Cutting Edge Data Analysis for Gravitational Wave Detection with LISA

by

Jonas Elias El Gammal

Thesis submitted in fulfilment of
the requirements for the degree of
PHILOSOPHIAE DOCTOR
(PhD)

Faculty of Science and Technology
Department of Mathematics and Physics
2025

University of Stavanger
NO-4036 Stavanger
NORWAY
www.uis.no

ISBN: 978-82-8439-397-1
ISSN: 1890-1387
PhD: Thesis UiS No. 883

Preface

This thesis is submitted in partial fulfillment of the requirements for the degree of Philosophiae Doctor (PhD) at the University of Stavanger, Faculty of Science and Technology, Norway. The research has been carried out at the University of Stavanger from January 2021 to March 2025.

First, I would like to express my gratitude to my supervisor, Germano Nardini, who has supported and guided me through the amazing time I had at UiS. I would also like to thank my co-supervisor, Anders Tranberg, for always being there for me.

My sincere gratitude goes to my long time collaborator, Jesús Torrado, who has been by my side throughout this whole journey, not only as a collaborator but also taking up a role as a mentor and a friend.

I would like to thank Riccardo Buscicchio for being a great support during my PhD and now commencing postdoc, and for accommodating me at Bicocca University during my stay in Milan.

Another special thanks goes to Gabriele Franciolini and Mauro Pieroni, who have been a great support during my time at CERN and afterwards, giving me very valuable academic guidance along the way.

I would like to extend my appreciation to all of the other people with whom I had the chance to collaborate throughout these four years, and in particular, I would like to thank everyone in the LISA Consortium for working towards an amazing goal.

The people who surrounded me during my time at UiS deserve a special thanks. I could not have hoped for a better environment to have spent my PhD years in this beautiful place. Furthermore, I would like to thank the IMF department for providing such an amazing working environment. Gaurang deserves a special shoutout here. Without him, my PhD would have been a lot more productive but infinitely less fun.

Lastly, I would like to thank my family and friends for their continuous support and encouragement and my partner Ellen, who has been a huge support during this last phase of my PhD journey.

Jonas
Stavanger, April 2025

Abstract

This thesis presents novel contributions to the field of gravitational wave data analysis, with a focus on the Laser Interferometer Space Antenna (LISA) mission. The goal is to advance Bayesian inference methods with the help of modern machine learning techniques and to develop efficient frameworks for analyzing astrophysical and cosmological gravitational wave signals in anticipation of the data expected from LISA.

The first part of the thesis develops a foundation in Bayesian statistics, exploring both traditional and approximate inference techniques and establishing how machine learning can be used to accelerate the inference process. A particular emphasis is placed on accelerating likelihood-based inference through the use of Gaussian Processes and active learning. This results in the `GPry` algorithm, a Gaussian Process-based posterior emulator that significantly reduces the computational cost for inference in cases where the posterior is expensive to compute. `GPry` is benchmarked against state-of-the-art Monte Carlo samplers on both toy models and real-world data from the Cosmic Microwave Background. Subsequent work extends this method to LISA-specific inference tasks. Using `GPry`, parameter estimation for three LISA source types–double white dwarfs, stellar-mass black hole binaries, and supermassive black hole binaries–is performed. The results demonstrate a substantial acceleration in the inference process.

The final part of the thesis assesses how scalar-induced gravitational waves arising from enhanced curvature perturbations during inflation can be constrained using LISA. A fast framework is developed for computing the second-order gravitational wave signal using `JAX`, enabling efficient inference. Three approaches to modeling the curvature power spectrum are explored: a model-agnostic method using binned spectra, a template-based method using phenomenological descriptions of the curvature power spectrum, and an ab initio approach for ultra-slow-roll inflation. This study demonstrates that LISA can probe a wide range of inflationary scenarios with high precision.

The findings of this thesis contribute to the understanding of gravitational wave data analysis and establish novel methods for maximizing the scientific output of LISA.

List of papers

Paper I

Fast and robust Bayesian Inference using Gaussian Processes with GPry.
Jonas El Gammal, Nils Schöneberg, Jesús Torrado, Christian Fidler.

JCAP 10(2023)021, arXiv:2211.02045, [1]

Paper II

Parallelized Acquisition for Active Learning using Monte Carlo Sampling.
Jesús Torrado, Nils Schöneberg, Jonas El Gammal.

Preprint, arXiv:2305.19267, [2]

Paper III

Accelerating LISA inference with Gaussian processes.
Jonas El Gammal, Riccardo Buscicchio, Germano Nardini, Jesús Torrado.

Submitted to PhysRevD, ArXiv:2503.21871, [3]

Paper IV

Reconstructing Primordial Curvature Perturbations via Scalar-Induced Gravitational Waves with LISA.
Jonas El Gammal, Aya Ghaleb, Gabriele Franciolini, Theodoros Papanikolaou, Marco Peloso, Gabriele Perna, Mauro Pieroni, Angelo Ricciardone, Robert Rosati, Gianmassimo Tasinato

Matteo Braglia, Jacopo Fumagalli, Jun'ya Kume, Enrico Morgante, Germano Nardini, Davide Racco, Sébastien Renaux-Petel, Hardi Veermäe, Denis Werth, Ivonne Zavala

(For the LISA Cosmology working Group).

Submitted to JCAP, arXiv:2501.11320, [4]

Table of Contents

Appendix

1 Introduction

This work covers the research that I have performed throughout the last four years as part of my PhD studies at the University of Stavanger. This research has covered several areas within machine learning accelerated inference and was partially done with the aim of improving source modeling and inference within the future LISA space mission. To tie the relatively different topics together, the first three chapters comprise an introduction of the methodology and state-of-the-art of the different areas of research. After this, the four papers that have been produced during the PhD are presented.

Chapter 2 covers the basics of Bayesian inference and machine learning and how these two can be combined to accelerate the inference process. As the `GPry` algorithm, which was developed as part of this work, is based on Gaussian processes and active learning, these topics are introduced in more detail.

Chapter 3 introduces the LISA mission with a particular focus on introducing the relevant quantities that are of interest for modeling the sources that are projected to be observed by LISA and the difficulties arising when inferring these quantities. Furthermore, this chapter gives a brief overview of the current state-of-the-art in machine learning accelerated inference for LISA.

Chapter 4 introduces a particular type of cosmological signal that might be observed by LISA: scalar-induced gravitational waves. This chapter gives a brief overview of the mechanism generating such a signal and the equations that govern the evolution of scalar perturbations with a particular focus on the single-field slow-roll inflationary model. Furthermore, the chapter explains how gravitational waves are generated at second order from the scalar perturbations and what contemporary bounds exist.

Paper I covers the `GPry` algorithm, which is a machine learning-accelerated inference algorithm based on Gaussian processes and active learning. The algorithm is tested on toy models and a real-world example of cosmic microwave background data. The results show that the algorithm can significantly accelerate the inference process compared to traditional methods.

Paper II builds on paper I by proposing an improved method for the active learning part of the `GPry` algorithm by using a nested sampling approach together with an efficient ranking of candidate samples to acquire batches of training points in parallel.

Paper III uses `GPry` to infer the source parameters of three types of astrophysical sources projected to be observed by LISA: double white dwarf systems, stellar mass black hole binaries, and supermassive black hole binaries, achieving a speedup of up to two orders of magnitude in wall clock time when compared to nested sampling.

Paper IV introduces a framework for inferring the source parameters of scalar-induced gravitational waves through a fast computation of the second-order gravitational wave signal with `JAX`. This allows for a fast computation of the likelihood function that is used in the inference process. We forecast the sensitivity of LISA to SIGWs using *a)* a model-agnostic approach that involves binning the power spectrum of scalar perturbations $\mathcal{P}_\zeta$, *b)* a template-based approach using analytical, phenomenological descriptions of $\mathcal{P}_\zeta$, and *c)* an ab initio approach solving for the equations of motion for a single-field ultra-slow roll inflationary model. The results show that LISA will be able to detect SIGWs at high confidence levels for a wide range of models.

A note regarding notation: In general, in this work we use bold font notation for vectors and tensors, where $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$ (lowercase letters) refer to vectors and $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ (uppercase letters) are higher-order tensors. We use this notation consistently; however, where quantities can be either scalars or higher-dimensional objects and the generalization is trivial, we simplify the notation by considering only the scalar case.

Whenever logarithms appear, if not specified otherwise by a subscript, they refer to the natural logarithm with base e.

$|\dots|$ refers to the absolute value for scalars, the Euclidean norm for vectors, and to the determinant for matrices, except for the absolute value of a determinant, which is denoted by $|\det(\dots)|$.

When discussing concepts in General Relativity, Greek letters refer to indices of the 4-dimensional spacetime. Latin letters indicate the 3-dimensional space components.

2 Bayesian inference with machine learning

This chapter covers the basis of Bayesian inference for parameter estimation and explores how modern computational methods, including machine learning, can accelerate this process. As modern Bayesian inference is mainly a numerically driven field, the focus is on the computational aspects of it. This field has evolved considerably in recent years, specifically due to the advent of machine learning techniques. Therefore, this section is divided into four parts:

Section 2.1 covers the basics of Bayesian inference, introducing Bayes' theorem, parameter inference, model selection, and touching on the role of priors and potential pitfalls when performing Bayesian inference.

Section 2.2 explains the use of Monte Carlo methods in Bayesian inference through what would nowadays be considered traditional methods. We focus on two of the most commonly used and versatile methods: Markov Chain Monte Carlo and Nested sampling.

Section 2.3 introduces the two most commonly used schemes for approximate Bayesian inference in physics: Variational inference and Simulation-based inference.

Section 2.4 gives an overview of some of the machine learning techniques that have been used in the past for Bayesian inference, focusing on some of the most commonly used methods: Normalizing Flows, Gaussian Processes, and Active Learning.

Lastly, Section 2.5 explains how machine learning techniques can be used in different aspects of Bayesian inference and how they can be combined with traditional and approximate inference methods to improve the efficiency, robustness, and accuracy of the inference process.

2.1 Bayesian inference

This section provides a brief overview of the rationale behind Bayesian inference, starting with Bayes' theorem, introducing all relevant quantities–namely the *prior*, *likelihood*, *posterior*, and *evidence*–and two measures of similarity between probability distributions in the form of the Kullback-Leibler and Jensen-Shannon divergence.

2.1.1 Bayes' theorem

At the core of Bayesian inference is Bayes' theorem, which is a mathematical formula that describes how to update the probability of a *hypothesis* A given new *information* (or data) B [5]

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \ , \tag{2.1}$$

where $P(A|B)$ is the probability of A given B, $P(B|A)$ is the probability of B given A, $P(A)$ is the prior probability of hypothesis A, and $P(B)$ is the marginal probability of the data B. In the context of parameter estimation, A is the quantity of interest. Written in a more illustrative form, Eq. (2.1) can be expressed as

$$P(\text{hypothesis}|\text{data}) = \frac{P(\text{data}|\text{hypothesis})P(\text{hypothesis})}{P(\text{data})} \ . \tag{2.2}$$

If our hypothesis is that *the data is generated (up to some numerical noise) by a model M with parameters θ*, then we can shorten Bayes' theorem to the common form that we use throughout this work:

$$p(\theta|D) = \frac{L(D|\theta)\pi(\theta)}{Z(D)} \ , \tag{2.3}$$

where $p(\theta|D)$ is the *posterior* probability of the parameter(s) θ given the data D, $L(D|\theta)$ is the *likelihood* of the data given the parameter(s) θ, $\pi(\theta)$ is the *prior* probability of the parameter(s), and $Z(D)$ is the *evidence*. If multiple parameters govern the model M, $\boldsymbol{\theta}$ becomes a vector.

In the transition from Eq. (2.2) to Eq. (2.3), we have transitioned from the discrete parameter space of P to the continuous parameter space of p, as parameters in most practical applications are modeled as continuous variables. This in turn means that p, L, and π are probability density functions (PDFs) as opposed to discrete probabilities, and Z is a normalization constant that is frequently ignored in parameter estimation when only relative probabilities are of interest.

2.1.2 Likelihood and log-likelihood

The likelihood $L(D|\theta)$ is a measure of how well the model with parameter(s) θ explains the data D. It requires calculating the probability of the data if coming from the evaluation of the model for a given set of parameters. Typically, this is done in two steps: First, the model is evaluated for a given set of parameters, and then the likelihood is computed as a measure of the agreement between the simulated data and the true data, assuming some noise σ in the data.

Hence, when either the model is computationally expensive to evaluate or the comparison step between $M(\theta)$ and D is expensive (typically due to the volume of data), the likelihood becomes the bottleneck of the inference pipeline.

In the case of independent data points, the likelihood is simply defined as the product of the likelihoods of each data point given the parameter(s):

$$L(D|\theta) = \prod_{i=1}^{N} L(D_i|\theta) \ . \tag{2.4}$$

This makes it convenient to define the log-likelihood $\mathcal{L}$ as

$$\mathcal{L}(D|\theta) = \log L(D|\theta) = \sum_{i=1}^{N} \log L(D_i|\theta) \ , \tag{2.5}$$

where D_i is the i-th data point, and N is the number of data points.

The exact prescription for computing $L(D_i|\theta)$ depends on the problem at hand and can range from simple analytical expressions to complex numerical simulations. In either case, it involves the evaluation of the model for a given set of parameters and the computation of some measure of the agreement between the simulated data of the model and the true data, assuming some noise σ_i in the data.

When the data is assumed to have independent Gaussian noise, the likelihood for a single data point is given by

$$L(D_i|\theta) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{(D_i - M(\theta))^2}{2\sigma_i^2}\right) \ . \tag{2.6}$$

2.1.3 Evidence and model selection

In practice, the evidence is a normalization constant that is often difficult to compute. Using some elemental algebra on Bayes' theorem, it is easy to show that the evidence is equal to the *marginal likelihood*, i.e., the integral of the likelihood over the prior density

$$Z(D) = \int L(D|\theta)\pi(\theta)\,\mathrm{d}\theta \ . \tag{2.7}$$

The role of the evidence is important in model comparison. Comparing two models M_1 and M_2 can be done using the Bayes factor, which is the ratio of the evidence of two models:

$$B_{12} = \frac{Z(D|M_1)}{Z(D|M_2)} \ , \tag{2.8}$$

where B_{12} is the Bayes factor between M_1 and M_2, $Z(D|M_1)$ is the evidence of model M_1, and $Z(D|M_2)$ is the evidence of model M_2. Typically, the Bayes factor is compared to Jeffreys' scale [6] to quantify the strength of evidence in favor of one model over another.

2.1.4 The role of priors

The computation of the likelihood $L(D|\theta)$, given a fixed theory and data, is independent of the choice of the model (and its parameterization) that is used to compute a prediction of the data. Conversely, the choice of the prior $\pi(\theta)$ is subjective, as it represents the prior knowledge or beliefs about the parameter before the data is observed. Even when attempting to make the prior as uninformative as possible, it is impossible to avoid some level of impact. It is, for example, already computationally impossible to define a prior with infinite support, as its value (and hence the value of the posterior density) vanishes.

Typical uninformative choices for priors are uniform distributions on the parameters (if their order of magnitude is more or less determined) or their logarithm (if it is not). Using such a parameterization, the posterior density function becomes proportional to the likelihood. Additionally, the prior can incorporate known physics, i.e., symmetries.

Another intuitive choice for the prior, commonly referred to as *Jeffreys' prior*, is a prior that is invariant under reparameterization. This is achieved by choosing a prior that is proportional to the square root of the Fisher information matrix. This prior is often used in the context of model comparison.

Lastly, a useful relation that is easy to see is the one obtained by extending the Bayesian framework to include a second independent dataset ($P(D|D') = P(D)$). In this case, Bayes' theorem states that

$$P(\theta|D, D') = \frac{P(D, D'|\theta)P(\theta)}{P(D, D')} = \frac{P(D|D', \theta)P(D'|\theta)P(\theta)}{P(D|D')P(D')} \ , \tag{2.9}$$

which can be rearranged into

$$P(\theta|D, D') = \frac{P(D|\theta)}{P(D)} \cdot \frac{P(D'|\theta)P(\theta)}{P(D')} \ , \tag{2.10}$$

where the posterior of D' (the second term) becomes the prior for D. This means that if there are known measurements of a parameter θ, and the new information is independent of the old, the posterior from previous data can be used as the prior for new data. In this way, one can build *hierarchical* models.

2.1.5 Parameter inference

As mentioned in Section 2.1.1, when trying to infer model parameters, one typically neglects the evidence and focuses on the proportionality

$$p(\theta|D) \propto L(D|\theta)\pi(\theta) \ , \tag{2.11}$$

or in terms of the log-probabilities

$$\log p(\theta|D) \propto \mathcal{L}(D|\theta) + \log \pi(\theta) \ . \tag{2.12}$$

For simple combinations of prior and likelihood distributions, the posterior can be computed analytically, but in most real-world scenarios, the posterior is a complex, multi-dimensional distribution that depends on the data in a non-linear way. Inference in this case is equivalent to computing a high-dimensional integral over the parameter space, which is a computationally expensive task. The following sections cover different methods to perform Bayesian inference in a computationally efficient way.

2.1.6 Measures of similarity between probability distributions

As this work covers a variety of methods for Bayesian inference and frequently compares them, it is important to be able to quantify the dissimilarity between probability distributions. This is particularly important when comparing the true posterior distribution to the approximations that are used in the inference process. The measures that are used in this work are the Kullback-Leibler (KL) divergence, Jeffreys' divergence, and the Jensen-Shannon (JS) divergence.

The KL divergence between two discrete probability distributions P and Q is defined as [7]

$$D_{\mathrm{KL}}(P||Q) = \sum_{x\in\mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \ , \tag{2.13}$$

where $\mathcal{X}$ is the sample space that P and Q are defined on. Likewise, for continuous probability density functions (PDFs) p and q, the KL divergence reads

$$D_{\mathrm{KL}}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x \ . \tag{2.14}$$

It is important to note that the KL divergence is not symmetric, i.e., $D_{\mathrm{KL}}(P||Q) \neq D_{\mathrm{KL}}(Q||P)$, and it is 0 if and only if $P = Q$. There is no upper limit to the KL divergence. A useful analytical formula is that of

the KL divergence between two d-dimensional Gaussian distributions with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariances $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$:

$$D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)) = \frac{1}{2}\Bigg(\mathrm{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1)+ \\ +(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) - d + \log\frac{|\boldsymbol{\Sigma}_2|}{|\boldsymbol{\Sigma}_1|}\Bigg) \ . \tag{2.15}$$

A symmetrized version of the KL divergence is *Jeffreys' divergence*[1] [6]:

$$D_{\mathrm{KL}}^{\mathrm{sym}}(P,Q) = \frac{1}{2}\left[D_{\mathrm{KL}}(P||Q) + D_{\mathrm{KL}}(Q||P)\right] \ . \tag{2.16}$$

The Jensen-Shannon (JS) divergence is a symmetric and bounded measure of dissimilarity between probability distributions. It is defined as the average of the KL divergences between the two distributions P and Q and their mixture distribution $M \equiv \frac{1}{2}(P+Q)$ [8]:

$$D_{\mathrm{JS}}(P,Q) = \frac{1}{2}\left[D_{\mathrm{KL}}(P||M) + D_{\mathrm{KL}}(Q||M)\right] \ . \tag{2.17}$$

The JS divergence is bounded by $0 \leq D_{\mathrm{JS}}(P,Q) \leq \log(2)$ (using natural logarithms).

A similar but approximate result to Eq. (2.15) can be derived for the JS divergence between two Gaussian distributions. Assume two d-dimensional Gaussian distributions with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariances $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. The mixture distribution is not a Gaussian. However, if the two distributions are sufficiently close together, it can be approximated as a Gaussian with mean $\boldsymbol{\mu}_M = \frac{1}{2}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)$ and covariance $\boldsymbol{\Sigma}_M = \frac{1}{2}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)+\frac{1}{4}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\otimes(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \approx \frac{1}{2}(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)$. The JS divergence between the two distributions is then given by

$$D_{\mathrm{JS}}(\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1),\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)) \approx \frac{1}{4}\Delta\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}\Delta\boldsymbol{\mu} - \frac{d}{2}\log 2 + \\ + \frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}}\right) \ , \tag{2.18}$$

where $\Delta\boldsymbol{\mu} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$.

As P and Q approach each other, the KL divergence, Jeffreys' divergence, and JS divergence all approach 0 and can be used interchangeably. Usually a value < 0.01 for either can be considered to be a very good agreement between the two probability distributions. Compared to the other two divergences, the JS divergence is more numerically stable because the mixture distribution covers the common support of P and Q, and since the JS

[1] There is a slight ambiguity regarding the factor of 1/2 in the definition of Jeffreys' divergence as this factor is sometimes omitted. As we want to recover the original KL divergence for symmetric distributions, we include the factor.

divergence is bounded from above. This comes at the disadvantage of being harder to compute, however. Furthermore, in gradient descent methods where the divergence is minimized, using the KL divergence over Jeffreys' divergence can be advantageous, as it provides steeper gradients.

2.2 Bayesian inference with Monte Carlo methods

Monte Carlo (MC) methods, in their definition, comprise a broad class of algorithms that rely on random sampling for numerical computation. In the context of Bayesian inference, sampling the posterior distribution is the objective.

In its most basic form, the philosophy of MC methods is that a numerically driven random approach can be more computationally efficient than a deterministic algorithm. This is especially true in the context of high-dimensional integrals, where the computational cost of a deterministic approach scales exponentially with the dimensionality of the problem.

Monte Carlo methods often work sequentially, where one or more previously drawn samples inform the next sample. This is in contrast to most deterministic integration methods, where the samples are drawn according to a fixed grid or a fixed rule.

This sequential procedure becomes particularly advantageous when the posterior distribution *(a)* has a non-trivial surface such as multiple modes, tight degeneracies, or very different scales in different directions, or *(b)* when the posterior distribution is high-dimensional.

Furthermore, Monte Carlo methods tend to rely on simple rules for the generation of samples, which makes them easy to implement and computationally efficient.

This section covers the use of MC methods in Bayesian inference, focusing on two of the most commonly used methods: Markov Chain Monte Carlo and Nested sampling.

2.2.1 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) is a class of algorithms that generate a Markov chain of samples from the posterior distribution. A Markov chain is defined as a sequence of samples $\theta_1, \theta_2, \ldots, \theta_n$ where each sample is drawn from a proposal distribution $q(\theta|\theta')$ that depends only on the previous sample θ'. The exact form of the proposal distribution depends on the specific MCMC algorithm used, and we explore some of them in the following.

The core idea behind MCMC is to tune this Markov chain in such a way that the stationary distribution of the chain approaches a distribution that is proportional to the posterior. In other words, the goal is to construct a chain that follows a path in the parameter space such that the normalized histogram of values approaches the posterior probability density function (PDF) [9].

While the notation above assumes a one-dimensional parameter space, in practice, the parameter space is typically high-dimensional, and the chain is a sequence of vectors $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n$. The generalization is trivial.

Although there are numerous implementations of MCMC, we present the two most relevant ones: Metropolis-Hastings and Hamiltonian Monte Carlo.

Metropolis-Hastings

The Metropolis-Hastings (MH) algorithm [10, 11] is a widely used MCMC algorithm that stands out for its simplicity. One proposes θ' from a simple proposal distribution $q(\theta'|\theta)$ and then accepts or rejects the new sample based on an acceptance probability $\alpha(\theta \to \theta')$, given by

$$\alpha(\theta \to \theta') = \min\left(1, \frac{p(\theta'|D)}{p(\theta|D)} \frac{q(\theta|\theta')}{q(\theta'|\theta)}\right) , \tag{2.19}$$

where $p(\theta|D)$ is the posterior probability of the parameter. In practice, this computation is performed in two steps by first drawing a candidate sample θ' from the proposal distribution $q(\theta'|\theta)$ and then drawing a random number u from a uniform distribution over $[0, 1)$. If $u < \alpha(\theta \to \theta')$ the candidate sample is accepted; otherwise it is rejected, and the chain stays in the previous location θ.

Typically (and to satisfy the detailed balance condition), the proposal distribution is symmetric, i.e., $q(\theta'|\theta) = q(\theta|\theta')$, which simplifies the acceptance probability to

$$\alpha(\theta \to \theta') = \min\left(1, \frac{p(\theta'|D)}{p(\theta|D)}\right) . \tag{2.20}$$

This formulation ensures that the chain always jumps towards higher values of p. The probability to jump to locations where $p(\theta'|D)$ is lower than $p(\theta|D)$ is proportional to the ratio of the posterior probabilities. Furthermore, note that q can only depend on the current value of the chain θ to conserve the Markov property.

MH-MCMC stands out for its simplicity and ease of implementation, but choosing an appropriate proposal distribution can be challenging and typically requires some fine-tuning. Especially when the posterior distribution

is multimodal, choosing a small step size makes it unlikely that the chain crosses the "valleys" between the modes. Conversely, a large step size can lead to a low acceptance rate and slow convergence.

Furthermore, in a true Monte Carlo chain, the samples should be *independent*, i.e., statistically uncorrelated. However, the next sample in the chain always depends on the previous one, which means that this requirement is not satisfied. In the limit of an infinitely long chain, we would eventually reach independence, but in practice, one typically assumes the samples to be approximately independent after a number of steps of the MCMC.

Lastly, there is the issue of *burn in*. If the chain starts in a random location, it typically requires some iterations to "climb" the mode and to converge to the stationary distribution. This necessitates discarding the first samples (usually around $20 - 30\%$ of the chain).

Techniques such as simulated tempering [12] try to alleviate the problem of underexploring the posterior by introducing a temperature parameter that slowly decreases as a function of the number of steps taken.

Hamiltonian Monte Carlo

Hamiltonian Monte Carlo[2] (HMC), is a variant of the Metropolis-Hastings algorithm that uses physics-inspired dynamics to propose new samples. The core idea behind HMC is to use the gradient of the log-posterior to propose new samples, which allows for more efficient exploration of the parameter space. The dynamics of the system are described by the Hamiltonian [13]

$$H(\boldsymbol{\theta}, \mathbf{p}) = -\log p(\boldsymbol{\theta}|D) + \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} \ , \tag{2.21}$$

where $\mathbf{p}$ (not to be confused with the posterior probability p) is the *momentum* and $\mathbf{M}$ is the *mass matrix*. The mass matrix is a positive definite, symmetric matrix that is usually chosen to be the identity matrix. The notation implies a multi-dimensional parameter space; however, the same principles apply to a one-dimensional parameter space. The negative log-posterior acts as a potential well, counteracted by the kinetic energy of the system (the quadratic term). As in classical mechanics, the energy of the system is conserved, and the dynamics of the system are described by the usual equations of motion:

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = \frac{\partial H}{\partial \mathbf{p}} \quad \text{and} \quad \frac{\mathrm{d}\mathbf{p}}{\mathrm{d}t} = -\frac{\partial H}{\partial \boldsymbol{\theta}} \ , \tag{2.22}$$

[2] In the literature, the naming of this algorithm is not consistent; it is also commonly referred to as Hybrid Monte Carlo

which can be solved numerically for each component using, e.g., a leapfrog integrator. A new sample is obtained by setting a momentum and solving the equations of motion to obtain a new sample θ'. The choice of both the momentum and the step size in the leapfrog integrator depend on the specific algorithm (see e.g., [14, 15, 16, 17]).

The acceptance probability is then given by

$$\alpha(\theta \to \theta') = \min\left(1, \frac{\exp(H(\theta, \mathbf{p}))}{\exp(H(\theta', \mathbf{p}'))}\right) , \tag{2.23}$$

where $\mathbf{p}'$ is the new momentum. HMC is particularly useful in high-dimensional problems where the posterior distribution is complex and multimodal, as it allows for more efficient exploration of the parameter space. However, it requires computing the gradient of the log-posterior, which can be computationally expensive or infeasible if not analytical.

2.2.2 Nested sampling

Nested sampling is an alternative to MCMC that produces an estimate of the evidence along with a Monte Carlo sample from the distribution. Introduced in [18], it estimates the evidence from Eq. (2.7) (dropping the explicit dependence on the data)

$$Z = \int L(\theta)\pi(\theta)\mathrm{d}\theta \ . \tag{2.24}$$

The nested sampling algorithm transforms this integral by defining

$$X(\lambda) = \int_{L(\theta)>\lambda} \pi(\theta)\mathrm{d}\theta \ , \tag{2.25}$$

where X decreases from 1 to 0 as λ increases. By defining $L(X)$ as the inverse function, the evidence becomes

$$Z = \int_0^1 L(X)\mathrm{d}X \ . \tag{2.26}$$

This transformation simplifies the integral, but inverting L is non-trivial. Instead, the integral is approximated using a weighted sum

$$\sum_{i=1}^{m} w_i L_i \to Z \ , \tag{2.27}$$

where $w_i = \Delta X$ is the distance between X_i and X_{i+1}.

To sample efficiently, a variable t is introduced such that

$$X_1 = t_1, \ X_2 = t_1 t_2, \ldots, X_i = t_1 t_2 \ldots t_i, \ldots X_m = t_1 t_2 \ldots t_m \ , \tag{2.28}$$

where each t_i is drawn from a uniform distribution over $[0, 1)$. The expectation value and variance of $\log(t)$ are

$$\mathbb{E}(\log(t)) = -1/N \ , \qquad \mathrm{var}(\log(t)) = \frac{1}{N^2} \ , \tag{2.29}$$

where N is the number of samples. In practice, the algorithm begins with a pool of samples called *live points* and iteratively selects the worst point (lowest L) and replaces it with a new point drawn from the prior with $L > L_i$. The discarded points, called *dead points*, are retained. This process continues until the desired precision is reached.

The uncertainty in Z can be estimated from the distribution of t_i values:

$$p(\boldsymbol{t})\mathrm{d}\boldsymbol{t} = \prod_i N t_i^{N-1} \ , \tag{2.30}$$

which defines a probability distribution for Z

$$P(Z) = \int \delta \left(Z - \sum_{i-1}^{m} L_i w_i(\boldsymbol{t}) \right) p(\boldsymbol{t})\mathrm{d}\boldsymbol{t} \ . \tag{2.31}$$

The moments of this distribution can be determined using integration algorithms, and the evidence is typically calculated in log space for numerical convenience.

Additionally, the distribution of dead points, weighted by the likelihood, provides a Monte Carlo estimate of the posterior distribution.

2.3 Approximate inference

While "traditional" Bayesian inference methods such as MCMC and Nested sampling are powerful algorithms for estimation and model selection, they can *(a)* be computationally expensive and slow to converge, requiring many samples to accurately estimate the posterior distribution, and *(b)* require the likelihood to be known, which is not always possible or feasible.

This has led to the development of inference methods designed to approximate the posterior distribution efficiently and effectively. This section covers two of the most relevant methods for approximate inference in physics, which address the two problems stated above: Variational inference and Simulation-based inference.

2.3.1 Variational inference

Variational inference (VI) is a class of methods that approximate intractable posterior distributions by recasting the posterior estimation into an optimization problem. Instead of sampling $p(\theta|d)$ directly, VI posits one

or multiple trial densities $q_\lambda(\theta)$, also called *variational distribution* (with learnable parameters λ) to approximate the true posterior [19].

This trial density, which is easier to evaluate than $p(\theta|d)$, is tuned to be a locally optimal approximation to the true posterior. This typically yields an analytically computable marginal posterior, thus eliminating the need for sampling, although not all variational models rely on this property.

In practice, VI is performed by defining a measure of dissimilarity (or *divergence*) between the proposal $q_\lambda(\theta)$ and the target $p(\theta|d)$ and then optimizing λ to minimize this divergence. A common and natural choice is the KL divergence $D_{\mathrm{KL}}(q_\lambda||p)$ as defined in Eq. (2.14). Note the order of the arguments in the divergence ($q_\lambda(\theta)$ is the proposal, $p(D|\theta)$ the target), which is "backwards", and thus in regions where $q_\lambda(\theta)$ has no support, the contribution to the KL divergence is small. This encourages better fitting towards the top of the mode at the risk of underestimating the tails of the posterior.

Building on the KL divergence as a measure of dissimilarity, one arrives at the evidence lower bound (ELBO) objective [20]. The ELBO can be derived from Eq. (2.7) by taking the logarithm and recasting the integral to be the expectation with respect to $q_\lambda(\theta)$:

$$\log Z(D) = \log \int L(D|\theta)\pi(\theta)\mathrm{d}\theta = \log \int q_\lambda(\theta)\frac{L(D|\theta)\pi(\theta)}{q_\lambda(\theta)}\mathrm{d}\theta. \tag{2.32}$$

Jensen's inequality [21] states that for any convex function f, $\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$. Applying this to the logarithm, one finds

$$\log Z(D) \geq \int q_\lambda(\theta) \log \frac{L(D|\theta)\pi(\theta)}{q_\lambda(\theta)}\mathrm{d}\theta \ . \tag{2.33}$$

This can be recast into the ELBO:

$$\mathrm{ELBO}(\lambda) = \mathbb{E}_{q_\lambda(\theta)}[\mathcal{L}(D|\theta)] - \mathrm{D}_{\mathrm{KL}}(q_\lambda(\theta) \,||\, \pi(\theta)) \ . \tag{2.34}$$

where $\mathcal{L}$ is the log-likelihood. We can examine the two terms separately: The first term is the expected log-likelihood of the data under the variational distribution ("fit" term), while the second term penalizes deviation from the prior ("complexity" term). The ELBO naturally acts as a lower bound as the KL divergence is non-negative. Maximizing the ELBO is therefore equivalent to minimizing the KL divergence between the variational distribution and the true posterior, hence leading to a q_λ that approximates the posterior as closely as possible. Typically, the ELBO is minimized using standard optimization techniques such as stochastic gradient descent. In principle, the variational distribution can be any distribution, such as a Gaussian or a mixture of Gaussians. However, in

modern VI implementations, Neural Networks (NNs) have seen increasing use, enabling highly flexible approximations. Popular choices include Normalizing Flows [22, 23], Variational Autoencoders [24], and Gaussian Processes [25].

2.3.2 Simulation-based inference

Simulation-based inference (SBI), also known as likelihood-free inference (see e.g., [26] for a review), refers to Bayesian inference methods that rely on a simulator (generative model) rather than an explicit likelihood function. In Bayesian terms, the goal is to obtain the posterior $p(\theta|D)$ without having to explicitly compute the likelihood $L(D|\theta)$. Instead, we can generate synthetic data $D \sim p(D|\theta)$ via a simulator. Bayes' theorem still holds in this approach, but the evaluation of the likelihood is bypassed by using forward simulations and comparisons to the observed data to approximate the posterior. Effectively, one replaces the likelihood computation with repeated simulation experiments.

SBI is particularly useful for complex models with intractable or computationally expensive likelihoods. This scenario arises in many fields where simulating the process (forward model) is achievable, but writing down $L(D|\theta)$ in closed form poses a challenge. SBI addresses this by using the simulator as a black-box generative model: One generates many (θ, D) pairs from the prior and simulator and uses these to infer the relationship between the data and parameters. The posterior is thus "learned" from simulations rather than computed from an analytic formula. In its simplest form, the idea of SBI is to sample $\theta \sim \pi(\theta)$, simulate $D \sim p(D|\theta)$, and accept those θ whose simulated data D closely match D_{obs}. This procedure yields the basic Approximate Bayesian Computation algorithm (discussed below). As the matching criterion tightens (likelihood approximation improves), the accepted θ's approximate draws from the true posterior. In its more complex form, SBI leverages machine learning methods such as NNs to learn the relationship between θ and D and to approximate the posterior more efficiently. This can involve learning a mapping from θ to a compressed representation of D (summary statistics) or directly from θ to D. SBI has seen huge developments in the last few years (see [26] for a review) with a plethora of algorithms. A select number of them is introduced in the following.

Approximate Bayesian Computation

Approximate Bayesian Computation (ABC) is the prototypical SBI approach. Pioneered in [27], it generates simulations from the model and

uses an acceptance criterion to approximate the posterior. The basic ABC rejection sampling algorithm requires only three ingredients: the simulator that generates data $D \sim p(D|\theta)$, a prior $\pi(\theta)$, and a distance metric $\rho(D, D_{\text{obs}})$ that quantifies the discrepancy between simulated and observed data. In its simplest form, this distance metric can be the Euclidean distance between the simulated and observed data. The algorithm then involves the following steps:

1. Sampling θ_i from the prior $\pi(\theta)$.
2. Simulating a data set $D_i \sim p(D|\theta_i)$ using the simulator.
3. Comparing D_i to the observed data D_{obs} using the distance measure $\rho(D_i, D_{\text{obs}})$.
4. Accept θ_i if $\rho(D_i, D_{\text{obs}}) \leq \epsilon$, where $\epsilon \geq 0$ ($\epsilon > 0$ in continuous spaces) is a tolerance threshold. Otherwise, reject θ_i.

The distribution of accepted θ-values approximates the desired posterior $p(\theta|D_{\text{obs}})$ (exactly so as $\epsilon \to 0$ for a sensible choice of ρ). Crucially, this procedure requires no evaluation of $p(D|\theta)$ and only uses simulated outcomes to decide acceptance. In practice, the value of ϵ must balance accuracy and computational cost: smaller values of ϵ yield more accurate posteriors but lower acceptance rates, while larger ϵ values lead to higher acceptance rates at the cost of less accurate posteriors. ABC yields an approximate posterior that is broader than the true posterior due to the non-zero tolerance.

One major and hard-to-control drawback is the difficulty in predicting the convergence speed of the ABC posterior to the true posterior as $\epsilon \to 0$. The choice of distance metric certainly plays a role, but it is not always trivial to find a good distance metric in the first place. Additionally, there is no way of determining whether the tail mass in the posterior is due to the choice of the distance metric or ϵ, or whether it comes from the posterior itself. Additionally, disentangling errors introduced by the approximation from those introduced through model misspecification remains challenging [28].

