and so it is better – in terms of the ACT – to assign more chains to lower temperatures in order to increase their acceptance ratios. In our case, the egg-box likelihood has 243 modes in 5 dimensions, while the sampler uses only 100 walkers, so these walkers tend to become isolated from one another. Since the sampler relies on clustering of walkers on an individual mode to inform jump proposals within that mode, jumps are instead proposed *between* modes when there are on average fewer than one walker per mode.

We anticipate that running the same tests on a traditional single-walker MCMC sampler, or reducing the number of modes of the likelihood distribution so that $n_w \gg n_m$, will dramatically increase $\tau_{\text{geo}}/\tau_{\text{acc}}$ in the low temperature regime. We should expect that $\tau_{\text{geo}} \gg \tau_{\text{acc}}$ when $T_{\max}(N) \ll T_{\text{prior}}$ for the geometric ladder and that $\tau_{\text{geo}} \approx \tau_{\text{acc}}$ when $T_{\max}(N) \approx T_{\text{prior}}$.

Fig. 3.15 therefore illustrates a very specific case for $N < 7$ that does not reflect the importance of choosing $T_{\max} \approx T_{\text{prior}}$. Nonetheless, the ACTs of the two temperature allocation strategies – geometric and uniform-A – are still fairly similar for $N \leq 7$ and there is a distinct improvement for $N > 7$.

3.5 Discussion

The temperature selection scheme set out in Section 3.3 solves two problems in the application of parallel tempering:

(i) It identifies $T_{\max} = \infty$ as a suitable temperature for the hot chain – such that it will sample from the prior – that is independent of the target distribution.

(ii) It allocates a fixed number of intermediate temperatures to ensure good communication between fixed extremal temperatures $T_{\min}$ and $T_{\max}$, and therefore efficient sampling of the target distribution – i.e., with few iterations between independent samples.

The intermediate temperatures are allocated so that acceptance ratios for swaps proposed between neighbouring pairs of chains are uniform across the temperature ladder. The dynamical algorithm that implements this scheme requires only two parameters: ν and t_0. These parameters, discussed in Section 3.3.2, describe only the initial dynamics of the temperatures, setting the time-scale for temperature adjustments, and do not determine the equilibrium uniform-A ladder.

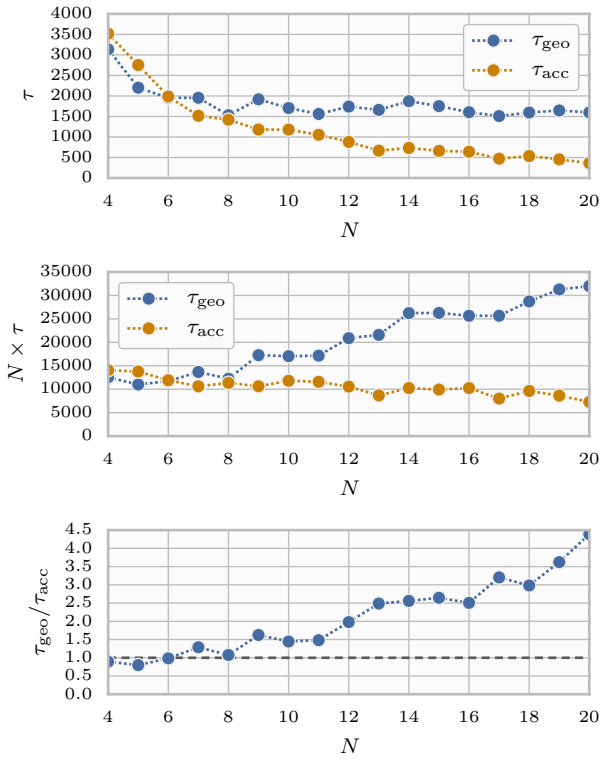

Figure 3.15: **Top:** the ACTs of the cold chain ($T = 1$) of a sampler exploring the egg-box likelihood in Eq. (3.22) with ladders of different sizes N, for both geometric temperature ladders and ladders dynamically adapted for uniform acceptance ratios. **Middle:** the total CPU time, $N \times \tau$, for the runs. **Bottom:** the relative improvement in ACT conferred by dynamically adapting for uniform acceptance ratios over a geometric ladder.

While a temperature configuration that is selected for uniform acceptance ratios between all chain pairs is not necessarily *optimal* in the ACT of the sampler, we have demonstrated that it is generally better than a conventional geometric temperature configuration and provides more consistent behaviour across different likelihood distributions and numbers of chains. Importantly, the dynamics that achieve such a temperature ladder are simple and easily implemented, requiring very little tuning or intervention.

The factor by which the ACT is reduced by the uniform-A scheme depends strongly on the likelihood distribution that is explored and on the specific geometric ladder against which the uniform-A scheme is being compared. For a geometric ladder, one must make an ad hoc choice of the maximum temperature $T_{\max}$; this is difficult and a poor guess can yield a very sub-optimal ladder. In particular, if $T_{\max}$ is not high enough that the sampler can efficiently migrate between modes, then the ACT will be significantly higher than it needs to be. On the other hand, if $T_{\max}$ is too high, then many of the chains will effectively sample from the prior, and CPU time will be wasted in sampling from redundant tempered likelihood distributions.

The uniform-A temperature dynamics guarantee that, for a given number of chains N, no such wastage of CPU time occurs and that there will always be precisely one chain sampling at $T_{\max} = \infty$ (i.e., sampling from the prior). Tests of the dynamics generally demonstrate lower ACTs when compared with geometric temperature ladders of the same number of chains, N.

In Section 3.4.2 we demonstrated that, even with a judicious choice of $T_{\max}$ that is close to T_{prior}, a traditional geometric ladder is outperformed by a ladder chosen for uniform acceptance ratios (with $T_{\max} = \infty$). Fig. 3.10 illustrates that, when T_{prior} is known, a uniform-A ladder confers the greatest reduction in ACT when N is small. In this case, the temperature ratio γ of the geometric ladder is large enough that phase transitions in the distribution of $\log L$ cause a bottleneck in the communication between hot and cold chains around a critical temperature, where $A \ll 1$. The uniform acceptance scheme allocates more chains over these temperature regimes in an effort to optimise the communication.

For larger N, $\tau_{\text{geo}}/\tau_{\text{acc}} \approx 1$, suggesting that – as long as there are no pairs of chains with prohibitively low swap acceptance ratios – a geometric spacing is adequate if $T_{\max}$ is chosen appropriately.

It is unclear how to determine the threshold A below which communication is impeded, but it is likely related to the time-scale of the intra-chain motion of the sampler. If intra-chain jumps are accepted seldom with respect to the rate of inter-chain swaps, then increasing the inter-chain swap acceptance ratio is unlikely to

make the sampler any more efficient.

In general, while $\tau_{\rm geo} > \tau_{\rm acc}$ for all N, the improvement fraction $\tau_{\rm geo}/\tau_{\rm acc}$ will asymptote to 1 as $N \to \infty$. The rate of decay will depend strongly on the target distribution. A system with a wide distribution of $\log L$ (e.g., with many dimensions) or with sharp phase transitions at certain temperatures (e.g., with many modes of various shapes and weights) will see the most benefit from having many chains, while a better-behaved distribution without such features can be efficiently sampled with fewer.

Meanwhile, from our tests on the 5-dimensional egg-box distribution discussed in Section 3.4.3, we can see the consequences of a poor choice of $T_{\rm max}$. While the egg-box distribution does not have as strong a phase transition as the double Rosenbrock function of Section 3.4.2, our ignorance of $T_{\rm prior}$ means that a geometric ladder (which in this case is constructed from a fixed temperature ratio γ) is mostly worse than a uniform-A ladder. Fig. 3.15 demonstrates this, specifically when N is large enough that for a given γ the geometric ladder places many temperatures redundantly above $T_{\rm prior}$. In this case, we see a dramatic improvement in ACT τ from using a uniform-A ladder when compared with a geometric ladder of the same number of chains N; indeed, the ratio $\tau_{\rm geo}/\tau_{\rm acc}$ becomes as large as ~ 4 for the values of N tested. Since $\tau_{\rm geo}$ is independent of N when $N \gtrsim 7$, we should expect that this ratio will saturate as $N \to \infty$, where $\tau_{\rm acc}$ reaches a minimum. Moreover, the CPU time, $N \times \tau$ of the uniform-A runs continues to decrease with N in the explored range, even as the CPU time of the geometric runs rises.

On the other hand, when N is too small for a geometric ladder to reach the prior (i.e., $T_N \ll T_{\rm prior}$), we notice that in fact $\tau_{\rm geo} < \tau_{\rm acc}$. As discussed in Section 3.4.3, this somewhat surprising result arises from a limitation of the ensemble sampler that was used to sample the distribution. We anticipate that if the number of walkers were increased to many times the number of modes – which is required for efficient sampling – the geometric ladder will fail dramatically in this regime of N, giving $\tau_{\rm geo} \gg \tau_{\rm acc}$.

3.5.1 Evidence calculations

The current paper focuses mainly on the efficiency of a parallel-tempered MCMC sampler in producing independent samples from its target distribution. Another important task in Bayesian statistical inference is to compute the evidence integral of the posterior distribution. At a given temperature, this is given by

$$Z(\beta) \equiv \int L(\vec{\theta})^{\beta} p(\vec{\theta})\, \mathrm{d}\vec{\theta}, \tag{3.23}$$

where $\beta \equiv 1/T$ is the inverse temperature.

Since we are interested in the untempered posterior, we wish to calculate $Z(1)$. From Eq. (3.23), we can use thermodynamic integration (Goggans & Chi, 2004; Lartillot & Philippe, 2006) to express the log evidence (relative to the prior) in terms of the mean $\log L$, such that

$$\Delta \log Z \equiv \log Z(1) - \log Z(0) = \int_0^1 \mathrm{E}[\log L]_\beta \, \mathrm{d}\beta, \tag{3.24}$$

to which the logarithm of the integral of the prior, $\log Z(0)$, can be added to give the absolute evidence $\log Z(1)$.

The log evidence can therefore be computed by a sampler through numerical integration of the mean $\log L$ values collected over all of the chains. In the same way that inter-chain communication is hindered by phase transitions in the system, numerical estimation of this integral is susceptible to sharp changes in $\log L$ with the temperature T. Such phase transitions are marked by a diverging specific heat C_V since, from Eq. (3.3), C_V is the derivative of $\log L$ with respect to T.

Since allocating temperatures for uniform acceptance ratios yields a logarithmic chain density η that appears to scale with $\sqrt{C_V}$, such a temperature ladder will naturally increase the accuracy of numerical estimates of (3.24) with respect to one that does not increase η around phase transitions.

We can test the degree of improvement conferred by a uniform-A ladder by returning to the truncated Gaussian discussed in Section 3.4.1. Normalising (3.17) so that $\max \log L = 0$, the log evidence is

$$\begin{aligned} \Delta \log Z &= \left(\frac{\sqrt{2}}{R} \operatorname{erf}\left(\frac{R}{\sqrt{2}} \right) \right)^n \Gamma\left(1 + \frac{n}{2} \right) \\ &\approx -55.1, \end{aligned} \tag{3.25}$$

with $R = 30$ and $n = 25$.

Fig. 3.16 illustrates the numerical estimates of $\Delta \log Z$ from a uniform-A ladder (with $T_{\max} = \infty$) and from geometric ladders with $T_{\max} = 10$ and $T_{\max} = 10^4$. The evidence quadratures for the geometric ladders are augmented with a copy of $\mathrm{E}[\log L]_{T_{\max}}$ placed at $T = \infty$ as a crude measure to cover the integration domain.

The evidence estimates recovered from these samplers are reported in Table 3.1 for 6 chains and 10, from which it is clear that selecting temperatures for uniform acceptance ratios can greatly increase the accuracy of the evidence estimate, bypassing the need to select a good initial temperature ladder. Note that the under- and over-estimates of $\Delta \log Z$ from the geometric ladders in this case are a conse-

Table 3.1: The evidence values of the truncated Gaussian of Section 3.4.1, estimated from a samplers of 6 and 10 temperatures allocated in three different ways, as compared to the analytical result.

	$\Delta \log Z$	
Temperature ladder	$N = 6$	$N = 10$
Uniform-A: $T_{\max} = \infty$	-58.0	-55.9
Geometric: $T_{\max} = 10^4$	-78.0	-61.8
Geometric: $T_{\max} = 10$	-42.3	-41.6
Analytical result	-55.1	

quence of poor choices of $T_{\max}$ rather than of sharp changes in $\mathrm{E}[\log L]$. While these comparisons are reasonable – since for a geometric ladder it is very difficult to pick an appropriate $T_{\max}$ in advance – we expect the presence of phase transitions to increase this disparity, and with it the advantages of adapting the ladder dynamically for uniform acceptance ratios.

3.5.2 Other measures of optimality

We have investigated the performance of a temperature ladder adapted for uniform acceptance ratios in reducing the ACT of a parallel-tempered MCMC sampler. The KL divergence discussed in Section 3.2.2 provides an alternative measure of the distance between two temperatures. In Section 3.4.1 we showed that uniform KL divergence in a temperature ladder does not correspond to uniform acceptance ratios beyond the special case of the ideal, unbounded Gaussian distribution described in Section 3.2.2.

When applied to the truncated Gaussian discussed in Section 3.4.1, for which $D_{\mathrm{KL}}(\pi_{T_i} \| \pi_{T_{i+1}})$ is analytically available, the D_{KL} and A measured between chains drop off at different rates as T approaches T_{prior} (see Fig. 3.5). Indeed, it is possible in principle to estimate the temperature-dependent normalising constants required to adapt on the KL divergence (Cameron & Pettitt, 2014; Geyer, 1994).

Meanwhile, Katzgraber et al. (2006) propose an optimisation scheme in which temperatures are chosen to minimise the round-trip time of a sample from $T_{\min}$ to $T_{\max}$, which they suggest will improve sampling performance on systems with strong phase transitions. Their algorithm is tested on simulations of the two-dimensional Ising model, and is shown to select a different temperature configuration than the uniform-A scheme that has been discussed so far.