ABC suffers from the curse of dimensionality: as the dimensionality of θ increases, the ratio of accepted to rejected simulations decreases exponentially. There are several avenues towards improving ABC, among which are the use of Sufficient Summary Statistics, ABC-MCMC, and ABC-Sequential Monte Carlo (ABC-SMC).

Sufficient Summary Statistics $S(D)$ compress the data into lower-dimensional representations. One matches these lower-dimensional summaries

$S(D_i) \approx S(D_{\text{obs}})$ that capture most of the information. This does not improve upon the problem of low acceptance rate but instead reduces the computational cost of calculating the distance metric. Furthermore, if the simulation is already performed in the compressed space, this speeds up the simulation process too. The choice of good summaries (ideally sufficient for θ) is critical, as poor summaries can bias the posterior [26].

ABC-MCMC (see e.g., [29]) combines ABC with a Metropolis-Hastings algorithm to achieve a higher acceptance rate of samples. Proposals θ' are drawn via an MH-MCMC chain, and an acceptance probability is defined analogously to Eqs. (2.19) and (2.20) that includes the ABC condition (simulating D' and accepting the step if $\rho(D', D_{\text{obs}}) < \epsilon$). By proposing θ' from a distribution that is more likely to yield accepted samples (i.e., mostly proposing samples nearby), the acceptance rate can be increased. The downside is that this method inherits the usual MCMC issues of correlation within the chain and fine-tuning of the proposal distribution.

ABC-SMC [30, 31] extends ABC-MCMC by constructing a sequence of intermediate distributions that gradually transition from the prior to the posterior, tightening the tolerance ϵ in steps. Starting with a loose tolerance, one obtains a pool of accepted θ "particles". The tolerance is then lowered, and the particle set evolved. Every particle is moved via some rule (usually MCMC) and reweighted, or resampled, to fit the new tolerance criterion. Over several iterations, the particles iteratively approach the posterior from the prior, automatically adjusting tolerance levels. ABC-SMC methods greatly improve efficiency in each step due to their gradual approach and do not require the user to hand-select ϵ beforehand as the algorithm can schedule it adaptively. However, the problem of judging the goodness of the approximation remains.

2.4 Machine-learning methods

The algorithms presented in Sections 2.2 and 2.3 have seen widespread adaptation in physics and Bayesian inference more broadly. However, they are not without their limitations: Both MCMC and Nested sampling can be slow to converge in high-dimensional problems. This is especially true if the likelihood has a complex shape[3] or if the posterior distribution is multimodal. VI and SBI can approximate the posterior distribution efficiently, but they are sensitive to the choice of the variational distribution

[3] In the context of likelihoods, *complex shape* refers to features in the likelihood surface that samplers struggle to map. These typically comprise tight degeneracies between parameters or deviations from multivariate Gaussians ("bananas"). In some cases, these can be mitigated by finding a suitable invertible transformation into a different parameter space, which is sampled instead.

or distance metric. Furthermore, they tend to suffer from the same issues as MCMC and Nested sampling in high-dimensional problems.

At the same time, the precision of modern experiments in some fields has outpaced increases in computational power, which has led to a situation where the likelihoods are becoming ever more computationally expensive to evaluate. In turn, physics faces a trade-off between the precision of the likelihood, the computational cost of evaluating it, and the human hours required to optimize the inference pipeline. This has made some inference problems prohibitively expensive.

Machine learning has rapidly developed in recent decades, driven by increased computational power and advances in the development of NNs. Training these NNs is often a Bayesian inference problem in itself. This has naturally led to the development of machine learning methods that can outcompete these classic algorithms in a set of scenarios. In this section, we cover three of the most commonly used machine learning techniques for Bayesian inference and relate them to the methods covered above.

It is important to note that many of these concepts borrow from one another and are used in combination with the aforementioned methods (such as e.g., Variational inference with Gaussian Processes, Nested sampling with Normalizing Flows) as explained in Section 2.5.

The use of these methods, specifically in the field of gravitational wave astronomy and LISA inference, is discussed in the next chapter.

2.4.1 Neural networks

Neural networks (see e.g., [32, 33] for reviews) are a class of machine learning algorithms that are broadly inspired by the structure of the human brain. They consist of layers of interconnected nodes (neurons). Each node receives input from the previous layer, applies a (nonlinear) transformation to the sum of its inputs through an *activation function*, and passes the output to the next layer. The complexity of the network is determined by the number of layers and the number of nodes in each layer. The first and last layers are called the *input layer* and *output layer*, respectively, and the layers in between are called *hidden layers*. Mathematically, the network is represented through a series of matrix multiplications and element-wise operations. The output of each layer is given by:

$$\boldsymbol{a}^{(l)} = \sigma(\mathbf{W}^{(l)}\boldsymbol{a}^{(l-1)} + \boldsymbol{b}^{(l)}) \ , \tag{2.35}$$

where $\boldsymbol{a}^{(l)}$ is the output of layer l, $\mathbf{W}^{(l)}$ is the weight matrix, $\boldsymbol{b}^{(l)}$ is the bias vector, and σ is the activation function. The weights $\mathrm{W}_{ij}^{(l)}$ and biases

$\mathrm{b}_i^{(l)}$ are learned from the training data. This is done by minimizing a distance (loss function in machine learning terms), which quantifies the difference between the predicted output and the true output. Typically, an optimization algorithm is used to adjust the weights and biases of the network.

Normalizing Flows

One type of NN particularly suited for applications in Bayesian inference is the Normalizing Flow (see e.g., [34] for a review). Normalizing Flows are a class of generative models that learn a mapping from a simple initial probability distribution $p^{(0)}(\boldsymbol{\theta}^{(0)})$ (e.g., a Gaussian) to a more complex target distribution $p^{(T)}(\boldsymbol{\theta}^{(T)})$ (e.g., the likelihood or posterior distribution). This mapping is achieved through a series of invertible transformations. For $i = 1, 2, \ldots, T$, the transformations are given by

$$\boldsymbol{\theta}_i = f^{(i)}(\boldsymbol{\theta}^{(i-1)}), \tag{2.36}$$

where the functions f_i are invertible and differentiable. The change-of-variable formula states that

$$p^{(i)}(\boldsymbol{\theta}^{(i)}) = p^{(i-1)}(\boldsymbol{\theta}^{(i-1)}) \left| \det \left(\frac{\mathrm{d}(f^{(i)}(\boldsymbol{\theta}^{(i-1)}))^{-1}}{\mathrm{d}\boldsymbol{\theta}^{(i-1)}} \right) \right| , \tag{2.37}$$

where the determinant term $\left|\det \left(\mathrm{d}(f^{(i)}(\boldsymbol{\theta}^{(i-1)}))^{-1} / \mathrm{d}\boldsymbol{\theta}^{(i-1)} \right)\right|$ is the Jacobian of the transformation. The transformations can be applied sequentially to map the initial distribution to the target distribution

$$p^{(T)}(\boldsymbol{\theta}^{(T)}) = p^{(0)}(\boldsymbol{\theta}^{(0)}) \prod_{i=1}^{T} \left| \det \left(\frac{\mathrm{d}(f^{(i)}(\boldsymbol{\theta}^{(i-1)}))^{-1}}{\mathrm{d}\boldsymbol{\theta}^{(i-1)}} \right) \right| . \tag{2.38}$$

Using the identity $\det(A^{-1}) = \det(A)^{-1}$ and taking the logarithm of Eq. (2.38), one arrives at the easier-to-compute formula

$$\log p^{(T)}(\boldsymbol{\theta}^{(T)}) = \log p^{(0)}(\boldsymbol{\theta}^{(0)}) - \sum_{i=1}^{T} \log \left| \det \left(\frac{\mathrm{d}f^{(i)}(\boldsymbol{\theta}^{(i-1)})}{\mathrm{d}\boldsymbol{\theta}^{(i-1)}} \right) \right| . \tag{2.39}$$

The mappings $f^{(i)}$ can take various forms, ranging from simple linear transformations such as $f^{(i)}(\boldsymbol{\theta}^{(i-1)}) = \mathbf{W}^{(i)}\boldsymbol{\theta}^{(i-1)} + \boldsymbol{b}^{(i)}$, to more complex transformations. These complex transformations often involve differentiable NNs. A comprehensive exploration of all possible transformations is beyond the scope of this work, but the interested reader is referred to [34] for a detailed review.

2.4.2 Gaussian Processes

Gaussian Processes (GPs) have become a widely used tool for regression and classification tasks over the past few decades. They are a mathematically simple, yet powerful, non-parametric Bayesian method leveraging the simple structure of multivariate Gaussian distributions such as analytical marginalization and conditioning. The key idea behind GPs is to assume that any finite set of function values of an arbitrary function $f(x)$ is jointly Gaussian distributed. As GPs constitute a cornerstone of the methods discussed in this work, they are introduced in more detail in the following.

Concept

To understand GPs, it is useful to first define a stochastic process. A stochastic process can be thought of as a function $\{Y(t) : t \in T\}$, Y being a random variable drawn from some probability measure P. T is often referred to as *index set* [35]. With this, one defines a GP:

Definition 2.4.1. A Gaussian Process is a collection of random variables, any finite number of which have a joint Gaussian distribution. [36]

This means that a GP is a stochastic process defined on any set $T = \{t_1, ..., t_n\}$ where the n values $\{y_1, ..., y_n\}$ are drawn from a joint Gaussian distribution:

$$\mathcal{N}(\boldsymbol{t}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^n|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{t}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{t}-\boldsymbol{\mu})\right) . \tag{2.40}$$

Here $\boldsymbol{t} = (t_1, ..., t_n)^\top$ is the vector of indices, $\boldsymbol{\mu}$ the *mean* vector (representing the expected values), and $\boldsymbol{\Sigma}$ the *covariance* matrix (capturing the relationships between the indices). Any set of indices T is allowed for a valid GP, but for our case, only the case where T is continuously defined on $\mathbb{R}^d$ is of interest, although we restrict ourselves to the case where $d = 1$ for simplicity and explain the (straightforward) extension to higher dimensions later. In the following, the index set of the continuous GP is denoted as X.

The trick of GPs is to make $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ functions of the index set X such that for any two points $x, x' \in X$, the mean and covariance can be expressed as

$$m(x) = \mathbb{E}[f(x)] , \tag{2.41}$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] . \tag{2.42}$$

This allows defining a GP as a distribution over functions $f(x)$, which can

be expressed as

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')) \ , \tag{2.43}$$

where $m(x)$ and $k(x, x')$ are the mean and covariance functions, respectively. The covariance function $k(x, x')$ is often called the *kernel*. The definition as a stochastic process furthermore implies a consistency requirement (also called Kolmogorov's extension theorem [37]), which demands that any GP that specifies $(y_1, \ldots, y_n) \sim \mathcal{N}(x|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ on any set X must equally specify $(y'_1, \ldots, y'_n) \sim \mathcal{N}(x'|\boldsymbol{\mu}', \boldsymbol{\Sigma}')$ for any subset $X' \subset X$ by taking the relevant parts of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. In other words, this means that predictions about a finite subset of indices can be made without knowing the full infinite-dimensional distribution.

Typically, this *unconditioned* version of the GP is called the prior GP[4] as it reflects the distribution of functions that one would draw if one had no knowledge about the function. This GP prior is fully specified by the mean and kernel functions.

Conditioning

To incorporate knowledge from a set of training points $\{(x_i, y_i = f_i)|i = 1, \ldots, n\}$, we condition the GP. The joint distribution of the training points $\boldsymbol{f}(\boldsymbol{x}) \equiv \boldsymbol{y}$ and test points $\boldsymbol{f}_*(\boldsymbol{x}_*) \equiv \boldsymbol{y}_*$ is given by:

$$\begin{bmatrix} \boldsymbol{f}(\boldsymbol{x}) \\ \boldsymbol{f}_*(\boldsymbol{x}_*) \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{y}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m}(\boldsymbol{x}) \\ \boldsymbol{m}(\boldsymbol{x}_*) \end{bmatrix}, \begin{bmatrix} \mathbf{K}(\boldsymbol{x}, \boldsymbol{x}) & \mathbf{K}(\boldsymbol{x}, \boldsymbol{x}_*) \\ \mathbf{K}(\boldsymbol{x}_*, \boldsymbol{x}) & \mathbf{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) \end{bmatrix} \right) \ , \tag{2.44}$$

where $\boldsymbol{m}_i = m(x_i)$ is the vector of the mean function and $\mathbf{K}_{i,j} = k(x_i, x_j)$ is called the *Gram matrix* of the training points. Conditioning on the observed values is straightforward for multivariate Gaussians:

$$\boldsymbol{f}_*(\boldsymbol{x}_*)|\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x}) \sim \mathcal{N}(\overline{\boldsymbol{f}}_*, \Sigma_{\boldsymbol{f}_*}) \ , \tag{2.45}$$

with

$$\boldsymbol{\mu}(\boldsymbol{x}_*) \equiv \overline{\boldsymbol{f}}_* = \boldsymbol{m}(\boldsymbol{x}_*) + \mathbf{K}(\boldsymbol{x}_*, \boldsymbol{x})\mathbf{K}(\boldsymbol{x}, \boldsymbol{x})^{-1}(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{m}(\boldsymbol{x})) \ , \tag{2.46}$$

and

$$\mathrm{cov}(\boldsymbol{f}_*(\boldsymbol{x}_*)) \equiv \Sigma_{\boldsymbol{f}_*} = \mathbf{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) - \mathbf{K}(\boldsymbol{x}_*, \boldsymbol{x})\mathbf{K}(\boldsymbol{x}, \boldsymbol{x})^{-1}\mathbf{K}(\boldsymbol{x}, \boldsymbol{x}_*) \ . \tag{2.47}$$

[4] As we are using GPs in the context of Bayesian inference, the terminology can be confusing as concepts such as priors, likelihoods, and posteriors exist both for the GP as well as for the target distribution of the inference. We stick to a notation where the quantities in the context of the GP are referred to as prior GP, GP-likelihood, etc., and the symbols get superscripts (e.g., $p^{\mathcal{GP}}, \mathcal{L}^{\mathcal{GP}}, \ldots$).

This conditioned GP is called the *GP-posterior*. For brevity, the explicit dependence on the inducing points $\boldsymbol{x}_*$ is dropped in the following.

Even with a prior zero-mean function $m(x) = 0$, the GP-posterior mean can be non-zero. This motivates the choice of $m(x) = 0$, simplifying the GP construction to the selection of an appropriate kernel function.

Figure 2.1 illustrates this conditioning process. The left side shows sample functions from a GP prior with $m(x) = 0$ and $k(x, x') = \exp\left(-(x - x')^2/2\right)$, while the right side shows sample functions from a GP conditioned on training points. The standard deviations are the square roots of the diagonal entries of the unconditioned and conditioned covariance matrices, respectively

$$\sigma(x_i) = \sqrt{\Sigma_{f(x_i),ii}} \ , \qquad \sigma_*(x_{*,i}) = \sqrt{\Sigma_{f_*,ii}} \ . \tag{2.48}$$

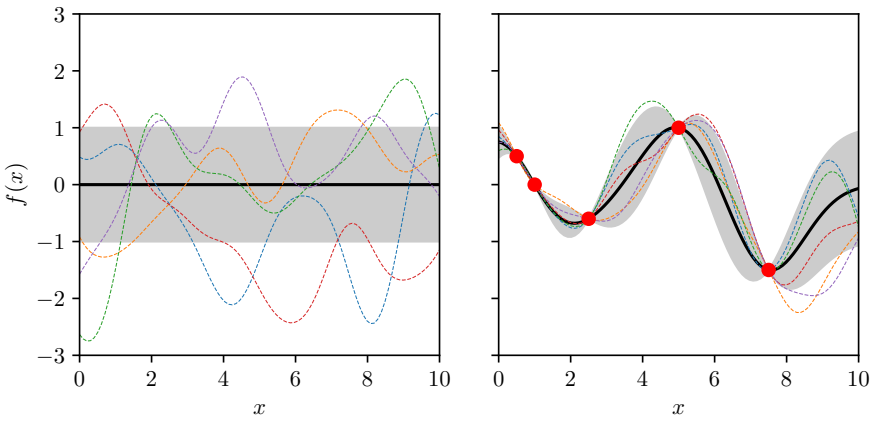


Figure 2.1: **Left:** Sample functions drawn from a GP with $m(x) = 0$ and $k(x, x') = \exp(-1/2(x - x')^2)$ (dashed lines) as well as the value of the prior GP mean function and the standard deviation $\sqrt{k(x, x)} = 1$ (solid black line and gray band respectively). **Right:** Sample functions drawn from the same GP (dashed lines), values of the posterior GP mean (solid black line), and standard deviation (gray band) after conditioning on five observations (red dots). Note how even with a zero prior mean function $m(x) = 0$, one obtains a non-zero posterior mean. Furthermore, after conditioning, only functions that pass through the training points are allowed.

The framework can be extended to include associated noise in the training data $y = f(x) + \epsilon$, where ϵ is a random variable with variance σ_n^2. This is done by adding a noise term to the kernel function

$$\tilde{k}(x, x') = k(x, x') + \sigma_n^2 \delta_{x,x'} \ , \tag{2.49}$$

where $\delta_{x,x'}$ is the Kronecker delta. In the conditioning step, this term is only applied to the gram matrix between training points $\tilde{K}(\boldsymbol{x}, \boldsymbol{x}) = \mathbf{K}(\boldsymbol{x}, \boldsymbol{x}) + \sigma_n^2 \mathbb{I}$, changing Eqs. (2.46) and (2.47) to

$$\boldsymbol{\mu}_* = \boldsymbol{m}(\boldsymbol{x}_*) + \mathbf{K}(\boldsymbol{x}_*, \boldsymbol{x})\tilde{\mathbf{K}}(\boldsymbol{x}, \boldsymbol{x})^{-1}(\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{m}(\boldsymbol{x})) \ , \tag{2.50}$$

and

$$\mathrm{cov}(\boldsymbol{f}_*) = \mathbf{K}(\boldsymbol{x}_*, \boldsymbol{x}_*) - \mathbf{K}(\boldsymbol{x}_*, \boldsymbol{x})\tilde{\mathbf{K}}(\boldsymbol{x}, \boldsymbol{x})^{-1}\mathbf{K}(\boldsymbol{x}, \boldsymbol{x}_*) \ . \tag{2.51}$$

The kernel function

As shown earlier, with the mean function typically assumed to be zero, the GP is fully characterized by its kernel, which means that the main challenge lies in choosing an appropriate kernel.

To choose a valid kernel, it is helpful to first establish the requirements it must satisfy. These are:

1. The kernel needs to map $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.
2. It needs to be symmetric: $k(x, x') = k(x', x)$.
3. The covariance matrix obtained from the kernel needs to be positive definite: $\boldsymbol{z}^T K(\boldsymbol{x}, \boldsymbol{x}')\boldsymbol{z} \geq 0$ for all $z \in \mathbb{R}^D \backslash 0$ and with $K(\boldsymbol{x}, \boldsymbol{x}')_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}'_j)$ being the Gram matrix of $\boldsymbol{x}$ and $\boldsymbol{x}'$. This condition is fulfilled if and only if $k(x, x') \geq 0$ for all $x, x' \in X$ [38].

In addition to these strict mathematical requirements, the kernel should also reflect the prior knowledge about the functional shape of f as well as possible. As GPs are non-parametric models, the choice of kernel function is not always trivial; however, there are some properties that translate from the kernel into the GP. Three important properties of kernels that influence the behavior of the GP are stationarity, differentiability, and periodicity.

1. *Stationarity* means that a kernel function is invariant to translations, i.e., $k(x + z, x' + z) = k(x, x) \quad \forall\ z$. To achieve this, typically the kernel is defined as a function of the distance between two points $r = |x - x'|$ (Euclidian distance if $d > 1$). If the kernel is stationary, the GP has the same property.
2. Likewise, *differentiability* directly translates to the GP: If the kernel is n times differentiable, the GP is n times differentiable as well. Typically it is desirable that the kernel be at least once differentiable to ensure that the GP is continuous.
3. A less commonly enforced property is *periodicity* $k(x, x') = k(x, x' + n \cdot z), \quad n \in \mathbb{Z}$ with periodicity z. This is less important for our case,

as posterior distributions rarely have periodic directions (although they do exist in the context of e.g., gravitational waves).

Having established these properties, we can now introduce some commonly used kernels. This list is by no means exhaustive. Perhaps, the most commonly used kernel function is the *Radial Basis Function* (RBF) kernel[5]. Typically, it is defined as

$$k^{\text{RBF}}(x, x') \equiv k^{\text{RBF}}(r) = C^2 \cdot \exp\left(-\frac{r^2}{2l^2}\right), \quad l \in \mathbb{R}\ , \tag{2.52}$$

where the output scale C^2 is commonly referred to as a *constant kernel*. The RBF kernel is infinitely differentiable, stationary, and not periodic. It produces very smooth GPs. It has been argued by some authors that this makes the kernel unsuitable for applications where real-world data is involved [36].

The Matérn is a generalization of the RBF kernel. It introduces an additional parameter ν which controls the differentiability [9]:

$$k_\nu^{\text{Matern}}(r) = C^2 \cdot \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}r}{l}\right)^\nu \cdot K_\nu\left(\frac{\sqrt{2\nu}r}{l}\right), \qquad l \in \mathbb{R}^+\ . \tag{2.53}$$

where Γ is the gamma function and K_ν the modified Bessel function of the second kind. The kernel is k times differentiable if $\nu > k$ [36]. For $\nu \to \infty$, the Matérn kernel approaches the RBF kernel. Typically, the Matérn kernel is used with $\nu = 3/2$ or $\nu = 5/2$, which correspond to once and twice differentiable functions, respectively. In this case, the kernel simplifies to

$$k_{\nu=3/2}^{\text{Matern}}(r) = C^2\left(1 + \frac{\sqrt{3}r}{l}\right)\exp\left(-\frac{\sqrt{3}r}{l}\right)\ , \tag{2.54}$$

$$k_{\nu=5/2}^{\text{Matern}}(r) = C^2\left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right)\exp\left(-\frac{\sqrt{5}r}{l}\right) \tag{2.55}$$

The exponential sine squared (ESS) kernel[6] is an example of a stationary, periodic, and infinitely differentiable kernel [39]:

$$k^{\text{ESS}}(r) = C^2 \exp\left(\frac{2\sin^2\left(\frac{\pi r}{p}\right)}{l^2}\right)\ , \tag{2.56}$$

where p controls the periodicity.

[5] This kernel has numerous different names, among which are *Squared Exponential* kernel and *Gaussian* kernel. We stick to RBF.

[6] This kernel is often simply referred to as *the periodic kernel* [39, 36, 40]. This is somewhat misleading, as there are many ways to construct a periodic kernel.

Examples of GPs conditioned with these kernels are shown in Fig. 2.2. The GPs are conditioned on the same data but with different kernels. The RBF kernel produces a very smooth GP, the Matérn kernels are less smooth, and the ESS kernel is periodic. It is clear that the choice of kernel drastically changes the interpretation of the data.

Kernels can be combined to create more complex kernels, allowing for greater flexibility in modeling diverse data patterns. For example, the sum and product of two valid kernels are also valid kernels [39]. This allows for greater flexibility in modeling diverse data patterns.

Extending the kernel (and hence the GP) to higher dimensions is straightforward and done by promoting the kernel to a function of $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, where d is the dimensionality of the data. For this we introduce an additional dimension such that $\mathbf{X}_{ik} = (\boldsymbol{x}_i)_k$ with $k = 1, \ldots, d$ [7]. For stationary kernels such as the RBF kernel, this can be achieved by either using the same length scale for all dimensions (promoting the scalar distance r to the Euclidean distance $\boldsymbol{r} = |\boldsymbol{x} - \boldsymbol{x}'|$) or by using a different length scale for each dimension. The latter is called an *anisotropic* kernel, which offers greater flexibility at the cost of additional hyperparameters. The anisotropic RBF kernel is defined as:

$$\begin{aligned} k^{\mathrm{RBF},d}(\mathbf{X}_i, \mathbf{X}'_i) &= C^2 \exp\left(-\sum_{i=1}^{N} \frac{(\mathbf{X}_{ik} - \mathbf{X}'_{ik})^2}{2l_k^2}\right) \\ &= \frac{1}{C^{2d-2}} \prod_{i=1}^{d} k^{\mathrm{RBF}}(\mathbf{X}_{ik}, \mathbf{X}'_{ik}) \ , \end{aligned} \tag{2.57}$$

where $\boldsymbol{l}$ is a vector of length scales for each dimension. The anisotropic Matérn and ESS kernels are defined analogously.

In Paper I, Paper II and Paper III we use anisotropic kernels.

Choosing the kernel's hyperparameters

The kernel typically has one or more free hyperparameters, henceforth denoted as $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_{n_\lambda}\}$. These can be determined in a Bayesian way

[7] It is worth taking a moment here to discuss what it means to promote x to a vector, as there are three different dimensionalities appearing in our equations now: *(i)* the dimensionality of the data d, which refers to the number of dimensions that each training and test point has (e.g., $d = 2$ if the goal is to approximate a function $f(x, y)$), *(ii)* the dimensionality of the index set n, which is the number of observations in the data set (sticking to our example from before, let's say we have the observations $f(0, 1), f(1, 2), f(2, 3), f(2, 2)$, then $n = 4$) (iii) the number of the kernel's hyperparameters n_λ, which can be seen as living in a n_λ-dimensional space (if we want to map our function $f(x, y)$ with the RBF kernel from Eq. (2.57), n_λ would be 3). For Eqs. (2.46) and (2.50), this means that $\boldsymbol{f}(\boldsymbol{x})$ and $\boldsymbol{m}(\boldsymbol{x})$ go from n-vectors to $n \times d$ tensors. To make this change clear, we henceforth call the training set $\mathbf{X}$ and index it with i, k.

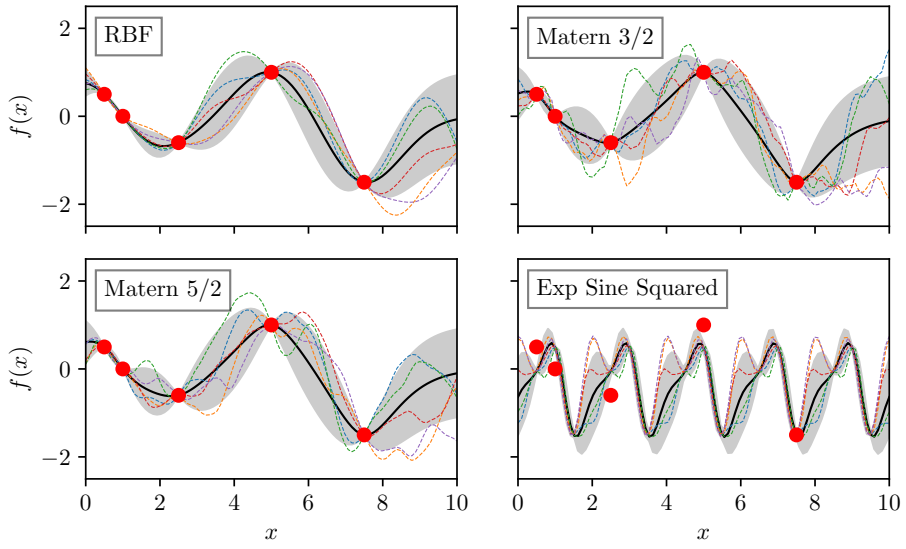


Figure 2.2: GPs with four different kernels: RBF (top left), Matérn $\nu = 3/2$ (top right), Matérn $\nu = 5/2$ (bottom left), and ESS (bottom right). The kernels are defined in Eqs. (2.52) and (2.54) to (2.56). The GPs are conditioned on the data shown as red dots. The solid black line is the GPs conditioned mean, and the standard deviation is depicted as the gray shaded region. The Matérn kernels are less smooth than the RBF kernel, and the ESS kernel is periodic.

through the likelihood of the data given the GP [36]

$$\mathcal{L}^{\mathcal{GP}}(\boldsymbol{y}|\mathbf{X},\boldsymbol{\lambda}) = \int p(\boldsymbol{y}|\boldsymbol{f},\mathbf{X},\boldsymbol{\lambda})p(\boldsymbol{f}|\mathbf{X},\boldsymbol{\lambda})\mathrm{d}\boldsymbol{f} \tag{2.58}$$

$$= -\frac{1}{2}\boldsymbol{y}^T(K+\sigma_n^2 I)^{-1}\boldsymbol{y} - \frac{1}{2}\log|\mathbf{K}+\sigma_n^2\mathbb{I}| - \frac{n}{2}\log 2\pi \ , \tag{2.59}$$

where $\boldsymbol{y}$ is the vector of training data and $\mathbf{K}$ is the Gram matrix of the training set $\mathbf{X}$. Unfortunately, evaluating this likelihood involves the computationally expensive matrix inversion of the Gram matrix. This makes full Bayesian inference of λ infeasible in practice. Instead, one typically finds the maximum likelihood estimate, otherwise known as *Maximum a Posteriori* (MAP) or *Maximum Likelihood type II* estimate of $\boldsymbol{\lambda}$ using a gradient-based optimization algorithm. An alternative technique, which performs approximate inference of the GP likelihood, has been proposed in [41].

Putting all the ingredients together, one arrives at the full GP regression algorithm. GP regression (sometimes also referred to as *Kriging*) involves fitting a GP with kernel $k(\boldsymbol{x},\boldsymbol{x}'|\lambda)$ to a training set $\mathbf{X},\boldsymbol{y}$ with associated noise σ_n. This is done by finding the MAP estimate of $\boldsymbol{\lambda}$ by maximizing

the marginal GP log-likelihood in Eq. (2.58). This approach is convenient because the only choice left to the user is that of a suitable kernel function. Furthermore, and in contrast to most NNs, the hyperparameters of the GP typically have an intuitive interpretation.

The algorithm consists of two simple steps:

1. In the *training* step, the hyperparameters of the model are optimized by maximizing Eq. (2.58). Additionally, $(K+\sigma_n^2 I)^{-1}$ is precomputed for use in the prediction step.
2. In the *prediction* step, the GP is conditioned according to Eqs. (2.50) and (2.51).

The algorithm is divided into these two steps to highlight the distinction between the computationally expensive $\mathcal{O}(N^3)$ training phase and the relatively cheap $\mathcal{O}(N^2)$ prediction phase, where N is the number of points in the training set.

2.4.3 Active Learning

Active Learning (see e.g., [42] for a review) is an umbrella term for methods that aim to reduce the number of training points needed for a model to achieve a desired performance. In machine learning lingo, the idea is to let the model decide which data points to label next, rather than providing a predetermined dataset. In physics terms, assume that the goal is to learn a mapping $f(\boldsymbol{x})$ from some input space $\mathcal{X}$ to some output space $\mathcal{Y}$. In passive learning, the model is fed a fixed set of training points $\mathbf{X}, \boldsymbol{y}$ where $\boldsymbol{y} = f(\mathbf{X})$ and trained on this data. In Active Learning, the model is allowed to (typically iteratively) modify the current set of training points to optimize the performance of the model. This is particularly useful when evaluating $f(\boldsymbol{x})$ is computationally expensive or when training on large datasets becomes slow. Active Learning can be done in a discrete space, e.g., when there is only a finite dataset to choose from, or in a continuous space, e.g., if $f(\boldsymbol{x})$ can be evaluated at any point. The case that is of interest for us (Bayesian inference of continuous parameters) is the latter.

Typically, the task of Active Learning strategies is to find the smallest set possible that still allows the model to reach a certain performance. This can be done in conjunction with the task of exploring the sampling space as efficiently as possible.

To illustrate the concept, consider MH-MCMC [8](defined in Section 2.2.1).

[8]MH-MCMC can be considered an Active Learning strategy, though not a good

A simple way to extend HM-MCMC to an Active Learning strategy would be to modify the proposal distribution as the sampling progresses as done in e.g., [43, 44] by using a covariance matrix that is estimated from the current state of the chain.

Active Learning typically relies on some measure to quantify the informativeness of a data point given the current state of the model. This measure can be optimized to find new samples to query, and its choice depends on what the model is supposed to achieve. In the context of GPs, this measure is called the *acquisition function* and is typically a function of the posterior GP. In the context of GP regression for surrogate inference of probability densities, two concepts are of particular interest that we borrow heavily from: *Bayesian optimization* and *Bayesian quadrature*.

Bayesian optimization

Bayesian optimization (BO) is a method for optimizing black-box functions that are expensive to evaluate. It uses the GP as a surrogate model and iteratively optimizes an acquisition function to determine the next point to evaluate. An example of an acquisition function is the *Expected Improvement* (EI), which is defined as

$$\mathrm{EI}(\boldsymbol{x}) = \mathbb{E}\left[\max(0, f(\boldsymbol{x}) - f(\boldsymbol{x}^+))\right] \ , \tag{2.60}$$

where $\boldsymbol{x}^+$ is the point with the highest value of f found so far. The goal is to identify the point expected to provide the greatest improvement over the current best point, given the model. For GPs, the EI is analytical and given by [45]

$$\mathrm{EI}(\boldsymbol{x}) = (f(\boldsymbol{x}) - \mu(\boldsymbol{x}^+))\Phi(z) + \sigma(\boldsymbol{x})\varphi(z) \ , \tag{2.61}$$

where $z = (f(\boldsymbol{x}) - \mu(\boldsymbol{x}^+))/\sigma(\boldsymbol{x})$ and $\Phi = 1/2\left(1 + \mathrm{erf}(z/\sqrt{2})\right)$ and $\varphi = 1/\sqrt{2\pi}\exp(-z^2/2)$ are the standard normal distribution CDF and PDF, respectively. The point that maximizes the EI is evaluated next. This process is repeated until a stopping criterion is reached.

From the structure of Eq. (2.61), one can see two competing terms: one term is proportional to the difference between the current best point and the point to be evaluated and encourages sampling near the current best point (exploitation). The other term is proportional to the uncertainty of the GP and encourages sampling in regions of high uncertainty (exploration). This is a common feature in Active Learning strategies and is known as the *exploration-exploitation trade-off*.

one. Its next move depends only on the last, and the query step must occur before rejection, offering no computational savings.

The EI is just one of many possible acquisition functions, with alternatives like the *Probability of Improvement* (PI) [45] and *Upper Confidence Bound* (UCB) [46] offering different balances between exploration and exploitation.

Bayesian quadrature

While BO focuses on optimizing a target function, Bayesian quadrature (BQ) applies similar principles to estimate the integral of a function [41]. Let

$$I = \int_{x_0}^{x_1} f(x)\mathrm{d}x \tag{2.62}$$

be the integral of some arbitrary function $f(\boldsymbol{x})$ over possibly multiple dimensions such that $\boldsymbol{x} \in \mathbb{R}$. We restrict ourselves to the one-dimensional case for simplicity, but the generalization to $\boldsymbol{x} \in \mathbb{R}^d$ is straightforward. The idea is to use the GP as a surrogate model for the integrand. The expectation value of the integral is then

$$\mathbb{E}[I|D] = \int_{x_0}^{x_1} \mu_{f|D}(x)\mathrm{d}x \ , \tag{2.63}$$

and the variance

$$\mathrm{var}[I|D] = \int_{x_0}^{x_1}\int_{x_0}^{x_1} \mathrm{cov}_{f|D}(x, x')\mathrm{d}x\mathrm{d}x' \ , \tag{2.64}$$

where $\mu_{f|D}$ and $\mathrm{cov}_{f|D}$ are the mean and covariance of the posterior GP (conditioned on the training data D), respectively. This not only provides an estimate of the integral but also an estimate of its uncertainty. With this, it is possible to construct an acquisition function that is optimized to find the next point to evaluate.

The disadvantage of this approach is that the d and $2d$ integrals in Eqs. (2.63) and (2.64) still have to be computed numerically. How this can be done in an efficient way to make BQ feasible for Bayesian inference is the main topic of Paper I and Paper II.

An illustration of BQ is shown in Fig. 2.3, showing how the GP formulation induces a normal distribution over the value of the integral.

2.5 The role of machine learning in Bayesian inference

In the previous sections, we have introduced "traditional" Monte Carlo methods, approximate inference methods, and Bayesian machine learning methods. These methods are frequently combined to address complex and expensive inference problems. This section is dedicated to discussing the

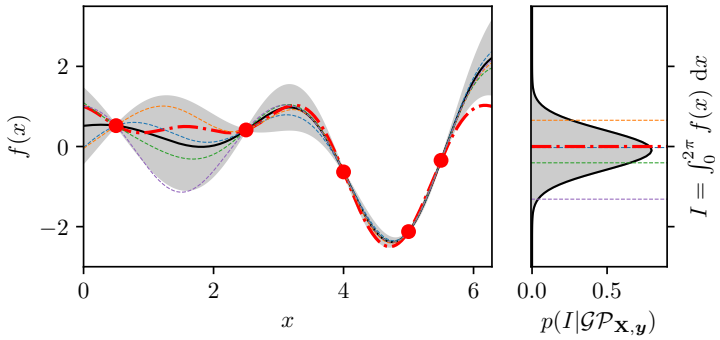


Figure 2.3: Illustration of the BQ procedure. **Left:** GP fit to the function $f(x) = \sin(x) + \cos(2x) - 1/2\sin(3x)$ (red) for. The GP has been conditioned on 5 training points (red dots). $\mu(x)$ and $\sigma(x)$ are shown as the black line and gray band, respectively. The dashed lines show four sample functions drawn from $\mathcal{GP}(x|X, y)$. **Right:** Normal distribution induced by BQ for the integral $I = \int_0^{2\pi} f(x)\mathrm{d}x = 0$ (gray distribution). The true value is shown in red. The four dashed lines correspond to the integrals of the sample functions.

role of machine learning in Bayesian inference and to relating the different algorithms to each other. We also give an overview of the current state of research. The specific application to LISA and gravitational wave science is discussed in Chapter 3.