However, the ACT of the sampler – what we are ultimately concerned with in

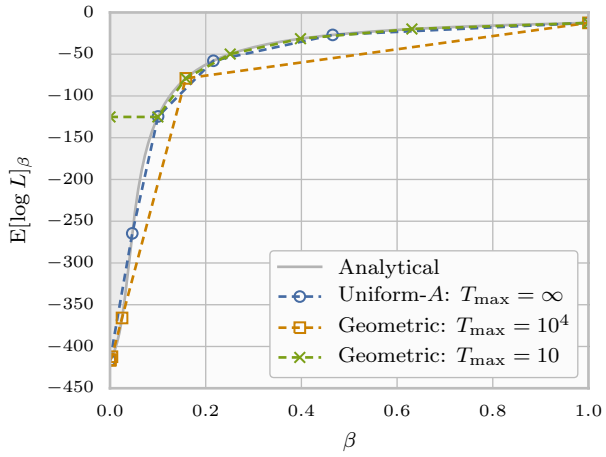

Figure 3.16: An illustration of the thermodynamic quadrature estimates of the log evidence of the truncated Gaussian discussed in Section 3.4.1. The shaded area shows the analytical mean $\log L$ as a function of β, while the dashed lines illustrate numerical approximations from the values of $\log L$ collected by samplers with different ladders, each of 6 temperatures. The resulting evidence estimates are reported in Table 3.1. Note the denser spacing of temperatures in the high-curvature region for the uniform-A ladder, and the errors incurred in extrapolating from $T_{\max}$ to $T = \infty$ ($\beta = 0$) for the geometric ladders.

efficient Bayesian inference – is not discussed, so it is unclear whether this strategy is better than selecting temperatures for uniform acceptance ratios. Their feedback optimisation method in fact prefers a higher density of chains per T across phase transitions of the system than the uniform-A scheme. We have shown, however, that the ACT yielded by a particular ladder is not critically sensitive to under-densities over phase transitions so long as the acceptance ratio is not prohibitively small in these temperature regimes (see Section 3.4.2). Indeed, increasing the density of chains over phase transitions too far might unnecessarily hinder inter-chain communication at other temperatures (by reducing A), leading to an overall rise in ACT.

These reservations, together with the complicated book-keeping involved in optimising for round-trip time, lead us to favour the dynamical method presented in Section 3.3. By comparison, this dynamical method is simple and guaranteed to produce an equilibrium ladder that yields efficient – if not perfectly optimal – sampling from any target distribution, with the proviso of many walkers per temperature.

Chapter 4

Adaptive parallel tempering for compact binary coalescences using LALInference

This chapter is an extension of Chapter 3, containing text from a section of Vousden et al. (2015) that is omitted from the previous chapter in favour of a more thorough discussion. Sections 4.3 and 4.4 contain text that is adapted from this paper, while Figs. 4.2 to 4.5 and 4.7 are reproduced directly.

4.1 Introduction

In Chapter 3 we developed an algorithm that dynamically selects a temperature ladder for a parallel-tempered Markov chain Monte Carlo (MCMC) sampler to achieve uniform acceptance ratios between pairs of neighbouring chains. We did so in the setting of an ensemble-based sampler, in which each chain comprises many individual walkers, providing a reference implementation on the *emcee* ensemble sampler (Foreman-Mackey et al., 2013; Vousden et al., 2015). I now present an example application of this algorithm to a challenging and computationally expensive astrophysical inference problem, along with an implementation for a conventional single-walker sampler.

For gravitational wave (GW) astrophysicists, recovery of the source parameters of a compact binary coalescence (CBC) event from its GW signature, as observed by ground-based interferometric detectors, is a significant challenge in the application of Bayesian statistics (e.g., Aasi et al., 2013c; Raymond et al., 2010; Rodriguez et al., 2014; Singer et al., 2014; Veitch et al., 2015; Vitale et al., 2014; van der Sluys et al., 2008a,b). Parallel-tempered MCMC is one method used within the LIGO Scientific

Collaboration (LSC) to perform this inference, so parameter estimation for CBC detections presents an ideal test for the scheme outlined in Chapter 3.

The remainder of this chapter is divided into three sections. Firstly, in Section 4.2, I will describe the *LALInference* software and my implementation of the dynamics developed in Chapter 3 in this setting.

Secondly, I will present in Section 4.3 example applications to physically motivated test cases involving binary neutron star (BNS) and binary black hole (BBH) CBC signals. I will describe the physical significance of the results and the utility of the dynamic temperature selection mechanism in producing them.

Finally, Section 4.4 will address the limitations of the dynamic parallel tempering scheme in the context of *LALInference* and the CBC events it is designed to analyse.

4.2 Adaptive parallel tempering in LALInference

LALInference is a software package developed and used by the LSC for Bayesian inference on interferometric GW data, intended primarily for the study of CBCs (Veitch et al., 2015). This package is a component of the larger LIGO Algorithm Library (LAL) software suite, and is implemented in a mixture of C and Python to provide

(i) access to GW detector data,

(ii) implementations of the likelihood function defined by Eq. (1.9) in Chapter 1,

(iii) several stochastic methods for sampling posterior distributions and computing their evidence, including parallel-tempered MCMC, and

(iv) post-processing tools for analysing the output of these samplers

(see Veitch et al., 2015 for a complete description).

For the bulk of this chapter, we shall restrict our attention to the MCMC sampler provided by *LALInference*. Unlike *emcee*, this sampler uses only one walker, with the generic stretch-move proposal of Goodman & Weare (2010) replaced with jump proposals that are tuned to the structure of the posterior distribution generated by a CBC signal.

Parallel tempering is implemented in the *LALInference* MCMC sampler by assigning chains to independent computing nodes – e.g., an individual core on a processor – that communicate with each other via the Message Passing Interface (MPI) library.

Each chain samples independently from its tempered target distribution and proposes a swap with its hot neighbour (if it isn't the hottest chain in the ladder)

every T_{skip} samples (100 by default). Their communication isn't synchronised at each iteration, so when chain (i) proposes a swap it must wait for chain $(i+1)$ for complete its current likelihood evaluation. Since the average cost of evaluating the likelihood function varies between chains according to their temperatures, synchronising chains at every iteration would limit the iteration rate to that of the longest likelihood evaluation. Chains must instead synchronise pair-wise when a swap is proposed, so that they can sample unhindered between proposals. This protocol is codified by Algorithm 1 and illustrated in Fig. 4.1.

Algorithm 1: Default *LALInference* PT protocol

```
i ← MPI rank;
t ← 0;
while sampling do
    perform Metropolis–Hastings step;
    t ← t + 1;
    if t mod T_skip = 0 then
        notify (i + 1);
        block;
        send swap proposal to (i + 1);
    if (i − 1) is blocking then
        receive swap proposal from (i − 1);
```

However, the pair-wise synchronisation of the chains in this scheme makes it difficult to implement the dynamics of Chapter 3. Since the dynamics are written in terms of $S \equiv \log \Delta T$, an adjustment requires the entire ladder to be reconstructed. Therefore, in order to implement the temperature dynamics as written, all chains must be synchronised on each temperature adjustment.

Since individual chains do not know in advance when to expect swap proposals from other chains, each chain must poll for a message on every iteration. When T_{skip} samples have elapsed on one chain, it will notify the others and wait; when all chains have finished their current likelihood evaluation, they join and begin swapping.

At this point, each chain (i) first waits for a swap proposal from chain $(i-1)$ (if $i > 0$) and then sends a swap proposal to chain $(i+1)$ (if $i < N-1$). Swapping therefore begins with the cold chain and propagates up the ladder to higher temperatures; this process is codified by Algorithm 2. Note that this synchronisation mechanism is not necessary in the *emcee*-based sampler used in Chapter 3, since its chains synchronise between iterations.

It should be noted that different chains will in fact iterate at different rates, depending on their temperatures. This synchronisation scheme will therefore result

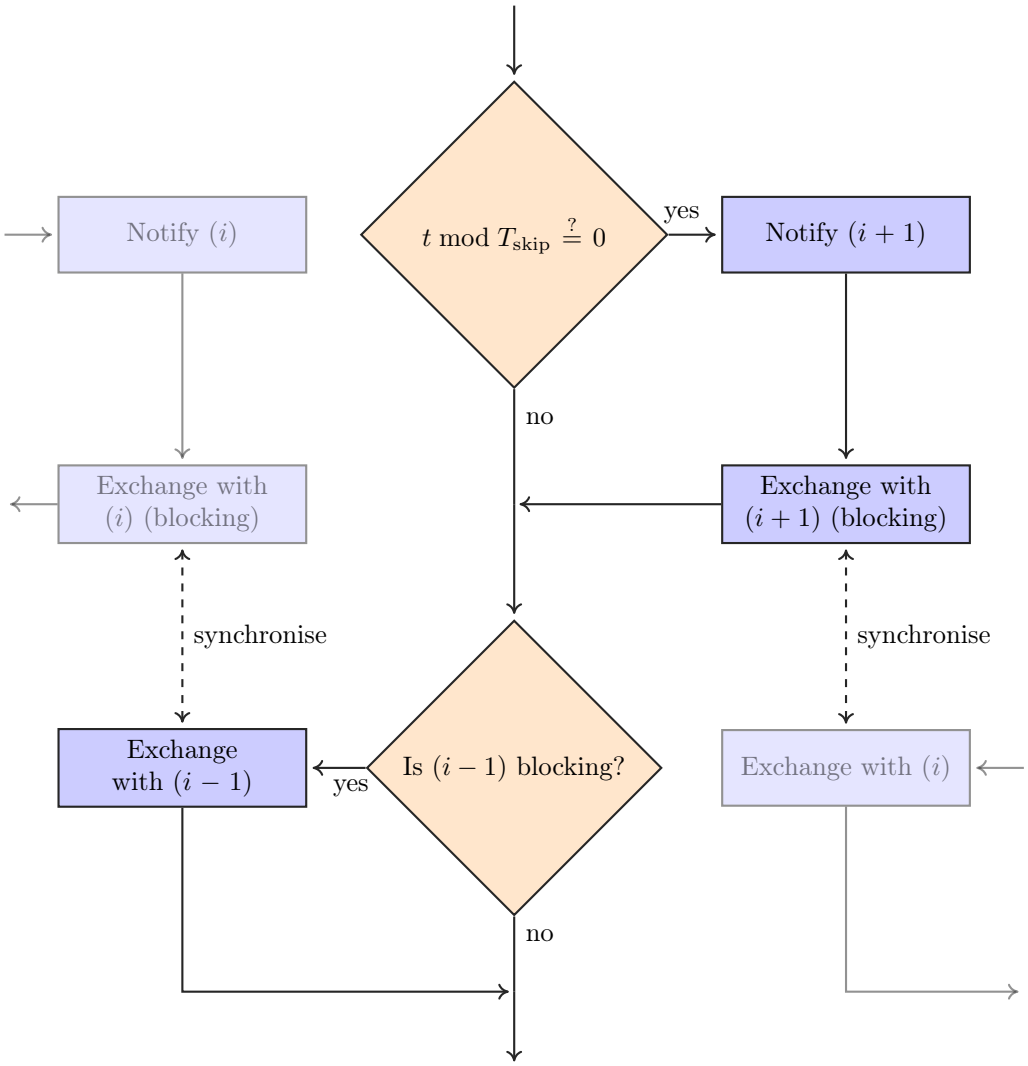

Figure 4.1: A schematic of the asynchronous MPI protocol used to co-ordinate swap proposals in the existing implementation of parallel tempering in *LALInference*. This control flow is invoked on each iteration t of chain (i) of *LALInference*'s MCMC sampler (see Algorithm 1 for a more precise description).

Algorithm 2: Adaptive parallel tempering protocol

```
i ← MPI rank;
t ← 0;                                          // Time (no. of iterations)
k ← 0;                                          // Time of last swap
while sampling do
    perform Metropolis–Hastings step;
    t ← t + 1;
    if notified then
        proposing ← true;
    else if (t − k) ≥ T_skip then
        proposing ← true;
        notify all other chains;
    else
        proposing ← false;
    if proposing then
        k ← t;                                  // Update time of last swap
        synchronise all chains;
        A_i ← 0;
        if i > 0 then
            receive swap proposal from (i − 1);
            if accepted swap then
                A_i ← 1;
        A_{i+1} ← 0;
        if i < N − 1 then
            send swap proposal to (i + 1);
            if accepted swap then
                A_{i+1} ← 1;
        update temperatures from A_i, A_{i+1};
        distribute new temperatures across ladder;
```

in a new swap proposal with every T_{skip} samples generated *by the fastest chain*. We can understand this as follows.

From Eq. (1.3) in Chapter 1, the time until coalescence from a given frequency (e.g., the low-frequency cut-off for the detector) scales inversely as the total mass of the binary. Therefore, higher-mass systems will have shorter in-band waveforms, and the likelihood evaluation from Eq. (1.9) will be less costly. Different chains will therefore iterate at different rates, depending on the marginal distribution of the total mass at each temperature. For example, if the true total mass is below the prior median, colder chains will run more slowly, while if it is above, the colder chains will run more quickly.

This synchronisation mechanism therefore incurs a slight loss in performance over the default mechanism detailed in Algorithm 1. In the worst case, all chains must wait for one complete iteration until the slowest chain has completed its likelihood evaluation. We should therefore expect a fractional slow-down of at most

$$\frac{1}{T_{\text{skip}}} \frac{\mathrm{d}N_{\text{fast}}}{\mathrm{d}N_{\text{slow}}},$$

where N_{fast} is the number of likelihood evaluations on the fastest chain and N_{slow} is that on the slowest chain. In the examples described in Section 4.3, $\mathrm{d}N_{\text{fast}}/\,\mathrm{d}N_{\text{slow}} \approx 2$, so the slow-down amounts to $\sim 2\,\%$ in *LALInference*'s default configuration.

However, the appropriate interval for swap proposals likely depends on the auto-correlation times (ACTs) of the sampler, since we would like each chain to be able to draw at least one independent sample between swap proposals. Given that the cold chain will generally have the longest ACT, a sensible strategy here is to pick T_{skip} to be at least the ACT of the cold chain and for swap proposals to be scheduled by the cold chain. This strategy will be relatively simple to implement in the parallel tempering dynamics described in Algorithm 2. In its current implementation, the *LALInference* sampler might propose temperature swaps before each chain has decorrelated from its position at the last round of swaps, which will likely reduce the efficiency of the temperature dynamics in achieving a uniform-A equilibrium (though in the worst case, it may inhibit convergence).

Finally, *LALInference* includes an adaptive jump proposal with a similar decay to the temperature dynamics of Chapter 3, which is reset when the sampler decides that it has found a new mode. This reset mechanism is also effective when applied to the temperature dynamics, so we adopt it in the *LALInference* implementation discussed above.

From Eq. (3.6), the variance of $\log L$ on the cold chain, for an ideal Gaussian

distribution, is half the number of parameters. Therefore, if we approximate the target distribution as a Gaussian, a jump of $n/2$ in the maximum $\log L$ observed so far suggests that a new mode has been found, and that temperature dynamics should begin afresh. This is implemented as follows.

Each chain records its maximum $\log L$ as it runs. For each iteration t, if

$$\log L_t > \log L_{\max} + \frac{n}{2},$$

the chain records $\log L_t$ as a new maximum $\log L_{\max}$, resets the decay in its dynamical driving term and instructs other chains to do the same. While the decay is reset each time *any* chain records a new maximum, the chains record their $\log L_{\max}$ independently, since a new mode should be found by all chains before the decay is allowed to suppress the temperature dynamics.

The additional command line options that control the temperature dynamics added to *LALInference* are described in Appendix B.

4.3 Tests

To test the implementation of Chapter 3 under *LALInference* – and to compare it with the default geometric temperature ladder – we test both schemes on a number of synthetic GW events simulating the signals received from two different compact binary sources.

We conduct tests against two non-spinning prototype GW sources: a BNS system and a BBH system, detailed in Table 4.1. For each of these prototypes, we simulate coherent detections by a network of GW detectors, for a range of network signal-to-noise ratios (SNRs), by injecting the computed GW signal into mock Gaussian noise generated from the noise power spectral densities (PSDs) of each detector. We simulate a network comprising the Advanced LIGO detectors in Hanford, Washington and Livingston, Louisiana and the Advanced Virgo detector in Cascina, Italy, using noise PSDs that approximate the detectors' design sensitivities (The LIGO Scientific Collaboration, 2010; The Virgo Collaboration, 2009).

Since, for the purposes of parallel tempering, we are concerned mostly with likelihood ratios (rather than absolute likelihoods), it is convenient to normalise the log likelihood so that it is zero in the absence of a signal (i.e., when $h = 0$). The log likelihood (1.9) can be expanded as the sum of a signal-dependent term and a constant that represents the likelihood of the noise-only model (the "null-likelihood"). We therefore subtract the null-likelihood term from the expansion of

Table 4.1: The CBC event prototypes used to test the adaptation scheme of Chapter 3. All prototypes are simulated at distances that yield 5 different SNRs: 10, 11, 15, 19, and 25.

Source	Injection waveform	q	$\mathcal{M}$ ($M_\odot$)	Recovery waveforms
BNS	*SpinTaylorT4*	0.970	1.30	*TaylorF2*
BBH	*IMRPhenomP*	0.996	4.82	*TaylorF2*, *IMRPhenomP*

(1.9) and instead define

$$\log L(\vec{\theta}; s) = \langle s|h(\vec{\theta})\rangle - \frac{1}{2}\langle h(\vec{\theta})|h(\vec{\theta})\rangle\,, \tag{4.1}$$

that is, the difference between the likelihoods of the signal model and the noise-only model.

The SNR ρ of a GW detection is a proxy for the maximum log likelihood, such that $\max_{\vec{\theta}} \log L(\vec{\theta})$ – with the above definition – scales as $\rho^2/2$. The SNR therefore indicates how sharply peaked the posterior distribution will be. Since the SNR can be estimated by the detection pipeline (e.g., Allen et al., 2012; Cannon et al., 2012), it can also be used to decide the $T_{\max}$ used in constructing a geometric temperature ladder for that run, against which we will compare a uniform-A ladder (see Appendix C for details).

We shall attempt to recover the parameters of the injected events with the likelihood function (1.9), using two families of frequency-domain waveform approximants:

(i) *TaylorF2*, which describes with 9 to 11 free parameters the post-Newtonian inspiral of two masses, optionally with spins aligned with the orbital axis (Buonanno et al., 2009), and

(ii) *IMRPhenomP*, which describes the full inspiral-merger-ringdown sequence of a CBC, allowing for arbitrary precessing spins and having 15 free parameters in the *LALInference* implementation (Hannam et al., 2014).

When recovering with *TaylorF2*, we allow for aligned spins in the system, while for both approximants, we analytically marginalise the reference phase ϕ_c out of the likelihood. For these runs, therefore, the *TaylorF2* approximant generates a 10-dimensional parameter space, while *IMRPhenomP* generates a 14-dimensional parameter space.

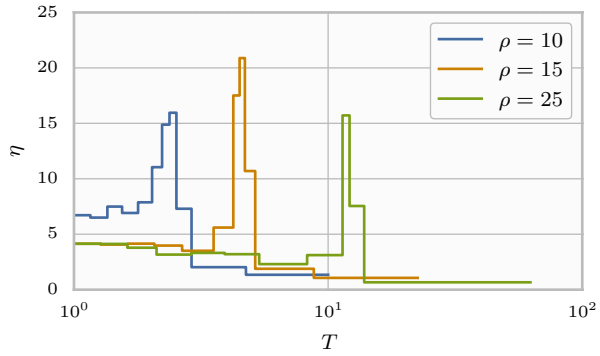

Figure 4.2: The equilibrium (uniform-A) density of chains per $\log T$, from Eq. (3.19), for the *TaylorF2* BBH runs described in Section 4.3 at various SNRs. Note how the features of the ladders scale to higher temperatures as the square of the SNR.

4.3.1 Results: sampling efficiency

Fig. 4.2 shows the effect of the SNR on the equilibrium (uniform-A) chain density η, from Eq. (3.19), selected by our dynamical scheme. While the structure of the temperature ladder is preserved, its features scale to higher temperatures as the SNR of the injected signal increases. Specifically, in the high-SNR limit, $T_{\text{prior}} \propto \rho^2$ from the argument in Appendix C. Indeed, we observe a scaling consistent with this approximation in Fig. 4.2, with features being scaled in temperature by factors of $\sim (15/10)^2$ between the SNR 10 and 15 runs, and $\sim (25/15)^2$ between the SNR 15 and 25 runs.

Meanwhile, Fig. 4.3 shows the ratios of ACTs for runs using uniform-A ladders versus those using the default geometric ladders selected by *LALInference* from the trigger SNRs. The lowest SNR that we simulate, 10, represents a signal that is on the threshold of detectability, where we expect most detections to occur, while the maximum, 25, represents a relatively loud signal (at around the 90$^{\text{th}}$ percentile of detectable events).

While there is significant variation in the ACT measurements between SNRs, there is on average a reduction in ACT of 26 % for the systems and SNRs tested. In general, a uniform-A ladder is at least as effective as a geometric ladder in all cases; that is, the ACT ratio $\tau_{\text{geo}}/\tau_{\text{acc}}$ is never less than one (within error bars). In some cases, this ratio is appreciably greater than one, e.g., for low-SNR BNS events.

However, as we shall discuss in more depth in Sections 4.3.3 and 4.4, the single-walker nature of the *LALInference* sampler inhibits communication between hot and cold chains. Consequently, the chains are instead partitioned into two independent,

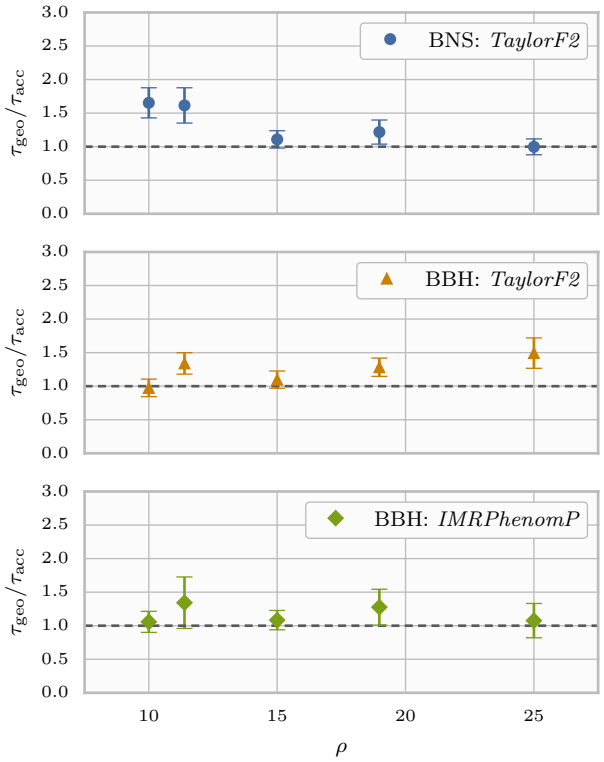

Figure 4.3: The fractional improvements in ACT conferred by a uniform-A temperature ladder over a geometric ladder for the CBC parameter estimation problem described in Section 4.3 at various SNRs.

non-communicating groups, separated by a critical temperature $T_{\rm crit}$ that defines a phase transition. The improvement we observe in Fig. 4.3 therefore arises in fact from more efficient allocation of the temperatures *below* $T_{\rm crit}$. Meanwhile, those chains above $T_{\rm crit}$ – which are sampling in the regime where the noise-only model is preferred over the presence of a GW signal – remain isolated.

4.3.2 Results: physical interpretation

The posterior distribution for one of these problems, a BNS source recovered with *TaylorF2* at an SNR of 25, is illustrated in Figs. 4.4 and 4.5. These show the one- and two-dimensional marginal distributions of the recovered samples, partitioned into intrinsic and extrinsic parameters. Some parameters, such as the chirp mass $\mathcal{M}$, are very accurately measured, while others show multiple modes (e.g., the polarisation angle ψ) or strong correlations (e.g., distance $d_{\rm L}$ and inclination $\theta_{\rm JN}$).

For example, in all of these runs, we observe the extremely high precision with which the chirp mass $\mathcal{M}$ is measured. In the case of the BNS run with SNR 15,

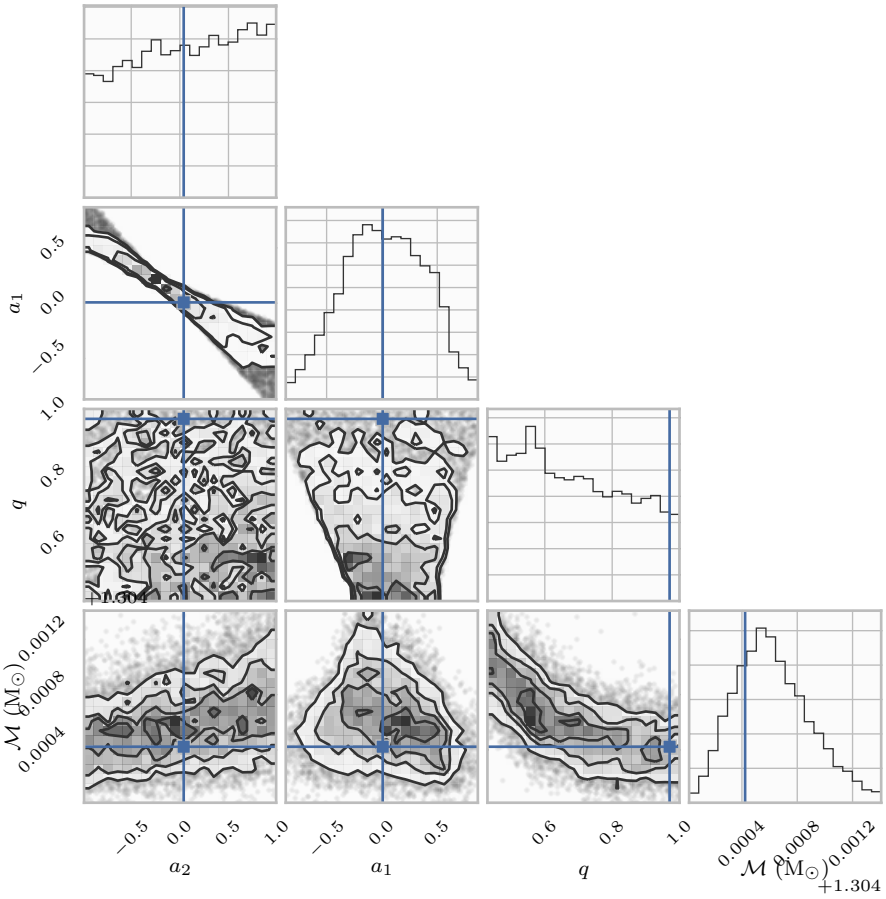

Figure 4.4: The one- and two-dimensional marginal distributions of the intrinsic parameters – described on page 10 – of a BNS event with SNR 25. The true values of parameters are indicated by the blue crosshairs. Note, in particular, the very accurate measurement of the chirp mass $\mathcal{M}$ (the plotted range is only $\sim 0.1\,\%$ of the true value) and how only the total spin, $a_1 + a_2$, is measured. This plot was produced with `triangle.py` (Foreman-Mackey et al., 2014).

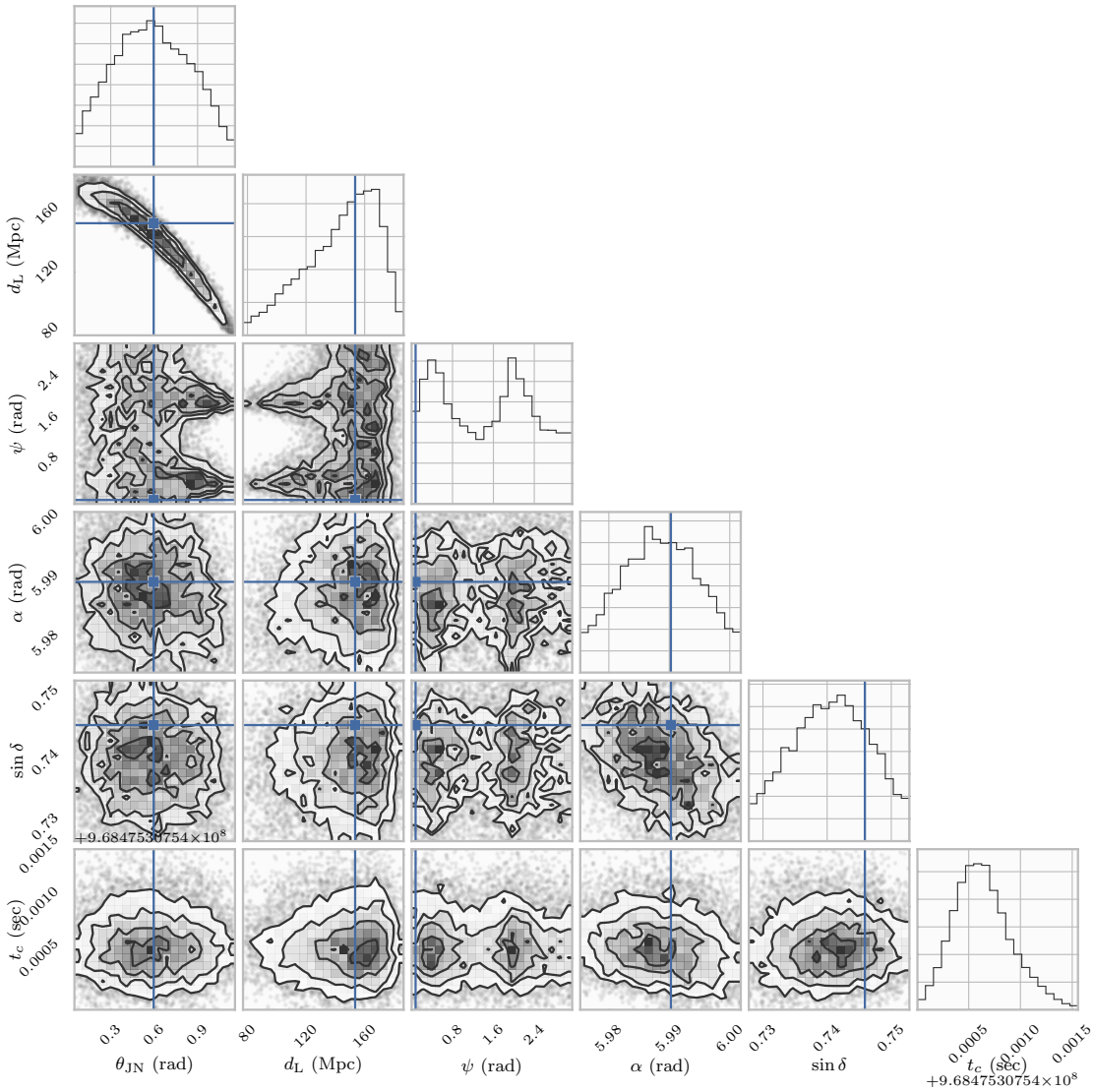

Figure 4.5: The one- and two-dimensional marginal distributions of the extrinsic parameters – described on page 10 – recovered with *TaylorF2* from a BNS event with SNR 25. The true values of the parameters are indicated by the blue crosshairs. Note the multiple modes for the polarisation angle ψ and the strong correlation between distance d_L and inclination θ_{JN}. The two modes observed in the marginal distribution of ψ are in fact identical since, from Fig. 1.1 on page 3, a rotation of $\pi/2$ in ψ is indistinguishable from a phase shift in the GW waveform of π. This plot was produced with `triangle.py` (Foreman-Mackey et al., 2014).