2.5.1 Speeding up aspects of the likelihood

A straightforward way to use machine learning in Bayesian inference–without modifying the inference algorithm–is by speeding up the likelihood computation itself. This can be achieved by *(i)* speeding up parts or the entirety of the forward model (simulator), see e.g., [47, 48, 49, 50], *(ii)* accelerating the likelihood computation itself (i.e., the comparison of model to data), or *(iii)* mapping both the simulator and data to a lower dimensional space in which the forward model and likelihood can be computed cheaply (see e.g., [51]).

This approach has seen extensive use in a variety of problems. Its main advantage is that it directly addresses the slowest parts of the computation, potentially incorporating physics knowledge (e.g., symmetries) into the model. The main disadvantage is that the model must be trained over a large parameter space to be effective and that the model is typically limited to the specific problem on which it was trained. Additionally, the model can introduce biases into the inference process if it is not trained carefully.

2.5.2 Enhancing traditional Monte Carlo methods

One of the simplest and most robust uses of machine learning in Bayesian inference is to modify and enhance traditional sampling methods. This can be done in a variety of ways, one approach being the use of a Normalizing Flow as a proposal distribution. The idea is to learn a representation of the posterior distribution as the sampling progresses, using this representation (or typically some enlarged distribution to avoid bias) as a proposal distribution. This has been done for different flavors of MCMC, see e.g., [52, 53, 54] and Nested sampling [55, 56, 57]. The goal is to improve the acceptance rate of the proposed samples as the learned proposal distribution approaches the target distribution. This approach is similar to providing simple information, such as the mean and covariance of the current samples, to the proposal distribution but offers greater flexibility. Furthermore, the hope for MCMC is that this approach mitigates the risk of missing modes if the posterior is multimodal with modes spaced far apart.

The key advantage of this approach is that it retains some of the robustness of traditional methods (such as the detailed balance of MCMC) while offering the flexibility of machine learning methods. The main disadvantage is that training the Normalizing Flow requires significant computational power. Additionally, it typically cannot drastically reduce the number of samples needed, as the samples are still drawn from the posterior distribution and, like all machine learning methods, it risks overfitting.

2.5.3 Simulation-based inference with machine learning

SBI, even more than MCMC and Nested sampling, naturally lends itself to accelerating the inference process. There are typically two ways in which this is achieved: by learning a lower-dimensional representation of the data or by learning the mapping from the prior to either the likelihood, likelihood ratio, or the posterior.

The first approach is to use an NN to learn a lower-dimensional representation of the data. This can be done in a supervised way by training the network on the data. The network is then used to project the data to a lower-dimensional space, speeding up the comparison step. Learning an optimal representation can typically be done with an autoencoder [58].

The second approach, often called *amortized*, refers to the idea that the expensive training of the simulator is offset by the low cost of inference at runtime. It aims to learn the mapping from the prior to the likelihood, likelihood ratio, or posterior [26]. They differ in which quantity is

learned:

- *Neural Posterior Estimation* (NPE) estimates the posterior density $p(\theta|D)$ directly by learning the mapping from the prior to the posterior, typically using some density estimator like a Normalizing Flow [59, 60, 61, 62].
- *Neural Likelihood Estimation* (NLE) estimates the likelihood $L(D|\theta)$ by learning a proxy of the likelihood function from data of the forward model [63, 64, 61]. This can be used in traditional samplers such as MCMC as a fast substitute for the likelihood.
- *Neural Ratio Estimation* (NRE) estimates the likelihood-to-evidence ratio $L(D|\theta)/Z(D)$ [65, 66, 67, 68, 69, 70, 71, 72, 73]. The advantage of this is that it can be used with different priors, enabling flexibility in Bayesian model comparison without retraining the model.

2.5.4 Machine learning for Variational inference

A machine learning-powered emulator like an NN, Normalizing Flow, or GP can be viewed as a variational model itself. This is because it serves as an approximate surrogate of a true function, making it natural to use in the context of VI. To this extend the machine learning model can be used at several "depths" again: either by emulating the likelihood or the posterior or by using it to get proposals for the variational distribution. A natural way of obtaining these points is Active Learning, although this is not the only way.

Like for SBI, a straightforward choice for the machine learning model is a Normalizing Flow, as it naturally maps a probability distribution [23, 22, 24] but also simple deep NNs are an option [74]. GPs can be powerful models too, as they have fewer tunable hyperparameters and are more interpretable [75, 25, 76, 77, 41, 78]. In this work, we focus on the use of GPs for generating a surrogate model of the posterior distribution with Active Learning.

Our approach stretches the definition of VI, as we do not learn a simpler representation of the data but instead build an emulator of the posterior. Nevertheless, the approach presented in Paper I, Paper II, and Paper III can broadly be seen as a VI approach.

3 Inference in the LISA mission

As two of the papers that have been produced in the context of this work are related to different aspects of inferring astrophysical and cosmological signals within the commissioned Laser Interferometer Space Antenna (LISA) mission, this chapter provides a brief overview of the mission itself, the sources that are expected to be observed, and the challenges that the data that the LISA mission will provide pose to correctly identifying and characterizing these sources.

The chapter is structured as follows: Section 3.1 provides a brief overview of the LISA mission and its objective as it stands today. Section 3.2 discusses the different types of sources that are expected (or theorized) to be observed by LISA. Furthermore, the role of the instrumental noise is discussed briefly. Section 3.3 discusses the challenges that the LISA data pose to correctly identifying and characterizing these sources in the overlapping signal-dominated datastream and how these challenges have been addressed so far.

3.1 The LISA mission

The LISA mission is a planned space-based gravitational wave observatory that is currently in the early implementation phase. The mission is a collaboration between the European Space Agency (ESA) and the National Aeronautics and Space Administration (NASA) and is expected to be launched in the mid-2030s [79, 80]. It is designed to detect gravitational waves in the frequency range of $\sim 10^{-4}$ to $\sim 1\,\mathrm{Hz}$, which is a frequency range that is yet unexplored. Compared to other gravitational wave detection experiments, LISA sits between the ground-based detectors like LIGO, Virgo, and Kagra (LVK), which are sensitive to higher frequency gravitational waves in the $\sim 100\,\mathrm{Hz}$ range [81, 82, 83], and pulsar timing arrays (PTAs), which are sensitive to lower frequency gravitational waves ($\sim 10^{-9}\,\mathrm{Hz}$) [84].

LISA will consist of three spacecraft in a triangular configuration, with each spacecraft separated by $\approx 2.5 \cdot 10^6\,\mathrm{km}$ in a heliocentric orbit, trailing the Earth by 20°. The constellation acts as a Michelson interferometer, where each spacecraft contains a free-falling test mass that is shielded from external forces. The test masses are monitored by laser interferometry, and the phase difference between the laser beams is used to infer the presence of gravitational waves.

Currently, the mission is expected to last a minimum of 4.5 years, with a

possible extension to 10 years. The mission is expected to detect a wide range of sources like double white dwarf systems, extreme mass ratio inspirals, and black hole mergers. The mission is also expected to detect an astrophysical stochastic background of gravitational waves, which is a superposition of many unresolved sources, and possibly a cosmological background of gravitational waves originating from the early Universe.

Part of the mission objective is not only to provide the raw data to the scientific community but also to provide a single catalog of gravitational wave source candidates that can be used by the community to perform follow-up studies. Furthermore, at the same time as the mission, low-latency data alerts will be generated from a faster pipeline that will be used to alert the community of potential gravitational wave events that can be used to search for electromagnetic counterparts [80].

As of 2025, the mission has been adopted by ESA and is currently in the early implementation phase. The current projected launch date is in 2035.

3.2 Source types and noise in LISA

The currently operational set of Michelson interferometers (LVK) are mainly sensitive to binary black hole, binary neutron star, and binary neutron star-black hole mergers and are noise-dominated experiments [81]. Conversely, LISA is expected to be a signal-dominated experiment containing a wide zoology of overlapping signals coming from different sources at once. The sources that are expected to be observed by LISA can be broadly divided into two categories: astrophysical sources and cosmological sources.

Astrophysical sources are sure to be observed by LISA, with some sources expected to be numbering in the thousands [80]. Some of these sources are expected to be observed as individual signals in the data, and the parameters of these sources can be inferred with high precision. Cosmological sources, on the other hand, may or may not be observed by LISA as there is currently no complementary experiment that could confirm the presence of such a signal. Furthermore, ΛCDM slow roll inflationary cosmology does not predict such a signal. Nevertheless, there are a wide variety of signals that could be generated by both standard model physics and beyond standard model physics during different periods of the Universe's history. These signals manifest themselves as a stochastic background of gravitational waves in the data. The amplitude of the cosmological background is highly model-dependent, ranging from a barely detectable signal to one that dominates the LISA band [85].

3.2.1 Astrophysical sources

In the context of LISA, astrophysical sources are localized emitters of gravitational waves, generated by compact objects. These sources can broadly be divided into two categories: resolved sources and unresolved sources. Resolved sources are sources that are expected to be observed as individual signals in the data, while unresolved sources contribute to a stochastic background of GWs but cannot be distinguished as individual signals.

Compact binary systems produce GWs through quadrupole radiation, which is emitted as the system loses energy and angular momentum. To understand this effect, it is instructive to linearize the Einstein equation [86]:

$$G_{\mu\nu} = 8\pi G T_{\mu\nu} \ , \tag{3.1}$$

where $G_{\mu\nu}$ is the Einstein tensor, G is the gravitational constant, and $T_{\mu\nu}$ is the energy-momentum tensor. We are using natural units where $c = 1$. Assuming that the metric is nearly Minkowski, $g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$ with $\eta_{\mu\nu} = \text{diag}(-1, 1, 1, 1)$ and a small perturbation $h_{\mu\nu}$, and momentarily introducing the trace-reversed perturbation $\bar{h}_{\mu\nu} = h_{\mu\nu} - \frac{1}{2}\eta_{\mu\nu}h$, the linearized Einstein equation can be written as

$$G_{\alpha\beta} = -\frac{1}{2}\left[\bar{h}_{\alpha\beta,\mu}{}^{,\mu} + \eta_{\alpha\beta}\bar{h}_{\mu\nu}{}^{,\mu\nu} - \bar{h}_{\alpha\mu,\beta}{}^{,\mu} - \bar{h}_{\beta\mu,\alpha}{}^{,\mu}\right] = 8\pi G T_{\alpha\beta} \ , \tag{3.2}$$

where $()_{,\mu} = \partial_\mu()$. In the Lorentz gauge ($\bar{h}_{\mu\nu}{}^{,\mu} = 0$), this simplifies to:

$$\Box\bar{h}_{\alpha\beta} = -16\pi G T_{\alpha\beta} \ , \tag{3.3}$$

where $\Box = \eta^{\mu\nu}\partial_\mu\partial_\nu$ is the d'Alembert operator. We see that away from the source, where $T_{\alpha\beta} = 0$, the solution to this is a plane wave, $\bar{h}_{\alpha\beta} = A_{\alpha\beta}e^{ik_\mu x^\mu}$, where $k_\mu k^\mu = 0$. This is the gravitational wave.

Typically, one introduces the transverse-traceless (TT) gauge, where $\bar{h} = 0$, which implies $\bar{h}_{\mu\nu} = h_{\mu\nu}$. The gauge is chosen such that

$$h^{0,\mu} = 0 \ , \qquad h^i_i = 0 \ , \qquad h^{ij}_{,j} = 0 \ , \tag{3.4}$$

meaning that the information about the GWs is contained in the spatial part of the perturbation. The GWs are described by the two polarizations h_+ and $h_\times$, which are the two independent solutions to the wave equation in the TT gauge. The GWs are then described by

$$h_{ij} = h_+\mathbf{e}_{ij,+} + h_\times\mathbf{e}_{ij,\times} \ , \tag{3.5}$$

where $\mathbf{e}_{ij,+}$ and $\mathbf{e}_{ij,\times}$ are the two polarization tensors.

For compact binary systems, computing the emission involves solving the Einstein Equation (3.1) for the energy-momentum tensor of the system. The energy-momentum tensor for a binary system of point masses can be written as [86]

$$T_{\mu\nu}(\boldsymbol{x},t) = \sum_{a=1}^{2} m_a \int \mathrm{d}s_a u_{a\mu} u_{a\nu} \delta^{(4)}(x - x_a(s_a)) \ , \tag{3.6}$$

where m_a is the mass of the a-th object, $u_{a\mu}$ is the four-velocity of the a-th object, and $x_a(s_a)$ is its worldline. If we introduce $\gamma = 1/\sqrt{1-v^2}$, where v is the velocity of the object, we can write the four-velocity as $u_{a\mu} = \gamma_a(1, \boldsymbol{v}_a)$. The energy-momentum tensor can then be written as

$$T_{\mu\nu}(\boldsymbol{x},t) = \sum_{a=1}^{2} m_a \gamma_a v_{a\mu} v_{a\nu} \delta^{(3)}(\boldsymbol{x} - \boldsymbol{x}_a(t)) \ , \tag{3.7}$$

where we have used that the objects are point masses and that the worldlines are parametrized by the time t.

Solving the Einstein Eq. (3.1) at the source with the matter tensor for the binary defined above is a hard task because it involves the full non-linear regime and is typically done using numerical relativity. This is particularly true close to the merger, where the objects are moving at relativistic speeds. However, in the weak field limit, where the objects are moving at non-relativistic speeds, the Einstein equation can be solved perturbatively. This is typically done using the post-Newtonian (PN) expansion, where the metric is expanded in powers of v/c, although other formalisms, like the effective-one-body formalism, also exist [86].

To compute the waveform at large distances, it is customary to use the multipolar post-Minkowskian formalism [87], which provides a systematic expansion of the metric in terms of multipole moments of the source. In this framework, the leading-order (quadrupolar) contribution to the spatial components of the metric perturbation in the harmonic gauge is given by

$$h_{ij}(t,\mathbf{x}) = \frac{2G}{c^4 r} \ddot{Q}_{ij}^{\mathrm{TT}}(t - r/c) \ , \tag{3.8}$$

where Q_{ij}^{TT} is the transverse-traceless part of the mass quadrupole moment of the source, and $r \equiv |\mathbf{x}|$ is the distance from the source.

The far-field approximation assumes that the observer is located at a much greater distance r than the characteristic wavelength λ of the GWs and the size R of the source, i.e., $r \gg \lambda \gg R$. Under this assumption, the waveform depends on the retarded time $t_{\mathrm{ret}} = t - r/c$, and spatial dependence

enters only through an overall $1/r$ decay and the angular dependence of the multipole moments.

Typically, one expresses the equations that arise in terms of the *chirp mass* $\mathcal{M}_c = (m_1 m_2)^{3/5}/(m_1 + m_2)^{1/5}$, the symmetric mass ratio $\eta = m_1 m_2/(m_1 + m_2)^2$, where m_1 and m_2 are the individual masses of the two objects in the binary, and the luminosity distance d_L from source to observer.

For binary systems in quasi-circular orbits, the polarization amplitudes depend on the inclination angle ι between the orbital angular momentum and the line of sight:

$$h_+ \propto (1 + \cos^2 \iota) \cos(2\phi) \ , \tag{3.9}$$

$$h_\times \propto 2 \cos \iota \sin(2\phi) \ , \tag{3.10}$$

where ϕ is the orbital phase.

A GW detector measures a linear combination of h_+ and $h_\times$, weighted by antenna pattern functions F_+ and $F_\times$ that depend on the detector orientation and source sky location:

$$h(t) = F_+(\theta, \varphi, \psi) h_+(t) + F_\times(\theta, \varphi, \psi) h_\times(t) \ , \tag{3.11}$$

where ψ is the polarization angle and (θ, φ) specify the source location in the sky.

For sources at cosmological distances, the propagation of GWs must be treated in a Friedmann-Lemaître-Robertson-Walker (FLRW) background (see Section 4.1.2 for an in-depth discussion). In this case, the gravitational wave equation is modified to include the expansion of the Universe. The tensor perturbation h_{ij} satisfies:

$$\ddot{h}_{ij} + 3H\dot{h}_{ij} + \frac{k^2}{a^2} h_{ij} = 0 \ , \tag{3.12}$$

where $a(t)$ is the scale factor and $H = \dot{a}/a$ is the Hubble parameter. In the subhorizon regime ($k/a \gg H$), the solution behaves approximately as a redshifted plane wave:

$$h_{ij}(t) \propto \frac{1}{a(t)} \cos\left(k \int \frac{dt'}{a(t')} \right) \ . \tag{3.13}$$

As a result, both the GW amplitude and GW frequency redshift $\propto 1/a(t)$, analogously to electromagnetic waves. In practice, observed waveforms from distant sources must be rescaled by the redshift z as:

$$f_{\rm obs} = \frac{f_{\rm em}}{1+z} \ , \quad h_{\rm obs} = \frac{h_{\rm em}}{1+z} \ . \tag{3.14}$$

Paper III, produced in the context of this work, focuses on inferring the parameters of three types of astrophysical sources: double white dwarf (DWD) systems, stellar mass black hole binaries (stBHBs), and supermassive black hole binaries (SMBHBs). Paper IV assumes that all resolvable sources have been subtracted from the data, with only the unresolved sources remaining as a stochastic background. In the context of a cosmological background, we call this the *astrophysical foreground*, which can be divided into two contributions: the *galactic foreground*, consisting of unresolved DWDs in the Milky Way, and the *extragalactic foreground*, consisting of unresolved stBHBs and SMBHBs and other extragalactic stellar-mass sources, such as binary neutron stars and neutron-star-black-hole binaries.

Some SMBHBs are projected to merge within the LISA band and therefore require a full numerical relativity treatment. On the other hand, DWDs within LISA's sensitivity are very far from merger and therefore can be treated at low order in post-Newtonian (PN) theory. stBHBs are expected to merge outside the LISA band and can also be treated using post-Newtonian approximations.

In the following, we briefly explain DWDs, stBHBs, and SMBHBs focusing on the main properties that are relevant for the inference of these sources in the context of LISA and only mention other astrophysical sources that may contribute to the LISA signal.

Double white dwarf systems

Double white dwarf (DWD) systems are expected to be the most common source of GWs within the LISA band, with expected numbers for resolvable systems being as high as 10^4 individual sources [80]. They form the bulk of the binary systems within the Milky Way (galactic binaries). DWD systems are binary systems consisting of two orbiting white dwarfs. White dwarfs are remnants of low-to-intermediate mass stars, weighing between 0.17 and 1.33 $\mathrm{M}_\odot$ [88, 89], and are supported against gravitational collapse by electron degeneracy pressure. DWDs within the LISA band are in the early phases of inspiral, thus slowly losing energy and angular momentum to gravitational waves. The signal from these systems is expected to be nearly monochromatic, with the signal frequency only increasing marginally over the mission duration.

This makes generating waveforms for DWDs computationally cheap as the evolution can be approximated very well by a simple sinusoidal with a slowly drifting frequency $\dot{f}$ that is projected to be in the range of 10^{-4} Hz to 10^{-3} Hz [90]. This reduces the parameter space for these sources to 8:

$f, \dot{f}, A, \phi, \iota, \psi, \theta, \phi$ where f is the frequency, $\dot{f}$ is the frequency derivative, A is the amplitude, ϕ is the phase, ι is the inclination angle, ψ is the polarization angle, and (θ, ϕ) are the sky location angles. As LISA follows a heliocentric orbit, the source location is typically expressed in ecliptic coordinates through the ecliptic longitude λ and latitude β.

Due to these properties, DWDs are typically searched for and analyzed in narrow frequency windows. A challenge that arises from this is that multiple sources that are close in frequency lead to so-called *confusion*, where the signals from the sources overlap and are hard to distinguish.

Lastly, some of the DWD systems in our galactic neighborhood are known from electromagnetic observations. These systems are called *verification binaries* and will be used to calibrate LISA.

Stellar mass black hole binaries

Stellar mass black hole binaries (stBHBs) are binary systems consisting of two black holes with masses in the range of $\sim 5\,\mathrm{M}_\odot$ to $\sim 100\,\mathrm{M}_\odot$. stBHBs are expected to merge outside the LISA band, and therefore the signal from these systems is expected to be a chirp signal with a frequency that increases over time, eventually leaving the LISA band and merging within the LVK band later [80]. The signal from these systems is expected to be more complex than the signal from DWDs as its frequency increases significantly over time. This makes generating waveforms for stBHBs computationally more expensive than for DWDs as they exhibit a more complicated evolution, which increases the number of parameters needed to describe the signal. Assuming quasi-circular orbits (neglecting eccentricity), neglecting environmental effects, and assuming that the spins of the black holes are aligned with the total angular momentum of the system, the full parameter space describing a BHB system is 11 dimensional, containing four intrinsic parameters: the masses m_1, m_2, which are typically expressed through the chirp mass and symmetric mass ratio (or reduced mass ratio $\delta\mu = (m_1 + m_2)/(m_1 + m_2)$), and two spins χ_1, χ_2, and the 7 extrinsic parameters: the luminosity distance d_L, the inclination angle ι, the polarization angle ψ, the position (λ, β), the initial phase ϕ_0, and the time to coalescence t_c.

Numerous tools for computing waveforms for stBHBs have been developed and optimized over the last decade of LVK observations (see e.g., [91, 92, 93])

Supermassive black hole binaries

Supermassive black hole binaries (SMBHBs) are binary systems consisting of two supermassive black holes with masses in the range of $\sim 10^5\,\mathrm{M}_\odot$ to $\sim 10^9\,\mathrm{M}_\odot$ [94, 95]. SMBHBs in the mass range between $\sim 10^5\,M_\odot$ and $\sim 10^7\,M_\odot$ (see Fig. 3.1) are expected to merge within the LISA band and therefore require a full numerical relativity treatment. The signal from these systems is expected to be a chirp signal, potentially merging within the LISA mission duration. If observed, the signal from these systems is expected to be the loudest in the LISA band, reaching a signal-to-noise ratio of $\sim 10^3$ for the loudest systems [96, 97]. Generating waveforms for SMBHBs is computationally expensive, as modeling the merger and ringdown is a computationally expensive task, and due to their high signal-to-noise ratio (SNR), inferring the source parameters of SMBHBs requires accurate waveforms [98]. As they furthermore evolve over a large range of frequencies, the waveform generation requires a large number of frequency bins. Under the same assumptions as for the stBHBs, the number of parameters describing a SMBHB system is also 11.

Other sources

Other types of sources that are expected to be observed by LISA include extreme mass ratio inspirals (EMRIs), which are systems consisting of a stellar mass compact object orbiting a supermassive black hole, and white dwarf-neutron star binaries. Furthermore, there is the possibility of detecting bursts from cosmic strings, signals from supernovae, and potentially exotic sources that have not yet been modeled [80]. Exploring these sources and their properties would be beyond the scope of this work.

Instead of expressing the signal amplitude as the strain amplitude h as defined in Eq. (3.11), one typically defines the characteristic strain amplitude h_c as

$$h_c(f)^2 = 4f^2 \left|\tilde{h}(f)\right|^2 \ , \tag{3.15}$$

where $\tilde{h}(f)$ is the Fourier transform of the strain amplitude $h(t)$. For nearly monochromatic sources like DWDs, the characteristic strain is given by [99]

$$h_c(f) = \sqrt{\frac{2f^2}{\dot{f}}} h_0 \ , \tag{3.16}$$

where h_0 is the root-mean-square amplitude of the source. Figure 3.1 shows a simulated population of sources that could be observed by LISA, containing the astrophysical source types that have been presented above.

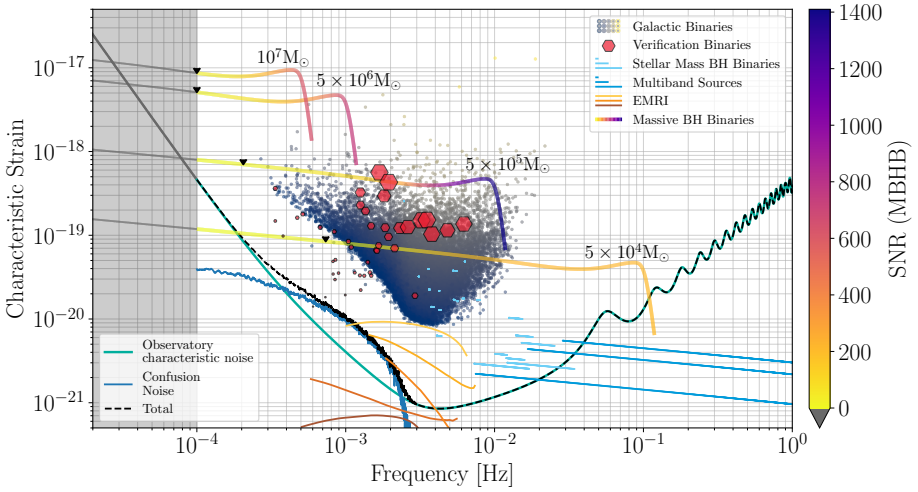


Figure 3.1: simulated population of sources that could be observed by LISA containing galactic binaries (with verification binaries highlighted), stBHBs, SMBHBs and one EMRI. Furthermore, the galactic background is depicted (denoted as confusion noise) where the threshold for detection is set at a SNR of 7. The teal line shows the strain sensitivity of LISA, with the dashed black line denoting the total noise curve if taking the confusion noise into account. A cosmological contribution to the stochastic GW background is not shown but would contribute to the total noise curve. The figure is taken in slightly adapted form from [80].

3.2.2 Cosmological sources

Cosmological sources are sources that may be observed by LISA but for which there is currently no complementary experiment that could confirm the presence of such a signal. These sources are expected to contribute to a stochastic background signal that is projected to be nearly isotropic in most scenarios that generate such a background. In particular, an enhancement of scalar perturbations during inflation sources gravitational waves at second order. Studying how such a signal can be detected and characterized by LISA is the topic of 4 and Paper IV. This mechanism can also generate primordial black holes, which could make up a significant fraction of the dark matter observed today. The two other scenarios most commonly discussed are cosmic strings [100, 101] and strong first-order phase transitions in the early Universe (typically around the electroweak scale) [102].

Although it is certainly possible to express the amplitude of gravitational waves from cosmological sources in terms of the characteristic strain amplitude h_c, the amplitude of the signal is typically expressed in terms of the energy density per log frequency of the gravitational waves Ω_{GW}, as a fraction

of the critical energy density of the Universe ρ_{crit} (see Section 4.1.2):

$$\Omega_{\mathrm{GW}}(f) = \frac{1}{\rho_{\mathrm{crit}}} \frac{\mathrm{d}\rho_{\mathrm{GW}}}{\mathrm{d}\ln f} \ , \tag{3.17}$$

where ρ_{GW} is the spectral energy density of the gravitational waves. The characteristic strain amplitude h_c can then be expressed in terms of Ω_{GW} as [99]

$$h_c(f)^2 = \frac{3H_0^2}{2\pi^2} \frac{\Omega_{\mathrm{GW}}(f)}{f^2} \ , \tag{3.18}$$

where H_0 is the Hubble constant.

3.2.3 Definition of noise and signal-to-noise ratio

The data that LISA will provide will be a superposition of the signals from the sources that are present in the data and the instrumental noise. The instrumental noise is expected to be a combination of white noise and a low-frequency noise component that is expected to be correlated between the different LISA arms [103, 104]; however, for simplicity, we assume stationary Gaussian noise. This is a common assumption in LISA data analysis preparation studies. Under this assumption the noise is fully characterized by the one-sided power spectral density (PSD) [99]

$$\langle \tilde{n}(t)\tilde{n}(t')\rangle = \frac{1}{2}\delta(t-t')S_n(f) \ , \tag{3.19}$$

where $\tilde{n}(t)$ is the noise in the Fourier domain, $S_n(f)$ is the one-sided PSD, and $\langle \dots \rangle$ denotes the ensemble average. Analogously to the characteristic strain amplitude h_c, the noise amplitude is typically expressed in terms of the square root of the PSD as

$$h_n(f) = \sqrt{f S_n(f)} \ . \tag{3.20}$$

The squared signal-to-noise ratio (SNR) ρ^2 of a signal with strain h can be expressed in terms of the characteristic strain amplitude h_c and the noise amplitude h_n as [99]

$$\rho^2 = \int_0^\infty \mathrm{d}f \frac{4|\tilde{h}(f)|^2}{S_n(f)} = \int_{-\infty}^\infty \mathrm{d}(\log f) \left[\frac{h_c(f)}{h_n(f)}\right]^2 \ , \tag{3.21}$$

3.3 Source inference in LISA

The signal-dominated source landscape consisting of many thousands of overlapping signals, with a partially unknown noise curve as well as a potential cosmological background of gravitational waves, poses a number of challenges to correctly identifying and characterizing the sources in the

data. On top of this, neither the existence of all source types nor the number of sources in the LISA data is known. This problem, sometimes referred to as the "cocktail party" problem [105], is a major challenge for the LISA mission.

Different strategies have been proposed to infer the source population of LISA from this single datastream. The most ambitious approach, referred to as the *global fit*, attempts to fit the entire datastream with a model that contains all possible sources that could be present in the data. Of course, doing this with classic Monte Carlo methods is computationally infeasible, as the parameter space for $\sim 10^4$ sources with ~ 10 parameters each is way too large to be explored by an algorithm like an MCMC in reasonable time. Furthermore, the unknown number of sources in the data calls for a flexible approach such as reversible jump MCMC or Gibbs sampling.

Steady progress is being made to implement a prototypical global fit [106, 107, 108, 109, 110] where sources are either inferred individually or in groups.

An additional problem that remains is distinguishing between the astrophysical foreground, the cosmological background, and instrumental noise. Under the assumption that the shapes of the foregrounds and noise are well known, Paper IV explores how well they can be distinguished in the scalar-induced gravitational wave case, both if the signal is known and if it is only assumed to be scalar-induced without any prior knowledge of the shape of the signal.

Machine learning in LISA inference

The aforementioned difficulties in inferring the source population of LISA from the data have led to numerous studies exploring the use of machine learning techniques at different stages of the inference pipeline.

The high dimensionality of GW parameter spaces and costly waveform evaluations motivated early adoption of VI and SBI, initially in the context of inference for LVK events. In [24], a conditional variational autoencoder is used to learn the mapping from the data to the posterior distribution of the source parameters. In [107, 111, 112], VI is used to infer the source population of galactic binaries in the LISA data. Furthermore, the approach that we use in Paper III could be classified as a VI approach.

SBI, particularly utilizing Normalizing Flows, has seen an even greater adoption in PTA, LVK, and LISA applications [113, 114, 115, 116].

For a more complete review of how ML has been used in LISA inference, see [117].

4 Scalar-induced gravitational waves from inflation

This chapter gives a theoretical introduction and explains how gravitational waves are produced from scalar perturbations generated during inflation. The chapter is structured as follows: Section 4.1 introduces the inflationary paradigm and the (background) dynamics of the inflaton field, followed by a discussion of how inflation generates perturbations of the FLRW metric and how these metric perturbations are decomposed into scalar and tensor perturbations and how they evolve at first order in Section 4.1.5. Next, Section 4.2 describes the mechanism by which gravitational waves are nonlinearly generated from first-order scalar perturbations. Lastly, Section 4.3 discusses the implications of these results for the observations of gravitational waves and the production of primordial black holes.

4.1 Inflation

4.1.1 Motivation and the connection to LISA

Inflation is a theoretical framework that describes a period of accelerated, (nearly) exponential expansion in the earliest phase of the Universe. Initially proposed to explain the near homogeneity of the Universe on large scales, inflation resolves issues such as the horizon, flatness, and monopole problems [118, 119]. Moreover, it offers a mechanism for the generation of primordial density fluctuations, which later evolve into the large-scale structure of the Universe observed today [120].

The central idea is that one or multiple field(s), known as the inflaton(s), drive this rapid expansion. Quantum fluctuations in the inflaton(s) are imprinted into the spacetime metric and stretched to macroscopic scales by the inflationary expansion, providing the seeds for the observed anisotropies in the Cosmic Microwave Background (CMB) and the formation of galaxies and other structures.

Inflation also predicts the generation of primordial gravitational waves, though they have yet to be detected, which leads us to the main focus of this chapter: While the dynamics of inflation and the amplitude of scalar perturbations are very well constrained and understood at scales around the CMB ($k \sim 0.05\,\mathrm{Mpc}^{-1}$) [121], the dynamics of the inflaton field at smaller scales are yet to be meaningfully constrained. This is particularly interesting, as enhancements at these scales can produce primordial black holes (PBHs), which could at least be a fraction of the dark matter observed

in the Universe today. LISA specifically would be sensitive to PBHs in the asteroid-mass range ($\sim 10^{-12}\,M_\odot$ or $\sim 10^{21}\,\mathrm{g}$), which is a yet almost unconstrained part in the PBH mass distribution and thus could represent a significant fraction of the dark matter present in the Universe.

To obtain a signal that is detectable by LISA, however, it is necessary to consider models that either predict a significant enhancement of first-order tensor perturbations or consider the second-order production of gravitational waves from scalar perturbations. We focus on the latter.

Generating *Scalar Induced Gravitational Waves* (SIGWs) strong enough to be detected by LISA requires either *(i)* a deviation from single field slow-roll (SR) inflation, which–at scales observed by Planck–is the preferred and generally accepted model, or *(ii)* modifications in the standard thermal history of the Universe, such as a period of early matter (possibly PBH) domination as predicted by some beyond the standard model scenarios.

It is important to stress that while the SR approximation is a very good approximation at CMB scales, there are no fundamental reasons (apart from some weak constraints arising from e.g., *Big Bang Nucleosynthesis*, see Section 4.3.2) to assume that the inflaton potential is such that SR occurs at all scales.

4.1.2 The Einstein- and Friedmann Equations in Cosmology

To understand the dynamics of the Universe during and after inflation, we start with the Einstein field equation, which describes the interaction between matter and spacetime through gravity. In natural units ($c = \hbar = 1$), the Einstein equation is given as [122]

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{1}{M_{\mathrm{pl}}^2} T_{\mu\nu} \ , \tag{4.1}$$

where $G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu}$ is the Einstein tensor ($R_{\mu\nu}$ is the Ricci tensor, R is the Ricci scalar, and $g_{\mu\nu}$ is the metric), Λ is the cosmological constant, and $M_{\mathrm{pl}} = 1/\sqrt{8\pi G}$ is the reduced Planck mass (G is Newton's gravitational constant). $T_{\mu\nu}$ is the stress-energy tensor, which can be decomposed into the energy density ρ and pressure p of a perfect fluid as

$$T_{\mu\nu} = (\rho + p) u_\mu u_\nu + p g_{\mu\nu} \ , \tag{4.2}$$

where u_μ is the four-velocity of the fluid. Assuming homogeneity and isotropy implies no bulk velocity, which means that the 4-velocity is given

by $u_\mu = (-1, 0, 0, 0)$, and the energy-momentum tensor simplifies to

$$T_{\mu\nu} = \mathrm{diag}(-\rho, p, p, p) \ . \tag{4.3}$$

The next step is to introduce a suitable metric to describe a homogeneous and isotropic expanding Universe. This metric is called the *Friedmann-Lemaître-Robertson-Walker* (FLRW) metric and reads (in spherical coordinates) [122]

$$\mathrm{d}s^2 = -\mathrm{d}t^2 + a(t)^2 \left(\frac{\mathrm{d}r^2}{1 - kr^2} + r^2 \mathrm{d}\Omega^2 \right) \ , \tag{4.4}$$

where $a(t)$ is the scale factor of the Universe, t is time, r is the comoving radial coordinate, $\mathrm{d}\Omega^2 = \mathrm{d}\theta^2 + \sin^2\theta \mathrm{d}\phi^2$ is the solid angle element, and k is the curvature of the Universe. The scale factor $a(t)$ describes the expansion of the Universe and relates physical distances to comoving distances. We use the $(-1, 1, 1, 1)$ signature for the metric as is standard in cosmology. As the Universe has no such thing as an absolute scale, $a(t)$ needs to be normalized to some arbitrary time, which is typically done by setting $a(t_0) = 1$ today.

Inserting this metric into the time component of the Einstein equation (4.1), we obtain the Friedmann equation, which describes the dynamics of an expanding Universe. It reads

$$H^2 \equiv \left(\frac{\dot{a}}{a} \right)^2 = \frac{1}{3M_{\mathrm{pl}}^2} \rho - \frac{k}{a^2} \ , \tag{4.5}$$

where we have introduced the *Hubble rate* H. Furthermore, the spatial component implies that

$$\dot{\rho} = -3H(\rho + p) \ . \tag{4.6}$$

This equation states that the *dilution* of the energy is proportional to the expansion rate and depends on the relation between energy and pressure, otherwise known as the *equation of state* (EOS). We define the *equation of state parameter* $w = p/\rho$, which gives the following useful relations:

$$(\rho + p) = \rho(1 + w) = 3(1 + w) H^2 M_{\mathrm{pl}}^2 \ . \tag{4.7}$$

The EOS is a well-known quantity for the different ingredients of the Universe:

- For a non-relativistic fluid (matter), the pressure is negligible ($p = 0$), and therefore

$$\dot{\rho} = -3\frac{\dot{a}}{a}\rho \quad \Rightarrow \quad \rho \propto a^{-3} \ . \tag{4.8}$$

- For a relativistic fluid (radiation), the pressure is $p = \frac{1}{3}\rho$, implying

$$\dot{\rho} = -4\frac{\dot{a}}{a}\rho \quad \Rightarrow \quad \rho \propto a^{-4} \ . \tag{4.9}$$

- A cosmological constant is defined as energy that does not dilute. This implies that $\rho = \text{const.}$ and $p = -\rho$.
- The curvature term dilutes as a^{-2}, as can be seen in Eq. (4.5).