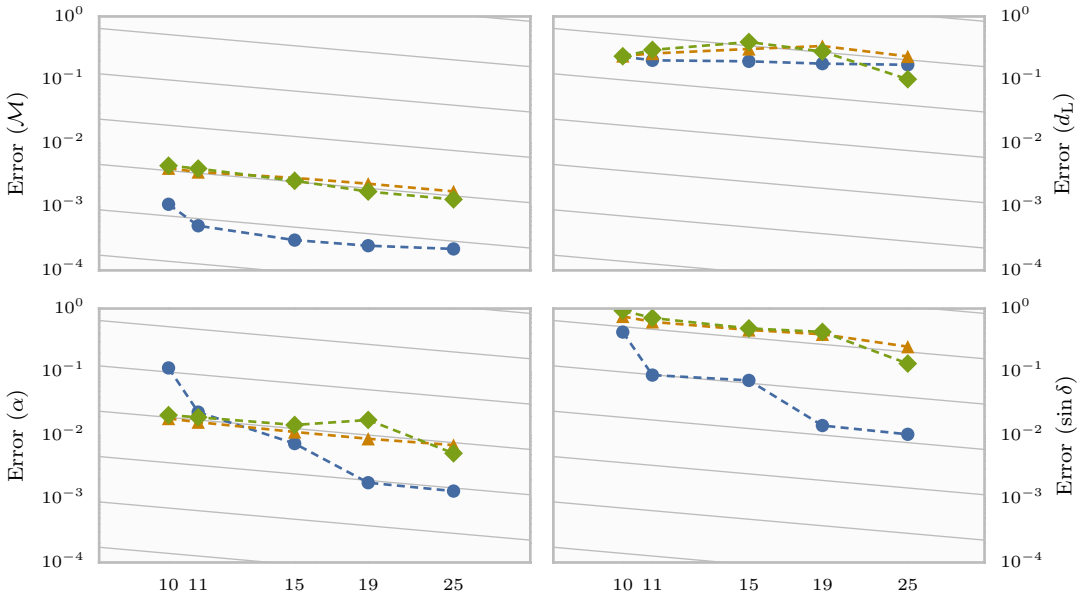

Figure 4.6: The relative errors on four of the parameters recovered from the runs described in Table 4.1. Overlaid are contours proportional to $1/\rho$, showing the dependence of relative errors on SNR predicted by the Fisher information approximation. The sources are depicted with the same colours and markers as in Fig. 4.3.

for instance, the measured $\mathcal{M}$ is precise (and accurate) to within $\sim 0.03\,\%$. This is because the post-Newtonian (PN) phase evolution for the inspiral part a CBC waveform, which is very well measured for sufficiently long waveforms, is controlled to leading order by the chirp mass (Buonanno et al., 2009; Maggiore, 2007).

We should hope that as the SNR of a signal increases, we will be able to measure its source parameters more accurately and precisely. We can quantify this uncertainty as the error in the Bayes estimator – that is, the mean of the posterior distribution – that is recovered by the sampler. Indeed, plotting the relative error, $\delta\theta_i/\theta_i$, on four parameters ($\mathcal{M}$, d_L, α, and $\sin\delta$) in Fig. 4.6, we can see a fall-off in the error that is roughly proportional to $1/\rho$ (in all cases but for d_L).

We can understand this observation by approximating the likelihood peak as a Gaussian distribution. In the high-SNR limit, where this approximation is valid, the Fisher information matrix of the maximum likelihood estimator (MLE) is

$$\Gamma_{ij} \sim \langle \partial_i h | \partial_j h \rangle \tag{4.2}$$

(Maggiore, 2007). Γ_{ij} therefore scales as h^2 and, since we can write the covariance matrix of the Gaussian as $\Sigma = \Gamma^{-1}$, the errors on the measurements of individual parameters from high-SNR signals scale as $1/\rho$.

The exception to this rule is the luminosity distance $d_{\rm L}$, whose uncertainty appears to be roughly constant with SNR. We can understand this as follows. From Eq. (1.6), the SNR of the signal scales as the amplitude of the GW waveform. Since the distance enters the waveform only as a factor of $1/d_{\rm L}$ in the amplitude (Maggiore, 2007), $d_{\rm L}$ is proportional to $1/\rho$. However, from Eq. (4.2), the error $\delta d_{\rm L}$ in luminosity distance also scales with $1/\rho$, and so the relative error $\delta d_{\rm L}/d_{\rm L}$ measured from a signal is independent of its SNR.

We also notice from Fig. 4.6 that the measurement uncertainties from BNS signals are up to an order of magnitude smaller than from BBH signals.

For example, the chirp mass is measured most accurately from the phase evolution of the GW waveform, so its measurement uncertainty is inversely proportional to the noise-weighted number of cycles of the waveform $\mathcal{N}$ – first defined by Damour et al. (2000) as "useful cycles". Since $\mathcal{N}$ is approximately proportional to $\mathcal{M}^{-5/3}$, the ratio of uncertainties in chirp mass between the BNS and BBH systems in Table 4.1 is ~ 8, consistent with Fig. 4.6.

Finally, the smaller errors in the sky location parameters α and $\sin\delta$ for the BNS are a consequence of the wider bandwidth of the detectable part of a BNS signal when compared with that of a BBH signal. Specifically, Fairhurst (2009) shows that the timing accuracy in a detector is

$$\sigma_t = \frac{1}{2\pi\rho\sigma_f},$$

where σ_f is the effective bandwidth obtained by computing the variance of the signal frequency with respect to the noise-weighted inner product defined by Eq. (1.7). Since more massive systems coalesce at lower frequencies (while the low-frequency cut-off remains fixed), the timing accuracy from their signals is poorer, and the uncertainty in sky location – estimated from timing triangulation between detector sites – is correspondingly greater (although Grover et al., 2014 demonstrate that a coherent Bayesian analysis can reduce uncertainty in sky location with respect to timing triangulation).

4.3.3 Results: temperature dynamics

In the test cases detailed in Table 4.1, equal (and large) acceptance rates between all chains do not guarantee good communication of walker positions between extremal temperatures.

This failure can be observed directly by tracking the progress of individual walkers as they are swapped between chains (remembering that there is only one per

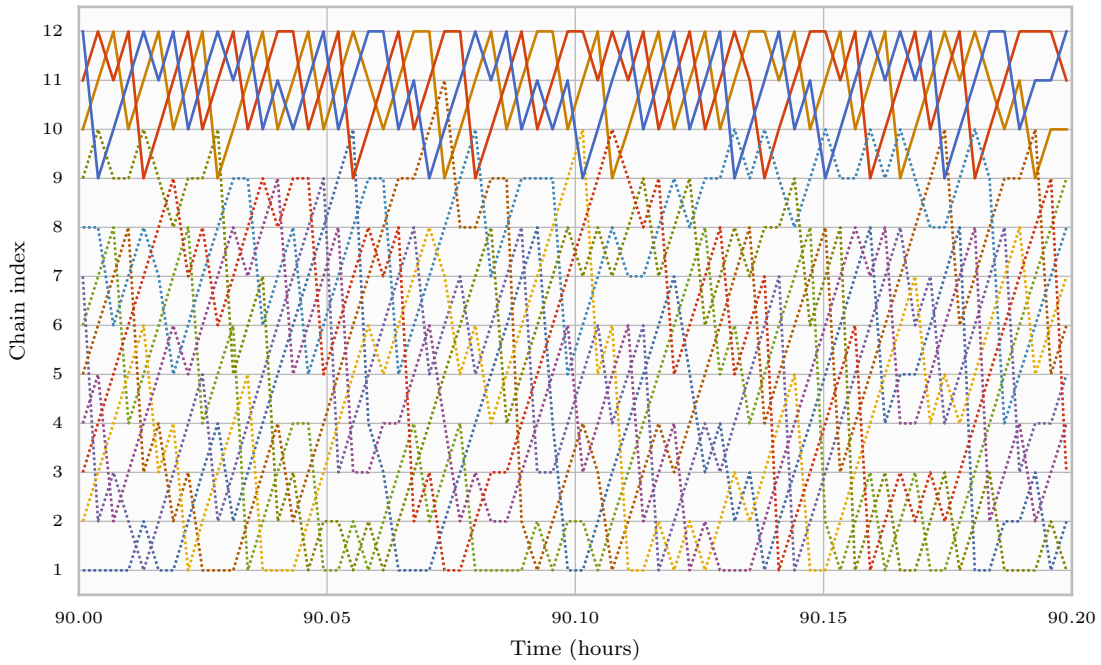

Figure 4.7: The paths traced out between chains by the 12 walkers in with the *LALInference* sampler on a BNS signal of SNR 15. Walkers are identified by their colour. While swap proposals between chains 9 and 10 are frequently accepted, there is no migration of walkers starting above chain 10 (solid lines) to chains below 9 (dotted lines), and vice versa.

chain). For example, Fig. 4.7 illustrates the paths taken by individual walkers, identified by colour, for the BNS test with SNR 15 detailed in Table 4.1. In this case, 9 walkers (whose paths are shown as dotted lines) occupy the high-likelihood part of the parameter space, while the remaining 3 (solid lines) occupy the low-likelihood part.

The progress of the sampler for this run is illustrated in Fig. 4.8. Two distinct likelihood regimes are clearly visible: one in which $\log L \approx 0$, occupied by high-temperature walkers, and one at $\log L \approx 120$, occupied by low-temperature walkers[1]. The high-likelihood peak represents the model in which a signal is present, and therefore depends on the SNR of the signal, while the low-likelihood peak represents the noise-only model, where the signal is weak or absent and the posterior is dominated by the prior.

Regardless of the temperature that the chains in this simulation select, these two likelihood regimes remain separate. The only way to satisfy the uniform-A criterion is therefore to drive the temperatures so that $S \to 0$, and so the temperatures

[1]Note how the SNR of this run predicts that $\max \log L \approx 15^2/2 \approx 110$, consistent with observations from this plot.

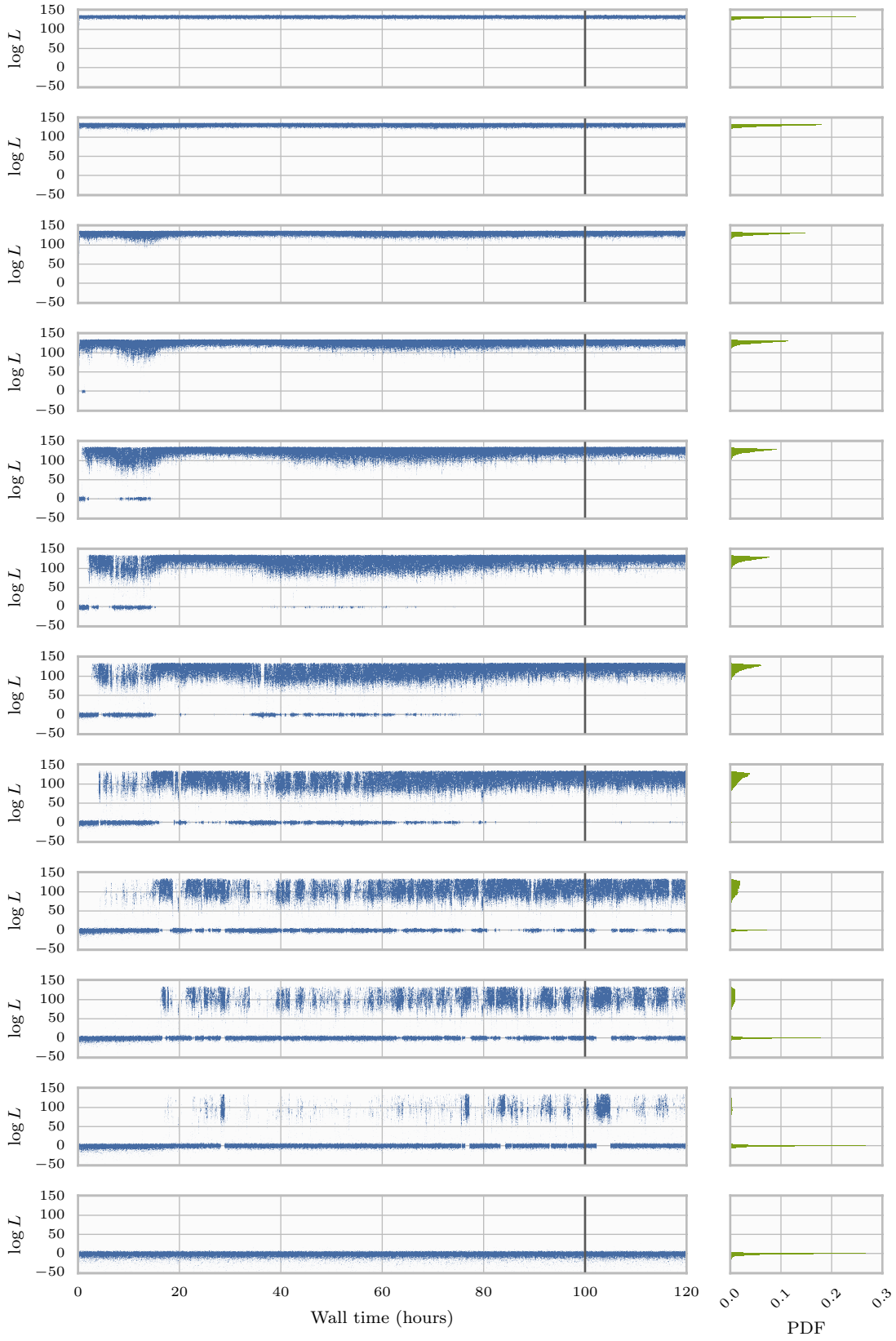

Figure 4.8: The $\log L$ of individual samples over a 5 day run of *LALInference*'s MCMC sampler, using the adaptive scheme of Chapter 3, for an SNR 15 BNS event. Temperatures are arranged in ascending order from top to bottom, while the vertical line denotes the burn-in time used to generate the histograms to the right.