This different scaling of each component motivates rewriting the Friedmann equation in terms of the energy densities Ω_r, Ω_Λ, and Ω_k of the different components of the Universe:

$$\frac{H^2}{H_0^2} = \left(\frac{\Omega_m}{a^3} + \frac{\Omega_r}{a^4} + \Omega_\Lambda + \frac{\Omega_k}{a^2}\right), \tag{4.10}$$

where H_0 is the *Hubble constant*, i.e., the Hubble rate today, and a is the scale factor at any given time. The energy densities are defined as:

$$\Omega_m = \frac{\rho_m}{\rho_{\text{crit}}}, \quad \Omega_r = \frac{\rho_r}{\rho_{\text{crit}}}, \quad \Omega_\Lambda = \frac{\rho_\Lambda}{\rho_{\text{crit}}}, \quad \Omega_k = -\frac{k}{H_0^2 a_0^2} \ , \tag{4.11}$$

where ρ_m, ρ_r, and ρ_Λ are the energy densities of matter, radiation, and the cosmological constant, respectively, and $\rho_{\text{crit}} = 3H_0^2/8\pi G$ is the critical density of the Universe.

Typically, the Universe is assumed to be spatially flat ($k = 0$)[1], which we also assume in the following, and the energy densities are normalized such that $\Omega_m + \Omega_r + \Omega_\Lambda = 1$. Note also that this equation implicitly includes gravitational waves, which contribute to the radiation density Ω_r.

Looking at the structure of Eqs. (4.5) and (4.6), it is easy to see that

$$\dot{a} = \sqrt{\frac{1}{3M_{\text{pl}}^2}\rho a^2} \quad \text{and} \quad \ddot{a} = \frac{1}{6M_{\text{pl}}^2}(\rho + 3p)a \ , \tag{4.12}$$

This is also a good place to introduce an alternative choice for the time coordinate. Instead of using the *cosmic time* (or proper time) t, which is defined as the time that a clock at rest in the Universe would measure, we can define the *conformal time* η as

$$\mathrm{d}\eta = \frac{\mathrm{d}t}{a(t)} \ . \tag{4.13}$$

Using this time has the advantage that the relation between the conformal time and the physical distances $\mathrm{d}x$ remains the same while the Universe

[1] Flatness is compatible with observations and is in fact a prediction of inflation, as any initial curvature is quickly diluted by the rapid expansion.

expands, meaning that we can define *comoving* quantities that are constant in time. We frequently use this time coordinate in the following sections.

Using conformal time, the FLRW metric Eq. (4.4) becomes

$$\mathrm{d}s^2 = a^2(\eta)\left[-\mathrm{d}\eta^2 + \frac{\mathrm{d}r^2}{1-kr^2} + r^2\mathrm{d}\Omega^2\right] , \tag{4.14}$$

and the Friedmann equation becomes

$$\mathcal{H}^2 = \frac{1}{3M_{\mathrm{pl}}^2}\rho \ - \frac{k}{a^2} , \tag{4.15}$$

where we have introduced the conformal Hubble rate $\mathcal{H} = a'/a = aH$. To keep the notation consistent, in the following, primes denote derivatives with respect to conformal time, and dots denote derivatives with respect to cosmic time.

To achieve accelerated expansion, Eq. (4.12) tells us that we need a negative pressure, i.e., $\rho + 3p < 0$. Simply assuming that the Universe is dominated by a cosmological constant, we can see that the Universe expands exponentially, i.e., $a(t) \propto e^{Ht}$, where $H = \sqrt{\Lambda/3}$, called a *de Sitter* Universe. Unfortunately, this would be too easy, as we would get a Universe that is exponentially expanding forever, meaning that modes that exit the Hubble horizon would never re-enter. This is incompatible with the expansion history we need to form galaxies and explain the Universe as we observe it today.

An elegant solution to this problem is inflation, which achieves a quasi-de Sitter expansion by introducing one or multiple scalar fields that drive the expansion by slowly rolling down a potential, eventually decaying into standard model particles through a process called *reheating*. This model is discussed in detail in Sections 4.1.3 and 4.1.4.

After inflation, the standard theory suggests that initially all particles are ultra-relativistic, leading to a period of radiation domination, followed by matter domination, and finally, today, we are in a period of dark energy domination. The implications of this on the generation of gravitational waves is discussed in Section 4.2.

4.1.3 Single field inflation

Although there exists a plethora of proposed models of inflation–which are briefly discussed in Section 4.1.4–the simplest and most widely studied model assumes that a single scalar field called the *inflaton* drives the expansion.

If we assume a scalar field with a canonical kinetic term, the Lagrangian for the inflaton ϕ reads (in the Einstein frame for convenience)

$$\mathcal{L} = \frac{1}{2}\partial_\mu\phi\partial^\mu\phi - V(\phi) \ , \tag{4.16}$$

where $V(\phi)$ is the potential energy function of the inflaton field. This is equivalent to the action

$$S = -\int \mathrm{d}^4x\sqrt{|g|}\left[\frac{1}{2M_{\rm pl}^2}R + \frac{1}{2}\partial_\mu\phi\partial^\nu\phi - V(\phi)\right] \ , \tag{4.17}$$

where R is the Ricci scalar, originating from the Lagrangian of gravitation, and we assume minimal coupling to gravity.

Varying the action with respect to the metric $g_{\mu\nu}$ gives us back the Einstein equation, where the inflaton field contributes to the energy momentum tensor

$$T_{\mu\nu} = \partial_\mu\phi\partial_\nu\phi - g_{\mu\nu}\left(\frac{1}{2}\partial_\alpha\phi\partial^\alpha\phi - V(\phi)\right) \ , \tag{4.18}$$

and varying with respect to the inflaton field gives us the Klein-Gordon equation

$$\frac{1}{\sqrt{|g|}}\partial_\mu\left[\sqrt{|g|}\partial^\nu\phi\right] - \frac{dV}{d\phi} = 0 \ . \tag{4.19}$$

As we are only interested in the background evolution for now, we assume the inflaton field to be homogeneous, i.e., $\phi = \phi(t)$ and a flat FLRW metric Eq. (4.4). This simplifies the stress energy tensor to $T_{\mu\nu} = \mathrm{diag}(-\rho, p, p, p)$ with

$$\rho = \frac{1}{2}\dot{\phi}^2 + V(\phi), \qquad p = \frac{1}{2}\dot{\phi}^2 - V(\phi) \ . \tag{4.20}$$

and the Klein-Gordon equation to

$$\ddot{\phi} + 3H\dot{\phi} + V(\phi)_{,\phi} = 0 \ , \tag{4.21}$$

where $\dot{\phi}$ is the time derivative of ϕ, and $V(\phi)_{,\phi} = \frac{dV}{d\phi}$ is the derivative of the potential $V(\phi)$ with respect to the inflaton field ϕ.

By injecting Eq. (4.20) into Eq. (4.5), we obtain an equation for the expansion of the Universe (we assume that only the inflaton injects energy into the Universe during inflation):

$$G_0^0 = 3H^2 = \frac{1}{M_{\rm pl}^2}\left(\frac{1}{2}\dot{\phi}^2 + V(\phi)\right) \ , \tag{4.22}$$

while the spatial part of the Einstein equation G_i^i provides the relation

$$G_i^i = -2\frac{\ddot{a}}{a} - H^2 = \frac{1}{M_{\rm pl}^2}\left(\frac{1}{2}\dot{\phi}^2 - V(\phi)\right) \ . \tag{4.23}$$

Combining Eqs. (4.22) and (4.23), we obtain the useful relation

$$\dot{H} = -\frac{\dot{\phi}^2}{2M_{\rm pl}^2} \ . \tag{4.24}$$

For inflation to occur, the inflaton field must be slowly rolling down its potential, i.e., the kinetic energy of the inflaton must be much smaller than the potential energy

$$\frac{1}{2}\dot{\phi}^2 \ll V(\phi) \ . \tag{4.25}$$

Likewise, the potential must be flat enough to allow for a long period of inflation, i.e.,

$$\left|\ddot{\phi}\right| \ll |V_{,\phi}| = 3H\left|\dot{\phi}\right| \ . \tag{4.26}$$

Using the Klein-Gordon and Friedmann equations, these conditions can be written in terms of the Hubble rate and its derivatives as

$$-\dot{H} \ll 3H^2 \quad \text{and} \quad |\ddot{H}| \ll -6H\dot{H} \ll 18H^3 \ . \tag{4.27}$$

This motivates defining the slow-roll parameters ϵ_H and η_H as [2]

$$\epsilon_H = -\frac{\dot{H}}{H^2} = \frac{(\phi')^2}{2M_{\rm pl}^2} \ , \qquad \eta_H = -\frac{\ddot{H}}{2H\dot{H}} = \epsilon_H - \frac{1}{2}(\log\epsilon_H)' \ . \tag{4.28}$$

We speak of slow-roll (SR) inflation when $\epsilon_H \ll 1, |\eta_H| \ll 1$. Note that the condition $\rho + 3p < 0$, which is necessary for accelerated expansion, is equivalent to $\epsilon_H < 1$ meaning that inflation ends when $\epsilon_H = 1$.

In practice, when computing the dynamics of inflation, we need to solve the Klein-Gordon Eq. (4.21) together with the Friedmann Eq. (4.22). This is typically done numerically, but as the expansion happens nearly exponentially, the computation spans a large range of scales, which leads to numerical issues. To avoid this, it makes sense to recast Eqs. (4.21) and (4.22) in terms of the number of e-folds N defined as

$$N = \int_{t_i}^{t} H(t')\mathrm{d}t' \ , \tag{4.29}$$

where t_i is some (arbitrary) initial time. We set $t = 0$ to the time of the initial conditions. This is equivalent to defining $\mathrm{d}N = H\mathrm{d}t$, which allows us to write the Klein-Gordon equation as

$$\phi_{,NN} + \phi_{,N}\left(3 - \frac{(\phi_{,N})^2}{2M_{\rm pl}^2}\right) + \frac{V_{,\phi}}{H^2} = 0 \ , \tag{4.30}$$

[2] Note that the way of defining the slow-roll parameters is not unique, and different conventions exist. This is both regarding whether they are defined in terms of the Hubble rate or the potential and up to some numerical factors.

where we used Eq. (4.24) and $()_{,N}$ denote derivatives with respect to efolds. Likewise, the Friedmann equation becomes

$$3H^2 + H_{,N}H - \frac{V(\phi)}{M_{\rm pl}^2} = 0 \ . \tag{4.31}$$

4.1.4 Alternative models of inflation

Inflationary models capable of generating detectable SIGWs in the LISA band are not limited to single-field scenarios with ultra-slow roll (USR) dynamics. A number of alternative mechanisms can lead to enhancements in the scalar curvature power spectrum $\mathcal{P}_\zeta(k)$ (see Eq. (4.66) for the definition) at small scales and hence produce observable SIGWs at second order in perturbation theory. These include multi-field models, particle production during inflation, or non-attractor dynamics outside USR. A full in-depth description is beyond the scope of this work, so we stick to briefly mentioning a few classes of such models (see e.g., [123] for a review).

Multi-field inflationary models

In multi-field inflation, additional scalar fields can source curvature perturbations through entropic (isocurvature) modes, which can convert into adiabatic fluctuations. This conversion can be highly efficient in the presence of sharp turns in field space or non-trivial kinetic couplings, leading to localized peaks or broad enhancements in $\mathcal{P}_\zeta(k)$ [124].

A prominent subclass is hybrid inflation models, where inflation ends via a tachyonic instability in an auxiliary field. Before this transition, the waterfall field can remain light and dynamically subdominant, sourcing a bump-like feature in $\mathcal{P}_\zeta(k)$ at small scales. Models of this kind typically lead to quasi-lognormal spectra:

$$\mathcal{P}_\zeta(k) \sim A_s \exp\left[-\frac{\ln^2(k/k_\star)}{2\Delta^2}\right], \tag{4.32}$$

which in turn source SIGW signals, potentially within LISA sensitivity [125].

In scenarios with curved inflationary trajectories, transient entropic-to-adiabatic conversion occurs during sudden field-space turns. These turns induce oscillations in the scalar power spectrum superimposed on a peaked envelope. The resulting SIGW signals inherit these features and are sensitive to the duration and sharpness of the turn [126, 127, 128].

Particle production during inflation

Another mechanism is particle production sourced by coupling the inflaton (or a spectator axion-like field) to gauge fields, typically via a Chern-

Simons term $\phi F\tilde{F}$. As the scalar field evolves, it induces a tachyonic instability for one helicity of the gauge field, leading to exponential amplification. These amplified gauge quanta act as a source of scalar (and tensor) fluctuations via inverse decay processes [129, 130].

The power spectrum in such cases acquires a blue tilt at small scales and may contain a distinct bump. The amplitude and shape of the feature depend on the axion velocity and coupling strength. These models often yield strongly non-Gaussian curvature perturbations (see 4.2.2) and hence the non-Gaussian contribution to SIGWs [131].

Non-attractor phases beyond USR

Although USR is the prototypical non-attractor phase, there exist other classes of non-attractor dynamics capable of producing an enhancement in $\mathcal{P}_\zeta$. For instance, constant-roll solutions with $\ddot{\phi}/H\dot{\phi} = \beta = \text{const.}$ generalize USR and can sustain growing super-Hubble curvature perturbations [132, 133]. The extent of enhancement and the slope of the power spectrum after the peak are sensitive to the specific value of β and the transition history back to slow-roll.

Another example is "transient non-attractor" evolution induced by a temporary violation of the null energy condition or modifications of the kinetic term. In k-inflation [134, 135] or G-inflation models [136, 137], a rapidly evolving sound speed c_s can lead to power spectrum features. If c_s becomes significantly less than unity at small scales, the curvature perturbation is enhanced by a factor $\sim 1/c_s^2$, and the resulting SIGWs are boosted accordingly.

In Paper IV we discuss the empirical signatures of these models in the context of LISA observations and the prospects for detecting the corresponding SIGW signals.

4.1.5 Scalar and Tensor perturbations at first order

In Section 4.1.3, we have established the background dynamics of inflation (we assumed a flat FLRW metric, which by definition is homogeneous and isotropic). However, we know that the Universe is not perfectly homogeneous and isotropic, as observable from the CMB and the large-scale structure of the Universe. These inhomogeneities are generated during inflation and are imprinted into the spacetime metric as perturbations. They naturally arise from inflation as quantum fluctuations in the inflaton field or of the relevant degrees of freedom in the case of multi-field inflation.

To allow for such perturbations, we perturb both the FLRW metric and the inflaton field at linear order:

$$g_{\mu\nu}(t,\boldsymbol{x}) = \overline{g}_{\mu\nu}(t) + \delta g_{\mu\nu}(t,\boldsymbol{x}), \quad \phi(t,\boldsymbol{x}) = \overline{\phi}(t) + \delta\phi(t,\boldsymbol{x}) \ . \tag{4.33}$$

where the overline denotes the background quantities. Note that the perturbed metric has 10 degrees of freedom, but we can remove 4 by gauge transformations, leaving 6 degrees of freedom. These can be decomposed into 2 scalar, 2 vector, and 2 tensor perturbations. The vector perturbations only yield decaying solutions, so they can be neglected safely.

This leaves the two scalar perturbations, which can be written in terms of the Bardeen potentials Φ and Ψ, and the tensor perturbations, which are described by the tensor modes h_{ij}. Choosing a gauge where the scalar perturbations are diagonal, called the longitudinal (or conformal Newtonian) gauge, the perturbed FLRW metric reads [122]

$$\mathrm{d}s^2 = a^2(\eta)\left[-(1+2\Phi)\mathrm{d}\eta^2 + (1-2\Psi)\delta_{ij}\mathrm{d}x^i\mathrm{d}x^j\right] + h_{ij}\mathrm{d}x^i\mathrm{d}x^j \ , \tag{4.34}$$

where $\mathrm{d}t = a\,\mathrm{d}\eta$ is the conformal time running from $-\infty$ in the asymptotic past to 0 in the asymptotic future, and the two degrees of freedom of the tensor perturbations are encoded in the tensor modes $h_{ij} = h_1 e^1_{ij} + h_2 e^2_{ij}$, where e^1_{ij} and e^2_{ij} are the two polarization tensors. We choose $\sum_{ij\lambda} e^{\lambda}_{ij} e^{\lambda'}_{ij} = 1$. At first order in perturbation theory, the scalar and tensor perturbations evolve independently and can therefore be treated separately. This also implies that the perturbations in the inflaton field affect only the scalar perturbations directly.

Lastly, in order to effectively find solutions to the equations of motion of the scalar and tensor modes, we need to discuss how to quantize these fields [3]. For this, recall the quantum harmonic oscillator in Minkowski space. Momentarily introducing the free, massless scalar field χ with the Lagrangian

$$\mathcal{L} = \frac{1}{2}\partial_\mu\chi\partial^\mu\chi \ , \tag{4.35}$$

one can write down the equation of motion (EOM) for the Fourier modes χ_k as

$$\ddot{\chi_k} + k^2\chi_k = 0 \ , \tag{4.36}$$

and the Hamiltonian as

$$H = \int \mathrm{d}^3k \left(\dot{\chi}^*_k\dot{\chi}_k + k^2\chi^*_k\chi_k\right) \ . \tag{4.37}$$

[3] Quantization is needed here, as deep in the past, quantum fluctuations generate the perturbations that classicalize only after Hubble crossing.

Each Fourier mode is a harmonic oscillator, and in the fundamental state, the wave functional of each mode is a Gaussian:

$$\Psi_0[\chi_k] \propto \exp\left(-\frac{1}{2}k|\chi_k|^2\right) \ . \tag{4.38}$$

This means that the expectation value of the field χ_k is zero, and its variance is

$$\langle\chi_k^2\rangle = \frac{1}{2k} \ . \tag{4.39}$$

Assuming that deep in the Hubble horizon (i.e., $k \gg aH$) the modes are in their ground state and the metric is approximately Minkowski, we can use this result to match our solutions of the quantum harmonic oscillator to the solutions of the equations of motion of the perturbations. This is done by requiring that in the sub-horizon limit, the variance of the field is given by Eq. (4.39) and the solution of the mode is of the form of a plane wave, called *Bunch-Davies vacuum* [138]

$$\chi_k = \frac{1}{\sqrt{2k}}e^{-ikt} \ . \tag{4.40}$$

Tensor perturbations

We first have a brief look at the tensor perturbations, as they are easier to understand and compute. Although the focus of this work lies on GWs induced by scalar perturbations, it is important to check how tensor perturbations evolve at first order, as we need to be able to predict how they change (or do not change) when introducing an ultra-slow-roll phase.

The action of the tensor perturbations only contains the gravitational part of Eq. (4.17) and no direct coupling to the inflaton field. The Lagrangian describing the tensor perturbations is given by [122]

$$\mathcal{L}_{\text{tensors}} = \frac{a^4 M_{\text{pl}}^2}{8}\left[\partial_\mu h_1 \partial^\mu h_1 + \partial_\mu h_2 \partial^\mu h_2 - (\nabla h_1)^2 - (\nabla h_2)^2\right] \ , \tag{4.41}$$

where we are discarding all terms $\mathcal{O}(\Phi^2), \mathcal{O}(\Psi^2), \mathcal{O}(\delta\phi^2)$ but we keep terms that are quadratic in h_{ij} (there are no linear terms in h_{ij}). The implications of keeping higher order terms in $\Phi, \Psi, \delta\phi$ are discussed in Section 4.2.

The divergence term ∇h_k can be gauged away by switching to the transverse-traceless (TT) gauge ($\partial_i h^{ij} = 0$ and $h^i{}_i = 0$). Furthermore, the two polarizations share the same EOM. We therefore concentrate on a single mode h_λ (the λ is to emphasize that we are in real space as opposed to Fourier space). In addition, we introduce the comoving, rescaled tensor mode $v_\lambda = a h_\lambda M_{\text{pl}}/2$, which simplifies the Lagrangian to

$$\mathcal{L}_{\text{tensors}} = \frac{1}{2}\left[\left(v_\lambda' - \frac{a'}{a}v_\lambda\right)^2 - (\partial_i v_\lambda)^2\right] \ . \tag{4.42}$$

We then introduce the Fourier modes v_k, which obey the equation of motion

$$v_k'' + \left(k^2 - \frac{a''}{a}\right) v_k = 0 \ . \tag{4.43}$$

Let us consider exact de Sitter inflation to see how the tensor modes evolve. In this case, the scale factor is given by $a(t) = a_0 e^{Ht}$ and therefore $\eta = -1/(aH)$ with $H = \text{const}$, and $a''/a = 2/\eta^2$. The EOM for the tensor modes becomes

$$v_k'' + \left(k^2 - \frac{2}{\eta^2}\right) v_k = 0 \ . \tag{4.44}$$

The solution to this equation is well known and can be expressed by the Hankel functions of the first and second kind $H_\nu^{(1)}(-k\eta)$ and $H_\nu^{(2)}(-k\eta)$ with $\nu = 3/2$ and the general solution is a linear combination of these two solutions $v_k = A_k H_{3/2}^{(1)}(-k\eta) + B_k H_{3/2}^{(2)}(-k\eta)$ but by choosing the initial conditions such that the mode is in its ground state in the asymptotic past, also known as *Bunch-Davies initial conditions* see Eq. (4.40), we find that $B_k = 0$.

The remaining term can be written down in closed form as

$$A_k H_{3/2}^{(1)} = A_k \sqrt{\frac{2}{\pi k\eta}} e^{-ik\eta} \left(1 - \frac{i}{k\eta}\right) \ . \tag{4.45}$$

and by matching to the mode function for $k \gg aH$ (see Eq. (4.40)), one arrives at

$$v_k = \frac{1}{\sqrt{2k}} \left(1 - \frac{i}{k\eta}\right) e^{-ik\eta} \ . \tag{4.46}$$

There are three important things to note about this solution:

1. In the deep sub-Hubble limit, the mode function is a plane wave, as expected.
2. In the super-Hubble limit ($k \ll aH$), the mode function is constant, meaning that the tensor modes are frozen in this regime.
3. As long as inflation does not significantly deviate from de Sitter, the tensor modes are nearly unaffected by the dynamics of the inflaton field.

The last point is crucial, as it ensures that the tensor modes are only marginally affected by the ultra-slow-roll phase. This is in stark contrast to the scalar modes, which are discussed below.

The last quantity that is useful to compute is the *power spectrum* of the tensor modes. This is defined as [4]

$$\mathcal{P}_h(k) \equiv \frac{k^3}{2\pi^2} \left\langle \sum_{ij} |h_{ij\,k}|^2 \right\rangle \ . \tag{4.47}$$

which, in the case of de Sitter inflation, is given by

$$\mathcal{P}_h(k) = \frac{2}{\pi^2} \frac{H^2}{M_{\rm pl}^2} \ . \tag{4.48}$$

where H is the Hubble rate during inflation. This result is independent of the wavenumber k and is a direct consequence of the scale invariance of the de Sitter space.

Scalar perturbations

Compared to the tensor perturbations, the scalar perturbations are somewhat trickier to compute as they are directly coupled to the inflaton field.

We recall the three degrees of freedom of the scalar perturbations, the two Bardeen potentials Φ and Ψ, and the inflaton field perturbation $\delta\phi$. In the longitudinal gauge, the perturbed Einstein equation provides the relation

$$\delta G_i^j = \partial_i \partial^j (\Phi - \Psi) = \frac{1}{M_{\rm pl}^2} \partial_i \delta\phi \partial^j \delta\phi \tag{4.49}$$

for $i \neq j$. As the right-hand side vanishes at first order in perturbation theory, in the absence of anisotropic stress, this means that $\Phi = \Psi$. We are left with only Φ and $\delta\phi$, which evolve according to the Einstein and Klein-Gordon equations. Skipping the lengthy but straightforward derivation and going directly to Fourier space, the Klein-Gordon equation reads [122]

$$\delta\phi_k'' + 2\mathcal{H}\delta\phi_k' + (k^2 + a^2 V_{,\phi\phi})\delta\phi_k - 4\overline{\phi}'\Phi_k' + 2a^2 V_{,\phi}\Phi_k = 0 \ , \tag{4.50}$$

where $\mathcal{H} = aH$ is the conformal Hubble parameter. The Einstein equation yields the equation of motion for the Bardeen potential Φ. The time $(0,0)$ component is given by

$$k^2\Phi_k + 3\mathcal{H}(\Phi_k' + \mathcal{H}\Phi_k) = -\frac{1}{2M_{\rm pl}^2}\left(\overline{\phi}'^2\Phi - \overline{\phi}'\delta\phi_k - a^2 V_{,\phi}\delta\phi_k\right) \ . \tag{4.51}$$

[4] The definition is not totally unique in literature. Some authors omit the factor of $2\pi^2$ in the denominator.

The time-space $(0,i)$ component gives rise to a momentum constraint

$$\Phi' + \mathcal{H}\Phi = -\frac{1}{2M_{\rm pl}^2}\overline{\phi}'\delta\phi \ , \tag{4.52}$$

and the space-space (i,i) component (the off-diagonal terms vanish) yields

$$\Phi_k'' + 3\mathcal{H}\Phi_k' + \left(2\mathcal{H}' + \mathcal{H}^2 + \frac{k^2}{2}\right)\Phi_k = \frac{1}{2M_{\rm pl}^2}\left(\overline{\phi}'\delta\phi_k' - \overline{\phi}'^2\Phi_k - a^2V_{,\phi}\delta\phi_k\right) \ . \tag{4.53}$$

Using any of the Einstein components (the Bianchi identities ensure self-consistency), one can eliminate either $\delta\phi$ or Φ from the equations, leaving us with only one equation of motion. A more clever way is to introduce the Mukhanov-Sasaki variable [139, 140]

$$\xi_k = a\left(\Phi_k + \frac{\overline{\phi}'}{\mathcal{H}}\delta\phi_k\right) \ , \tag{4.54}$$

and

$$z = \frac{a\overline{\phi}'}{\mathcal{H}} \ , \tag{4.55}$$

which yields a very similar equation of motion to the one of the tensor perturbations:

$$\xi_k'' + \left(k^2 - \frac{z''}{z}\right)\xi_k = 0 \ . \tag{4.56}$$

By comparing this equation to Eq. (4.43) and Eq. (4.41), it becomes easy to reverse-engineer the Lagrangian in terms of the Mukhanov-Sasaki variable:

$$\mathcal{L}_{\rm scalar} = \frac{1}{2}\left[\left(\xi' - \frac{z'}{z}\xi\right)^2 - (\partial_i\xi)^2\right] \ . \tag{4.57}$$

We can make our lives even easier by directly choosing a favorable gauge where $\Phi = 0$ and all scalar perturbations are naturally encoded in the perturbation of the inflaton field. This gauge is called the spatially flat gauge, and the metric reads

$$\mathrm{d}s^2 = a^2(\eta)\left[-(1+2A)\mathrm{d}\eta^2 + 2\partial_i B\mathrm{d}x^i\mathrm{d}\eta + \delta_{ij}\mathrm{d}x^i\mathrm{d}x^j\right] \ , \tag{4.58}$$

where the new variables A and B have been introduced. With this metric, the Klein-Gordon equation reads

$$\delta\phi_k'' + 2\mathcal{H}\delta\phi_k' + (k^2 + a^2V_{,\phi\phi})\delta\phi_k - \overline{\phi}'(A' + 3\mathcal{H}A) + 2a^2V_{,\phi}A = 0 \ , \tag{4.59}$$

and the Einstein equations give the following equations of motion for A and B:

$$3\mathcal{H}(\mathcal{H}A - B') + k^2 B = -\frac{\kappa^2}{2}\left(\overline{\phi}'\delta\phi' + a^2\overline{V}_{,\phi}\delta\phi + \overline{\phi}'^2 A\right) \ , \tag{4.60}$$

$$\mathcal{H}A - B' = \frac{\kappa^2}{2}\overline{\phi}'\delta\phi \ . \tag{4.61}$$

Introducing the much simpler Mukhanov-Sasaki variable $v_k = a\delta\phi_k$, one finds

$$v_k'' + \left(k^2 - \frac{a''}{a} + a^2\overline{V}_{,\phi\phi}\right)v_k - a\overline{\phi}'(A' + 3\mathcal{H}A) + 2a^3\overline{V}_{,\phi}A = 0. \tag{4.62}$$

which simplifies to the same Mukhanov-Sasaki equation as before:

$$v_k'' + \left(k^2 - \frac{z''}{z}\right)v_k = 0 \ . \tag{4.63}$$

As the quantity of interest for this work is the power spectrum of the scalar perturbations, we define the comoving curvature perturbation ζ as

$$\zeta = \Phi + \frac{\mathcal{H}}{\overline{\phi}'}\delta\phi \ . \tag{4.64}$$

This means that at first order in perturbation theory

$$\zeta_k = \frac{v_k}{z} \ . \tag{4.65}$$

Solving the Mukhanov-Sasaki equation for the scalar perturbations then follows the same logic as for the tensor perturbations. The details of how to set the initial conditions and how to make this calculation computationally efficient are explained in Paper IV.

Once the Mukhanov-Sasaki equation has been solved, the power spectrum of the scalar perturbations can be computed analogously to the tensor perturbations. The power spectrum is defined as

$$\mathcal{P}_\zeta(k) = \frac{k^3}{2\pi^2}|\zeta_k|^2 \ . \tag{4.66}$$

Note, however, that it matters at which time we compute the power spectrum. Just after Hubble crossing, $\mathcal{P}_\zeta(k)$ remains conserved and does not evolve. However, during and after reheating, the modes re-enter the horizon, and the power spectrum evolves until linearity breaks down. Discussing the implications of this is beyond the scope of this work, as the focus is on second-order tensors generated from scalar perturbations that decouple long before non-linearities take over.

Let us now explicitly write down the solution to the Mukhanov-Sasaki equation for a quasi-de Sitter, slow roll period of inflation. One uses the

same Ansatz as for the tensor perturbations, but now for the Mukhanov-Sasaki variable v_k instead of the metric perturbation directly. Going directly to the large wavelength limit, v_k becomes

$$v_{k\ll aH} = \frac{i}{\sqrt{2k}k\eta} = \frac{iaH}{\sqrt{2k^3}} \ . \tag{4.67}$$

and it is possible to use the first slow-roll parameter ϵ_H to write z as

$$z = \frac{a\dot{\phi}}{H} = \sqrt{2\epsilon_H} H M_{\rm pl} \ . \tag{4.68}$$

This allows writing the power spectrum of the scalar perturbations as

$$\mathcal{P}_\zeta(k) = \frac{H^2}{4\pi^2(\phi')^2} = \frac{H^2}{8\pi^2 M_{\rm pl}^2 \epsilon_H} \ . \tag{4.69}$$

Notice that–unlike the tensor modes–the scalar perturbations are directly affected by the dynamics of the inflaton field. In particular, Eq. (4.69) implies that the scalar perturbations can be enhanced by decreasing the velocity of the inflaton field ϕ' (or equivalently decreasing the slope of the potential $V_{,\phi}$). At the same time, this modification only marginally affects the tensor perturbations. This is the key idea behind USR inflation.

Tensor-to-scalar ratio and spectral index

A useful quantity to define in the context of inflation is the tensor-to-scalar ratio r. This is defined as the ratio of the power spectrum of the tensor perturbations to the power spectrum of the scalar perturbations at horizon crossing:

$$r = \frac{\mathcal{P}_h(k)}{\mathcal{P}_\zeta(k)} \ . \tag{4.70}$$

For quasi-de Sitter inflation, this ratio is given by

$$r = 16\epsilon_H \ . \tag{4.71}$$

Additionally, assuming slow roll inflation, it is easy to compute the amplitude and spectral tilt of the scalar perturbations. For this, one deviates from scale invariance by introducing the spectral index n_s, which changes the power spectrum to

$$\mathcal{P}_\zeta(k) = A_s \left(\frac{k}{k_*}\right)^{n_s - 1} \ , \tag{4.72}$$

where k_* is some pivot scale, which is typically chosen to be around the scale of the CMB ($k_* = 0.05\,\mathrm{Mpc}^{-1}$). At linear order in the slow-roll parameters, the spectral index is given by

$$n_s - 1 = -6\epsilon_H + 2\eta_H \ . \tag{4.73}$$

The amplitude of the scalar perturbations is the scale-invariant power spectrum from Eq. (4.69)

$$A_s = \frac{H^2}{8\pi^2 M_{\rm pl}^2 \epsilon_H} \,, \tag{4.74}$$

as shown above.

Currently, the CMB offers the most competitive bounds on the inflationary parameters around the CMB pivot scale $k_* = 0.05\,\mathrm{Mpc}^{-1}$. The Planck collaboration has measured the amplitude of the scalar perturbations to be $A_s = (2.10 \pm 0.03) \times 10^{-9}$, the scalar spectral index to be $n_s = 0.9649 \pm 0.0042$ [121, 141], while BICEP2/Keck offers the most competitive upper bound on the tensor-to-scalar ratio $r < 0.036$ at 95% confidence level [142]; however, these constraints only exist in a narrow range of scales around the pivot scale.

This result implies that the primary tensor modes are suppressed compared to the scalar modes, and unlike the CMB, which offers a very sensitive and precise measurement of the scalar perturbations, the tensor modes are much harder to detect. They do leave an imprint in the B-modes of the CMB (hence the upper bound), but direct detection with gravitational wave detectors is still very far out of reach. Furthermore, we do not expect to see a significant signal coming from primary tensor modes at scales outside the CMB if inflation is driven by a single scalar field.

All of this motivates looking at gravitational waves generated at second order from the scalar perturbations. These are not only a potential source of primordial black holes but also a source of gravitational waves that could be detected by future experiments.

4.2 Second-order production of gravitational waves

In the last section, we have seen that primordial tensor modes at first order are suppressed compared to the scalar modes. However, the scalar modes are directly coupled to the inflaton field and can be enhanced without significantly changing the primary tensor modes. The question arises whether the scalar modes can generate tensor modes at second order, which are detectable by future experiments such as LISA.

4.2.1 Basic equations for second-order gravitational waves

Before we get into the details of second-order gravitational waves, we need to quickly digress to explain the meaning of "order" when talking of second-order scalar-induced gravitational waves.

The perturbed FLRW metric in the longitudinal gauge was established in Eq. (4.33), and we saw that at leading order in perturbations, the scalars and tensors evolve separately and according to the Lagrangians Eqs. (4.42) and (4.57). To arrive at these equations, the metric was perturbed to linear order in scalar perturbations, introducing the Bardeen potentials Φ and Ψ. The equations of motion were computed by only keeping linear terms in Φ, Ψ, and $\delta\phi$.

Going higher in the order of perturbation theory, ($h_{ij} = h_{ij}^{(1)} + \frac{1}{2}h_{ij}^{(2)} + \dots$), the source for h gets terms that are quadratic in the (first order) scalar, vector, and tensor perturbations, which motivates studying the effects of these quadratic source terms in Φ, Ψ, and $\delta\phi$.

To find the equation of motion of the tensor perturbations at second order in Φ, Ψ (dropping the explicit dependence on $\delta\phi$ for now), it is instructive to start from the perturbed metric in the longitudinal gauge from Section 4.1.5.

Deriving the equation of motion is a very cumbersome task, and repeating it is beyond the scope of this work (for a full derivation, see e.g., [143]), but it makes sense to give an outline of the procedure to understand the terms appearing. We follow the notation and reasoning of [144].

We have already seen the first-order (linear) equation of motion for the scalar perturbations Eq. (4.56), obtained by perturbing the Einstein equation to first order

$$G^{(1)i}{}_{j} = \frac{1}{M_{\rm pl}^2} T^{(1)i}{}_{j} \,. \tag{4.75}$$

while at quadratic order we get the independently evolving tensor perturbations (see Eq. (4.43)).