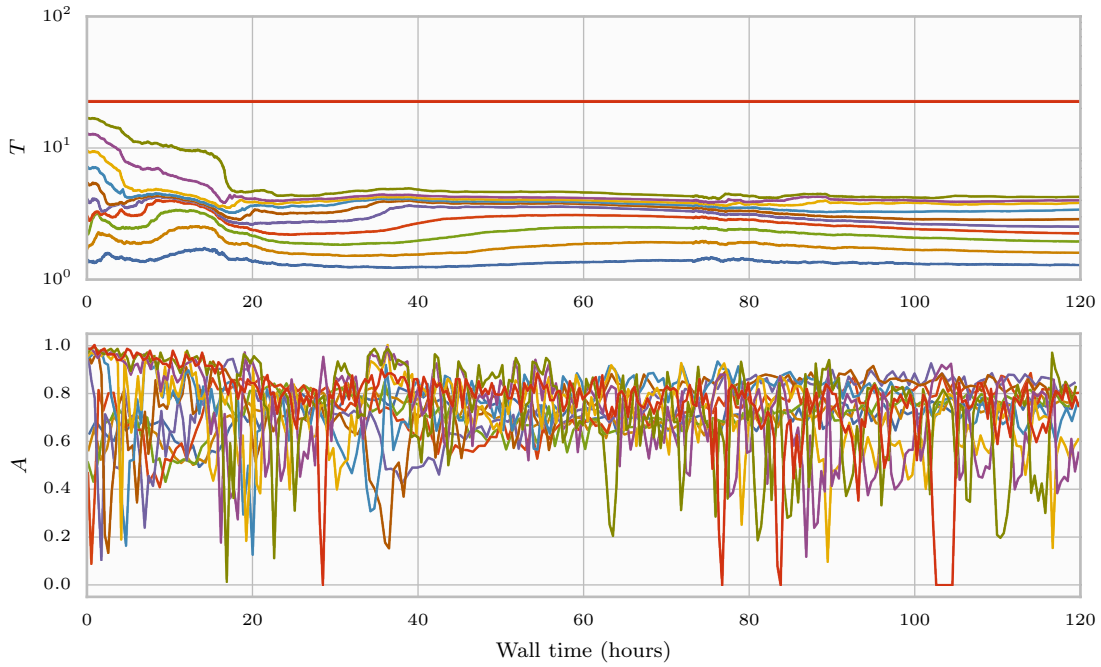

Figure 4.9: The evolution of temperatures T_i and accompanying swap acceptance ratios A_i for the *LALInference* MCMC sampler on the BNS event with SNR 15. Chain 1 is not shown, having fixed temperature $T_1 = 1$.

converge as $t \to \infty$. This coalescence is visible in Fig. 4.9, and corresponds to the "barrier" observed in Fig. 4.7 that prohibits transfer of walkers between the cold and hot groups of chains. While this achieves large acceptance rates between chains, it does not allow walkers on the high-likelihood peak to access the low-likelihood peak, and vice versa.

This barrier represents a phase transition in the likelihood distribution of the system as the temperature is increased from $T = 1$ to $T = T_{\text{prior}}$, and occurs at the temperature at which the evidence of the signal model is equal to that of the noise-only model. In the case of the BNS event with SNR 15 discussed above, the temperature evolution depicted in Fig. 4.9 suggests that the critical temperature of this phase transition is $T_{\text{crit}} \approx 4$. Equivalently, if the models have equal evidence, we should expect equally many samples in the low-likelihood peak as in the high-likelihood peak. Figure 4.8 shows that this is approximately true for the 3$^{\text{rd}}$ hottest chain, whose temperature is indeed ~ 4.

For higher SNRs, where the posterior distribution is more strongly peaked, the phase transition becomes yet more problematic. For example, the SNR 25 run illustrated by Fig. 4.10, exhibits mode-hopping between the high- and low-likelihood peaks on time-scales of tens of hours of wall time (millions of iterations). The ACT

for each walker's exploration of the parameter space is therefore extremely long, and the dynamics are no longer adiabatic with respect to changes in $\mathrm{E}[\log L]$. Indeed, this run demonstrates long-term features in the temperature evolution in response to sharp changes in the acceptance ratio induced by this mode-hopping, visible in Fig. 4.11.

A useful diagnostic for this problem is to compare the distribution of likelihood ratios for intra-chain jumps with that for inter-chain jumps. If the likelihood peaks are mixing correctly between temperatures, then we should expect these distributions to overlap.

For instance, considering the BNS event with SNR 10, Fig. 4.12 exhibits the same separation of peaks in the likelihood distribution as observed in other runs. However, Fig. 4.14 shows that an individual walker will never (or at least very rarely) find its way from the low-likelihood peak (the noise-only model) to the high likelihood peak (the signal model), since there is no intermediate regime in $\log L$ to bridge this gap. While the inter-chain swaps shown in Fig. 4.13 swap chains between likelihood peaks, they don't help the walkers to move between them, severely limiting the benefits of parallel tempering in this application.

4.4 Discussion

In this chapter I have described an implementation of the dynamic temperature selection scheme developed in Chapter 3 for the parallel tempered MCMC sampler used in *LALInference*. This implementation has conferred benefits in sampling efficiency of tens of percent, as measured by the sampler's ACT. Indeed, within error bars, the ACT of the sampler measured under the uniform-A temperature ladder selected by this scheme is never greater than that measured under the default geometric ladder (chosen by *LALInference* according to the estimated SNR of the event).

However, while we can improve the performance of parallel tempering, these tests have also exposed problems in its use for CBC parameter estimation problems. In these test cases, a temperature ladder with uniform acceptance ratios between neighbouring chains does not in fact correspond to efficient transfer of walkers between hot and cold chains. We can interpret this failure as follows.

When $T = 1$, the posterior is dominated by the likelihood, peaked around the true parameter values; when $T = \infty$, it is dominated by the much larger prior volume far away from the parameter values, corresponding to a weak or absent signal; and at a critical temperature T_{crit} that defines a phase transition, the two

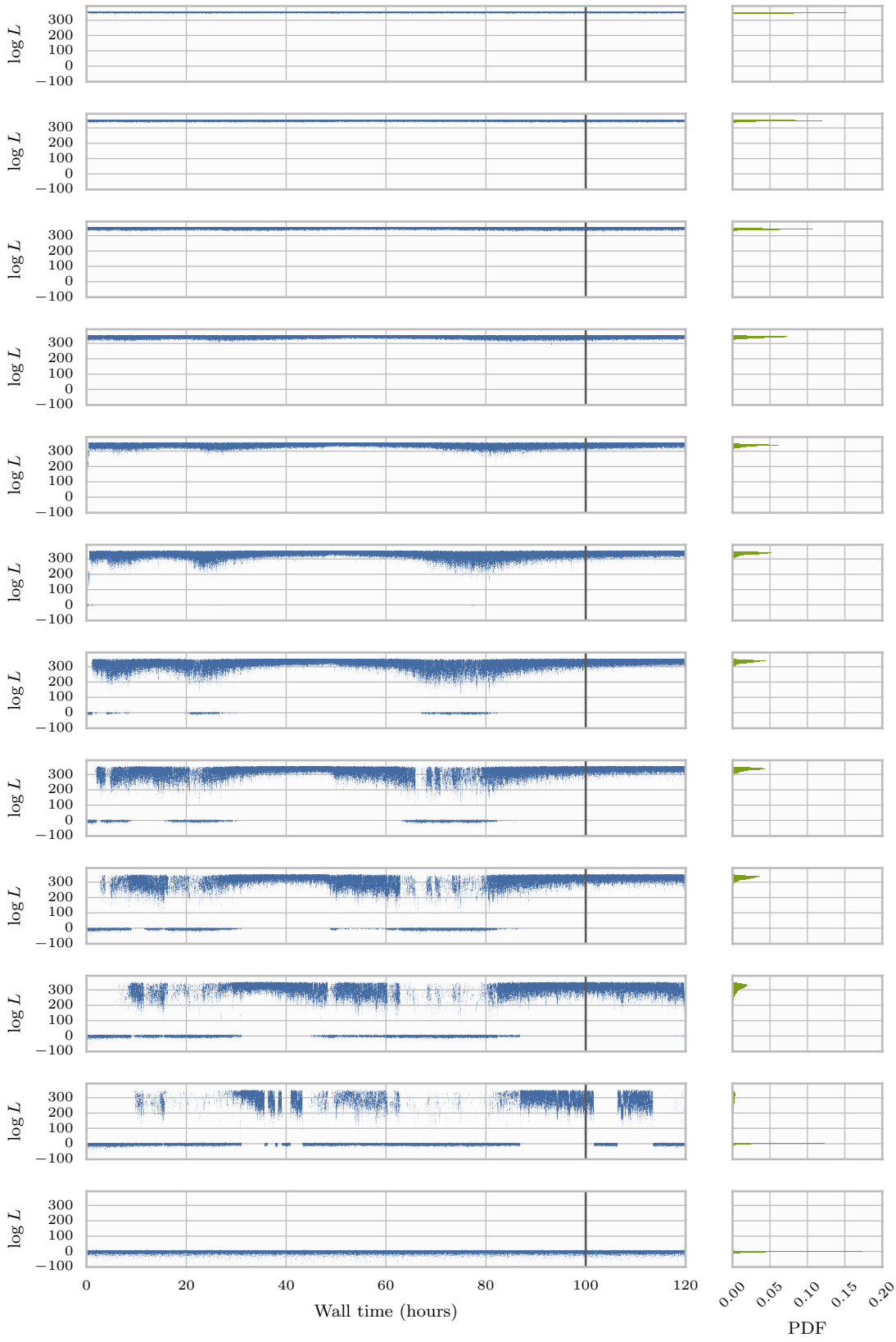

Figure 4.10: The log L samples for an SNR 25 BNS event, as described in Fig. 4.8.

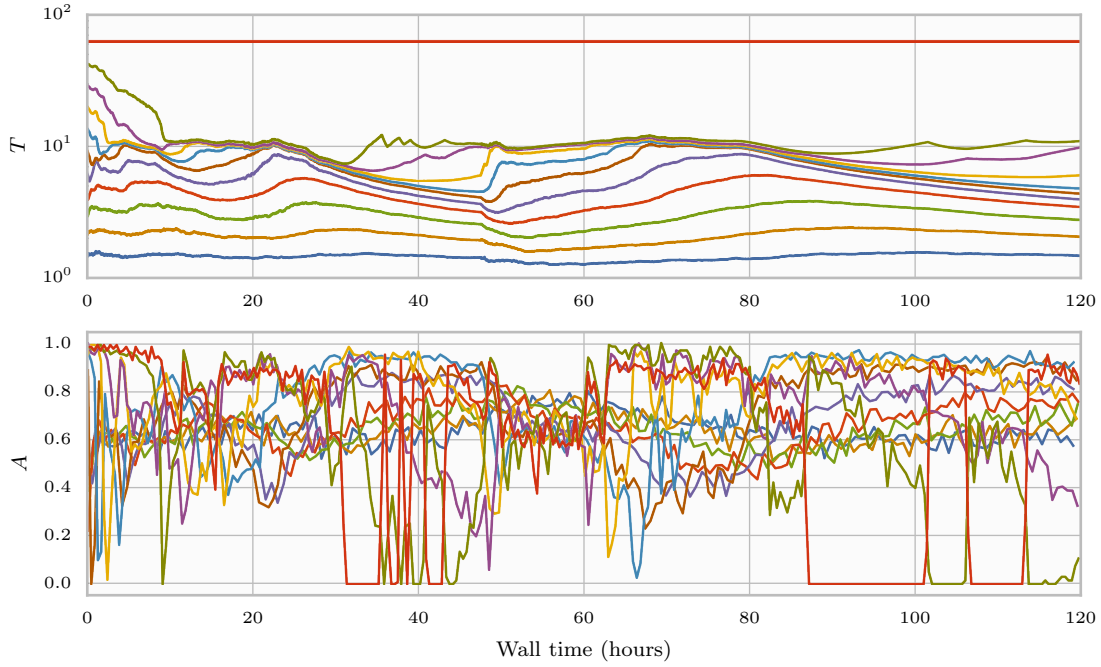

Figure 4.11: The evolution of temperatures T_i and accompanying swap acceptance ratios A_i for the *LALInference* MCMC sampler on the BNS event with SNR 25. Chain 1 is not shown, having fixed temperature $T_1 = 1$, while chain 12 is fixed at $T = 25^2/10$, calculated from the SNR.

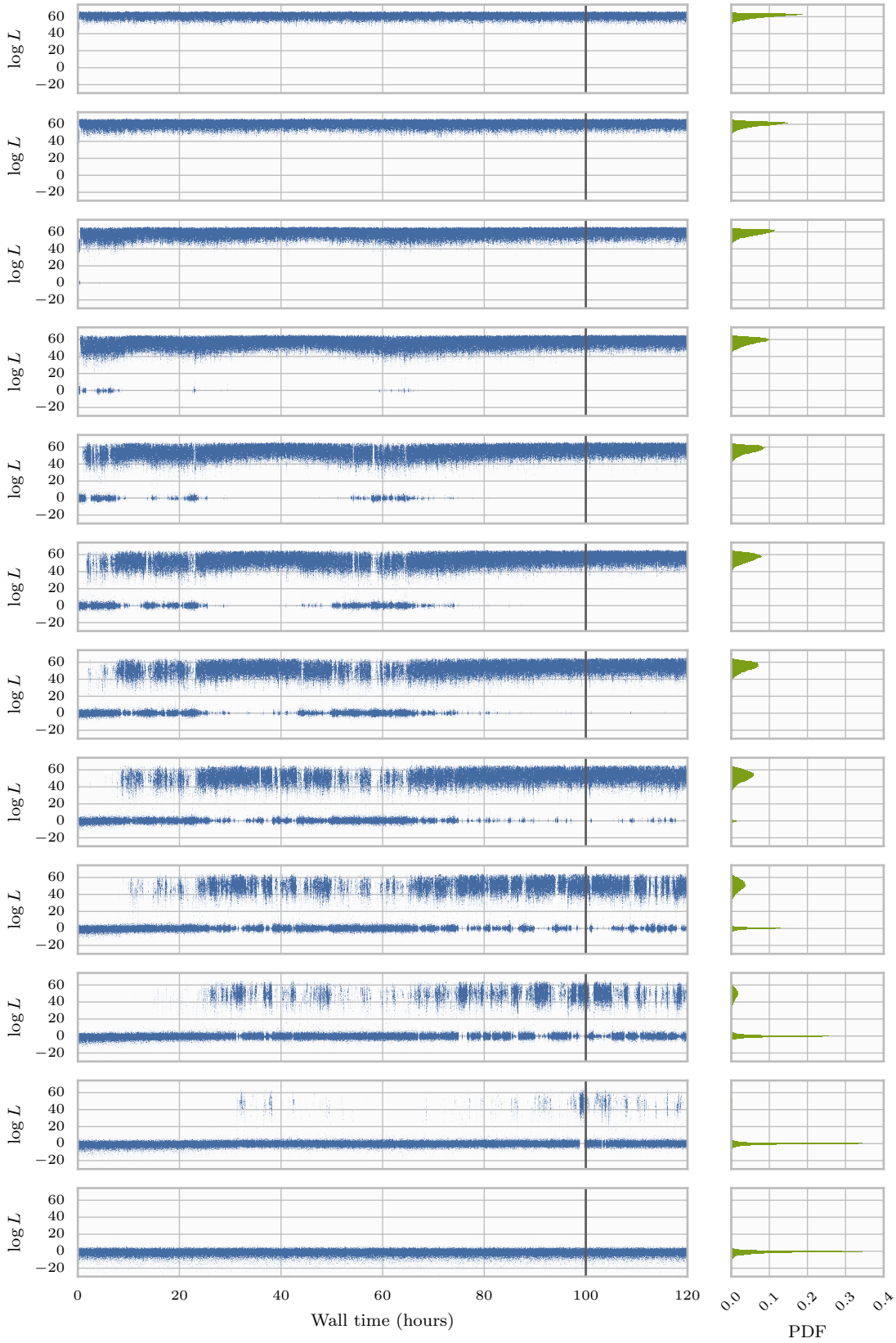

Figure 4.12: The $\log L$ samples for an SNR 10 BNS event, as described in Fig. 4.8.

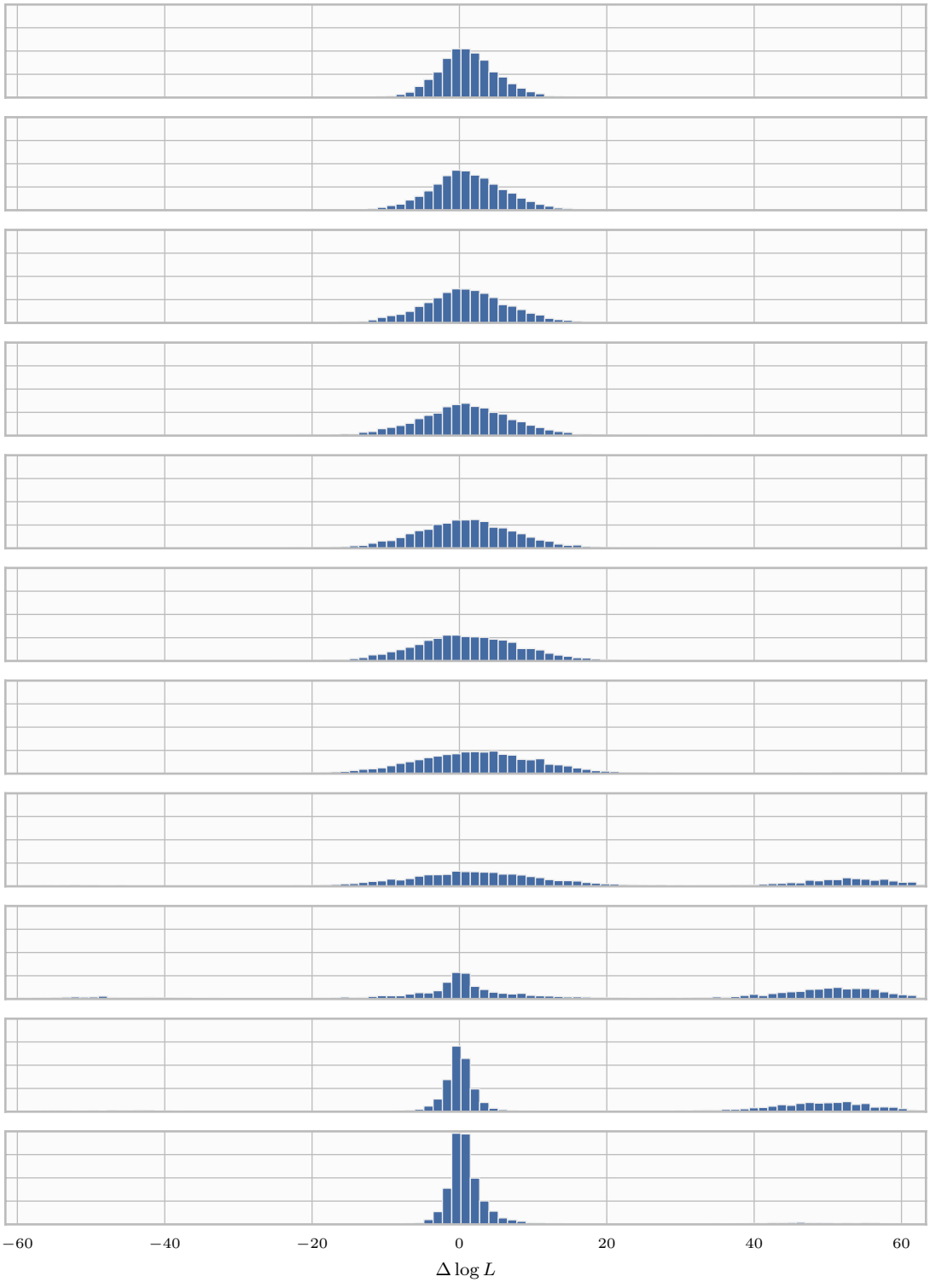

Figure 4.13: A histogram showing the distribution of differences in $\log L$ for accepted inter-chain swap proposals for the SNR 10 BNS event detailed in Table 4.1. Note the wings at $\Delta \log L \approx \pm 50$ that correspond to jumps between the signal and noise-only models. The vertical scale is arbitrary but uniform across plots; plots are arranged in ascending order of temperature from top to bottom.

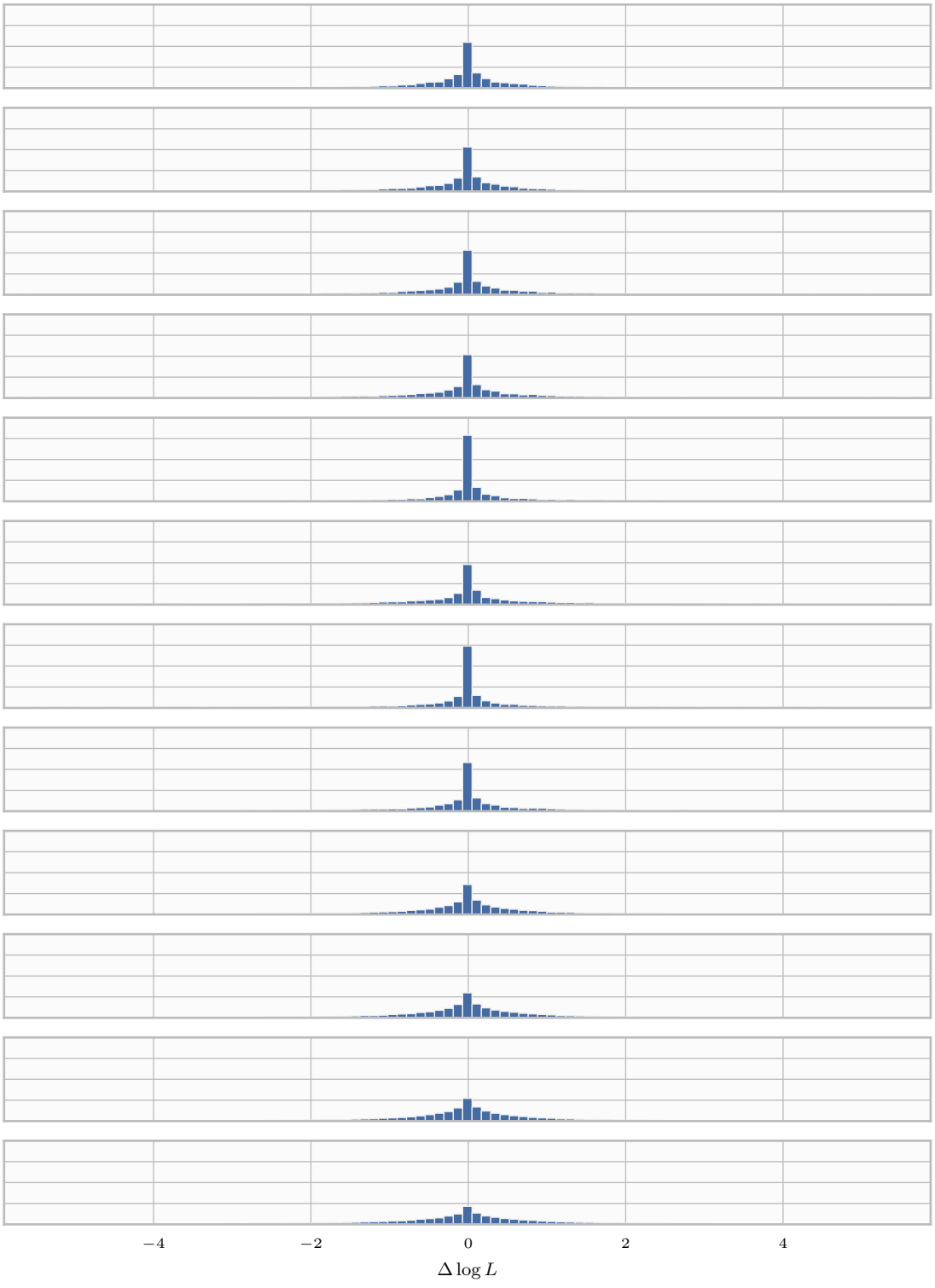

Figure 4.14: A histogram showing the distribution of differences in $\log L$ for accepted intra-chain jump proposals for the SNR 10 BNS event detailed in Table 4.1. Note the narrow width of the distribution relative to that of inter-chain jumps in Fig. 4.13. The vertical scale is arbitrary but uniform across plots; plots are arranged in ascending order of temperature from top to bottom.

contribute comparably to the posterior. The posterior is therefore described by two distinct peaks in the likelihood distribution: a high-likelihood peak near the signal parameters, and a low-likelihood peak in the region of significant prior support. In effect, this becomes a reversible-jump MCMC problem with two distinct modes: the high-likelihood signal model, and the low-likelihood noise-only model.

In this situation it is difficult engineer efficient swap proposals between models since they occupy very different likelihood regimes, between which there is no bridge. The dynamical algorithm therefore has a tendency to select very small temperature gaps around phase transitions representing, since $A \to 1$ as $\Delta\beta \to 0$ regardless of the likelihoods of the chains, from Eq. (3.2). However, the likelihood distributions in each of the two models described above will remain distinct enough that there is no intra-chain migration of walkers between them. Consequently, despite efficient swapping between chains, the higher temperatures do not help low-temperature walkers to efficiently jump between the two modes.

The usefulness of parallel tempering relies on the ability of individual walkers both to explore the entire parameter space and to explore individual modes in detail. The parallel tempering formalism achieves this by allowing walkers to move between temperatures, under the assumption that the scale of a walker's movement through the parameter space itself scales with the temperature at which it is exploring.

The CBC tests discussed in Section 4.3 have shown that this assumption is not always justified. In these test cases, there is no temperature at which the high-likelihood peak corresponding to the signal model broadens enough that it mixes with the low-likelihood peak that represents the noise-only model. Consequently, an individual walker must accept a jump proposal directly from the high-likelihood region of parameter space to the low-likelihood region, which is highly improbable. For example, in the SNR 25 BNS case detailed in Table 4.1, this corresponds to a change in $\log L$ of ~ 300.

An effective metric by which to judge how well a parallel tempering implementation is working is the rate of round trips of an individual walker between the cold chain (at $T = 1$) and the hot chain. In the SNR 25 BNS case, each walker recorded ~ 3 such trips on average over the course of a run of $\sim 4 \times 10^7$ iterations, for a rate of $\sim 7 \times 10^{-8}$. It is therefore clear that the chains above the "barrier" do not contribute to the samples recorded on the $T = 1$ chain in which we are ultimately interested, and are instead wasting CPU time. Lower SNRs yielded slightly higher round trip rates, owing to the milder phase transitions between the signal and noise-only models. For example, the walkers in the SNR 10 and 15 runs recorded round trip rates of $\sim 4 \times 10^{-7}$ and $\sim 2 \times 10^{-7}$ respectively.

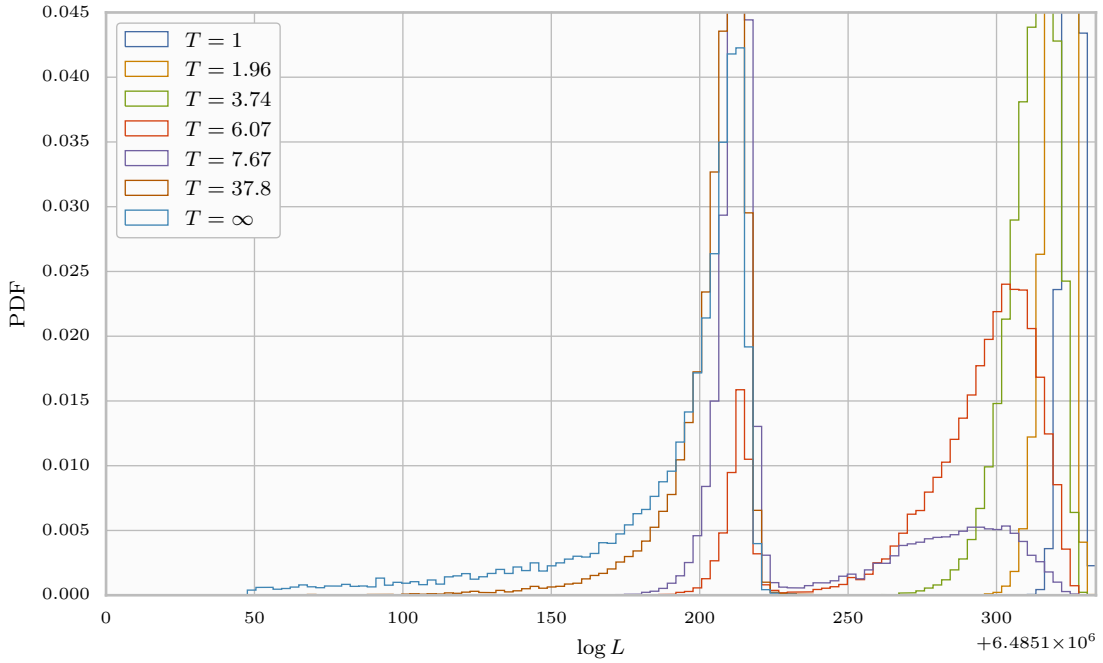

Figure 4.15: A histogram of the log likelihoods collected by *ptemcee* while sampling from the BNS system in Table 4.1 at an SNR of ~ 14. Note the lack of mixing between the low- and high-likelihood peaks. The log likelihood in this plot is absolute, without having had the null log likelihood subtracted; see Eq. (1.9).

It is possible that an ensemble-based parallel tempered MCMC sampler would not experience these difficulties, since an ensemble at the critical temperature can jointly occupy both of these likelihood peaks. The intra-chain proposals used by ensemble samplers, such as the stretch move proposal (Goodman & Weare, 2010), are informed by the shape of the entire ensemble. The efficiency of intra-chain jumps between the signal and noise-only models might therefore be improved when the ensemble is split between them by inter-chain swaps.