We can now go to second order in perturbations by expanding the perturbed FLRW metric to second order (denoting the order i with $()^{(i)}$):

$$\begin{aligned} \mathrm{d}s^2 = a^2(\eta)\Big[&- (1 + 2\Phi^{(1)} + 2\Phi^{(2)})\mathrm{d}\eta^2 + 2V_i^{(2)}\mathrm{d}\eta\mathrm{d}x^i \\ &+ \left[(1 - 2\Psi^{(1)} - 2\Psi^{(2)})\delta_{ij} + \frac{1}{2}h_{ij}^{(2)}\right]\mathrm{d}x^i\mathrm{d}x^j\Big] \,, \end{aligned} \tag{4.76}$$

where $V_i^{(2)}$ is the vector perturbation at second order, and we are neglecting first-order vector and tensor perturbations. The Einstein equation at second order reads

$$G^{(2)i}{}_{j} = \frac{1}{M_{\rm pl}^2} T^{(2)i}{}_{j} \,. \tag{4.77}$$

We ignore second-order terms $\mathcal{O}(\Phi^{(2)}, \Psi^{(2)}, V^{(2)})$ (They are eliminated by the projection tensor defined below in Eq. (4.83)) but keep the terms in $h_{ij}^{(2)}$, which we call h_{ij} from now on. Furthermore, going back to our previous notation of $\Phi \equiv \Phi^{(1)}, \Psi \equiv \Psi^{(1)}$, we can write the Einstein tensor at second order as

$$\begin{aligned} G^{(2)i}{}_j(\mathcal{O}(\Phi,\Psi,h)) = a^{-2}\Bigg[&\frac{1}{4}\left((h_j^i)'' + 2\mathcal{H}(h_j^i)' + \nabla^2 h_j^i\right) + \\ &- 2\Phi\partial^i\partial_j\Phi - 2\Psi\partial^i\partial_j\Phi + 4\Psi\partial^i\partial_j\Psi + \partial^i\Phi\partial_j\Phi + \\ &- \partial^i\Phi\partial_j\Psi - \partial^i\Psi\partial_j\Phi + 3\partial^i\Psi\partial_j\Psi + \delta^i_j(\text{diagonal terms})\Bigg] , \end{aligned} \tag{4.78}$$

which, using the first order relation $\Phi = \Psi$, simplifies to

$$\begin{aligned} G^{(2)i}{}_j(\mathcal{O}(\Phi,h)) = a^{-2}\Bigg[&\frac{1}{4}\left((h_j^i)'' + 2\mathcal{H}(h_j^i)' + \nabla^2 h_j^i\right) + \\ &+ 4\Phi\partial^i\partial_j\Phi + 2\partial^i\Phi\partial_j\Phi + \delta^i_j(\text{diagonal terms})\Bigg] . \end{aligned} \tag{4.79}$$

Likewise, the energy-momentum tensor at second order reads

$$T^{(2)i}{}_j(\mathcal{O}(\Phi)) = (\overline{\rho} + \overline{p})v^i v_j + \delta p \Pi^i_j \ , \tag{4.80}$$

where v, Π, and δp are the velocity, anisotropic stress, and pressure perturbations at first order, respectively. Neglecting anisotropic stress implies $\Pi = 0$, and the velocity perturbation is given by

$$v_i = -\frac{2M_{\rm pl}^2}{a^2(\overline{\rho} + \overline{p})}\partial_i\left(\Phi' + \mathcal{H}\Phi\right) \ , \tag{4.81}$$

which implies that

$$T^{(2)i}{}_j(\mathcal{O}(\Phi)) = \frac{M_{\rm pl}^2}{a^2}\frac{4}{3(w+1)\mathcal{H}^2}\partial^i\left(\Phi' + \mathcal{H}\Phi\right)\partial_j\left(\Phi' + \mathcal{H}\Phi\right) \ , \tag{4.82}$$

where we have used the background EOS from Eq. (4.7). If one chooses to look at the GWs in the TT gauge, one can define a projection tensor $\mathcal{T}_{ij}^{lm}$, which acts on the spatial part of the Einstein equation and selects the transverse-traceless part of the tensor perturbations

$$\mathcal{T}_{ij}^{lm} G^{(2)i}{}_j = \frac{1}{M_{\rm pl}^2}\mathcal{T}_{ij}^{lm} T_{lm}^{(2)} \ . \tag{4.83}$$

Putting Eq. (4.79) and Eq. (4.82) together, one finds the equation of motion for the second-order tensor perturbations:

$$h_{ij}''(\eta,\boldsymbol{x}) + 2\mathcal{H}h_{ij}'(\eta,\boldsymbol{x}) - \nabla^2 h_{ij}(\eta,\boldsymbol{x}) = -4\mathcal{T}_{ij}^{lm}\mathcal{S}_{lm}(\eta,\boldsymbol{x}) \ , \tag{4.84}$$

where there is a source term $\mathcal{S}_{ij}$ on the right that is a quadratic function of the first order-scalar perturbations

$$\mathcal{S}_{ij}(\eta,\boldsymbol{x}) = 4\Phi\partial_i\partial_j\Phi + 2\partial_i\Phi\partial_j\Phi \ - \frac{4}{3(w+1)}\partial_i\left(\frac{\Phi'}{\mathcal{H}} + \Phi\right)\partial_j\left(\frac{\Phi'}{\mathcal{H}} + \Phi\right) \ . \tag{4.85}$$

4.2.2 Solving the equations of motion

Inspecting Eq. (4.84), we see that the EOM is of the same type as the one for the tensor perturbations at first order, with the addition of the source term $\mathcal{S}_{ij}$. As before, it is therefore useful to introduce the Fourier modes h_k:

$$h_{ij}(\eta, \boldsymbol{x}) = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3} e^{i\boldsymbol{k}\cdot\boldsymbol{x}} [\mathbf{e}_{ij\,+}(\boldsymbol{k}) h_{k\;+}(\eta) + \mathbf{e}_{ij\,\times}(\boldsymbol{k}) h_{k\;\times}(\eta)] \;, \tag{4.86}$$

where $\mathbf{e}_{ij\,+,\times}$ are the polarization tensors for the $+$ and $\times$ modes, respectively. As the polarizations have the same EOM, it is sufficient to consider either, and denote it with the subscript s. The projected source term defines the Fourier-transformed quantity analogously

$$\mathcal{T}_{ij}^{lm} \mathcal{S}_{lm\,\mathrm{s}} = \int \frac{\mathrm{d}^3\boldsymbol{k}}{(2\pi)^3} \mathbf{e}_{ij\,\mathrm{s}}(\boldsymbol{k}) \mathcal{S}_s(\eta, \boldsymbol{k}) \;. \tag{4.87}$$

This turns Eq. (4.84) into

$$h_k''(\eta, \boldsymbol{k}) + 2\mathcal{H} h_k'(\eta, \boldsymbol{k}) + k^2 h_k(\eta, \boldsymbol{k}) = 4\mathcal{S}_s(\eta, \boldsymbol{k}) \;. \tag{4.88}$$

The source term $\mathcal{S}_s$ is constructed from the first-order scalar perturbation and its derivative and reads

$$\begin{aligned} \mathcal{S}_s(\eta, \boldsymbol{k}) = \int \frac{\mathrm{d}^3\boldsymbol{p}}{(2\pi)^3} \mathbf{e}_{lm\,\mathrm{s}} p^l p^m \frac{1}{3(1+w)} \times \\ \times \left[2(5w+3)\Phi_{\boldsymbol{p}}\Phi_{\boldsymbol{k}-\boldsymbol{p}} + \frac{4}{\mathcal{H}} \left(\Phi_{\boldsymbol{p}}\Phi_{\boldsymbol{k}-\boldsymbol{p}}' + \Phi_{\boldsymbol{p}}'\Phi_{\boldsymbol{k}-\boldsymbol{p}} \right) + \frac{4}{\mathcal{H}^2} \Phi_{\boldsymbol{p}}'\Phi_{\boldsymbol{k}-\boldsymbol{p}}' \right] \;, \end{aligned} \tag{4.89}$$

where we switched to spherical coordinates (p, θ, ϕ) and aligned the axes such that they form a basis $(\mathbf{e}_+, \mathbf{e}_\times, \mathbf{e}_z)$ with $\mathbf{e}_z$ pointing in the direction of $\boldsymbol{k}$. This equation can be compactified by introducing the *transfer function* $T(\eta, k)$, which encapsulates the time evolution of Φ, converging to 1 in the asymptotic past, and which relates the time evolution of comoving scalar perturbations to the Bardeen potential via

$$\Phi(\eta, k) = \frac{3+3w}{5+3w} T(\eta, k) \zeta(k) \;. \tag{4.90}$$

Defining the function $f(|\boldsymbol{k}-\boldsymbol{p}|, p, \eta)$, which absorbs the transfer functions, and $Q_{\mathrm{s}}(\boldsymbol{k}, \boldsymbol{p}) = \mathbf{e}_{lm\,\mathrm{s}} p^l p^m$, which absorbs the polarization tensor, allows us to re-write the source term in terms of the comoving scalar perturbations

$$\mathcal{S}_s(\eta, \boldsymbol{k}) = \int \frac{\mathrm{d}^3\boldsymbol{p}}{(2\pi)^3} Q_{\mathrm{s}}(\boldsymbol{k}, \boldsymbol{p}) f(|\boldsymbol{k}-\boldsymbol{p}|, p, \eta) \zeta(\boldsymbol{p}) \zeta(\boldsymbol{k}-\boldsymbol{p}) \;. \tag{4.91}$$

We see that the scalar perturbations enter the source term at quadratic order, just as expected. Furthermore, f is a function of T, T' and $\mathcal{H}$ and is given by

$$f(|\boldsymbol{k}-\boldsymbol{p}|,p,\eta) = \frac{3(1+w)}{(5w+3)^2}\Bigg[2(5w+3)T(p,\eta)T(|\boldsymbol{k}-\boldsymbol{p}|,\eta)+ \\ + \frac{4}{\mathcal{H}^2}T'(p,\eta)T'(|\boldsymbol{k}-\boldsymbol{p}|,\eta) + \frac{4}{\mathcal{H}}\Bigg(T(p,\eta)T'(|\boldsymbol{k}-\boldsymbol{p}|,\eta)+ \\ + T'(p,\eta)T(|\boldsymbol{k}-\boldsymbol{p}|,\eta)\Bigg)\Bigg] \ . \tag{4.92}$$

The function f encapsulates the time evolution of the scalar perturbations and therefore encodes the equation of state that the second-order gravitational waves see as they re-enter the Hubble sphere. The implications of this are discussed in Paper IV.

The solution to the Fourier-transformed inhomogeneous equation (4.88) is most easily found by using the Green's function method. The goal is to find the Green's function $G(\eta,\bar{\eta})$ solving the homogeneous equation

$$G''(\eta,\bar{\eta}) + \left(k^2 - \frac{a''}{a}\right)G(\eta,\bar{\eta}) = \delta(\eta-\bar{\eta}) \ , \tag{4.93}$$

with the boundary conditions $G(\eta,\eta)=0$ and $G'(\eta,\eta)=1$. The Green's function is well known and, in the general case, yields

$$G_k(\eta,\bar{\eta}) = k\cdot\eta\bar{\eta}[j_\nu(k\bar{\eta})y_\nu(k\eta) - j_\nu(k\eta)y_\nu(k\bar{\eta})] \ , \tag{4.94}$$

where j_ν and y_ν are the spherical Bessel functions of the first and second kind, respectively, $\nu = 3(1-w)/(2+6w)$, and $k=|\boldsymbol{k}|$. The solution to the inhomogeneous equation can be written as

$$h_{\mathrm{s}}(\boldsymbol{k},\eta) = \int_{\eta_i}^{\eta} \mathrm{d}\bar{\eta}\, G_k(\eta,\bar{\eta})\frac{a(\bar{\eta})}{a(\eta)}\mathcal{S}_{\mathrm{s}}(\bar{\eta}) \ , \tag{4.95}$$

where η_i is the time of emission.

Like for the first-order tensor modes, the energy density of GWs is related to the amplitude of the tensor perturbations via the two-point function

$$\langle h_{ij}(\boldsymbol{k},\eta)h_{ij}(\bar{\boldsymbol{k}},\eta)\rangle = (2\pi)^3\delta^{(3)}(\boldsymbol{k}+\bar{\boldsymbol{k}})\frac{2\pi^2}{k^3}\mathcal{P}_h(k,\eta) \ , \tag{4.96}$$

which explicitly written out yields

$$\langle h_{ij}(\boldsymbol{k},\eta)h_{ij}(\bar{\boldsymbol{k}},\eta)\rangle = 16\int\frac{\mathrm{d}^3\boldsymbol{p}_1}{(2\pi)^3}\frac{\mathrm{d}^3\boldsymbol{p}_2}{(2\pi)^3}Q_{\mathrm{s}_1}(\boldsymbol{k}_1,\boldsymbol{p}_1)Q_{\mathrm{s}_2}(\boldsymbol{k}_2,\boldsymbol{p}_2)\times \\ \times f(|\boldsymbol{k}_1-\boldsymbol{p}_1|,p_1,\eta)f(|\boldsymbol{k}_2-\boldsymbol{p}_2|,p_2,\eta)\langle\zeta(\boldsymbol{p}_1)\zeta(\boldsymbol{k}_1-\boldsymbol{p}_1)\zeta(\boldsymbol{p}_2)\zeta(\boldsymbol{k}_2-\boldsymbol{p}_2)\rangle \ . \tag{4.97}$$

The last term is the four-point function of the scalar perturbations ζ, which can be decomposed into disconnected (products of two-point functions) and connected parts. In the Gaussian case, the connected parts vanish, and the power spectrum of the tensor perturbations can be expressed as a function of the power spectrum of the scalar perturbations $\mathcal{P}_\zeta$. If the connected parts are taken into account as well, there is an additional term called the *trispectrum* [145, 146, 147]. For most inflationary models, this part is subdominant with respect to the disconnected part. The implications of including this part are discussed in Paper IV.

To simplify the equations, it makes sense to introduce the *interaction kernel* $I(|\boldsymbol{k}-\boldsymbol{p}|, p, \eta)$, which absorbs the transfer functions and the polarization tensor

$$I(|\boldsymbol{k}-\boldsymbol{p}|, p, \eta) = \int_{\eta_i}^{\eta} \mathrm{d}\bar{\eta}\, G_k(\eta, \bar{\eta}) \frac{a(\bar{\eta})}{a(\eta)} f(|\boldsymbol{k}-\boldsymbol{p}|, p, \bar{\eta}) \ , \tag{4.98}$$

and the dimensionless momentum variables

$$u = \frac{|\boldsymbol{k}-\boldsymbol{p}|}{k}, \qquad v = \frac{p}{k}. \tag{4.99}$$

The power spectrum of the tensor perturbations can then be written as

$$\overline{\mathcal{P}_h(k,\eta)} = 4\int_0^\infty \mathrm{d}v \int_{|1-v|}^{1+v} \mathrm{d}u \left(\frac{4v^2-(1+v^2-u^2)^2}{4vu}\right)^2 \times \\ \times \overline{I^2(v,u,k,\eta)} \mathcal{P}_\zeta(kv)\mathcal{P}_\zeta(ku) \ , \tag{4.100}$$

where the overline denotes the time average over multiple oscillations. The need for this comes from the definition of GW energy density. Computing this double integral efficiently enough to be able to infer the parameters governing $\mathcal{P}_\zeta$ is one of the main objectives of Paper IV.

Once the power spectrum of the tensor perturbations is known, it is easy to compute the fractional energy density per logarithmic wavenumber interval Ω_{GW} at some time after emission η_f:

$$\Omega_{\mathrm{GW}}(k,\eta_f) = \frac{\rho_{\mathrm{GW}}(k,\eta_f)}{\rho_c(\eta_f)} = \frac{1}{24}\left(\frac{k}{\mathcal{H}(\eta_f)}\right)^2 \overline{\mathcal{P}_h(k,\eta_f)} \ , \tag{4.101}$$

where ρ_c is the critical energy density of the Universe.

As GWs are massless they decouple early and then redshift as radiation. The fractional energy density today is given by accounting for entropy injections as particles become non-relativistic:

$$\Omega_{\mathrm{GW}}(k)h^2 = \Omega_{r,0}h^2 \frac{g_*(\eta_f)}{g_*^0}\left(\frac{g_{*,s}^0}{g_{*,s}(\eta_f)}\right)^{4/3} \Omega_{\mathrm{GW}}(k,\eta_f) \\ = \frac{c_g(\eta_f)}{24}\Omega_{r,0}h^2 \left(\frac{k}{\mathcal{H}(\eta_f)}\right)^2 \overline{\mathcal{P}_h(k,\eta_f)} \ , \tag{4.102}$$

where g_* and $g_{*,s}$ are the effective number of relativistic degrees of freedom in energy and entropy, respectively. The subscript 0 denotes the value today. $\Omega_{r,0}$ is the fractional energy density of radiation today, and $h = H_0/100\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ is the reduced Hubble constant that we are using due to the large uncertainty (and tension) in the value of H_0. The radiation density today has been measured very accurately by Planck and is $\Omega_{r,0}h^2 = 4.2 \times 10^{-5}$ assuming massless neutrinos [141] and $c_g(\eta_f) \simeq 0.39$ for η_f corresponding to signals in LISA assuming standard model degrees of freedom [122].

4.2.3 Effect of re-entry during different cosmological eras

The production of second-order gravitational waves depends on when the scalar perturbations re-enter the horizon. There are two key scenarios to consider: re-entry during the radiation-dominated era and re-entry during an early matter-dominated era.

Radiation-dominated Universe In the radiation-dominated (RD) Universe, the Green's function solution for the gravitational wave (GW) equation is [148]

$$\begin{aligned} kG_{\boldsymbol{k}}(\eta,\bar{\eta}) &= x\bar{x}\left[j_0(x)y_0(\bar{x}) - j_0(\bar{x})y_0(x)\right] \\ &= \sin(x-\bar{x})\ , \end{aligned} \tag{4.103}$$

where $x = k\eta$ and $\bar{x} = k\bar{\eta}$. The solution for the gravitational potential Φ is given by:

$$\Phi(x) = \frac{9}{x^2}\left(\frac{\sin(x/\sqrt{3})}{x/\sqrt{3}} - \cos(x/\sqrt{3})\right). \tag{4.104}$$

This potential decays as x^{-2} at large x. The factor $x/\sqrt{3}$ corresponds to the sound horizon in the RD era.

The source function f_{RD} in the RD era is derived using the gravitational potential Φ and its derivatives:

$$\begin{aligned} f_{\mathrm{RD}}(v,u,x) =& \frac{12}{u^3v^3x^6}\Bigg[18uvx^2\cos\frac{ux}{\sqrt{3}}\cos\frac{vx}{\sqrt{3}} \\ &+ \left(54 - 6(u^2+v^2)x^2 + u^2v^2x^4\right)\sin\frac{ux}{\sqrt{3}}\sin\frac{vx}{\sqrt{3}} \\ &+ 2\sqrt{3}ux(v^2x^2-9)\cos\frac{ux}{\sqrt{3}}\sin\frac{vx}{\sqrt{3}} \\ &+ 2\sqrt{3}vx(u^2x^2-9)\sin\frac{ux}{\sqrt{3}}\cos\frac{vx}{\sqrt{3}}\Bigg], \end{aligned} \tag{4.105}$$

which is equal to $4/3$ at $x = 0$ and decays like $\sim 1/(uvx^2)$ at large x. Using the Green's function, the integral $I_{\rm RD}(v, u, x)$ is given by:

$$I_{\rm RD}(v,u,x) = \int_0^x \mathrm{d}\bar{x}\frac{\bar{x}}{x}\sin(x-\bar{x})f_{\rm RD}(v,u,\bar{x}). \tag{4.106}$$

To evaluate this integral, one uses trigonometric identities and integration by parts. After simplification, the result is:

$$\begin{aligned} I_{\rm RD}(v,u,x) = \frac{3}{4u^3v^3x}\Bigg[&-\frac{4}{x^3}\Big(uv(u^2+v^2-3)x^3\sin x - 6uvx^2\cos\tfrac{ux}{\sqrt{3}}\cos\tfrac{vx}{\sqrt{3}} \\ &+6\sqrt{3}ux\cos\tfrac{ux}{\sqrt{3}}\sin\tfrac{vx}{\sqrt{3}} + 6\sqrt{3}vx\sin\tfrac{ux}{\sqrt{3}}\cos\tfrac{vx}{\sqrt{3}} \\ &-3(6+(u^2+v^2-3)x^2)\sin\tfrac{ux}{\sqrt{3}}\sin\tfrac{vx}{\sqrt{3}}\Big) \\ &+(u^2+v^2-3)^2\Bigg(\sin x\Big(\mathrm{Ci}\left(\left(1-\tfrac{v-u}{\sqrt{3}}\right)x\right) + \mathrm{Ci}\left(\left(1+\tfrac{v-u}{\sqrt{3}}\right)x\right) \\ &-\mathrm{Ci}\left(\left|1-\tfrac{v+u}{\sqrt{3}}\right|x\right) - \mathrm{Ci}\left(\left(1+\tfrac{v+u}{\sqrt{3}}\right)x\right) + \log\left|\tfrac{3-(u+v)^2}{3-(u-v)^2}\right|\Big) \\ &+\cos x\Big(-\mathrm{Si}\left(\left(1-\tfrac{v-u}{\sqrt{3}}\right)x\right) - \mathrm{Si}\left(\left(1+\tfrac{v-u}{\sqrt{3}}\right)x\right) \\ &+\mathrm{Si}\left(\left(1-\tfrac{v+u}{\sqrt{3}}\right)x\right) + \mathrm{Si}\left(\left(1+\tfrac{v+u}{\sqrt{3}}\right)x\right)\Big)\Bigg)\Bigg], \end{aligned} \tag{4.107}$$

where $\mathrm{Si}(x)$ and $\mathrm{Ci}(x)$ are the sine and cosine integral functions, respectively:

$$\mathrm{Si}(x) = \int_0^x \frac{\sin\bar{x}}{\bar{x}}d\bar{x}, \qquad \mathrm{Ci}(x) = -\int_x^\infty \frac{\cos\bar{x}}{\bar{x}}d\bar{x}. \tag{4.108}$$

At late times ($x \to \infty$), $I_{\rm RD}$ simplifies to:

$$\begin{aligned} I_{\rm RD}(v,u,x\to\infty) \approx \frac{3(u^2+v^2-3)}{4u^3v^3x}\Bigg[\sin x\Big(&-4uv + (u^2+v^2-3)\times \\ &\times\log\left|\frac{3-(u+v)^2}{3-(u-v)^2}\right|\Big) - \pi(u^2+v^2-3)\Theta(v+u-\sqrt{3})\cos x\Bigg]. \end{aligned} \tag{4.109}$$

The oscillation-averaged value of $I^2_{\rm RD}$ is given by:

$$\begin{aligned} \overline{I^2_{\rm RD}(v,u,x\to\infty)} = \frac{1}{2}\left(\frac{3(u^2+v^2-3)}{4u^3v^3x}\right)^2\Bigg[\Big(&-4uv + (u^2+v^2-3)\times \\ &\times\log\left|\frac{3-(u+v)^2}{3-(u-v)^2}\right|\Big)^2 + \pi^2(u^2+v^2-3)^2\,\Theta(v+u-\sqrt{3})\Bigg]. \end{aligned} \tag{4.110}$$

Matter domination In the matter-dominated (MD) era, the Green's function solution for the gravitational wave (GW) equation is [148]

$$\begin{aligned} kG_{\boldsymbol{k}}(\eta,\bar{\eta}) &= x\bar{x}\left[j_1(x)y_1(\bar{x}) - j_1(\bar{x})y_1(x)\right] \\ &= \frac{1}{x\bar{x}}\left[(1+x\bar{x})\sin(x-\bar{x}) - (x-\bar{x})\cos(x-\bar{x})\right] \end{aligned} \tag{4.111}$$

and the Bardeen potential is simply constant

$$\Phi(x) = 1 \ , \tag{4.112}$$

which, after some manipulation, yields

$$\overline{I^2_{\mathrm{MD}}(v,u,x\to\infty)} = \frac{18}{25} \tag{4.113}$$

Note that modes within the LISA band can only reenter during matter domination if it precedes radiation domination, as could be the case if initially the energy density is dominated by some pressureless quantity. This could, e.g., be the case if *(a)* at the end of inflation the inflaton becomes massive and slow ($H \ll m$), behaving like a pressureless fluid before reheating, or *(b)* some weakly coupled heavy particles dominate the early Universe in an extension of the standard model [149]. Furthermore, it is clear from Eq. (4.101) that if $\overline{\mathcal{P}_h}$ does not drop faster than $1/k^2$ at large k, the energy density of the GWs diverges. Interestingly, the MD to RD transition produces an enhancement of the GWs if it happens fast enough. The details are discussed in Paper IV.

4.3 Constraints on gravitational waves from inflation

Inflationary models predict a stochastic background of GWs spanning a wide range of frequencies, depending on the scale of inflation and the nature of the transition to the post-inflationary Universe. Various observational probes place stringent constraints on the amplitude and energy density of these GWs. The primary constraints arise from the CMB B-mode polarization, bounds on the effective number of extra neutrino species from the CMB temperature anisotropies and Big Bang nucleosynthesis, and primordial black holes. Further constraints are put by direct GW detection experiments. This section discusses these constraints in detail.

4.3.1 Cosmic Microwave Background

As already discussed in Section 4.1.5, the CMB puts strong bounds on the first-order scalar and tensor perturbations at a scale of $\sim 0.05\,\mathrm{Mpc}^{-1} \sim 5\cdot 10^{-16}\,\mathrm{s}^{-1}$. The scalar perturbations enter the CMB as temperature anisotropies, while the tensor modes enter as B-mode polarization. Planck

tightly constrains the amplitude of the scalar perturbations to be $A_s = (2.10 \pm 0.03) \times 10^{-9}$ [141]. The tensor-to-scalar ratio r is constrained to be less than 0.036 at 95% confidence level by BICEP2/Keck [142]. This yields an upper bound on first-order tensor perturbations and an estimate of the second-order contribution at this scale. Reminding ourselves of Eq. (4.102), the first-order tensor contribution is given by

$$\Omega_{\mathrm{GW}}^{(1)}(k)h^2 = \frac{c_g(\eta_f)}{24}\Omega_{r,0}h^2 \cdot \mathcal{P}_\zeta(k)r \lesssim 10^{-16} \,, \tag{4.114}$$

whereas the second-order contribution reads [148]

$$\Omega_{\mathrm{GW}}^{(2)}(k)h^2 = c_g(\eta_f)\Omega_{r,0}h^2 Q(n_s) \cdot A_s^2 \left(\frac{k}{k_*}\right)^{2n_s-2} \sim 10^{-22} \,, \tag{4.115}$$

where $Q(n_s) = 0.8149$ for the Planck value of $n_s = 0.9655$. The second-order contribution is suppressed by a factor of $\sim 10^6$ compared to the first order contribution and is therefore entirely irrelevant if there is no enhancement in $\mathcal{P}_\zeta$. Furthermore, if we assume a pure power law spectrum as shown in Fig. 4.1, both first- and second-order contributions remain well outside the sensitivity of Pulsar Timing Arrays, space-based observatories such as LISA, as well as current and next-generation ground-based observatories. On the other hand, if there is an enhancement of the scalar perturbations at small scales (i.e., due to a USR phase), the second-order contribution can generate signals that are strong enough to be detected by either of those observatories. The first-order tensor perturbations remain virtually unphased by such an enhancement.

Figure 4.1 indicates the BICEP2/Keck constraint on the tensor-to-scalar ratio r (gray arrow) and shows the sensitivity curve for the planned LiteBIRD space mission [150] that will offer better measurements of the B-mode polarization, thus providing tighter bounds on r.

4.3.2 Effective number of neutrino species

As discussed previously, GWs are relativistic species and therefore contribute to the energy density of the Universe as radiation. This makes processes that depend on the background evolution of the Universe sensitive to the abundance of GWs after their production. The energy density of radiation in the Universe is given by

$$\rho_r = \frac{\pi^2}{30} g_*(T) T^4, \tag{4.116}$$

where g_* is the effective number of relativistic degrees of freedom. Photons contribute $g_{*,\gamma} = 2$ and neutrinos (together with their antiparticles) contribute $g_{*,\eta} = 7/8 \cdot 2N_\nu$ to g_*, where $N_\nu = 3$ is the number of neutrino

species in the standard model. Therefore, if one treats the GW radiation like an addition of ΔN_ν to the number of neutrino species[5], the added density $\Delta\rho_r$ is given by [151]

$$\Delta\rho_r = \frac{\pi^2}{30}\frac{7}{4}\Delta N_\nu T^4 \ . \tag{4.117}$$

Assuming that only GWs contribute to extra radiation, this means that

$$\left(\frac{\rho_{\rm GW}}{\rho_\gamma}\right)_{T\sim{\rm MeV}} \leq \frac{7}{8}\Delta N_\nu \ , \tag{4.118}$$

where $\rho_{\rm GW}$ is the energy density of the GWs and ρ_γ is the energy density of the photons. The temperature of $T \sim$ MeV stems from the fact that it is easier to define the quantity before neutrinos decouple. Evolving Eq. (4.118) to today, one arrives at

$$h^2\left(\frac{\rho_{\rm GW}}{\rho_c}\right)_0 \leq h^2\Omega_{\gamma,0}\left(\frac{g_s(T_0)}{g_s(T\sim{\rm MeV})}\right)^{4/3}\frac{7}{8}\Delta N_\nu = 5.6\cdot 10^{-6}\Delta N_\nu \ , \tag{4.119}$$

and

$$\left(\frac{\rho_{\rm GW}}{\rho_c}\right)_0 = \int_{f_H}^{\infty}\frac{{\rm d}f}{f}h^2\Omega_{\rm GW}(f) \ , \tag{4.120}$$

where f_H is the frequency corresponding to the Hubble horizon at the time t where the imprint of GWs is happening $f_H = (H_t/2\pi)(a_t/a_0)$, as GWs can only contribute to radiation if their wavelength is within the Hubble sphere.

ΔN_ν is best constrained by two measurements:

- The abundance of light elements produced during Big Bang Nucleosynthesis (BBN) happening at $T_{\rm BBN} \sim 0.1\,{\rm MeV}$ puts a constraint of $(h^2\rho_{\rm GW}/\rho_c)_0 < 1.12\cdot 10^{-6}$ for $f \gtrsim 1.5\cdot 10^{-12}\,{\rm Hz}$ [151, 152].
- The CMB, which decouples at $T_{\rm CMB} \sim 0.3\,{\rm eV}$, offers less stringent but broader bounds at $(h^2\rho_{\rm GW}/\rho_c)_0 < 6.9\cdot 10^{-6}$ for $f \gtrsim 1.5\cdot 10^{-15}\,{\rm Hz}$ as CMB decoupling happens later than BBN [151, 153, 154, 155].

Except for cases in which the spectrum of GWs is a very narrow peak, we can interpret the bound on $(h^2\rho_{\rm GW}/\rho_c)$ as a bound on the fractional energy density of GWs $\Omega_{\rm GW}h^2$. These limits are indicated as the light blue band in Fig. 4.1.

[5]The treatment as an extra number of neutrino species as opposed to an extra number of, say, photons is purely due to convention. We could just as well define a new parameter ΔN_γ and treat the GWs as an addition to the photon density. That would in fact be closer to the truth as gravitons are massless while neutrinos eventually freeze out.

4.3.3 Direct detection

Of course it is possible to constrain primary and secondary tensor modes by directly detecting the associated GWs. Currently, the two methods for achieving this are Pulsar Timing Arrays (PTAs), which use the timing residuals of millisecond pulsars to detect GWs in the nano-Hertz regime [84], and ground-based detectors such as LIGO, Virgo, and Kagra, which are sensitive to GWs in the audio band ($\sim 100\,\mathrm{Hz}$). By searching for continuous, stochastic signals in these detectors, it is possible to put constraints on the energy density of GWs at the frequencies they are sensitive to [159, 160]. In particular, as observation time increases, the power law integrated sensitivity of Michelson interferometer experiments like LVK and LISA scales as $T^{1/2}$, where T is the observation time. For PTAs, this scaling is time-dependent [161].

Recently, PTA measurements have hinted at a measurement of stochastic GWs in the nHz regime, but it is yet to be determined whether this signal is of astrophysical or cosmological origin [162].

Figure 4.1 shows the sensitivity curves for the most competitive contemporary detectors, the International Pulsar Timing Array (IPTA) [163, 164, 165, 166, 167] and the Advanced Laser Interferometer GW Observatory (aLIGO) [168]. Future detectors such as the Square Kilometre Array (SKA) [169], the Einstein Telescope (ET) [170], and LISA [80] will further extend the reach of GW detection. To create this figure, a 4-year mission duration for LISA and a 10-year observation period for aLIGO, ET, and SKA were assumed. The sensitivity for IPTA corresponds to the observation time to date.

Determining the sensitivity of LISA to SIGWs coming from different models of inflation and determining whether they can be distinguished from other sources of stochastic gravitational waves is the purpose of Paper IV.

4.3.4 Primordial Black Holes

Primordial black holes (PBHs) [171, 172] are one of the main motivators for assuming that the scalar perturbations are enhanced at small scales, as if the $\mathcal{P}_\zeta$ generated during inflation is large enough, the perturbations generate overdensities as they re-enter the Hubble horizon. If these overdensities are dense enough with respect to the surrounding region, their gravitational pull overcomes their pressure, and they collapse into black holes. The mass of these black holes is directly related to the scale of the perturbations that generated it, meaning that SIGWs can provide a direct

constraint on the abundance of PBHs if measured and coming from large overdensities.

A full discussion of the details that go into calculating the abundance of PBHs and their measurements is far outside the scope of this work (see e.g., [173, 174] for reviews), but we very schematically outline the main points. The abundance of PBHs is determined by the amplitude of the curvature perturbations at scales corresponding to the horizon size at re-entry. To see how a curved region can collapse into a black hole, it is useful to look at Eq. (4.5). Since we are looking at the moment when the perturbation re-enters the Hubble horizon, we can define a locally curved universe with curvature $K(r)$ (not to be confused with the wavenumber k), which gives us the Friedmann equation:

$$H^2 = \frac{1}{3M_{\rm pl}^2}\rho - \frac{K}{a^2} \ . \tag{4.121}$$

By introducing the density contrast of a perturbed region on the comoving hypersurface with respect to the background

$$\delta \equiv \frac{\rho - \bar{\rho}}{\bar{\rho}} \ , \tag{4.122}$$

it is easy to relate this quantity to the curvature through

$$\delta = \frac{3M_{\rm pl}^2 K}{\bar{\rho}a^2} = \frac{K}{a^2H^2} \ . \tag{4.123}$$

Assuming radiation domination $\bar{\rho} \propto a^{-4}$, and ignoring the spatial dependence of K, this "mini universe" would eventually collapse in on itself if $K > 0$, which happens when $3K/a^2 = \rho/M_{\rm pl}^2$, i.e., when the comoving scale of this region is equal to the Hubble horizon and when $\delta = 1$[6]. Translating this to the wavenumber of the perturbation k and denoting the time at which a curvature perturbation with wavenumber k reenters as t_k, one arrives at

$$\delta(t_k) = \frac{K}{H^2(t_k)a^2(t_k)} \gtrsim \delta_c \tag{4.124}$$

where δ_c is the critical threshold for PBH formation. If the curvature perturbations exceed this threshold, PBHs form with a mass at formation of [173]

$$m_{\rm PBH}(k = aH) \sim \gamma \frac{4\pi}{3}\frac{\rho}{H^3}\bigg|_{k=aH} = \gamma m_H \ , \tag{4.125}$$

[6]The approximation as a mini universe breaks down when the scale of the perturbation becomes smaller than the Hubble sphere, but for an intuitive order of magnitude estimate, this treatment is still instructive.

where γ is an efficiency factor and m_H is the mass contained in the Hubble sphere. We assume that the overdensities collapse soon after they reenter the horizon.

The Hubble mass can be related to the scale of the comoving wavenumber k through [175]

$$m_H \simeq 10 M_\odot \left(\frac{g_*}{106.75}\right)^{1/2} \left(\frac{106.75}{g_{*,s}}\right)^{2/3} \left(\frac{10^6\,\mathrm{Mpc}^{-1}}{k/\kappa}\right)^2 , \tag{4.126}$$

where $\kappa = k/aH$ and aH is evaluated at Hubble crossing. At scales corresponding to LISA $k \sim$ mHz, the mass of the PBHs is in the range of $\sim 10^{-12}\,M_\odot$ or $\sim 10^{21}$ g, which corresponds to the asteroid mass range.

The abundance of PBHs at formation is then given by the fraction of the Universe's energy density contained in PBHs,

$$\beta = \frac{\rho_{\mathrm{PBH}}}{\rho_{\mathrm{tot}}} , \tag{4.127}$$

where ρ_{tot} is the total energy density of the Universe.

The abundance of PBHs is typically measured by constraining the ratio of the PBH mass density to the total dark matter density today $f_{\mathrm{PBH}} = \Omega_{\mathrm{PBH}}/\Omega_{\mathrm{DM}}$. The constraints on Ω_{PBH} can then be translated into constraints on the amplitude of the power spectrum of curvature perturbations, $\mathcal{P}_\zeta(k)$, and subsequently into a constraint on the SIGWs $\mathcal{P}_h^{(2)}(k)$, at horizon crossing, assuming a certain formation mechanism, evolution, and a fixed thermal history.

After PBHs form, they behave like any other black hole population, which, on large scales, acts like a pressureless non-relativistic fluid. They therefore contribute to the energy density of the Universe as matter, which makes PBHs compelling candidates for dark matter, which do not require any modification of the standard model of particle physics.

Figure 4.2 gives an overview of the current constraints on the abundance of PBHs as a function of mass as constrained by different types of measurements: Hawking evaporation, microlensing, direct GW detection, black hole accretion, and galaxy dynamics. The PBH masses that roughly correspond to scales that LISA is sensitive to are indicated by the gray shaded region. It is clear that LISA is well situated in a "sweet spot" where currently there are no constraints on the abundance of PBHs, and they therefore could make up a large fraction of the dark matter.

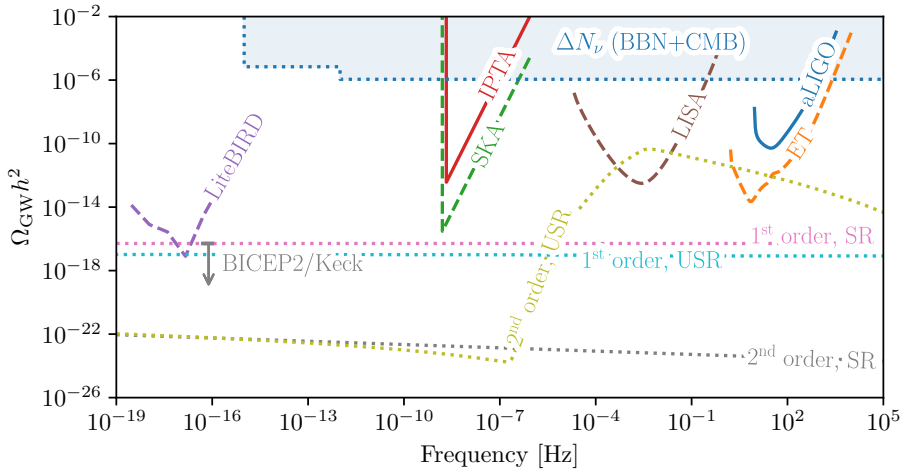


Figure 4.1: Fractional energy density of GWs $\Omega_{\rm GW}h^2$ as a function of frequency for gravitational wave sources generated by inflation and approximate sensitivity curves for various detectors. The dotted lines represent the first- and second-order tensor (scalar-induced) perturbations sourced for two inflationary scenarios: slow roll (SR) assumes a power spectrum for $\mathcal{P}_\zeta$ and a flat $\mathcal{P}_h^{(1)}$ across all scales shown using the Planck 2018 best fit values for A_s, n_s and the BICEP2/Keck upper bound on r. The induced fractional GW energy density at first- and second order is shown in pink and gray, respectively. The ultra-slow roll (USR) scenario assumes an enhancement of the scalar perturbations at small scales using the potential from Paper IV. The region that is excluded by ΔN_ν is marked as the light blue band. The solid lines represent the sensitivity curves for the currently existing IPTA (red) and aLIGO (blue) detectors. The BICEP2/Keck upper bound at $k = 0.05\,\mathrm{Mpc}^{-1}$ is indicated by the gray arrow. The dashed lines represent the sensitivity curves for the future LiteBIRD (purple), SKA (green), LISA (green), and ET (orange) experiments. The detector sensitivity curves have been taken from [156, 157] (LiteBIRD), [99] (IPTA, SKA, aLIGO, and ET), and [158] (LISA). We are assuming a 4-year mission duration for LISA and a 10-year observation period for aLIGO, ET, and SKA.