Preliminary attempts at such an implementation, using *ptemcee* (Vousden et al., 2015), showed the same bimodal likelihood distribution as observed in the *LALInference* runs discussed in Section 4.3 when exploring the same CBC target distributions (see Fig. 4.15). However, while the temperature evolution (see Fig. 4.16) is markedly more stable than in the single-walker tests in Section 4.3, these runs did not converge to the *correct* equilibrium distribution (see Fig. 4.17). It is not yet clear why the sampler failed to identify the true parameter values.

Finally, in Chapter 3 we discussed the possibility of constructing a temperature ladder for equal Kullback–Leibler (KL) divergence between neighbouring chains, rather than equal acceptance ratios. Since the KL divergence depends on the tem-

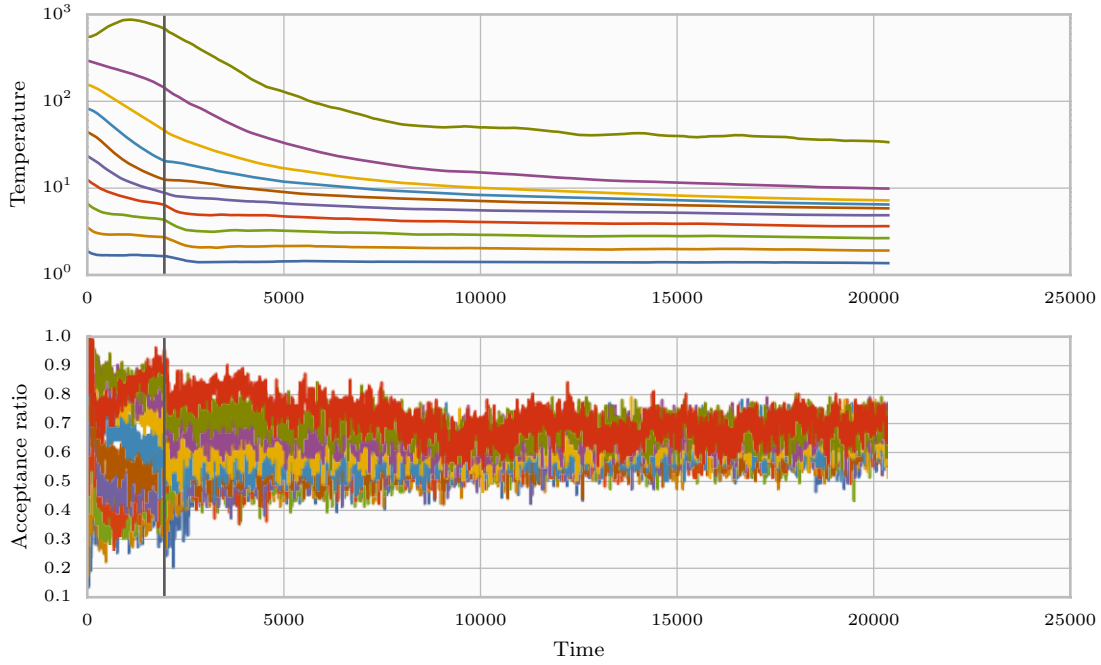

Figure 4.16: The evolution of temperatures under *ptemcee* while sampling from the BNS system in Table 4.1 at an SNR of ~ 14.

perature through the evidence, a scheme that adapts in response to it will be more difficult to implement than that developed in Chapter 3. Nonetheless, Cameron (2015) has suggested that the KL divergence between chains can in fact be estimated from the samples accumulated at each temperature. It may be interesting to investigate such a scheme for resilience to the problems encountered in the *LALInference* tests.

However, with the exceptionally long ACTs observed at higher SNRs (e.g., Fig. 4.10), caused by the dichotomy between the signal and noise-only models, it appears that an MCMC implementation such as that in *LALInference* cannot reliably explore the tempered posterior distribution around the phase transition. It will therefore be difficult to implement any scheme that selects a temperature ladder that bridges this phase transition in response to the shape of the tempered target distribution. Instead, the adaptive scheme described in Chapter 3 and tested here is most useful for lower-SNR events where the inter-model phase transition is less extreme.

Farr et al. (2015) have suggested an adaptive scheme that allows a reversible-jump MCMC sampler to efficiently explore the multi-model parameter spaces, with applications to similar CBC parameter estimation problems. Such a technique may alleviate the two-model problem observed here, and merits further study.

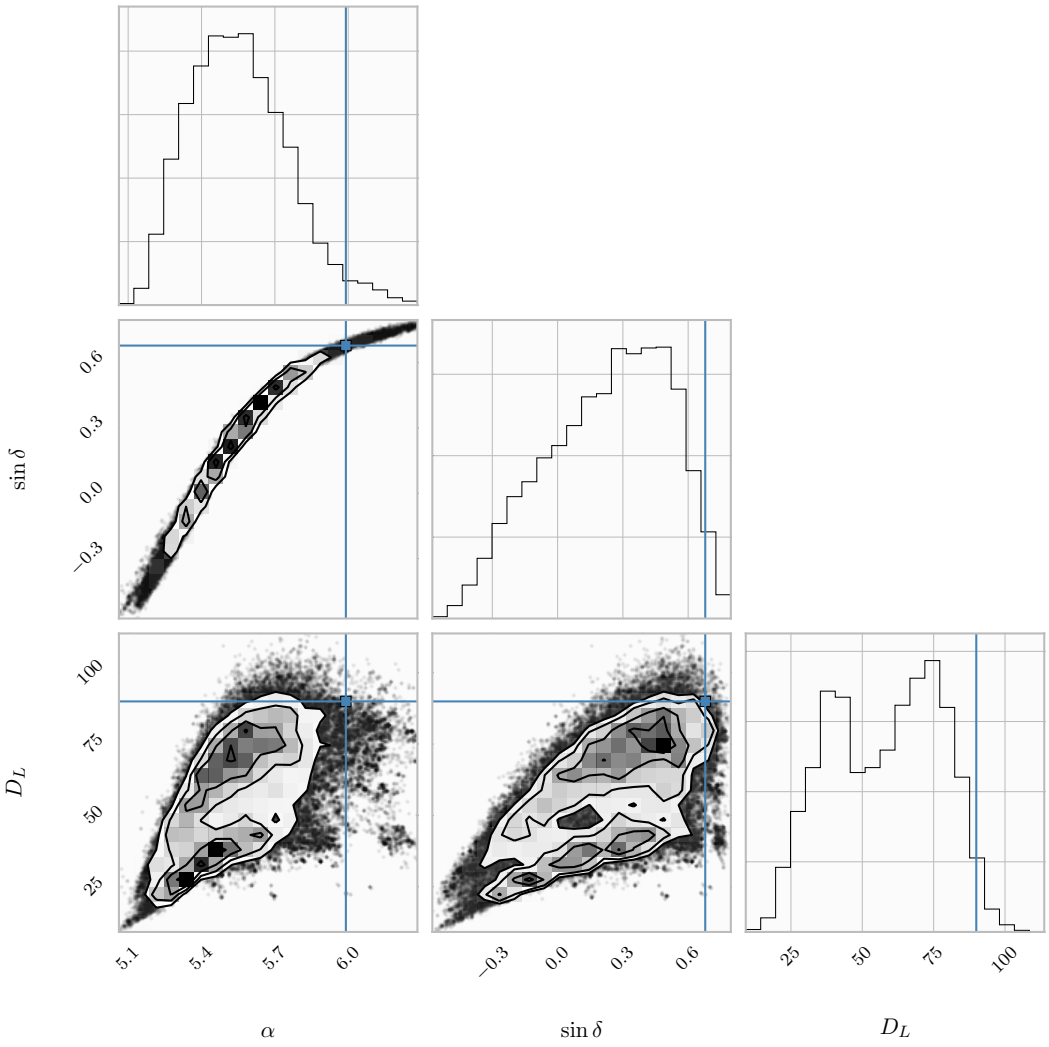

Figure 4.17: The one- and two-dimensional marginal posterior distributions for sky location and luminosity distance, as recovered by *ptemcee*, for the BNS system in Table 4.1 at an SNR of ~ 14. The blue lines denote the true parameter values; note the poor accuracy of the sampler for these parameters. This plot was produced with `triangle.py` (Foreman-Mackey et al., 2014).

Chapter 5

Conclusions

This thesis has considered, broadly, two themes: (i) joint observations of compact binary coalescence (CBC) events through electromagnetic (EM) and gravitational wave (GW) astronomy, and (ii) extraction of the source parameters from a GW signal through computational Bayesian inference.

Firstly, in Chapter 2 we considered the problem of observing – with wide-field telescopes – the EM counterpart to a CBC detection by a network of ground-based GW interferometers. The observational resources available for such an effort are severely limited with respect to the large sky localisation uncertainty for these events.

In particular, we demonstrated in Sections 2.3 and 2.4 that galaxy catalogues can aid such searches by prioritising the most luminous parts of the sky for observation with a follow-up telescope. The usefulness of this approach is limited by the modesty of fluctuations in the spatial density of galaxies, and therefore depends on the conical "pointing volume" captured by a telescope pointing (itself a function of the telescope's field-of-view (FOV) and the GW distance measurement). We found that direction from galaxy catalogues will be most fruitful at low pointing volumes, e.g., for early, low-sensitivity configurations of LIGO and for follow-up of loud, nearby events. In these cases, we reported up to a four-fold increase in the probability of successfully imaging an EM counterpart.

Follow-up searches for quieter and more distant detections, expected from advanced detectors in their design configurations, will benefit less from galaxy catalogue direction. However, even for events at the median distance of 200 Mpc for advanced detector networks, catalogue-directed follow-up searches will still be tens of percent more likely to image the event than an uninformed search.

Meanwhile, the effect of incompleteness in a galaxy catalogue that is used for follow-up searches is most pronounced at small pointing volumes and large follow-up fractions (see Section 2.5 and Fig. 2.5), and is likely small for realistic observing

scenarios. For a catalogue that is $\sim 33\,\%$ complete within the Advanced LIGO sensitivity volume, for instance, approximately $5\,\%$ of the improvement in success fraction over an uninformed search is lost, relative to a search with a complete catalogue.

There are several interesting avenues of continuation for this work, discussed in Section 2.6. Foremost among these is a more careful consideration of the astrophysical assumptions made in using a galaxy catalogue for follow-up searches. For example, the luminosity band in which to rank pointing targets (e.g., B- vs. K-band), the true completeness of the catalogue (vs. the completeness assumed by us), and the degree of spatial coincidence of mergers with their host galaxies might all affect the utility and use of a galaxy catalogue. Future studies might parameterise these effects as priors in order to quantify the effect of our astrophysical ignorance on EM follow-up searches.

In the second part of this thesis, we developed and applied a method of optimising Bayesian inference through parallel tempered Markov chain Monte Carlo (MCMC) sampling.

The performance of a parallel tempered MCMC sampler is highly sensitive to the choice – or ladder – of temperatures at which the individual Markov chains sample. In Chapter 3 we investigated the criteria for an effective temperature ladder in terms of the shape of the likelihood distribution as a function of temperature. We demonstrated that a geometrically spaced temperature ladder is a good starting point, and developed a method for dynamically adapting the temperatures while sampling to minimise the time between effective samples, i.e., the autocorrelation time (ACT) of the sampler. This method drives the temperatures in the simulation toward a configuration that yields uniform swap acceptance ratios between all adjacent pairs of chains in order to maximise the rate of transfer of information between cold and hot chains.

We have provided a reference implementation for the *emcee* ensemble sampler (Foreman-Mackey et al., 2013) and tested it on a number of simple test distributions, reporting improvements in performance of up to a factor of four. This reference implementation is now available in a separate package called *ptemcee*, available at `https://github.com/willvousden/ptemcee`.

Finally, Chapter 4 considered the problem of parameter estimation for CBC detections, with the aid of the dynamical temperature selection scheme developed in Chapter 3. In it, I described an implementation of this scheme in the *LALInference* library and the modifications that were needed. Tests of this implementation on astrophysically motivated GW data analysis problems consistently demonstrated

improvements in efficiency of tens of percent.

However, these tests also exposed problems in the implementation and application of the dynamical temperature selection scheme. Specifically, the MCMC sampler used by *LALInference* becomes extremely inefficient (i.e., its ACT becomes very large) around a critical temperature $T_{\rm crit}$ that determines a phase transition in the tempered posterior distribution – regardless of the temperatures selected. It is therefore difficult to adaptively select temperatures in response to the distributions sampled by each chain, and similarly to estimate an evidence value for the model. We also observed extremely limited communication of replicas across this barrier – suggesting that the chains at temperatures higher than $T_{\rm crit}$ do not in fact contribute to the sampler's output at $T = 1$.

These problems are so far unresolved and deserve further study. An important question is to identify the exact conditions under which a parallel tempered MCMC sampler fails in spite of a close temperature spacing. A more thorough comparison of parallel tempering implementations for ensemble-based and single-walker samplers is also merited.

Appendix A

Autocorrelation time estimation

The autocorrelation time (ACT) discussed in this thesis refers to the *integrated autocorrelation time* described by Sokal (1997). It is estimated in the following way.

If $x(t)$ is a time series with a normalised autocorrelation function $\rho(t)$, such that $\rho(0) = 1$, then the integrated ACT of x is defined by

$$\begin{aligned}\tau &\equiv \sum_{t=-\infty}^{\infty} \rho(t) \\ &= 1 + 2\sum_{t=1}^{\infty} \rho(t).\end{aligned}$$

Since, when $t \gg \tau$, $\rho(t) \approx 0$, there is little contribution to the integral at large lags, except through noise in the measured autocorrelation function ρ. We can therefore approximate the ACT as

$$\tau \approx 1 + 2\sum_{t=1}^{M\tau} \rho(t).$$

We estimate the ACT over a window that is $M = 5$ ACTs long, subject to the constraint that $M\tau < N/2$, where N is the number of samples in x. If this constraint is violated, the result is probably not trustworthy, since there are too few samples for a meaningful estimate.

Appendix B

LALInference command line options

Table B.1: The new command line options for the *LALInference* MCMC sampler, `lalinference_mcmc`, that control its temperature dynamics.

Name	Description	Default
`--adaptLadder`	Dynamically adapt parallel tempering ladder for equal inter-chain swap acceptance ratios. This option is incompatible with `--anneal`.	(off)
`--adaptLadderTimeScale`	The time-scale for temperature adaptations, in multiples of `--tempSkip`.	`100`
`--adaptLadderDecayLag`	The time-scale for the decay of temperature adaptations, in multiples of `--tempSkip`.	`1000`
`--tempInf`	If `--adaptLadder`, use an additional chain at $T = \infty$ to sample from the prior.	(off)

Appendix C

LALInference default temperatures

Since the search pipeline provides an estimate of the signal-to-noise ratio (SNR) of a detection (e.g., Allen et al., 2012; Cannon et al., 2012), *LALInference* can use this estimate to inform its default choice of temperature ladder (Veitch et al., 2015). The method is as follows.

The SNR ρ is a proxy for the maximum log likelihood, such that $\log L_{\max}$ scales as $\rho^2/2$, where

$$L_{\max} \equiv \max_{\vec{\theta}} L(\vec{\theta}),$$

with the likelihood defined as in Eq. (4.1).

Our aim in selecting a default temperature ladder is to pick the maximum temperature so that the hottest chain is effectively sampling from the prior; i.e. $T_{\max} \approx T_{\text{prior}}$. If the likelihood is sharply peaked with respect to the width of the prior, then the expected $\log L$ under the prior will be approximately zero (for a normalised prior). We therefore select $T_{\max}$ so that $\mathrm{E}[\log L]_{T_{\max}} \approx 0$.

In the high-SNR limit, the likelihood is indeed sharply peaked and is well-approximated by an n-dimensional Gaussian. With reference to Eq. (3.6)[1], we can then express this condition as

$$T_{\max} \approx \frac{2}{n} \log L_{\max} \approx \frac{\rho^2}{n}.$$

With this estimate of a suitable $T_{\max}$, *LALInference*'s default ladder is a geometric spacing of temperatures between $T = 1$ and $T = T_{\max}$ for a given number of chains.

[1]Note that this expression is valid for a Gaussian that is instead normalised so that $\log L_{\max} = 0$, while in this case $\mathrm{E}[\log L]_{T_{\max}} = 0$.

List of acronyms

ACT	autocorrelation time, denoted τ
ASD	amplitude spectral density
BBH	binary black hole
BH	black hole
BNS	binary neutron star
CBC	compact binary coalescence
EM	electromagnetic
FOV	field-of-view
GW	gravitational wave
GWGC	Gravitational Wave Galaxy Catalog
IMR	inspiral-merger-ringdown
ISCO	innermost stable circular orbit
ISM	interstellar medium
KL	Kullback–Leibler
LAL	LIGO Algorithm Library
LSC	LIGO Scientific Collaboration
MCMC	Markov chain Monte Carlo
MLE	maximum likelihood estimator
MPI	Message Passing Interface
NR	numerical relativity
NS	neutron star
NSBH	neutron star/black hole
PDF	probability density function
PN	post-Newtonian
PSD	power spectral density
PTA	pulsar timing array
SGRB	short γ-ray burst
SNR	signal-to-noise ratio, denoted ρ

List of references

Aasi J., et al., 2013a, preprint (arXiv:1304.0670)

Aasi J., et al., 2013b, Phys. Rev. D, 87, 022002

Aasi J., et al., 2013c, Phys. Rev. D, 88, 062001

Aasi J., et al., 2014, Astrophys. J. Suppl. Ser., 211, 7

Aasi J., et al., 2015, Class. Quant. Grav., 32, 074001

Abadie, J. et al., 2012, Astron. Astrophys., 539, A124

Abadie J., et al., 2010a, Class. Quant. Grav., 27, 173001

Abadie J., et al., 2010b, Phys. Rev. D, 82, 102001

Abadie J., et al., 2010c, Astrophys. J., 715, 1453

Abadie J., et al., 2011, Phys. Rev. D, 83, 122005

Abadie J., et al., 2012a, Phys. Rev. D, 85, 082002

Abadie J., et al., 2012b, Astron. Astrophys., 541

Abbott B. P., et al., 2009a, Rep. Prog. Phys., 72, 076901

Abbott B. P., et al., 2009b, Phys. Rev. D, 80, 062001

Accadia T., et al., 2012, J. Instrum., 7, P03012

Acernese F., et al., 2015, Class. Quant. Grav., 32, 024001

Allen B., Anderson W. G., Brady P. R., Brown D. A., Creighton J. D. E., 2012, Phys. Rev. D, 85, 122006

Amaro-Seoane P., et al., 2012, Class. Quant. Grav., 29, 124016

Babak S., et al., 2013, Phys. Rev. D, 87, 024033

Balasubramanian R., Sathyaprakash B. S., Dhurandhar S. V., 1996, Phys. Rev. D, 53, 3033

Berger E., 2010, Astrophys. J., 722, 1946

Berger E., Fong W., Chornock R., 2013, Astrophys. J. Lett., 774, L23

Berti E., Cardoso V., 2006, Phys. Rev. D, 74, 104020

Bloom J. S., et al., 2009, preprint, (arXiv:0902.1527)

Buonanno A., Iyer B. R., Ochsner E., Pan Y., Sathyaprakash B. S., 2009, Phys. Rev. D, 80, 084043

Burrows D. N., et al., 2006, Astrophys. J., 653, 468

Cameron E., 2015, On RJMCMC and Adaptive PT, https://astrostatistics.wordpress.com/2015/01/27/on-rjmcmc-and-adaptive-pt/

Cameron E., Pettitt A., 2014, Statist. Sci., 29, 397

Cannon K., et al., 2012, Astrophys. J., 748, 136

Chernoff D. F., Finn L. S., 1993, Astrophys. J., 411

Cutler C., Thorne K. S., 2002, in Bishop N. T., Maharaj S., eds, General Relativity and Gravitation: Proceedings of the 16th International Conference on General Relativity and Gravitation. pp 72–112, doi:11858/00-001M-0000-0013-5422-4

Damour T., Iyer B. R., Sathyaprakash B. S., 2000, Phys. Rev. D, 62, 084036

Earl D. J., Deem M. W., 2005, Phys. Chem. Chem. Phys., 7, 3910

Eichler D., Livio M., Piran T., Schramm D. N., 1989, Nat., 340, 126

Evans P. A., et al., 2012, Astrophys. J. Suppl. Ser., 203, 28

Fairhurst S., 2009, New J. Phys., 11, 123006

Falcioni M., Deem M. W., 1999, J. Chem. Phys., 110, 1754

Farr W. M., Mandel I., Stevens D., 2015, R. Soc. Open Sci., 2

Finn L. S., Chernoff D. F., 1993, Phys. Rev. D, 47, 2198

Finn L. S., Mohanty S. D., Romano J. D., 1999, Phys. Rev. D, 60, 121101

Finn L. S., et al., 2015, Gravitational Wave Interferometer Noise Calculator, v3, https://awiki.ligo-wa.caltech.edu/aLIGO/GWINC

Fong W., Berger E., 2013, Astrophys. J., 776, 18

Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, Pub. Astron. Soc. Pac., 125, 306

Foreman-Mackey D., Price-Whelan A., Ryan G., Emily Smith M., Barbary K., Hogg D. W., Brewer B. J., 2014, triangle.py v0.1.1, doi:10.5281/zenodo.11020

Freeman M., 2006, J. Epidemiol. Community Health, 60, 6

Fryer C., Kalogera V., 1997, Astrophys. J.

Geyer C. J., 1991, in Keramidas E. M., Kaufman S. M., eds, Computing Science and Statistics, Proceedings of the 23rd Symposium on the Interface. Interface Foundation of North America, New York, pp 156–163

Geyer C. J., 1994, Technical Report 568, Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Univ. Minnesota, Minneapolis, http://purl.umn.edu/58433

Goggans P. M., Chi Y., 2004, AIP Conf. Proc., 707, 59

Goodman J., 2009, Acor, statistical analysis of a time series, http://www.math.nyu.edu/faculty/goodman/software/acor/

Goodman J., Weare J., 2010, Commun. Appl. Math. Comp. Sci., 5, 65

Grover K., Fairhurst S., Farr B. F., Mandel I., Rodriguez C., Sidery T., Vecchio A., 2014, Phys. Rev. D, 89, 042004

Hanna C., Mandel I., Vousden W., 2014, Astrophys. J., 784, 8

Hannam M., Schmidt P., Bohé A., Haegel L., Husa S., Ohme F., Pratten G., Pürrer M., 2014, Phys. Rev. Lett., 113, 151101

Hobbs G., et al., 2010, Class. Quant. Grav., 27, 084013

Holz D. E., Hughes S. A., 2005, Astrophys. J., 629, 15

Hukushima K., Nemoto K., 1996, J. Phys. Soc. Jpn., 65, 1604

Iyer B., Souradeep T., Unnikrishnan C. S., Dhurandhar S., Raja S., Sengupta A., 2012, Technical report, LIGO-India: Proposal of the Consortium for Indian Initiative in Gravitational-wave Observations. LIGO Scientific Collaboration

Kanner J., Huard T. L., Márka S., Murphy D. C., Piscionere J., Reed M., Shawhan P., 2008, Classical and Quantum Gravity, 25, 184034

Kanner J., Baker J., Blackburn L., Camp J., Mooley K., Mushotzky R., Ptak A., 2013, Astrophys. J., 774, 63

Kasen D., Badnell N. R., Barnes J., 2013, Astrophys. J., 774, 25

Kasliwal M. M., Nissanke S., 2014, Astrophys. J. Lett., 789, L5

Katzgraber H. G., Trebst S., Huse D. A., Troyer M., 2006, J. Stat. Mech: Theory Exp., 2006, P03018

Kelley L. Z., Ramirez-Ruiz E., Zemp M., Diemand J., Mandel I., 2010, Astrophys. J. Lett., 725, L91

Kelley L. Z., Mandel I., Ramirez-Ruiz E., 2013, Phys. Rev. D, 87, 123004

Kochanek C. S., Piran T., 1993, Astrophys. J., 417, L17

Kofke D. A., 2002, J. Chem. Phys., 117, 6911

Kofke D. A., 2004, J. Chem. Phys., 120, 10852

Kopparapu R. K., Hanna C., Kalogera V., O'Shaughnessy R., González G., Brady P. R., Fairhurst S., 2008, Astrophys. J., 675, 1459

Lartillot N., Philippe H., 2006, Syst. Biol., 55, 195

Luan J., Hooper S., Wen L., Chen Y., 2012, Phys. Rev. D, 85, 102002

Maggiore M., 2007, Gravitational Waves: Volume 1: Theory and Experiments. Gravitational Waves, Oxford University Press, Oxford

Metzger B. D., Berger E., 2012, Astrophys. J., 746, 48

Metzger B. D., Kaplan D. L., Berger E., 2013, Astrophys. J., 764, 149

Mohanty S. D., Marka S., Rahkola R., Mukherjee S., Leonor I., Frey R., Cannizzo J., Camp J., 2004, Class. Quant. Grav., 21, S765

Nakar E., Piran T., 2011, Nat. Lett., 478, 82

Nakar E., Gal-Yam A., Fox D. B., 2006, Astrophys. J., 650, 281

Narayan R., Paczynski B., Piran T., 1992, Astrophys. J. Lett., 395, L83

Nissanke S., Holz D. E., Hughes S. A., Dalal N., Sievers J. L., 2010, Astrophys. J., 725, 496

Nissanke S., Sievers J., Dalal N., Holz D., 2011, Astrophys. J., 739, 99

Nissanke S., Kasliwal M., Georgieva A., 2013, Astrophys. J., 767, 124

Nuttall L. K., Sutton P. J., 2010, Phys. Rev. D, 82, 102002

O'Shaughnessy R., Kalogera V., Belczynski K., 2010, Astrophys. J., 716, 615

Ohme F., 2012, Class. Quant. Grav., 29, 124002

Owen B. J., 1996, Phys. Rev. D, 53, 6749

Perley D. A., et al., 2009, Astrophys. J., 696, 1871

Peters P. C., Mathews J., 1963, Phys. Rev., 131, 435

Phinney E. S., 1991, Astrophys. J. Lett., 380, L17

Rathore N., Chopra M., de Pablo J. J., 2005, J. Chem. Phys., 122

Raymond V., van der Sluys M. V., Mandel I., Kalogera V., Röver C., Christensen N., 2010, Class. Quant. Grav., 27, 114009

Roberts G. O., Rosenthal J. S., 1998, Can. J. Stat., 26, 5

Roberts G. O., Rosenthal J. S., 2007, J. App. Prob., 44, 458

Rodriguez C. L., Farr B., Raymond V., Farr W. M., Littenberg T. B., Fazi D., Kalogera V., 2014, Astrophys. J., 784, 119

Sanbonmatsu K., García A., 2002, Proteins: Struct., Funct., Bioinf., 46, 225

Schechter P., 1976, Astrophys. J., 203, 297

Schneider P., 2006, Extragalactic Astronomy and Cosmology. Springer-Verlag Berlin Heidelberg, doi:10.1007/978-3-642-54083-7

Schug A., Herges T., Wenzel W., 2004, Proteins: Struct., Funct., Bioinf., 57, 792

Schutz B. F., 1986, Nature, 323, 310

Sidery T., et al., 2014, Phys. Rev. D, 89, 084060

Singer L., Price L., Speranza A., 2012, preprint (arXiv:1204.4510)

Singer L. P., et al., 2014, Astrophys. J., 795, 105

Skilling J., 2006, Bayesian Anal., 1, 833

Soderberg A. M., et al., 2006, Astrophys. J., 650, 261

Sokal A., 1997, in DeWitt-Morette C., Cartier P., Folacci A., eds, NATO ASI Series, Vol. 361, Functional Integration. Springer US, pp 131–192, doi:10.1007/978-1-4899-0319-8_6

Somiya K., 2012, Class. Quant. Grav., 29, 124007

Sugita Y., Okamoto Y., 1999, Chem. Phys. Lett., 314, 141

Swendsen R. H., Wang J.-S., 1986, Phys. Rev. Lett., 57, 2607

Tanvir N. R., Levan A. J., Fruchter A. S., Hjorth J., Hounsell R. A., Wiersema K., Tunnicliffe R. L., 2013, Nat. Lett., 500, 547

The LIGO Scientific Collaboration 2010, LIGO Document T0900288-v3, Advanced LIGO anticipated sensitivity curves. The LIGO Scientific Collaboration

The Virgo Collaboration 2009, Virgo Technical Report VIR-0027A-09, Advanced Virgo Baseline Design. The Virgo Collaboration

Tunnicliffe R. L., et al., 2013, Mon. Not. R. Astron. Soc.,

Veitch J., et al., 2012, Phys. Rev. D, 85, 104045

Veitch J., et al., 2015, Phys. Rev. D, 91, 042003

Vitale S., Lynch R., Veitch J., Raymond V., Sturani R., 2014, Phys. Rev. Lett., 112, 251101

Vousden W., Farr W. M., Mandel I., 2015, preprint (arXiv:1501.05823)

Weisberg J., Taylor J., 2005, in Rasio F., Stairs I., eds, Astronomical Society of the Pacific conference series Vol. 328, Binary radio pulsars: proceedings of a meeting held in Aspen, Colorado, USA, 11-17 January 2004. Astronomical Society of the Pacific, p. 25, http://aspbooks.org/a/volumes/table_of_contents/?book_id=32

White D. J., Daw E. J., Dhillon V. S., 2011, Class. Quant. Grav., 28, 085016

de Freitas Pacheco J. A., Regimbau T., Vincent S., Spallicci A., 2006, Int. J. Mod. Phys., D15, 235

ter Braak C. J., Vrugt J. A., 2008, Stat. Comput., 18, 435

van Eerten H. J., MacFadyen A. I., 2011, Astrophys. J. Lett., 733, L37

van der Sluys M., Raymond V., Mandel I., Röver C., Christensen N., Kalogera V., Meyer R., Vecchio A., 2008a, Class. Quant. Grav., 25, 184011

van der Sluys M. V., et al., 2008b, Astrophys. J. Lett., 688, L61