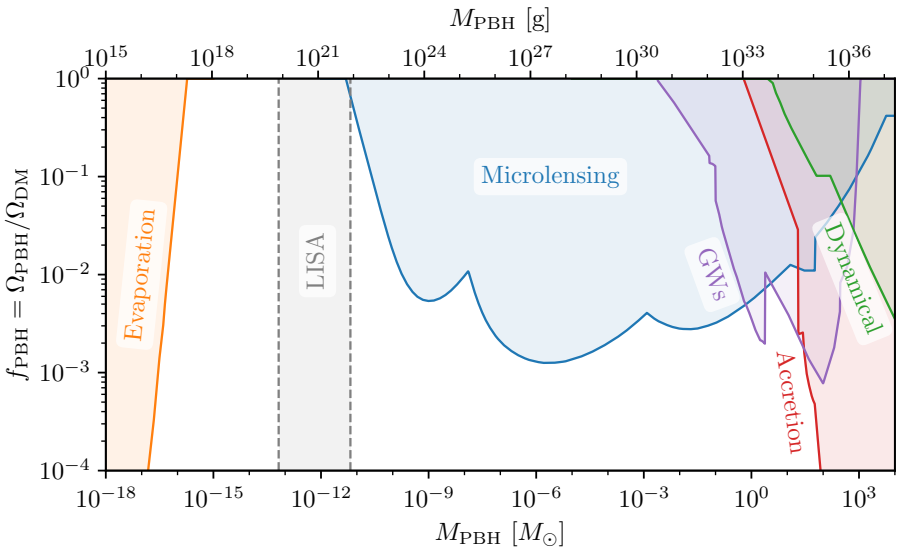


Figure 4.2: Bounds on the PBH abundance as a function of mass. The solid lines represent current constraints from Hawking evaporation (orange), microlensing (blue), direct GW detection (purple), black hole accretion (red), and galaxy dynamics (green). LISA will probe PBH masses at around the $10^{-12}M_\odot$ scale, and the rough scale is represented as the gray shaded region. The bounds are taken from [176].

Bibliography

[1] J. El Gammal, N. Schöneberg, J. Torrado and C. Fidler, *Fast and robust Bayesian inference using Gaussian processes with GPry*, *Journal of Cosmology and Astroparticle Physics* **2023** (2023) 021.

[2] J. Torrado, N. Schöneberg and J. El Gammal, *Parallelized Acquisition for Active Learning using Monte Carlo Sampling*, May, 2023. 10.48550/arXiv.2305.19267.

[3] J. El Gammal, R. Buscicchio, G. Nardini and J. Torrado, *Accelerating LISA inference with Gaussian processes*, Mar., 2025. 10.48550/arXiv.2503.21871.

[4] J. El Gammal, A. Ghaleb, G. Franciolini, T. Papanikolaou, M. Peloso, G. Perna et al., *Reconstructing Primordial Curvature Perturbations via Scalar-Induced Gravitational Waves with LISA*, Jan., 2025. 10.48550/arXiv.2501.11320.

[5] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin, *Bayesian data analysis*, Texts in statistical science series. CRC Press, Taylor and Francis Group, Boca Raton London New York, third edition ed., 2014.

[6] S. H. Jeffreys and S. H. Jeffreys, *The Theory of Probability*, Oxford Classic Texts in the Physical Sciences. Oxford University Press, Oxford, New York, third edition, third edition ed., Aug., 1998.

[7] S. Kullback and R. A. Leibler, *On Information and Sufficiency*, *The Annals of Mathematical Statistics* **22** (1951) 79.

[8] J. Lin, *Divergence measures based on the Shannon entropy*, *IEEE Transactions on Information Theory* **37** (1991) 145.

[9] K. P. Murphy, *Machine Learning - A Probabilistic Perspective*, Adaptive Computation and Machine Learning. MIT Press, Cambridge, 2014.

[10] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, *Equation of State Calculations by Fast Computing Machines*, *The Journal of Chemical Physics* **21** (1953) 1087.

[11] W. K. Hastings, *Monte Carlo Sampling Methods Using Markov Chains and Their Applications*, *Biometrika* **57** (1970) 97.

[12] E. Marinari and G. Parisi, *Simulated Tempering: A New Monte Carlo Scheme*, *Europhysics Letters* **19** (1992) 451.

[13] R. M. Neal, *MCMC Using Hamiltonian Dynamics*, in *Handbook of Markov Chain Monte Carlo*, Chapman and Hall/CRC, (2011).

[14] M. Girolami and B. Calderhead, *Riemann manifold Langevin and Hamiltonian Monte Carlo methods*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73** (2011) 123.

[15] M. Betancourt, *A General Metric for Riemannian Manifold Hamiltonian Monte Carlo*, in *Geometric Science of Information*, F. Nielsen and F. Barbaresco, eds., (Berlin, Heidelberg), pp. 327–334, Springer, 2013, DOI.

[16] M. D. Hoffman and A. Gelman, *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*, *Journal of Machine Learning Research* **15** (2014) 1593.

[17] M. Betancourt, *Identifying the Optimal Integration Time in Hamiltonian Monte Carlo*, Jan., 2016. 10.48550/arXiv.1601.00225.

[18] J. Skilling, *Nested Sampling*, in *AIP Conference Proceedings*, vol. 735, (Garching (Germany)), pp. 395–405, AIP, 2004, DOI.

[19] D. M. Blei, A. Kucukelbir and J. D. McAuliffe, *Variational Inference: A Review for Statisticians*, *Journal of the American Statistical Association* **112** (2017) 859.

[20] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola and L. K. Saul, *An Introduction to Variational Methods for Graphical Models*, in *Learning in Graphical Models*, M. I. Jordan, ed., (Dordrecht), pp. 105–161, Springer Netherlands, (1998), DOI.

[21] J. L. W. V. Jensen, *Sur les fonctions convexes et les inégalités entre les valeurs moyennes*, *Acta Mathematica* **30** (1906) 175.

[22] D. Rezende and S. Mohamed, *Variational Inference with Normalizing Flows*, in *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1530–1538, PMLR, June, 2015, https://proceedings.mlr.press/v37/rezende15.html.

[23] G. Papamakarios, E. Nalisnick, D. J. Rezende, S. Mohamed and B. Lakshminarayanan, *Normalizing Flows for Probabilistic Modeling and Inference*, *Journal of Machine Learning Research* **22** (2021) 1.

[24] H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini and R. Murray-Smith, *Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy*, *Nature Physics* **18** (2022) 112.

[25] L. Acerbi, *Variational Bayesian Monte Carlo*, in *Advances in Neural Information Processing Systems*, vol. 31, Curran Associates, Inc., 2018.

[26] K. Cranmer, J. Brehmer and G. Louppe, *The frontier of simulation-based inference*, *Proceedings of the National Academy of Sciences* **117** (2020) 30055.

[27] D. B. Rubin, *Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician*, *The Annals of Statistics* **12** (1984) 1151.

[28] M. A. Beaumont, *Approximate Bayesian Computation in Evolution and Ecology*, *Annual Review of Ecology, Evolution, and Systematics* **41** (2010) 379.

[29] P. Marjoram, J. Molitor, V. Plagnol and S. Tavar\'e, *Markov chain Monte Carlo without likelihoods*, *Proceedings of the National Academy of Sciences* **100** (2003) 15324.

[30] G. W. Peters, Y. Fan and S. A. Sisson, *On sequential Monte Carlo, partial rejection control and approximate Bayesian computation*, *Statistics and Computing* **22** (2012) 1209.

[31] S. A. Sisson, Y. Fan and M. M. Tanaka, *Sequential Monte Carlo without likelihoods*, *Proceedings of the National Academy of Sciences* **104** (2007) 1760.

[32] I. Goodfellow, Y. Bengio and A. Courville, *Deep learning*, Adaptive computation and machine learning. The MIT press, Cambridge, Mass, 2016.

[33] M. T. Hagan, H. B. Demuth, M. H. Beale and O. D. Jesús, *Neural Network Design*. Martin Hagan, 2014.

[34] I. Kobyzev, S. J. D. Prince and M. A. Brubaker, *Normalizing Flows: An Introduction and Review of Current Methods*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43** (2021) 3964.

[35] J. Lamperti, *Stochastic Processes: A Survey of the Mathematical Theory*, no. v.23 in Applied Mathematical Sciences Ser. Springer New York, New York, NY, 1997.

[36] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, Adaptive computation and machine learning. MIT Press, Cambridge, Mass, 2006.

[37] B. K. Øksendal, *Stochastic differential equations: an introduction with applications*, Universitext. Springer, Berlin Heidelberg New York Dordrecht London, sixth edition, sixth corrected printing ed., 2013.

[38] N. A. C. Cressie, ed., *Statistics for spatial data*, Wiley series in probability and mathematical statistics Applied probability and statistics. Wiley, New York, rev. ed ed., 2010.

[39] D. Duvenaud, *Automatic model construction with Gaussian processes*, Ph.D. thesis, University of Cambridge, Nov., 2014.

[40] A. Wilson and R. Adams, *Gaussian Process Kernels for Pattern Discovery and Extrapolation*, in *Proceedings of the 30th International Conference on Machine Learning*, pp. 1067–1075, PMLR, May, 2013, https://proceedings.mlr.press/v28/wilson13.html.

[41] M. Osborne, R. Garnett, Z. Ghahramani, D. K. Duvenaud, S. J. Roberts and C. Rasmussen, *Active Learning of Model Evidence Using Bayesian Quadrature*, in *Advances in Neural Information Processing Systems*, vol. 25, Curran Associates, Inc., 2012.

[42] B. Settles, *Active Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning. Springer International Publishing, Cham, 2012, 10.1007/978-3-031-01560-1.

[43] A. Lewis and S. Bridle, *Cosmological parameters from CMB and other data: A Monte Carlo approach*, *Physical Review D* **66** (2002) 103511.

[44] A. Lewis, *Efficient sampling of fast and slow cosmological parameters*, *Physical Review D* **87** (2013) 103529.

[45] D. R. Jones, *A Taxonomy of Global Optimization Methods Based on Response Surfaces*, *Journal of Global Optimization* **21** (2001) 345.

[46] T. Desautels, A. Krause and J. W. Burdick, *Parallelizing Exploration-Exploitation Tradeoffs in Gaussian Process Bandit Optimization*, *Journal of Machine Learning Research* **15** (2014) 4053.

[47] J. Albers, C. Fidler, J. Lesgourgues, N. Schöneberg and J. Torrado, *CosmicNet. Part I. Physics-driven implementation of neural networks within Einstein-Boltzmann Solvers*, *Journal of Cosmology and Astroparticle Physics* **2019** (2019) 028.

[48] S. Günther, J. Lesgourgues, G. Samaras, N. Schöneberg, F. Stadtmann, C. Fidler et al., *CosmicNet II: emulating extended cosmologies with efficient and accurate neural networks*, *Journal of Cosmology and Astroparticle Physics* **2022** (2022) 035.

[49] Y. Setyawati, M. Pürrer and F. Ohme, *Regression methods in waveform modeling: a comparative study*, *Classical and Quantum Gravity* **37** (2020) 075012.

[50] P. J. Easter, P. D. Lasky, A. R. Casey, L. Rezzolla and K. Takami, *Computing fast and reliable gravitational waveforms of binary neutron star merger remnants*, *Physical Review D* **100** (2019) 043005.

[51] S. E. Field, C. R. Galley, J. S. Hesthaven, J. Kaye and M. Tiglio, *Fast prediction and evaluation of gravitational waveforms using surrogate models*, *Physical Review X* **4** (2014) 031006.

[52] M. Hoffman, P. Sountsov, J. V. Dillon, I. Langmore, D. Tran and S. Vasudevan, *NeuTra-lizing Bad Geometry in Hamiltonian Monte Carlo Using Neural Transport*, Mar., 2019. 10.48550/arXiv.1903.03704.

[53] A. Matthews, M. Arbel, D. J. Rezende and A. Doucet, *Continual Repeated Annealed Flow Transport Monte Carlo*, in *Proceedings of the 39th International Conference on Machine Learning*, pp. 15196–15219, PMLR, June, 2022, https://proceedings.mlr.press/v162/matthews22a.html.

[54] R. Abbott, M. S. Albergo, A. Botev, D. Boyda, K. Cranmer, D. C. Hackett et al., *Aspects of scaling and scalability for flow-based sampling of lattice QCD*, *The European Physical Journal A* **59** (2023) 257.

[55] M. J. Williams, J. Veitch and C. Messenger, *Importance nested sampling with normalising flows*, *Machine Learning: Science and Technology* **4** (2023) 035011.

[56] M. J. Williams, J. Veitch and C. Messenger, *Nested Sampling with Normalising Flows for Gravitational-Wave Inference*, *Physical Review D* **103** (2021) 103006.

[57] M. J. Williams, J. Veitch, C. Chapman-Bird and R. Tenorio, *nessai*, Jan., 2025. 10.5281/zenodo.14627250.

[58] C. Albert, S. Ulzega, F. Ozdemir, F. Perez-Cruz and A. Mira, *Learning Summary Statistics for Bayesian Inference with Autoencoders*, *SciPost Physics Core* **5** (2022) 043.

[59] A. Tejero-Cantero, J. Boelts, M. Deistler, J.-M. Lueckmann, C. Durkan, P. J. Gonçalves et al., *sbi: A toolkit for simulation-based inference*, *Journal of Open Source Software* **5** (2020) 2505.

[60] M. Dax, S. R. Green, J. Gair, M. Deistler, B. Schölkopf and J. H. Macke, *Group equivariant neural posterior estimation*, Oct., 2021, https://openreview.net/forum?id=u6s8dSporO8.

[61] G. Papamakarios and I. Murray, *Fast ϵ-free inference of simulation models with Bayesian conditional density estimation*, in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, (Red Hook, NY, USA), pp. 1036–1044, Curran Associates Inc., Dec., 2016.

[62] J. Zeghal, F. Lanusse, A. Boucaud, B. Remy and E. Aubourg, *Neural Posterior Estimation with Differentiable Simulators*, July, 2022. 10.48550/arXiv.2207.05636.

[63] J. Alsing, T. Charnock, S. Feeney and B. Wandelt, *Fast likelihood-free cosmology with neural density estimators and active learning*, *Monthly Notices of the Royal Astronomical Society* **488** (2019) 4440.

[64] K. Lin, M. von wietersheim Kramsta, B. Joachimi and S. Feeney, *A simulation-based inference pipeline for cosmic shear with the Kilo-Degree Survey*, *Monthly Notices of the Royal Astronomical Society* **524** (2023) 6167.

[65] A. Cole, B. K. Miller, S. J. Witte, M. X. Cai, M. W. Grootes, F. Nattino et al., *Fast and credible likelihood-free cosmology with truncated marginal neural ratio estimation*, *Journal of Cosmology and Astroparticle Physics* **2022** (2022) 004.

[66] N. Anau Montel, A. Coogan, C. Correa, K. Karchev and C. Weniger, *Estimating the warm dark matter mass from strong lensing images with truncated marginal neural ratio estimation*, *Monthly Notices of the Royal Astronomical Society* **518** (2023) 2746.

[67] N. Anau Montel, J. Alvey and C. Weniger, *Scalable inference with autoregressive neural ratio estimation*, *Monthly Notices of the Royal Astronomical Society* **530** (2024) 4107.

[68] F. Rozet and G. Louppe, *Arbitrary Marginal Neural Ratio Estimation for Simulation-based Inference*, Nov., 2021. 10.48550/arXiv.2110.00449.

[69] A. Delaunoy, J. Hermans, F. Rozet, A. Wehenkel and G. Louppe, *Towards Reliable Simulation-Based Inference with Balanced Neural Ratio Estimation*, Aug., 2022. 10.48550/arXiv.2208.13624.

[70] B. K. Miller, C. Weniger and P. Forré, *Contrastive Neural Ratio Estimation*, Oct., 2022, https://openreview.net/forum?id=kOIaB1hzaLe.

[71] B. K. Miller, A. Cole, P. Forré, G. Louppe and C. Weniger, *Truncated Marginal Neural Ratio Estimation*, in *35th Conference on Neural Information Processing Systems*, July, 2021, DOI.

[72] J. Hermans, V. Begy and G. Louppe, *Likelihood-free MCMC with Amortized Approximate Ratio Estimators*, in *Proceedings of the 37th International Conference on Machine Learning*, pp. 4239–4248, PMLR, Nov., 2020, https://proceedings.mlr.press/v119/hermans20a.html.

[73] C. Durkan, I. Murray and G. Papamakarios, *On Contrastive Learning for Likelihood-free Inference*, in *Proceedings of the 37th International Conference on Machine Learning*, pp. 2771–2781, PMLR, Nov., 2020, https://proceedings.mlr.press/v119/durkan20a.html.

[74] I. Gómez-Vargas and J. A. Vázquez, *Deep learning and genetic algorithms for cosmological Bayesian inference speed-up*, *Physical Review D* **110** (2024) 083518.

[75] Z. Ghahramani and C. Rasmussen, *Bayesian Monte Carlo*, in *Advances in Neural Information Processing Systems*, vol. 15, MIT Press, 2002.

[76] L. Acerbi, *An Exploration of Acquisition and Mean Functions in Variational Bayesian Monte Carlo*, in *Proceedings of The 1st Symposium on Advances in Approximate Bayesian Inference*, pp. 1–10, PMLR, Jan., 2019, https://proceedings.mlr.press/v96/acerbi19a.html.

[77] L. Acerbi, *Variational bayesian monte carlo with noisy likelihoods*, *Advances in neural information processing systems* **33** (2020) 8211.

[78] T. Gunter, M. A. Osborne, R. Garnett, P. Hennig and S. J. Roberts, *Sampling for Inference in Probabilistic Models with Fast Bayesian Quadrature*, in *Advances in Neural Information Processing Systems*, vol. 27, Curran Associates, Inc., 2014.

[79] P. Amaro-Seoane, H. Audley, S. Babak, J. Baker, E. Barausse, P. Bender et al., *Laser Interferometer Space Antenna*, Feb., 2017. 10.48550/arXiv.1702.00786.

[80] M. Colpi, K. Danzmann, M. Hewitson, K. Holley-Bockelmann, P. Jetzer, G. Nelemans et al., *LISA Definition Study Report*, Feb., 2024. 10.48550/arXiv.2402.07571.

[81] J. Aasi and others, *Advanced LIGO*, *Class. Quant. Grav.* **32** (2015) 074001.

[82] F. Acernese and others, *Advanced Virgo: a second-generation interferometric gravitational wave detector*, *Class. Quant. Grav.* **32** (2015) 024001.

[83] T. Akutsu and others, *KAGRA: 2.5 Generation Interferometric Gravitational Wave Detector*, *Nature Astron.* **3** (2019) 35.

[84] G. Hobbs, A. Archibald, Z. Arzoumanian, D. Backer, M. Bailes, N. D. R. Bhat et al., *The International Pulsar Timing Array project: using pulsars as a gravitational wave detector*, *Classical and Quantum Gravity* **27** (2010) 084013.

[85] P. Auclair, D. Bacon, T. Baker, T. Barreiro, N. Bartolo, E. Belgacem et al., *Cosmology with the Laser Interferometer Space Antenna*, *Living Reviews in Relativity* **26** (2023) 5.

[86] M. Maggiore, *Gravitational Waves: Volume 1: Theory and Experiments*. Oxford University PressOxford, 1 ed., Oct., 2007, 10.1093/acprof:oso/9780198570745.001.0001.

[87] L. Blanchet, *Gravitational Radiation from Post-Newtonian Sources and Inspiralling Compact Binaries*, *Living Reviews in Relativity* **5** (2002) 3.

[88] M. Kilic, C. Allende Prieto, W. R. Brown and D. Koester, *The Lowest Mass White Dwarf*, *The Astrophysical Journal* **660** (2007) 1451.

[89] S. O. Kepler, S. J. Kleinman, A. Nitta, D. Koester, B. G. Castanheira, O. Giovannini et al., *White dwarf mass distribution in the SDSS: White dwarf mass distribution in the SDSS*, *Monthly Notices of the Royal Astronomical Society* **375** (2007) 1315.

[90] M. Vallisneri, *A LISA data-analysis primer*, *Classical and Quantum Gravity* **26** (2009) 094024.

[91] C. García-Quirós, M. Colleoni, S. Husa, H. Estellés, G. Pratten, A. Ramos-Buades et al., *Multimode frequency-domain model for the gravitational wave signal from nonprecessing black-hole binaries*, *Physical Review D* **102** (2020) 064002.

[92] S. A. Usman, A. H. Nitz, I. W. Harry, C. M. Biwer, D. A. Brown, M. Cabero et al., *The PyCBC search for gravitational waves from compact binary coalescence*, *Classical and Quantum Gravity* **33** (2016) 215004.

[93] L. Lindblom, B. J. Owen and D. A. Brown, *Model waveform accuracy standards for gravitational wave data analysis*, *Physical Review D* **78** (2008) 124020.

[94] M. Volonteri, *Formation of supermassive black holes*, *The Astronomy and Astrophysics Review* **18** (2010) 279.

[95] J. L. Johnson and F. Haardt, *The Early Growth of the First Black Holes*, *Publ. Astron. Soc. Austral.* **33** (2016) e007.

[96] A. Sesana, F. Haardt, P. Madau and M. Volonteri, *The gravitational wave signal from massive black hole binaries and its contribution to the LISA data stream*, *Astrophys. J.* **623** (2005) 23.

[97] R. Valiante, M. Colpi, R. Schneider, A. Mangiagli, M. Bonetti, G. Cerini et al., *Unveiling early black hole growth with multifrequency gravitational wave observations*, *Mon. Not. Roy. Astron. Soc.* **500** (2020) 4095.

[98] S. Yi, F. Iacovelli, S. Marsat, D. Wadekar and E. Berti, *Systematic biases from the exclusion of higher harmonics in parameter estimation on LISA binaries*, .

[99] C. J. Moore, R. H. Cole and C. P. L. Berry, *Gravitational-wave sensitivity curves*, *Classical and Quantum Gravity* **32** (2014) 015014.

[100] C.-F. Chang and Y. Cui, *Gravitational waves from global cosmic strings and cosmic archaeology*, *Journal of High Energy Physics* **2022** (2022) 114.

[101] M. Sakellariadou, *Cosmic Strings and Cosmic Superstrings*, *Nuclear Physics B - Proceedings Supplements* **192-193** (2009) 68.

[102] A. Kosowsky, M. S. Turner and R. Watkins, *Gravitational waves from first-order cosmological phase transitions*, *Physical Review Letters* **69** (1992) 2026.

[103] O. Hartwig, M. Lilley, M. Muratore and M. Pieroni, *Stochastic gravitational wave background reconstruction for a nonequilateral and unequal-noise LISA constellation*, *Phys. Rev. D* **107** (2023) 123531.

[104] M.-S. Hartig, S. Paczkowski, M. Hewitson, G. Heinzel and G. Wanner, *Postprocessing subtraction of tilt-to-length noise in LISA in the presence of gravitational wave signals*, *Phys. Rev. D* **111** (2025) 043048.

[105] J. Crowder and N. Cornish, *A Solution to the Galactic Foreground Problem for LISA*, *Phys. Rev. D* **75** (2007) 043008.

[106] T. B. Littenberg and N. J. Cornish, *Prototype global analysis of LISA data with multiple source types*, *Physical Review D* **107** (2023) 063004.

[107] S. H. Strub, L. Ferraioli, C. Schmelzbach, S. C. Stähler and D. Giardini, *Global analysis of LISA data with Galactic binaries and massive black hole binaries*, *Physical Review D* **110** (2024) 024005.

[108] S. Deng, S. Babak, M. L. Jeune, S. Marsat, \. Plagnol and A. Sartirana, *Modular global-fit pipeline for LISA data analysis*, Jan., 2025. 10.48550/arXiv.2501.10277.

[109] M. L. Katz, N. Karnesis, N. Korsakova, J. R. Gair and N. Stergioulas, *Efficient GPU-accelerated multisource global fit pipeline for LISA data analysis*, *Physical Review D* **111** (2025) 024060.

[110] A. D. Johnson, J. Roulet, K. Chatziioannou, M. Vallisneri, C. G. Trejo and K. A. Gersbach, *PETRA: From the global fit for LISA's Galactic binaries to a catalog of sources*, .

[111] S. H. Strub, L. Ferraioli, C. Schmelzbach, S. C. Stähler and D. Giardini, *Bayesian parameter estimation of Galactic binaries in LISA data with Gaussian process regression*, *Physical Review D* **106** (2022) 062003.

[112] S. H. Strub, L. Ferraioli, C. Schmelzbach, S. C. Stähler and D. Giardini, *Accelerating global parameter estimation of gravitational waves from Galactic binaries using a genetic algorithm and GPUs*, *Physical Review D* **108** (2023) 103018.

[113] U. Bhardwaj, J. Alvey, B. K. Miller, S. Nissanke and C. Weniger, *Peregrine: Sequential simulation-based inference for gravitational wave signals*, July, 2024. 10.48550/arXiv.2304.02035.

[114] S. R. Green, C. Simpson and J. Gair, *Gravitational-wave parameter estimation with autoregressive neural network flows*, *Physical Review D* **102** (2020) 104057.

[115] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno and B. Schölkopf, *Real-time gravitational-wave science with neural posterior estimation*, *Physical Review Letters* **127** (2021) 241103.

[116] J. Alvey, U. Bhardwaj, V. Domcke, M. Pieroni and C. Weniger, *Simulation-based inference for stochastic gravitational wave background data analysis*, *Physical Review D* **109** (2024) 083008.

[117] E. Cuoco, J. Powell, M. Cavaglià, K. Ackley, M. Bejger, C. Chatterjee et al., *Enhancing gravitational-wave science with machine learning*, *Machine Learning: Science and Technology* **2** (2020) 011002.

[118] A. Linde, *A new inflationary universe scenario: A possible solution of the horizon, flatness, homogeneity, isotropy and primordial monopole problems*, *Physics Letters B* **108** (1982) 389.

[119] A. H. Guth, *Inflationary universe: A possible solution to the horizon and flatness problems*, *Physical Review D* **23** (1981) 347.

[120] V. Mukhanov, *Theory of cosmological perturbations*, *Physics Reports* **215** (1992) 203.

[121] P. Collaboration, Y. Akrami, F. Arroja, M. Ashdown, J. Aumont, C. Baccigalupi et al., *Planck 2018 results. X. Constraints on inflation*, *Astronomy & Astrophysics* **641** (2020) A10.

[122] D. Baumann, *Cosmology*. Cambridge University Press, Cambridge, 1st ed ed., 2022.

[123] O. Özsoy and G. Tasinato, *Inflation and Primordial Black Holes*, *Universe* **9** (2023) 203.

[124] D. Wands, *Multiple Field Inflation*, in *Inflationary Cosmology*, M. Lemoine, J. Martin and P. Peter, eds., vol. 738, (Berlin, Heidelberg), pp. 275–304, Springer Berlin Heidelberg, (2008), DOI.

[125] L. Iacconi and D. J. Mulryne, *Multi-field inflation with large scalar fluctuations: non-Gaussianity and perturbativity*, 2023. 10.48550/ARXIV.2304.14260.

[126] G. A. Palma, S. Sypsas and C. Zenteno, *Seeding Primordial Black Holes in Multifield Inflation*, *Physical Review Letters* **125** (2020) 121301.

[127] J. Fumagalli, S. Renaux-Petel and L. T. Witkowski, *Oscillations in the stochastic gravitational wave background from sharp features and particle production during inflation*, *Journal of Cosmology and Astroparticle Physics* **2021** (2021) 030.

[128] J. Fumagalli, S. Renaux-Petel, J. W. Ronayne and L. T. Witkowski, *Turning in the landscape: A new mechanism for generating primordial black holes*, *Physics Letters B* **841** (2023) 137921.

[129] D. Racco, S. Verner and W. Xue, *Gravitational production of heavy particles during and after inflation*, *Journal of High Energy Physics* **2024** (2024) 129.

[130] N. Barnaby and Z. Huang, *Particle production during inflation: Observational constraints and signatures*, *Physical Review D* **80** (2009) 126018.

[131] N. Barnaby, *Nongaussianity from Particle Production During Inflation*, *Adv. Astron.* **2010** (2010) 156180.

[132] H. Motohashi, A. A. Starobinsky and J. Yokoyama, *Inflation with a constant rate of roll*, *Journal of Cosmology and Astroparticle Physics* **2015** (2015) 018.

[133] E. Silverstein and D. Tong, *Scalar speed limits and cosmology: Acceleration from D-cceleration*, *Physical Review D* **70** (2004) 103505.

[134] Y. Ageeva and P. Petrov, *K -inflation: The legitimacy of the classical treatment*, *Physical Review D* **110** (2024) 043527.

[135] C. Armendáriz-Picón, T. Damour and V. Mukhanov, *k-Inflation*, *Physics Letters B* **458** (1999) 209.

[136] T. Kobayashi, M. Yamaguchi and J. Yokoyama, *Inflation Driven by the Galileon Field*, *Physical Review Letters* **105** (2010) 231302.

[137] T. Kobayashi, M. Yamaguchi and J. Yokoyama, *Generalized G-Inflation: –Inflation with the Most General Second-Order Field Equations–*, *Progress of Theoretical Physics* **126** (2011) 511.

[138] T. S. Bunch, P. C. W. Davies and R. Penrose, *Quantum field theory in de Sitter space: renormalization by point-splitting*, *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* **360** (1997) 117.

[139] V. F. Mukhanov, *Quantum Theory of Gauge Invariant Cosmological Perturbations*, *Sov. Phys. JETP* **67** (1988) 1297.

[140] M. Sasaki, *Large Scale Quantum Fluctuations in the Inflationary Universe*, *Progress of Theoretical Physics* **76** (1986) 1036.

[141] N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini et al., *Planck 2018 results - VI. Cosmological parameters*, *Astronomy & Astrophysics* **641** (2020) A6.

[142] BICEP/Keck Collaboration, P. Ade, Z. Ahmed, M. Amiri, D. Barkats, R. B. Thakur et al., *Improved Constraints on Primordial Gravitational Waves using Planck, WMAP, and BICEP/Keck Observations through the 2018 Observing Season*, *Physical Review Letters* **127** (2021) 151301.

[143] V. Acquaviva, N. Bartolo, S. Matarrese and A. Riotto, *Second order cosmological perturbations from inflation*, *Nucl. Phys. B* **667** (2003) 119.

[144] D. Baumann, P. Steinhardt, K. Takahashi and K. Ichiki, *Gravitational wave spectrum induced by primordial scalar perturbations*, *Physical Review D* **76** (2007) 084019.

[145] P. Adshead, K. D. Lozanov and Z. J. Weiner, *Non-Gaussianity and the induced gravitational wave background*, *JCAP* **10** (2021) 080.

[146] S. Renaux-Petel, *The Trispectrum as a Diagnostic of Primordial Orthogonal non-Gaussianities*, *JCAP* **07** (2013) 005.

[147] S. Garcia-Saenz, L. Pinol, S. Renaux-Petel and D. Werth, *No-go theorem for scalar-trispectrum-induced gravitational waves*, *JCAP* **03** (2023) 057.

[148] K. Kohri and T. Terada, *Semianalytic Calculation of Gravitational Wave Spectrum Nonlinearly Induced from Primordial Curvature Perturbations*, *Physical Review D* **97** (2018) 123532.

[149] H. Assadullahi and D. Wands, *Gravitational waves from an early matter era*, *Physical Review D* **79** (2009) 083511.

[150] P. Campeti and others, *LiteBIRD science goals and forecasts. A case study of the origin of primordial gravitational waves using large-scale CMB polarization*, *JCAP* **06** (2024) 008.

[151] C. Caprini and D. G. Figueroa, *Cosmological backgrounds of gravitational waves*, *Classical and Quantum Gravity* **35** (2018) 163001.

[152] R. H. Cyburt, B. D. Fields, K. A. Olive and E. Skillman, *New BBN limits on physics beyond the standard model from 4He*, *Astroparticle Physics* **23** (2005) 313.

[153] L. Pagano, L. Salvati and A. Melchiorri, *New constraints on primordial gravitational waves from Planck 2015*, *Physics Letters B* **760** (2016) 823.

[154] I. Sendra and T. L. Smith, *Improved limits on short-wavelength gravitational waves from the cosmic microwave background*, *Physical Review D* **85** (2012) 123002.

[155] T. L. Smith, E. Pierpaoli and M. Kamionkowski, *New Cosmic Microwave Background Constraint to Primordial Gravitational Waves*, *Physical Review Letters* **97** (2006) 021301.

[156] P. Campeti, *pcampeti/SGWBProbe*, Aug., 2023.

[157] P. Campeti, E. Komatsu, D. Poletti and C. Baccigalupi, *Measuring the spectrum of primordial gravitational waves with CMB, PTA and laser interferometers*, *Journal of Cosmology and Astroparticle Physics* **2021** (2021) 012.

[158] C. Caprini, D. G. Figueroa, R. Flauger, G. Nardini, M. Peloso, M. Pieroni et al., *Reconstructing the spectral shape of a stochastic gravitational wave background with LISA*, *Journal of Cosmology and Astroparticle Physics* **2019** (2019) 017.

[159] P. D. Lasky, C. M. Mingarelli, T. L. Smith, J. T. Giblin, E. Thrane, D. J. Reardon et al., *Gravitational-Wave Cosmology across 29 Decades in Frequency*, *Physical Review X* **6** (2016) 011035.

[160] A. I. Renzini, B. Goncharov, A. C. Jenkins and P. M. Meyers, *Stochastic Gravitational-Wave Backgrounds: Current Detection Efforts and Future Prospects*, *Galaxies* **10** (2022) 34.

[161] S. Babak, M. Falxa, G. Franciolini and M. Pieroni, *Forecasting the sensitivity of pulsar timing arrays to gravitational wave backgrounds*, *Phys. Rev. D* **110** (2024) 063022.

[162] J. Antoniadis, Z. Arzoumanian, S. Babak, M. Bailes, A.-S. Bak Nielsen, P. T. Baker et al., *The International Pulsar Timing Array second data release: Search for an isotropic gravitational*

wave background, *Monthly Notices of the Royal Astronomical Society* **510** (2022) 4873.

[163] H. Xu and others, *Searching for the Nano-Hertz Stochastic Gravitational Wave Background with the Chinese Pulsar Timing Array Data Release I*, *Res. Astron. Astrophys.* **23** (2023) 075024.

[164] J. Antoniadis and others, *The second data release from the European Pulsar Timing Array - III. Search for gravitational wave signals*, *Astron. Astrophys.* **678** (2023) A50.

[165] D. J. Reardon and others, *Search for an Isotropic Gravitational-wave Background with the Parkes Pulsar Timing Array*, *Astrophys. J. Lett.* **951** (2023) L6.

[166] G. Agazie and others, *The NANOGrav 15 yr Data Set: Evidence for a Gravitational-wave Background*, *Astrophys. J. Lett.* **951** (2023) L8.

[167] M. T. Miles and others, *The MeerKAT Pulsar Timing Array: the first search for gravitational waves with the MeerKAT radio telescope*, *Mon. Not. Roy. Astron. Soc.* **536** (2024) 1489.

[168] R. Abbott and others, *Upper limits on the isotropic gravitational-wave background from Advanced LIGO and Advanced Virgo's third observing run*, *Phys. Rev. D* **104** (2021) 022004.

[169] T. J. W. Lazio, *The Square Kilometre Array pulsar timing array*, *Class. Quant. Grav.* **30** (2013) 224011.

[170] A. Abac, R. Abramo, S. Albanesi, A. Albertini, A. Agapito, M. Agathos et al., *The Science of the Einstein Telescope*, Mar., 2025. 10.48550/arXiv.2503.12263.

[171] Y. B. Zel'dovich and I. D. Novikov, *The Hypothesis of Cores Retarded during Expansion and the Hot Cosmological Model*, *Sov. Astron.* **10** (1967) 602.

[172] S. Hawking, *Gravitationally collapsed objects of very low mass*, *Mon. Not. Roy. Astron. Soc.* **152** (1971) 75.

[173] M. Sasaki, T. Suyama, T. Tanaka and S. Yokoyama, *Primordial black holes—perspectives in gravitational wave astronomy*, *Class. Quant. Grav.* **35** (2018) 063001.

[174] C. Byrnes, G. Franciolini, T. Harada, P. Pani and M. Sasaki, eds., *Primordial Black Holes*, Springer Series in Astrophysics and Cosmology. Springer, Mar., 2025.

[175] G. Franciolini, I. Musco, P. Pani and A. Urbano, *From inflation to black hole mergers and back again: Gravitational-wave data-driven constraints on inflationary scenarios with a*

first-principle model of primordial black holes across the QCD epoch, *Phys. Rev. D* **106** (2022) 123526.

[176] B. J. Kavanagh, *bradkav/PBHbounds: Release version*, Nov., 2019. 10.5281/zenodo.3538999.

Appendix

Paper I

Fast and robust Bayesian Inference using Gaussian Processes with GPry

Journal of Cosmology and Astroparticle Physics
An IOP and SISSA journal

Fast and robust Bayesian inference using Gaussian processes with `GPry`

Jonas El Gammal,[a] **Nils Schöneberg,**[b] **Jesús Torrado**[c,d] **and Christian Fidler**[e]

[a]Department of Mathematics and Physics, University of Stavanger,
Kristine Bonnevies vei 22, Stavanger 4021, Norway
[b]Institut de Ciències del Cosmos, Universitat de Barcelona,
Martí i Franquès 1, Barcelona E08028, Spain
[c]Dipartimento di Fisica e Astronomia "G. Galilei", Università degli Studi di Padova,
Via Marzolo 8, Padova I-35131, Italy
[d]INFN, Sezione di Padova,
Via Marzolo 8, Padova I-35131, Italy
[e]Institute for Theoretical Particle Physics and Cosmology (TTK), RWTH Aachen University,
Sommerfeldstraße 16, Aachen 52074, Germany
E-mail: jonas.e.elgammal@uis.no, nils.science@gmail.com,
jesus.torrado@pd.infn.it, fidler@physik.rwth-aachen.de

Received December 20, 2022
Revised June 3, 2023
Accepted July 6, 2023
Published October 6, 2023

Abstract. We present the `GPry` algorithm for fast Bayesian inference of general (non-Gaussian) posteriors with a moderate number of parameters. `GPry` does not need any pre-training, special hardware such as GPUs, and is intended as a drop-in replacement for traditional Monte Carlo methods for Bayesian inference. Our algorithm is based on generating a Gaussian Process surrogate model of the log-posterior, aided by a Support Vector Machine classifier that excludes extreme or non-finite values. An active learning scheme allows us to reduce the number of required posterior evaluations by two orders of magnitude compared to traditional Monte Carlo inference. Our algorithm allows for parallel evaluations of the posterior at optimal locations, further reducing wall-clock times. We significantly improve performance using properties of the posterior in our active learning scheme and for the definition of the GP prior. In particular we account for the expected dynamical range of the posterior in different dimensionalities. We test our model against a number of synthetic and cosmological examples. `GPry` outperforms traditional Monte Carlo methods when the evaluation time of the likelihood (or the calculation of theoretical observables) is of the order of seconds; for evaluation times of over a minute it can perform inference in days that would take months using traditional methods. `GPry` is distributed as an open source Python package (`pip install gpry`) and can also be found at `https://github.com/jonaselgammal/GPry`.

Keywords: Machine learning, Statistical sampling techniques, Bayesian reasoning
ArXiv ePrint: 2211.02045

https://doi.org/10.1088/1475-7516/2023/10/021

JCAP10(2023)021

Contents

1 Introduction

One of the fundamental tools of science is the comparison of observations with theory. In many situations, this involves inference on the parameters of a model (or on models themselves) given some observed data. This is often realised using Bayesian statistics, where

JCAP10(2023)021

one synthesises the probability of some data having been acquired into a *likelihood function*, assumes some *a priori* distribution for the model parameters, and samples from the product of both (proportional to the so-called *posterior*) using Monte Carlo methods, the most common ones in Cosmology being based on Markov Chain Monte Carlo [1–4] or Nested Sampling [5–12].

The new era of cosmological surveys will produce data in rapidly increasing amount and quality [13, 14]. This will in turn raise the computational costs of traditional Monte Carlo pipelines: data quality will call for an increase in the precision of theoretical computations of the observables that are compared against the data (e.g. including physical effects that could have been so far neglected), and likelihood computations will involve operations on ever larger data vectors. This can and will eventually result in traditional Bayesian inference becoming prohibitively slow, further increasing the already damaging carbon footprint of scientific computations in computer clusters [15, 16]. In order to keep being able to exploit cosmological data for parameter inference, we need to develop more advanced algorithms that significantly reduce the computational costs of performing inference, and machine-learning based methods are one of the most promising tools for that.

So far, a number of different solutions have been proposed. A family of them focus on substituting the theoretical computation of observables (or intermediate quantities to arrive at them) by appropriately-trained, usually Neural Network-based, *emulators* that cheaply map the theoretical parameters onto the space of vectors of observables. For applications to Cosmology and Astrophysics, see e.g. [17–40]. These methods are robust in the sense that they are guaranteed to reproduce the true posterior distribution, as long as the emulator is properly trained, which is easy to check a posteriori. Unfortunately their utility is limited by the need to retrain them whenever the theoretical model under investigation is varied. Additionally, as experiments become ever more precise, in order to achieve sufficient accuracy a larger number of systematic effects needs to be accounted for, which requires ever more costly experimental likelihoods, which cannot be easily accelerated by emulators.

Another proposed solution are simulation-based *likelihood-free* approaches, inspired by Approximate Bayesian Computation, but accelerated by Neural Networks [41, 42]. There, Neural Networks are used to learn a mapping between sets of model parameters and their corresponding simulated data, so that they can automatically extract features, marginalise over nuisance parameters, learn a likelihood function, or ultimately produce a posterior distribution of the model parameters when fed real experimental data. Recent development and applications in Cosmology and Astrophysics can be found in [43–54]. The claimed advantages are that they may discover or take into account features in the data that are not captured by summary statistics or observables, and the lack of need to formulate a likelihood, which can be complex or prohibitively expensive in some cases. On the other hand, they tend to require expensive training and the reusability of the trained networks is limited when considering model extensions. The need to accurately account for modelling uncertainty and possible biases has also been highlighted recently [55, 56].

The solution presented in this work differs from the previous ones in that it retains the full computation of the observable and data likelihood, but minimises the number of points in the parameter space where this full pipeline needs to be computed; it uses these points to create a model of the posterior, and to iteratively predict the next optimal evaluation locations. For the emulation of the posterior we use Gaussian Processes (GP) [57], which have a small number of hyperparameters that are easily interpretable in terms of properties of the posterior, and thus make it easier to incorporate prior information about its functional

JCAP10(2023)021

form. Furthermore, due to their simplicity, Gaussian Processes generally require smaller training sets than for example Neural Networks. We combine the GP model of the posterior with a support vector machine (SVM) [58, 59] to restrict the parameter space to a region of reasonable posterior values.

Our approach expands on previous work applying Bayesian quadrature and active sampling to statistical inference [60–64], which we improve upon by incorporating the expected scaling of the log-posterior with dimensionality, the definition of a cheap and consistent convergence criterion and the treatment of extreme log-posterior values with an SVM classifier. A previous attempt at a similar approach to inference in Cosmology with a GP surrogate of the posterior can be found in [65], and in the context of emulator-training in [66]. Alternative emulator-based approaches, relying on Variational Inference, have also been proposed, e.g. combined with a GP surrogate model to reduce the number of posterior evaluations [67–70], or targeted towards high dimensionalities but allowing for numbers of evaluations similar to MCMC [71, 72].

The result of our work is the development of the `GPry` algorithm. An open source implementation is available as a Python package (`pip install gpry`) and at https://github.com/jonaselgammal/GPry. `GPry` does not need any pre-training or parameter tuning, so it can be used as a *drop-in* replacement for traditional Monte Carlo algorithms for dimensionalities $N_d \lesssim 20$ (since the computational cost of the algorithm makes it impractical for larger problems in its current implementation). Unlike neural networks it also does not require any specialised hardware such as GPUs. As we will show, it allows for accurate and fast emulation of posteriors for moderate dimensionalities, including non-Gaussian distributions, by using just a few hundred or thousand evaluations of the posterior distribution. Especially when individual likelihood evaluations are computationally expensive, this can result in large speedups of typically two orders of magnitude.

This paper is structured as follows: in section 2 we review the basic concepts and useful notation. We continue in section 3 presenting the modelling choices involved in the construction of the GP surrogate model. The learning strategy for acquiring new sampling locations as well as a criterion for deciding on convergence are discussed in section 4. In section 5 we put together all the pieces and present the full algorithm, and comment on its general performance. We discuss the performance of `Gpry` on different synthetic and cosmological problems in section 6, and we present our conclusions and discuss possible future development in section 7. Appendix A is dedicated to discussing the inclusion of prior information on the dynamical range of the posterior into the surrogate model at different stages of the algorithm.

2 Basic concepts

In order to establish a consistent notation and a deeper understanding of the underlying concepts, we quickly summarize some of the theory, which we are going to use in the detailed description of section 3.

2.1 Bayesian inference of model parameters

A usual Bayesian inference problem is that of estimating the probability distribution $p(\mathcal{M}(\boldsymbol{x})|\mathcal{D})$ of the parameters $\boldsymbol{x}$ of a model $\mathcal{M}$ given some experimental data $\mathcal{D}$, also known as *posterior*. Following Bayes' theorem, this is proportional to the product of the *likelihood*

JCAP10(2023)021

$p(\mathcal{D}|\mathcal{M}(\boldsymbol{x}))$ (the probability of $\mathcal{D}$ having being measured given the model with these parameter values), and the *prior* probability of the parameter values $\boldsymbol{x}$ given the model, $p(\boldsymbol{x}|\mathcal{M})$, assigned before (or idependently of) the experiment that measured $\mathcal{D}$.[1] Fixing the model $\mathcal{M}$ and the data $\mathcal{D}$, we can drop their explicit dependence to simplify notation. With that Bayes' theorem reads

$$p(\boldsymbol{x}) \propto \mathcal{L}(\boldsymbol{x})\pi(\boldsymbol{x})\,, \tag{2.1}$$

where $p(\boldsymbol{x})$ is the posterior, $\mathcal{L}(\boldsymbol{x})$ the likelihood, and $\pi(\boldsymbol{x})$ the prior.

In Cosmology, likelihoods are typically provided by experimental collaborations, are generally non-analytic, or analytic but non-differentiable, and usually also costly to evaluate. Even when they are well-behaved, they sometimes depend on cosmological quantities whose computation in terms of the parameters to be inferred has the same undesirable properties. In these cases, the targeted solution to the inference problem is obtaining a Monte Carlo sample of the posterior, often using MCMC- or nested-sampling-based methods.

This work focuses on reducing the number of evaluations of the posterior (and thus the likelihood) needed to solve the inference problem. We do that by creating a surrogate model of the posterior using a Gaussian Process, and developing an active learning algorithm that decides sequentially on a small optimal set of parameter values where to evaluate the true likelihood, so that the surrogate model is accurate enough. One can then e.g. extract the usual Monte Carlo sample from the resulting surrogate model of the posterior (which, as a bonus, is differentiable) at a very low computational cost.

If the goal is to obtain 1D/2D posteriors (and their corresponding CLs) from the GP, one could wonder if there would be alternative efficient methods of computing the required marginalization integrals. However, generally the integrals involved are not solvable analytically and due to the high dimensionality of these integrals in most applications, the most efficient ways of computing them numerically are usually Monte-Carlo methods. We discuss the computational costs of this choice in section 5.3.

2.2 Gaussian processes

We briefly present the relevant GP notation and formulae that we will need for this work. For a more thorough review, see [57].

Gaussian Processes are useful to emulate a sufficiently smooth[2] function $f(\boldsymbol{x})$ at an arbitrary point $\boldsymbol{x}$ (within a certain domain) given a set of sampling locations $\boldsymbol{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_s}\}$ and their corresponding function values $y_i = f(\boldsymbol{x}_i)$ for $i = 1 \ldots N_s$. This last equation can be abbreviated to $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{X})$ (notice the bold symbol for $\boldsymbol{y}$ and $\boldsymbol{f}$ and the dependence on $\boldsymbol{X}$) following the usual notation in GP literature, where the number of samples is treated as an additional vector space of dimension N_s, with components denoted by a subscript. This means that $\boldsymbol{X}$ becomes a $N_s \times N_d$ matrix, where N_d is the dimensionality of the parameter space. This way, we write for a scalar function $s(\boldsymbol{x})$ evaluated at the N_s different sampling locations the vector $\boldsymbol{s}(\boldsymbol{X})$ with components $[\boldsymbol{s}(\boldsymbol{X})]_i = s(\boldsymbol{X}_i)$, and similarly for scalar functions of two arguments the tensor $\boldsymbol{s}(\boldsymbol{X}, \boldsymbol{X})$ with components $[\boldsymbol{s}(\boldsymbol{X}, \boldsymbol{X})]_{ij} = s(\boldsymbol{X}_i, \boldsymbol{X}_j)$.

A Gaussian Process posits that the function $f(\boldsymbol{x})$ in question is a random draw from a family of functions, informed by the sampling locations. For a given position $\boldsymbol{x}$ such random

[1] The missing proportionality constant is the inverse of the *evidence* $p(\mathcal{D}|\mathcal{M})$, which can usually be ignored in parameter estimation and will hence be omitted in all subsequent calculations. Note though, that the evidence is important when performing model selection.

[2] Here, "sufficiently smooth" refers to an underlying function which n-times continuously differentiable where $n \geq 1$. The function may still have some statistical or numerical noise added on top of it.

JCAP10(2023)021

draw of a function $f(\boldsymbol{x})$ is assumed to be Gaussian-distributed (hence the name) around a mean function $m(\boldsymbol{x})$ with a covariance between the functional value at two different points given by some function $k(\boldsymbol{x}, \boldsymbol{x}')$, often called the *kernel* of the GP.

$$\widehat{f} \sim \mathcal{GP}(m, k) \qquad \Leftrightarrow \qquad \widehat{f}(\boldsymbol{x}) \sim \mathcal{N}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}))\,, \tag{2.2}$$

where $\widehat{f}$ denotes a random function draw from the GP and $\sim$ means "is distributed according to". As a multivariate Gaussian distribution, the GP is completely defined by its mean and kernel functions. Their precise choice only aids in faster and more predictive emulation, but they do not in general restrict the shape of the functions being modeled, which can be complete arbitrary as long as the kernel function fulfills a number of weak conditions [73] (that all kernels considered in this work do). Importantly, while the *correlation* of the function value at two points is assumed to be Gaussian, this neither means that the function is itself assumed to be Gaussian, nor that the mean of the family of functions is presumed to be Gaussian.

We usually restrict the GP so that it agrees with the given set of sampling locations for all draws, $\widehat{\boldsymbol{f}}(\boldsymbol{X}) \stackrel{!}{=} \boldsymbol{f}(\boldsymbol{X}) = \boldsymbol{y}$, sometimes up to some uncorrelated Gaussian noise. This information modifies the value of the drawn function's *predictions* $\widehat{\boldsymbol{f}}(\boldsymbol{X}_*)$ away from the sampled values $\boldsymbol{X}$. The joint distribution for sampled and non-sampled locations is

$$\begin{bmatrix} \widehat{\boldsymbol{f}}(\boldsymbol{X}) \\ \widehat{\boldsymbol{f}}(\boldsymbol{X}_*) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{m}(\boldsymbol{X}) \\ \boldsymbol{m}(\boldsymbol{X}_*) \end{bmatrix}, \begin{bmatrix} \boldsymbol{k}(\boldsymbol{X}, \boldsymbol{X}) & \boldsymbol{k}(\boldsymbol{X}, \boldsymbol{X}_*) \\ \boldsymbol{k}(\boldsymbol{X}_*, \boldsymbol{X}) & \boldsymbol{k}(\boldsymbol{X}_*, \boldsymbol{X}_*) \end{bmatrix} \right). \tag{2.3}$$

This defines the *conditional* probability for the predictions $\widehat{\boldsymbol{f}}(\boldsymbol{X}_*)$ given the observations $(\boldsymbol{X}, \boldsymbol{y})$ as

$$\widehat{\boldsymbol{f}}|_{\boldsymbol{f}(\boldsymbol{X})=\boldsymbol{y}} \sim \mathcal{GP}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \Leftrightarrow \qquad \widehat{\boldsymbol{f}}(\boldsymbol{X}_*)|_{\boldsymbol{f}(\boldsymbol{X})=\boldsymbol{y}} \sim \mathcal{N}(\boldsymbol{\mu}(\boldsymbol{X}_*), \boldsymbol{\Sigma}(\boldsymbol{X}_*))\,. \tag{2.4}$$

with mean vector and covariance matrix

$$\boldsymbol{\mu}(\boldsymbol{X}_*) = \boldsymbol{m}(\boldsymbol{X}_*) + \boldsymbol{k}(\boldsymbol{X}_*, \boldsymbol{X})\boldsymbol{k}(\boldsymbol{X}, \boldsymbol{X})^{-1}\left[\boldsymbol{y} - \boldsymbol{m}(\boldsymbol{X})\right]\,, \tag{2.5}$$

$$\boldsymbol{\Sigma}(\boldsymbol{X}_*) = \boldsymbol{k}(\boldsymbol{X}_*, \boldsymbol{X}_*) - \boldsymbol{k}(\boldsymbol{X}_*, \boldsymbol{X})\boldsymbol{k}(\boldsymbol{X}, \boldsymbol{X})^{-1}\boldsymbol{k}(\boldsymbol{X}, \boldsymbol{X}_*)\,. \tag{2.6}$$

This conditioned GP for new sample predictions is then called the *posterior GP*. Comparing equations (2.2) and (2.4) we notice that the drawn samples $\widehat{\boldsymbol{f}}$ differ between the unconditioned and the conditioned GP, because the latter includes the additional information from the sampling locations. The algorithm described in section 5 will sequentially add new samples to the GP. These will be incorporated by *updating* the mean and covariance of this conditioned GP (μ, Σ) using sequentially enlarged sample sets $(\boldsymbol{X}, \boldsymbol{y})$.

From here on, we will use the scalar versions of equations (2.5) and (2.6) evaluated at an arbitrary single location $\boldsymbol{x}$ as $\mu(\boldsymbol{x})$ and $\Sigma(\boldsymbol{x})$, as well as $\sigma(\boldsymbol{x}) = \sqrt{\Sigma(\boldsymbol{x})}$ as the uncertainty of the GP at a location $\boldsymbol{x}$ to simplify notation, implicitly assuming it has been conditioned on the samples $\boldsymbol{X}$. As is standard in the literature (and as discussed without loss of modeling power for the GP), we will assume a zero-mean function $m(\boldsymbol{x}) = 0$ in all cases.

Kernel functions are usually chosen from a particular family of functions (such as squared exponentials, *Matérn* kernels, ...),[3] parameterized by some *hyperparameters* θ. Their value

[3]The kernel function is typically chosen according to the differentiability and smoothness of the given target function, see also section 3.1 for more details.

JCAP10(2023)021

is commonly chosen so that they maximize the likelihood that the GP would have produced the given sampled values $\boldsymbol{y}$ at the sampled locations $\boldsymbol{X}$. In practice, one marginalizes the evidence of the training data given the Gaussian Process [57]:

$$-\log p(\boldsymbol{y}|\boldsymbol{X},\theta) = \frac{1}{2}\boldsymbol{y}^T(\boldsymbol{k}(\boldsymbol{X},\boldsymbol{X}) + \sigma_n^2\boldsymbol{I})^{-1}\boldsymbol{y} + \frac{1}{2}\log|\boldsymbol{k}(\boldsymbol{X},\boldsymbol{X}) + \sigma_n^2\boldsymbol{I}| - \frac{N_s}{2}\log 2\pi\,, \quad (2.7)$$

where $\boldsymbol{I}$ is the identity matrix, and σ_n is an arbitrary small level of uncorrelated *noise* included to make the algorithm more numerically stable (possibly in addition to a noise term added into the kernel function to model stochasticity of the original function). Using Bayes' theorem, the product of this likelihood and some prior can then be sampled or (more commonly) simply maximized with respect to the hyperparameters θ.

3 Surrogate model of the posterior

Our goal is to interpolate an unknown, possibly multi-dimensional log-posterior distribution with a GP, using the mean prediction $\mu(\boldsymbol{x})$ of the GP as a best estimate for the distribution's value. Furthermore we want to achieve an accurate estimate for the standard deviation $\sigma(\boldsymbol{x})$ in order to compute where to sample next. The nature of the posterior distribution being an (un-normalized) probability distribution implies certain properties/restrictions, that can be incorporated into the GP surrogate model in order to increase the performance of the algorithm and reduce the risk of numerical issues. These will be discussed in the following.

3.1 Choice of kernel function

As discussed in section 2.2, a kernel function with a minimal set of properties will ensure that the GP converges towards the target function (the log-posterior) given a large enough set of samples. However, in order the keep the computational costs low, we aim to use as few samples as possible, and this can be achieved by choosing a kernel function that encapsulates our prior information on the posterior distribution.

The prior information that we aim to encode is that the log-posterior distribution is deterministic,[4] and smooth over a characteristic correlation length-scale, that possibly differs between dimensions and is a fraction of the prior size (as we cannot resolve length-scales much larger than the prior). Our default choice in GPry is an anisotropic quadratic RBF kernel multiplied by a constant:

$$k(\boldsymbol{x},\boldsymbol{x}') = c^2 \cdot \exp\left(-\sum_{i=1}^{d}\frac{|x_i - x_i'|^2}{2L_i^2}\right)\,, \quad (3.1)$$

where c is usually called the output-scale, and $L_{i=1,\dots,N_d}$ are the length-scales.[5] On top of the choice of the kernel function itself, prior knowledge on the target function is also incorporated

[4] It would be easy to extend this to stochastic functions by adding a noise component to equation (3.1), but posterior density functions of physical data are most commonly deterministic.

[5] If the covariance matrix of the posterior mode that is modelled is approximately known, and that mode is Gaussian enough, one could transform the parameter space using that covariance matrix so as to normalise the Gaussian, in which case the target function is isotropic and we can use a single common length scale, significantly reducing the computational cost of fitting the hyperparameters. In practice, this approach has its own difficulties: even at late stages of learning, the set of training points is too small to compute the covariance matrix via simple Monte Carlo (weighting by their posterior value), and one needs to resort to other approaches, such as fitting a Gaussian to the training, or MC-sampling from the GP (see e.g. [65]), the cost of which would likely compensate for the time saved by fitting a single isotropic correlation length-scale.

JCAP10(2023)021

in the priors for the hyperparameters. The fundamental assumption is that the length scales should be of an order of magnitude close to that of the posterior modes, and that the latter would be of an order of magnitude not much smaller than that of the prior ranges for the parameters of the posterior. We express this belief as the length-scales being between 0.01 and 1 in units of the prior length in each direction. The lower bound ensures that the GP does not overfit during early stages of the learning by fitting each sample individually as a peak on top of the mean of the GP,[6] while the upper bound represents the fact that the size of the prior box should prevent drawing any conclusions on the characteristic length-scale far beyond the region that can be sampled. The prior of the output scale c is chosen to be very broad and allows for values between 0.001 and 10000. The $N_d + 1$ free hyperparameters $\{c, L_i\}$ are then chosen such that they maximize equation (2.7).[7]

3.2 Parameter space transformations

As a un-normalized probability density, the posterior is a positive function ($p(\boldsymbol{x}) \geq 0$ everywhere), and even for a simple one-dimensional Gaussian, it varies over multiple orders of magnitude. Both enforcing positivity and reducing the dynamic range of function values can be achieved by modeling the result of a power-reduction operation $P\left(p(\boldsymbol{x})\right)$ on the posterior (e.g. a logarithm [61] or a square root [60]). We use a log-transformation, since in physics it is very common for likelihoods to belong to the *exponential family* of probability distributions [74] and in practice many likelihood codes usually return log-probabilities.

Another advantage of modelling in log-space, that was pointed out in [61], is that the characteristic length scale of isotropic kernels (e.g. Radial Basis Function (RBF) or Matérn) tends to be larger, which implies that the GP surrogate better generalizes to distant parts of the function, making the GP more predictive.

In practice, we construct a surrogate model for $\log p(\boldsymbol{x})$ given some training samples $\boldsymbol{y} = \log p(\boldsymbol{X})$. In addition, at every iteration of the algorithm, we *internally* re-scale the modeled function using the mean and standard deviation of the current samples set as

$$\log \widetilde{p}(\boldsymbol{X}) = \frac{\log p(\boldsymbol{X}) - \overline{\boldsymbol{y}}}{s_{\boldsymbol{y}}}\,, \tag{3.2}$$

where $\overline{\boldsymbol{y}}$ and $s_{\boldsymbol{y}}$ are the sample mean and standard deviation respectively. This re-scaling acts like a non-zero mean function, causing the GP to return to the mean value far away from sampling locations. This in turn encourages exploration when most samples are close to the mode and exploitation when most samples have low posterior values. This effect can be seen in figure 3 where the mean of the GP is pushed to higher values close to the edge of the prior. The variance reduction through division by $s_{\boldsymbol{y}}$ aids in ensuring numerical stability by restricting the range of values in the training set.

[6] This condition assumes that the size of the mode is larger than about 1/100th of the prior width in each dimension, which we find reasonably permissive. If this is not the case, either the prior dimensions or the allowed range for the length scales can be re-adjusted.

[7] In a full hierarchical Bayesian treatment, instead of maximising we would have to generate a family of GPs with hyperparameters following the likelihood of equation (2.7), each of them giving different predictions according to equations (2.5) and (2.6). Unfortunately, generating a MCMC sample in order to marginalize over equation (2.7) as function of θ is intractable. There have been some attempts at approximate methods [61] however even those introduce some computational overhead which we want to avoid. Luckily, as the number of training points of the GP increases we expect equation (2.7) to get narrower (for sufficiently tame distributions) so that the difference becomes negligible.

JCAP10(2023)021

As for the space of parameters $\boldsymbol{x}$, we transform the samples such that the prior boundary becomes a unit-length hypercube. For unbounded priors, such as Gaussian or half-Gaussian, we choose the prior boundary such that it contains a large fraction of the prior probability mass (99.95% by default, which is usually sufficient for the usual few-σ CL contours).

This parameter transformation aims at forcing posterior modes to have similar sizes in all dimensions. This usually leads to comparable correlation length scales of the GP across dimensions, which increases the effectiveness of the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS-B) constrained optimizer [75], used to optimize both the GP hyperparameters and the acquisition function.

Henceforth, if not specified otherwise, in the context of the training set we will refer to $\boldsymbol{X}$ (or $\boldsymbol{x}$) as the un-transformed values of the sampling locations while $\boldsymbol{y}$ refers to the un-transformed values of the log-posterior distribution at $\boldsymbol{X}$.

3.3 Treatment of infinities and extreme values

In realistic inference scenarios the prior is often chosen to be much larger than the posterior mode, since very little initial information is usually known. In these scenarios the log-posterior function is bound to return minus infinity for parameter values far away from the region of interest (the posterior modes): the negative log-posterior can be too large to be represented as a floating point number, or the physics code used to compute the likelihood could fail and report a zero-valued likelihood.

This is valuable information but unfortunately we cannot simply add those infinite values to the GP as equations (2.5) and (2.7) become ill-defined. Hence we are forced to find some numerically stable way of incorporating this information. Naively one could simply swap out the infinities with some large negative value. This approach turns out to be rather problematic as it introduces a discontinuity in the posterior shape or at least one of it derivatives thus modifying the hyperparameters of the GP's kernel. If we instead ignore these points, the learning algorithm will repeatedly try to acquire points in their vicinity, hence getting stuck.

Our solution to this problem is to simultaneously exclude these *infinities* from the GP, and to use them to divide the parameter space into a *finite* and an *infinite* region using a support vector machine (SVM) classifier [58, 59].[8] A SVM defines a hyperplane which maximizes the separation between samples with locations $\boldsymbol{x}_i$ belonging to one of two classes $y \in \{-1, 1\}$. By defining the distance between points through a kernel function $k(\boldsymbol{x}, \boldsymbol{x}')$ the separating hyperplane is drawn in a higher-dimensional space which is connected to the sample space by a non-linear transformation. This effectively transforms the separating hyperplane into more complex hypersurfaces which are better suited to the classification problem at hand.

The categorical predictions $\hat{y}(\boldsymbol{x})$ of the SVM are then given by

$$\hat{y}(\boldsymbol{x}) = \operatorname{sgn}\left(b + \sum_{i=1}^{N_s} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x})\right) \tag{3.3}$$

where the hyperparameters b and α_i are optimized in the training procedure.

We simply use the prediction of the SVM of whether a point is classified as being finite ($\hat{y} = +1$) or infinite ($\hat{y} = -1$) to "correct" the prediction of the GP. Compared to

[8] A similar "safe exploration space" approach, using different tools, has also been used e.g. [76, 77] in the context of Bayesian optimization.

JCAP10(2023)021

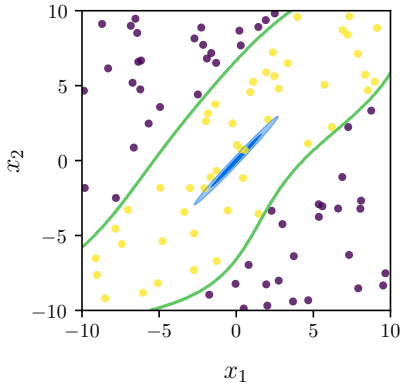


Figure 1. Illustration of the SVM classification. Yellow dots correspond to uniformly sampled locations where the log-posterior distribution is finite while purple dots correspond to infinite log-posterior samples. The green lines are the boundary found by the SVM separating the finite and infinite regions. The blue contours show the 1- and 2-σ contours of the posterior distribution (in this case a correlated 2-d Gaussian). In our construction this finite region is designed to roughly correspond to the 10σ volume of the Gaussian distribution.

equations (2.5) and (2.6) we can explicitly write

$$\mu_{\text{GP+SVM}}(\boldsymbol{x}) = \mu(\boldsymbol{x}) \cdot \begin{cases} 1 & \text{if } \hat{y}(\boldsymbol{x}) = +1 \\ -\infty & \text{if } \hat{y}(\boldsymbol{x}) = -1\,. \end{cases} \tag{3.4}$$

For now, we assert such classification from the SVM with absolute certainty, and set

$$\Sigma_{\text{GP+SVM}} = \Sigma(\boldsymbol{x}) \cdot \begin{cases} 1 & \text{if } \hat{y}(\boldsymbol{x}) = +1 \\ 0 & \text{if } \hat{y}(\boldsymbol{x}) = -1\,. \end{cases} \tag{3.5}$$

The precise way of cutting the covariance is irrelevant in our case.[9] Figure 1 shows a two-dimensional toy example of such a classification for a Gaussian distribution in a comparatively much larger prior region.

Aside from making the acquisition procedure more efficient by ignoring unimportant regions, this approach also keeps the overhead cost of the algorithm lower than including a regularized version of the inifinities in the GP. This is because the computational expense of training a SVM scales as N_s^2, which is smaller than the N_s^3 scaling for the GP.

It is important to recognize that the same arguments can be made for very low posterior values which are far away from the top of the mode, even for well-behaved posterior distributions in high dimensionality. While the SVM is not strictly needed here, adding these values to the surrogate GP model is undesirable as they can dramatically change the scale of the emulation problem even though they do not provide a large amount of additional

[9]Still, one could imagine using the SVM output before sgn function (the classification step) to more smoothly suppress both mean and covariance, possibly combined with a sigmoid function.

JCAP10(2023)021

information. In this sense, the algorithm also benefits from a regularization of forwarding too small log-posterior values to the SVM.

We accomplish that by treating all values where $\log p(\boldsymbol{x})$ is smaller than some (sufficiently low) threshold as infinities. However, one has to be careful about the un-normalized nature of the posterior when applying the threshold. In practice, we compare against the maximum of the posterior in the training sample (corresponding to point $\boldsymbol{x}_{\rm max}$) and only treat values as infinite when $\log p(\boldsymbol{x}) < \log p(\boldsymbol{x}_{\rm max}) - T$. We provide a default value for T based on the prescription of appendix A, also giving the user the option to set it manually.

Lastly, we stress that the additional modelling presented in this section is only used in practice if the log-posterior distribution ever returns either negative infinities or values below the proposed threshold. Otherwise, only the bare GP model described in the previous sections is used.

4 Learning strategy

In section 3 we have described the process of constructing a Gaussian process to emulate the log-posterior distribution once a given set of samples are known. As discussed, a sufficiently large naive set of samples (e.g. prior samples) will in general lead to an accurate model. Unfortunately the computational cost of the algorithm scales with the number of samples N_s, both directly as the number of times a possibly-costly true posterior needs to be evaluated, and indirectly by increasing the computational cost of the Gaussian Process itself (as $\sim N_s^2$ at evaluation, and $\sim N_s^3$ when fitting). In practice, samples are chosen so that their location maximises an *acquisition function*, representing some measure of how valuable they would be for the emulation when added to the GP. We discuss this approach in section 4.1. A further reduction in computational costs can be achieved by taking advantage of the number of machines/CPUs in computing clusters (and of CPU cores in user-level CPUs). Thus, we discuss the parallelization of the algorithm in section 4.2. Finally, in section 4.3 we discuss the vital question of when to end the acquisition of further samples automatically. Together, this allows `GPry` to tackle the emulation of arbitrary distributions in a highly parallelized way without relying on the end-user to optimize the number or locations of the samples.

4.1 Acquisition function

As discussed above, in order to find a small, but informative set of sampling locations, we will look for locations that maximize an *acquisition function* $a(\boldsymbol{x})$. This function will be constructed using a combination of the mean and variance of the GP estimate, in such a way that it balances exploration of the full parameter space (typically where the uncertainty in the prediction is high) with exploitation of areas of high posterior values (which should be more precisely modeled).

4.1.1 Choice of the acquisition function

A simple ansatz for an acquisition function that balances exploration and exploitation could be the product of the estimated posterior $p(\boldsymbol{x})$ (which is always positive) and its uncertainty $\sigma_p(\boldsymbol{x})$:[10]

$$a_p(\boldsymbol{x}) = p(\boldsymbol{x}) \cdot \sigma_p(\boldsymbol{x})\,. \tag{4.1}$$

[10]This is by no means the only ansatz one could make to arrive at a suitable acquisition function. For a thorough investigation see [78].

JCAP10(2023)021

Given that the GP models $\log p$, we have to convert the GP's mean $\mu(\boldsymbol{x})$ and uncertainty $\sigma(\boldsymbol{x})$ into those of the linear $p(\boldsymbol{x})$. Since the transformation from $\log p$ to p is non-linear, the corresponding prediction for p from the GP is not a Gaussian distribution and the computation of its mean and standard deviation is non-trivial. However, in practice these details are irrelevant since the acquisition function only needs to approximate the most beneficial sampling location. Then, we can simply write for the mean $p(\boldsymbol{x}) \approx \exp[\mu(\boldsymbol{x})]$ and for the uncertainty $\sigma_p \approx \exp[\mu(\boldsymbol{x}) + \sigma(\boldsymbol{x})] - \exp[\mu(\boldsymbol{x})]$. With this, the acquisition function above becomes:

$$a_p(\boldsymbol{x}) \approx \exp[2\mu(\boldsymbol{x})] \left\{\exp[\sigma(\boldsymbol{x})] - 1\right\} , \tag{4.2}$$

which is similar to the acquisition functions used in [79, 80]. This approximation can be further linearized assuming $\sigma(\boldsymbol{x}) \ll 1$ to give $a_p^{\rm lin}(\boldsymbol{x}) = \exp[2\mu(\boldsymbol{x})]\sigma(\boldsymbol{x})$.

As discussed in the next section, we found it beneficial to boost the exploratory behaviour of the acquisition function, especially in high dimensions. To achieve that, we include a relaxation factor $\zeta \in (0, 1]$ multiplying the mean to discourage exploitation (similar to what was done in e.g. [65]). This yields the final acquisition function:

$$a_p(\boldsymbol{x}) \approx \exp[2\zeta\mu(\boldsymbol{x})] \left\{\exp[\sigma(\boldsymbol{x}) - \sigma_n] - 1\right\} \tag{4.3}$$

which can again be linearized as $a_p^{\rm lin}(\boldsymbol{x}) = \exp[2\zeta\mu(\boldsymbol{x})][\sigma(\boldsymbol{x}) - \sigma_n]$. Notice also that from the $\sigma(\boldsymbol{x})$ term we have subtracted a possible uncorrelated noise term proportional to σ_n^2 in the kernel function (equivalently, a constant term added to the diagonal of the kernel covariance matrix). This is because for acquisition purposes we only care about the uncertainty coming from decorrelation from the sampled locations.

The logarithm of this acquisition function is maximized at every acquisition step which yields a candidate for the next sampling location. The computational overhead of the acquisition procedure is dominated by the prediction of $\mu(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$ by the GP which scales as $\sim N_s^2$.

4.1.2 Acquisition hyperparameter

The effect of ζ in equation (4.3) is that of balancing exploitation and exploration. Values of ζ that are too high make the algorithm focus too much on the top of a posterior mode, so that samples in the tails are unlikely to be proposed, and during large numbers of iterations the GP model is mostly stable (only adding high-posterior but low-information samples). This leads to unnecessarily high computational costs, and often to false positives in assessing convergence. These effects are more dramatic in higher dimensions. On the other hand, values of ζ that are too low would produce more regular but slower convergence, neglecting information about the expected value of the true function that could have been exploited to converge faster. In general, a sub-optimal choice of ζ will increase the amount of samples necessary for convergence, sometimes quite significantly.

To select appropriate values, we have conducted a series of experiments on degenerate Gaussian posterior distributions in 2, 4, 8 and 16 dimensions (for $N_d < 4$ the effect of ζ is small), generated as explained in section 6.1. In order to isolate the effects of ζ, in these tests we have not used the parallelization scheme described in section 4.2. The results of these experiments in terms of KL divergence (see appendix B) are shown in figure 2, and have led us to propose the empirical formula $\zeta = N_d^{-0.85}$ as a default value for ζ (users can override it if prior knowledge of the posterior shape suggests that exploration should be favored over exploration or vice versa). Preliminary tests in higher dimensions (up to $N_d = 27$) have

JCAP10(2023)021

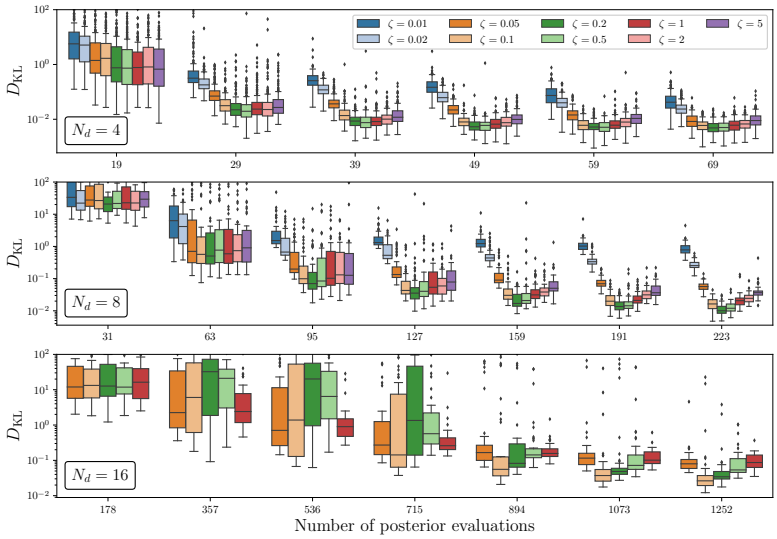


Figure 2. Distribution of Kullback-Leibler divergences between the GP prediction and the true distribution at various learning stages (i.e., N_s samples) for random correlated Gaussian posteriors with dimensionality $N_d = 4$, 8, and 16 (150, 50, and 50 realizations, respectively). The boxes represent inter-quartilic ranges, the black line inside them the median, and the whiskers and dots represent the tails of the distributions. For each dimensionality there is a visible trend towards an optimal trade-off between exploration and exploitation in terms of ζ.

shown this formula to produce good results. The fact that a fixed ζ becomes greedier as dimensionality goes up should not come as a surprise, as discussed in appendix A.

4.1.3 Optimization of the acquisition function

For the maximization of the acquisition function we use the L-BFGS-B optimizer [75] included in the `scipy` Python package. Since this optimization problem is highly non-convex, with the acquisition function often having many disconnected maxima, the numerical optimization is performed multiple times from different randomly-drawn starting locations. In high dimensions drawing an initial point with a non-vanishing value of the acquisition function becomes increasingly unlikely as the prior volume with vanishing posterior increases as a power of the dimension (curse of dimensionality).

Because of this problem, optimal proposals most likely fall in the vicinity of the current sampling locations. In order to generate such points we, by default, use a *centroid* algorithm: take the average location of $N_d + 1$ randomly selected samples, and perturb them in each dimension by the coordinate difference to one the samples multiplied by a draw from an exponential distribution with parameter $1/\lambda$. Here a lower λ increases the spread of the proposed locations. A fraction of the locations are drawn from a uniform distribution within the original prior boundaries, in case a region of high posterior has not yet been captured by the current samples.

JCAP10(2023)021

For highly non-Gaussian distributions this method of proposing points tends not to be exploratory enough. In these cases we resort to drawing proposals uniformly within the prior volume.

Lastly also provide a method to generate Gaussian-distributed proposals given an estimate of the mean and covariance matrix of the posterior, if such information happens to be known.

We notice that alternative approaches to maximizing the acquisition function exist. In [65], the sampling locations with high acquisition function value are picked out of an MCMC of the mean GP model.

4.2 Parallelization

The naive approach of using the acquisition function presented above is to acquire and evaluate sampling locations in sequence, with each acquisition step consisting of the evaluation of the true posterior distribution (and updating the GP model) in order to obtain the next candidate for a sampling location. However, as often multiple processing units (either on the same or across different machines) are available, we can make this algorithm more efficient by attempting to propose *batches* of sampling locations, so that the true posterior, which is expected to be the largest source of computational cost, can be evaluated in parallel.

There have been many different proposals for batch acquisition for GPs in the past which can broadly divided into two categories: algorithms like [81–84] construct an acquisition function which can be optimized for several points at once. However, for a d-dimensional posterior distribution acquiring q points at once involves global optimization in $d{\cdot}q$ dimensions which obviously becomes computationally prohibitive even if d and q are not extremely large.

The second category [83–85] works by sequentially acquiring multiple points without having to sample from the posterior distribution in between and afterwards evaluating the true posterior at the gathered locations in parallel. We will be using one of these methods called the *Kriging believer* method [84].

The Kriging believer method. The fundamental assumption of the Kriging believer method (similarly to our assumption when constructing the acquisition function) is that the value of the posterior distribution in any point roughly equals the predicted mean of the GP. We can therefore acquire a batch of points by sequentially (1) obtaining a maximum of the acquisition function at $\boldsymbol{x}_*$, (2) assuming for it a log-posterior evaluation equal to $\mu(\boldsymbol{x}_*)$, (3) adding it to an intermediate *augmented* GP (thereby producing a different new maximum of the *augmented* acquisition function), and repeating until the desired number of locations has been proposed. This method will be increasingly accurate as more samples are added to the GP so that $\mu(\boldsymbol{x}_*)$ approaches the true $\log p(\boldsymbol{x}_*)$. An illustration of the Kriging believer algorithm sampling on the log of a normal distribution is shown in figure 3.

The obvious advantage of this method, as discussed above, is that the true posterior can be evaluated in parallel for the acquired locations. This is beneficial as we expect the true posterior evaluations to dominate the computational cost in most scenarios. In addition, there is another source of speedup: since adding new mean-valued samples does not change the optimal hyperparameters of the GP according to equation (2.7), there is no point to re-fitting them (see section 5.3).[11]

[11] On top of that, the necessary step of inverting the kernel matrix after adding new points, in order to get predictions using equations (2.5) and (2.6), could be accelerated by taking advantage of the fact that the inverse of the previous kernel matrix is known, using a fast, blockwise matrix inversion formula. To our knowledge this has not been pointed out in the past. In our case, the amount of possible time savings is small.

JCAP10(2023)021

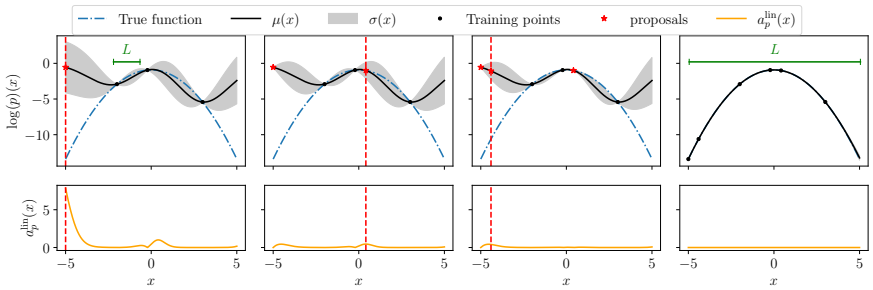


Figure 3. Illustration of the Kriging believer method. Three points are acquired sequentially (three left plots) by using the prediction from the GP instead of evaluating the posterior at each iteration. After the three samples have been acquired the posterior function can be evaluated at these points (right). The hyperparameters of the GP regressor only need to be refit at the last step. Obviously using this approach comes at the expense of requiring more points to converge (e.g. the third point did not add much information and is unlikely to have been selected after the second one if using sequential acquisition). This can however be compensated by the computation time that is saved by both acquiring points faster and evaluating the posterior in parallel. The characteristic length-scale L of the kernel increases as more samples are added, which aids the better fit in the right panel.

In terms of the precise size of the batch, there is evidently a trade-off between the speedup gained by not refitting the GP's hyperparameters at every iteration, and the inaccuracy of the GP's mean prediction making the active sampling less efficient as the number of Kriging believer steps grow. Due to the loss of accuracy in the predictions, more samples are required to converge to the true distribution, but this is compensated by the speedup achieved through parallel evaluations of the posterior. Overall this results in a smaller number of iterations (hence a smaller wall-clock run time) than sequential learning, as long as the size of the batches is kept reasonable.

We find that a batch size corresponding to at most the number of dimensions of the inference problem N_d works reasonably well. We therefore set the standard number of Kriging believer steps to the minimum between N_d and the number of parallel processes.

4.3 Convergence criterion

The last component of our algorithm is its *convergence criterion*, which should terminate it as soon as (or at least not much later than) the GP has reached sufficient precision at modelling the log-posterior. Precision could be assessed as the reduction in the variance of a GP-predicted global posterior quantity such as the evidence $\int p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. Analytical computation of these quantities in terms of the GP are usually not possible, e.g. in our case because of the modelling of the log-posterior instead of the posterior itself, or because the product of a GP times arbitrary priors does not have a closed-form integral in general. Numerical approaches would involve MC samples of the GP-modelled posterior, which come at a reasonably-small computational cost, but whose use for the convergence criterion would involve obtaining them at (nearly) every iteration.

A much cheaper convergence criterion would involve computations using the much smaller set of current and/or proposed GP samples. We propose one such criterion, that we call `CorrectCounter`, based on observing the accuracy of the learning process and stop-

JCAP10(2023)021

ping when the model does not seem to learn any new information. We will show how that the speedup in this case does not necessarily come at the cost of precision.

The assumption here is that our algorithm stops learning if the GP's predictions at newly acquired sampling locations $\boldsymbol{x}$ repeatedly match the value of the true log-posterior $\log p(\boldsymbol{x})$ distribution to close approximation. We set a threshold using relative and absolute tolerances $\epsilon_{\rm abs}$, $\epsilon_{\rm rel}$ such that

$$|\mu_{\rm GP+SVM}(\boldsymbol{x}) - \log p(\boldsymbol{x})| \overset{!}{<} \epsilon_{\rm abs} + |y_{\rm max} - \mu_{\rm GP+SVM}(\boldsymbol{x})| \cdot \epsilon_{\rm rel}\,, \tag{4.4}$$

where $y_{\rm max}$ is the largest log-posterior from the current GP sample, and $\mu_{\rm GP+SVM}(\boldsymbol{x})$ is the GP's prediction at $\boldsymbol{x}$ before the GP has been fit to this point. This criterion can be computed at virtually no cost, since both $\mu_{\rm GP+SVM}(\boldsymbol{x})$ and $\log p(\boldsymbol{x})$ have been computed as part of the acquisition procedure. If this condition is satisfied a few times in a row we consider the model converged and stop the algorithm. Convergence in this case means a guarantee that (on average) new evaluations of the GP will at least approximately comply with the true posterior at the same location (as opposed to convergence meaning stability of some global quantity).

Similarly to the discussion in sections 3.3 and 4.1.2, the behaviour of this convergence criterion is sensitive to the dimensionality N_d of the problem. As explained in appendix A, since the dynamic range of a log-posterior enclosing a given probability mass grows with dimensionality, the effect of a constant $\epsilon_{\rm abs}$ will become more stringent as dimensionality increases, making the criterion fail to report as converged GP models that already very precisely characterise the posterior. In appendix A we propose a way to relax $\epsilon_{\rm abs}$ in a dimensionally-consistent way. The relative thresold $\epsilon_{\rm rel}$ should not be affected by dimensionality, and it is fixed to 0.01. In both cases, we also give the user the option to set their own values for the convergence criterion.

On the other hand, as the number of dimensions N_d increases, correctly mapping the tails of the distribution becomes increasingly more important (for a detailed discussion see appendix A), while the surrogate model tends to converge first around the maximum of true posterior distribution. The tails usually remain underrepresented at first and only get explored later in the acquisition procedure. A higher dimensionality therefore makes it likelier to acquire a batch of consecutive correctly-predicted points in a non-converged GP model around the top of the mode. We account for this by increasing the number of times points have to be predicted correctly to claim convergence to $n = N_d/2$ (with the exception of fixing $n = 4$ for low dimensionality, $N_d < 8$). This reduces the risk of neglecting convergence at the tails.

We tested the `CorrectCounter` criterion on a set of correlated Gaussians in 2, 4, 8, 12 and 16 dimensions, generated as explained in section 6.1. We target a KL divergence with respect to the true Gaussian distribution of less than 5%. As shown in figure 4, we achieve such threshold with the settings described above for the tolerances and the number of consecutive correct predictions, at least for the range of dimensionality targeted in this study. Towards higher dimensionality there is a trend to converge before the convergence curve flattens out entirely, which hints at the need for more sophistication in dealing with dimensional consistency. We leave this for future work.

We also note that we have written a criterion based on the costlier KL divergence (see appendix B), which we provide as an alternative option. This alternative criterion is based on the posterior emulation stabilizing over multiple subsequent steps (defined through the KL divergence being below some critical threshold). This criterion comes with its own sets of

JCAP10(2023)021

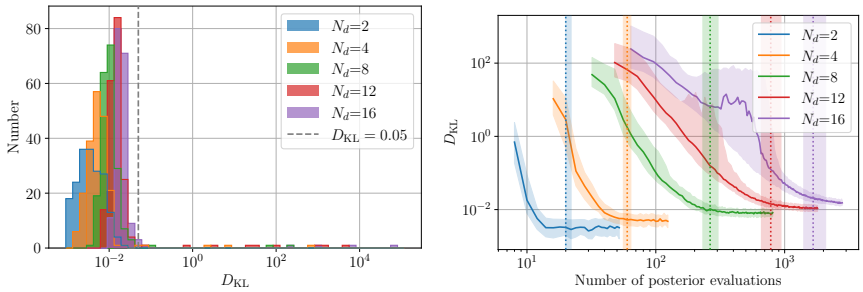


Figure 4. *Left*: distribution of KL divergences between GP models and their correspoding true posterior at `CorrectCounter`-reported convergence, for $N_d = 2, 4, 8, 12, 16$-dimensional random correlated Gaussians (200 draws per dimensionality). Only a small fraction ($< 5\%$) surpass our target value of $D_{\rm KL}^{\rm sym} = 0.05$. *Right*: medians (solid lines) and interquartile ranges (shaded bands) of the KL divergences between GP models and their correspoding true posterior, for the same sets of Gaussians, as function of their number of accepted (finite) samples. The dashed vertical lines indicate the median number of accepted steps at which `CorrectCounter` reports convergence, and the shaded vertical bands the respective interquartilic ranges. As expected there is a trend towards higher values of $D_{\rm KL}^{\rm sym}$ visible as N_d increases, but it is well under control for the dimensionalities targeted in this study.

challenges, such as incorrectly detecting convergence when non-informative points are added to the GP or the costly nature of its computation. Nonetheless it can be preferable when the log-posterior function is extremely expensive to evaluate or when the posterior distribution exhibits unusual features, as this convergence criterion is not only sensitive to the acquired samples but also to the hyperparameters of the GP.

5 The full algorithm

In this section we present the full structure of the algorithm, entailing the generation of the initial set of training samples (section 5.1), the main acquisition loop that sequentially looks for optimal samples and checks convergence (section 5.2), and the final generation of a Monte Carlo sample of the trained GP surrogate model of the posterior, which can be used to get marginalised quantities (section 5.3), together with a comparison of computational costs of this algorithm against those of classic Monte Carlo.

5.1 Initial training set

In order to start the sequential acquisition of points we need an initial training set containing samples from our posterior distribution. These do not have to be very informative samples but need to be finite (according to the definition in section 3.3) and uncorrelated, in order to generate some very crude but meaningful initial interpolation of the log-posterior distribution.

Of course we want to choose this sample such that the ratio of finite to infinite log-posteriors is reasonably high, in order not to waste too many posterior evaluations on the initial point generation. In low dimensions and with small priors compared to the size of the mode, any random generator (such as draws from the prior itself, or from a uniform distribution within the prior bounds) would produce initial samples satisfying the requirement above. As the ratio of the prior to posterior volume grows with the number of dimensions,

JCAP10(2023)021

randomly drawing a finite point from the prior becomes increasingly unlikely. In this case, prior knowledge of the posterior can be incorporated, usually in the form of a "reference" distribution which is a rough guess of where the mode might be (the same that is commonly used to generate initial points for MCMC). In general any guess for reasonable parameter values that lead to a finite posterior can be used, which can be obtained from physical considerations of the underlying model.

5.2 Main algorithm

In algorithm 1 we show the main algorithm used within the `GPry` tool in pseudo-code, consisting mostly of the optimization and acquisition loops (the latter based on the Kriging believer approach). This pseudo-code mostly summarizes the ideas which are explained in the corresponding sections 3 and 4.

Note that the $n_{r,\mathrm{GP}}$ starting locations for the optimization of the hyperparameters in line 4 are sampled logarithmically in the hypervolume. The step of line 4 is currently the most expensive step, scaling as N_s^3 due to the required repeated matrix inversion required for computing $\log p(\theta|\boldsymbol{X}, \boldsymbol{y})$. This is why we only perform this step every n_{opt}-th time, and otherwise we optimize the hyperparameters starting only from the previous best fit. The next most expensive step is the acquisition function optimization in line 12, and scales approximately as N_s^2 due to the repeated evaluation of the acquisition function requiring the evaluation of $a(\boldsymbol{x})$, which itself requires matrix multiplications.

5.3 Modelling the marginalized posterior

As mentioned above, to compute marginalized 1D/2D posteriors, we have to compute a high-dimensional integral of our emulated posterior (see section 2.1). This can be achieved by integrating the GP numerically through the creation a Monte Carlo sample, either based on nested sampling, Metropolis Hastings sampling, or (using the backward differentiable nature of the GP) even Hamiltonian sampling. As `GPry` is interfaced with the `Cobaya` package [86], its standard samplers can also be used to generate the final MCMC sample. Currently, this sampling is performed using the GP's mean prediction according to equation (3.4) as the posterior distribution to sample.[12]

One important question that such an approach poses, however, is whether the emulation of the posterior with the GP with subsequent sampling of the surrogate posterior will be computationally more efficient than the direct sampling of the true posterior. For this, let us use a simple back-of-the-envelope computation. Consider the time to run a full sampling of the true posterior as $N_t t_t$, where t_t is the approximate time for a single evaluation and N_t the total number of samples required. Instead, the time to run a sampling of the GP posterior can be estimated as $N_g t_g$, where t_g is the average time for a single GP evaluation and N_g the total number of required GP samples. Additionally, and crucially, there is the additional overhead of constructing the GP in the first place, which we will denote simply as T_o for now (we will discuss this in more detail later). In that case, the construction of a GP is advantageous if

$$T_o + N_g t_g < N_t t_t \,. \tag{5.1}$$

Typically it can be assumed that $t_g \ll t_t$ except for very simple toy models. Furthermore, typically $N_g \simeq N_t$ if one uses MCMC/nested sampling methods to sample the GP, or even

[12] Technically, the information that is available through the covariance of the GP could be used to obtain an estimate of the uncertainty of emulation on our final posterior sample. As the acquisition procedure only stops if the posterior mode is mapped accurately enough, this assures that at convergence this variance is sufficiently small to safely be neglected.

JCAP10(2023)021

Input: $\boldsymbol{X}$ (initial samples), $\boldsymbol{y}$ (initial log-posterior values)
[1] **for** $n < N_{\max}$ **do**
[2] | fit SVM with $\boldsymbol{X}$, $\boldsymbol{y}$
[3] | **every** $n_{\text{opt}} - - -$ th **time**
[4] | | find $\theta_{\text{MAP}} = \text{argmax}[\log p(\theta|\boldsymbol{X}, \boldsymbol{y})]$ from $n_{r,\text{GP}}$ starting locations — equation (2.7)
[5] | **otherwise**
[6] | | find $\theta_{\text{MAP}} = \text{argmax}[\log p(\theta|\boldsymbol{X}, \boldsymbol{y})]$ from last best-fit — equation (2.7)
[7] | **end**
[8] | `GP_fit`($\boldsymbol{X}$, $\boldsymbol{y}$)
[9] | $\boldsymbol{X}_{\text{new}} = [\,]$
[10] | $\boldsymbol{X}_{\text{lie}} = \boldsymbol{X}$ and $\boldsymbol{y}_{\text{lie}} = \boldsymbol{y}$
[11] | **repeat** M **times**
[12] | | find $\boldsymbol{x}_{\text{add}} = \text{argmax}[a(\boldsymbol{x})]$ starting from $n_{r,\text{acq}}$ starting locations
[13] | | $\boldsymbol{X}_{\text{lie}}$ *append* $\boldsymbol{x}_{\text{add}}$ and $\boldsymbol{X}_{\text{new}}$ *append* $\boldsymbol{x}_{\text{add}}$
[14] | | $\boldsymbol{y}_{\text{lie}}$ *append* $\boldsymbol{\mu}(\boldsymbol{x}_{\text{add}})$ — Kriging believer
[15] | | `GP_fit`($\boldsymbol{X}_{\text{lie}}$, $\boldsymbol{y}_{\text{lie}}$)
[16] | **end**
[17] | $\boldsymbol{y}_{\text{true}} = \log \mathcal{L}(\boldsymbol{X}_{\text{new}}) + \log \pi(\boldsymbol{X}_{\text{new}})$ — parallelizable
[18] | $\boldsymbol{X}$ *append* $\boldsymbol{X}_{\text{new}}$
[19] | $\boldsymbol{y}$ *append* $\boldsymbol{y}_{\text{true}}$
[20] | **if** *`is_converged`* (e.g. equation (4.4)) **then** *break*
[21] **end**
[22] Sample $\mu(\boldsymbol{x})$ with MC sampler
[23] **return** *MC sample*

[24] **Function** *`GP_fit`*$(\boldsymbol{X}, \boldsymbol{y})$
[25] | Compute $K^{-1} = \boldsymbol{k}(\boldsymbol{X}, \boldsymbol{X}|\theta_{\text{MAP}})^{-1}$ — matrix inversion
[26] | $\mu(\boldsymbol{x}) = \mu_{\text{GP+SVM}}(\boldsymbol{x})$ — equations (2.5) and (3.4)
[27] | $\sigma(\boldsymbol{x}) = \sqrt{\Sigma_{\text{GP+SVM}}(\boldsymbol{x})}$ — equations (2.6) and (3.5)
[28] | $a(\boldsymbol{x}) = \exp[2\zeta\mu(\boldsymbol{x})]\{\exp[\sigma(\boldsymbol{x}) - \sigma_n] - 1\}$ — equation (4.3)
[29] **end**

Algorithm 1. The `GPry` algorithm in a condensed format, omitting the internal transformations that are made to the data. M is the number of Kriging believer steps made in each iteration. The overhead of the algorithm is dominated by the computations performed in lines 12 and 4.

$N_g \ll N_t$ if one can use Hamiltonian MC methods on the GP but not on the true posterior. Thus, as long as T_o remains reasonably lower than $N_t t_t$ (the total runtime of the MCMC), the use of a GP would always be advantageous. It is thus crucial to obtain a precise estimate for the overhead time T_o. This overhead depends strongly on the dimensionality of the problem, the non-Gaussianity of the posterior, and the underlying machine executing the code.

Looking at the timing information from the multivariate Gaussian cases of section 6.1, the overhead was dominated by the numerical optimization of the acquisition function (line 12 of algorithm 1), taking very roughly $100s \cdot (N_s/100)^{2.4}$ (to give an order-of-magnitude estimate). The next most important factor, the optimization of hyperparameters (line 4 of al-

JCAP10(2023)021

gorithm 1) only takes around $3s \cdot (N_s/100)^{3.2}$ (order of magnitude) in total. It has a smaller pre-factor since it is only performed every n_{opt}-th iteration, while the acquisition optimization is performed $M \cdot n_{\text{r,acq}}$ times per iteration, see algorithm 1. It is thus comparatively irrelevant for $N_s \ll 10^4$, which is almost always the case for the range of dimensionalities considered in this study.

In figure 5 we report the approximate expected total runtime of `GPry` compared to the `Cobaya` implementation of the MCMC sampler `CosmoMC` [1, 2, 86] and the nested sampler `PolyChord` [9, 10] (via its `Cobaya` interface). For each dimensionality, these estimates were generated by drawing a large set of random multivariate Gaussians, and computing the distrubution of total evaluations needed for convergence (according to their respective default criteria) for MCMC, `PolyChord` and `GPry`. We multiply these numbers of posterior evaluations with the posterior evaluation times on the x-axis and add the overhead of each algorithm to get the total runtimes on the y-axis. For `GPry`, the computational overhead is caused by the optimization of the acquisition function and the fitting of the GP hyperparameters, and it is constant with respect to the posterior evaluation time, producing the particular shape of the curve. We neglect the overhead of MCMC and `PolyChord` as it is tiny compared to the overhead of `GPry` [86].

For example, in the case of a $N_d = 12$ multivariate Gaussian, `GPry` would outperform the MCMC (which requires $\approx 1.5 \cdot 10^5$ evaluations) for posterior evaluation times larger than ~ 0.1 seconds. Comparing to the average runtime of a cosmological code such as `CLASS`, on average we find a significant speedup all the way up to 16 dimensions.

Note that in figure 5 we show single-core performance with as many Kriging believer steps as dimensions (while still evaluating the posterior sequentially). The curves shown for MCMC and especially for `PolyChord` would drop almost proportionally to the number of cores available, while `GPry` does not scale quite as well. However, for a similar amount of computational resources, up to a number of processes similar to the dimensionality of the problem, these results are expected to hold in order of magnitude. While the runtime of MCMC and `PolyChord` is dominated by the posterior evaluations, the overhead of `GPry` is considerable and might scale differently depending on the underlying architecture. Further improvements in runtime could be made by optimizing the underlying GP implementation.

6 Examples

After having discussed the design of the `GPry` code in sections 3 to 5, we now demonstrate the performance of the code using a variety of examples, both Gaussian and non-Gaussian distributions considered in the literature, as well as examples from cosmological applications.

For each of the examples in this section, we will analyze the performance of `GPry` in terms of convergence by producing a number of runs with identical `GPry` settings (same choices of kernel functions, acquisition function and other training settings) but different random seeds, so that they start from different initial training samples (uniformly drawn from the prior) and find generally different optima for the acquisition function and the GP hyperparameters (maximizations are started from random starting positions). On top of the intrinsic variability between runs, the covariance matrices and means of the Gaussian examples in section 6.1 and the log-Gaussian ones in section 6.2.1 are drawn randomly for every run to make the tests more robust, whereas for the rest of non-Gaussian and multimodal examples, as well as the cosmological ones, the posteriors are fixed.

JCAP10(2023)021

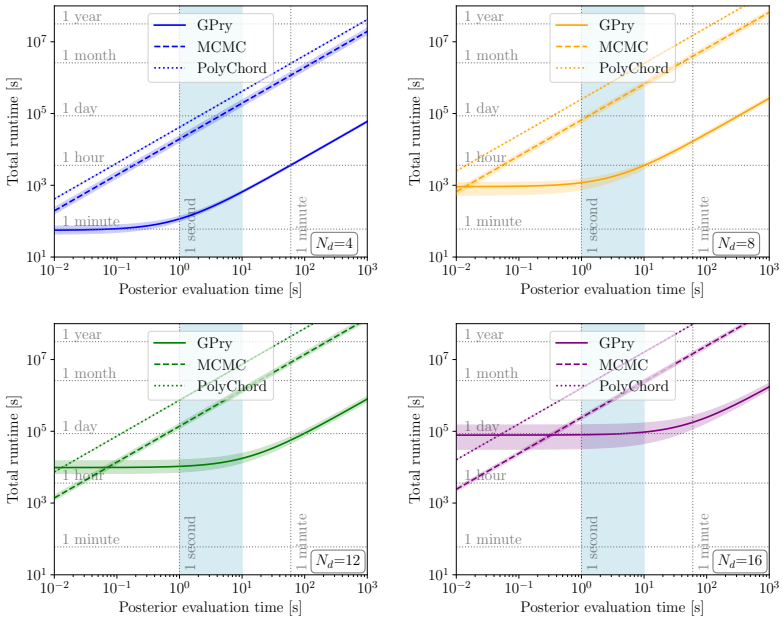


Figure 5. Order of magnitude estimate of total runtime comparison of `GPry` with the MCMC sampler `CosmoMC/Cobaya` and the nested sampler `PolyChord` (via its `Cobaya` interface). The comparison is done for multivariate Gaussians of various dimensionalities, and shows the median as a line and the 25% and 75% quantiles as a shaded area. The comparison is run with only a single CPU, but the orders of magnitude hold for similar computational resources for all three methods. The light blue band gives an approximate range of computation times of standard cosmological codes (like `camb` or `CLASS`) which depend strongly on the considered model and observables. Note that while MCMC and `PolyChord` are dominated by the posterior evaluation time everywhere, `GPry` is dominated by overhead for small posterior evaluation times.

6.1 Multivariate Gaussians

The example of a multivariate Gaussian distribution is enlightening as a benchmark for the average performance of the GP, as it can quite trivially be scaled with dimensionality and many likelihood functions can — at least around their maximum — be reasonably well approximated by Gaussian distributions. We can thus use it as a benchmark for performance and accuracy as a function of dimensionality, as well as to model critical scalings such as that of the ζ parameter from section 4.1, the factors involved in equation (4.4), and the timings relevant for section 5.3 (see appendix A).

We generate correlated multidimensional Gaussians with log-likelihood function

$$\log \mathcal{L}(x_0, \ldots, x_n) = -\frac{(\boldsymbol{x} - \mathfrak{m})^T \boldsymbol{C}^{-1} (\boldsymbol{x} - \mathfrak{m}) + \log((2\pi)^n |\boldsymbol{C}|)}{2} \tag{6.1}$$

by drawing a random covariance matrix that satisfies

$$\boldsymbol{C}_{i,j} = \sigma_i \sigma_j \mathrm{corr}_{i,j} \tag{6.2}$$

JCAP10(2023)021

where $\text{corr}_{i,j}$ is a randomly drawn correlation matrix with uniformly drawn eigenvalues,[13] and the standard deviations are uniformly drawn as $\sigma_i \in [0, 1]$. The mean vector $\mathfrak{m}$ is set to 0 and the prior fixed to $5\sigma_i$ in each direction. This ensures that the mode is centered within the prior. The case in which parts of the mode are cut off by the prior is discussed in section 6.2. We then conducted tests in $\{2, 4, 8, 12, 16\}$ dimensions recording the Gaussian KL divergence of equation (B.3), the number of posterior evaluations, and the overall overhead. The final results were already shown in figures 2, 4 and 5.

6.2 Non-Gaussian distributions

One of the main goals of our algorithm is to be robust with regards to the functional shape of the posterior distribution. We therefore tested the code also on non-Gaussian distributions with varying degrees of pathological features. All adopted priors are flat in the respective parameters.

6.2.1 Log-transformations

Our first example of a non-Gaussian feature is motivated by a common occurrence in Physics. In many applications, there are free scales in the problem which are not known across one or more dimensions in the parameter space. For these parameter one usually samples their logarithm with a flat prior (which is equivalent to imposing a logarithmic prior), distributing the prior probability density evenly across multiple orders of magnitude. If the likelihood is Gaussian in the (linear) parameter, this typically leads to a log-Gaussian distribution of the form

$$10^x \sim \mathcal{N}(\mu, \sigma) \tag{6.3}$$

across some dimensions.

To test whether our algorithm is robust with respect to these kind of likelihoods we drew randomly correlated 4-dimensional Gaussians according to equation (6.1) where the first two dimensions $\{x_0, x_1\}$ are sampled in log-space. The performance of the algorithm in this case is shown in figure 6. We recover the correct posterior shape and manage to sample the posterior accurately with only around 200 samples. An additional benefit of this test is that it shows that our algorithm is robust with respect to cases where the mode has a hard prior cutoff ($|x_i| < 2$ in this example).

In order to explore the limits of our algorithm, we perform the same test in 8 dimensions, with 4 of them being sampled in log-space. We set a budget of at most 2000 posterior evaluations. Figures 7 and 8 show the distribution of $D_{\text{KL}}^{\text{sym}}$ as a function of the number of posterior samples and at convergence for the default settings of `CorrectCounter` proposed in appendix A ($\epsilon_{\text{abs}} = 0.01[\Delta\chi^2](1)$, $\epsilon_{\text{rel}} = 0.01$) and for five times more accurate settings ($\epsilon_{\text{abs}} = 0.002[\Delta\chi^2](1)$, $\epsilon_{\text{rel}} = 0.002$). With default settings convergence tends to be declared prematurely while the more accurate settings mitigate this problem. Figure 9 shows corner plots of two example runs at declared convergence by `CorrectCounter` with default settings, one where convergence is declared prematurely while the mode is still being explored, and one where the mode has been characterized correctly (our target $D_{\text{KL}}^{\text{sym}}$ of 0.05 has not been reached in either case). This higher-dimensional example highlights two limitations of our algorithm. On one hand, the overhead of the GP regressor after such a large number of samples

[13] They are uniformly drawn between 0 and 1, then multiplied by a normalization constant such that their sum equals the number of dimensions, in order to avoid cases where many of the eigenvalues are close to zero simultaneously.

JCAP10(2023)021

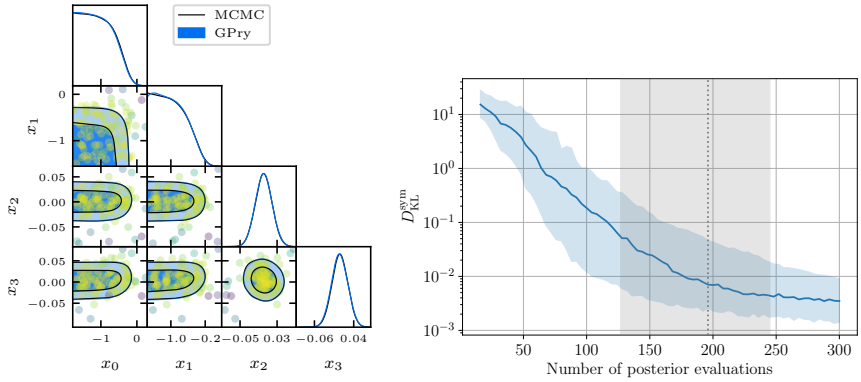


Figure 6. 2d and 1d posterior distributions of a typical four-dimensional log-gaussian distribution (left) at convergence (180 posterior evaluations), and convergence with respect to the true model against number of accepted steps for 200 realizations, where the blue band shows the $\{25, 50, 75\}$%-quantiles for the KL-divergence, and the grey band does the same for convergence as defined by the `CorrectCounter` criterion. (Right) The posterior distribution is cut off in x_0 and x_1, which is correctly captured by `GPry`.

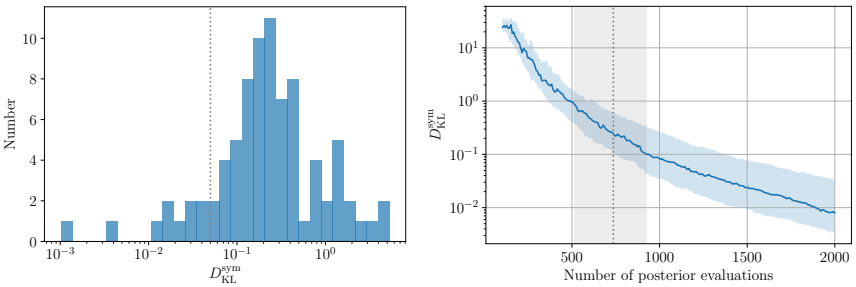


Figure 7. *Left*: distribution of KL divergences at convergence, according to `CorrectCounter` with default settings, of the 8-dimensional log-gaussian draws. *Right*: convergence against number of accepted steps, where the blue and grey bands are defined as in figure 6. Even though most of the runs converge within an acceptable accuracy, our convergence criterion declares convergence prematurely. 83 of the 84 runs we performed were declared as converged. Figure 9 shows the contour plots for a two examples of a prematurely and a correctly reported convergence with these default settings of `CorrectCounter`.

(~ 2000) reduces the advantage of our algorithm with respect to traditional MC samplers. On the other hand, the prematurely reported convergence seems to be a consequence of the combination of long tails and higher-dimensionality: the sequential optimization algorithm fails to propose points at the tails, which occupy a small fraction of the hypervolume in higher dimensionality. An active sampling scheme that explores the parameter space more thoroughly may mitigate this problem [87].

JCAP10(2023)021

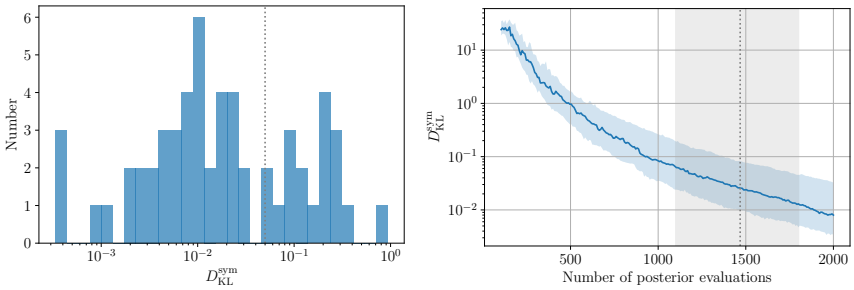


Figure 8. Same as figure 7, with the required accuracy of the `CorrectCounter` convergence criterion increased by a factor of 5 ($\epsilon_{\mathrm{abs}} = 0.002[\Delta\chi^2](1)$, $\epsilon_{\mathrm{rel}} = 0.002$). The KL-divergence of the converged runs is much better than in figure 7. However, only 58 of 84 runs are declared as converged by the convergence criterion when the evaluation budget has been exhausted.

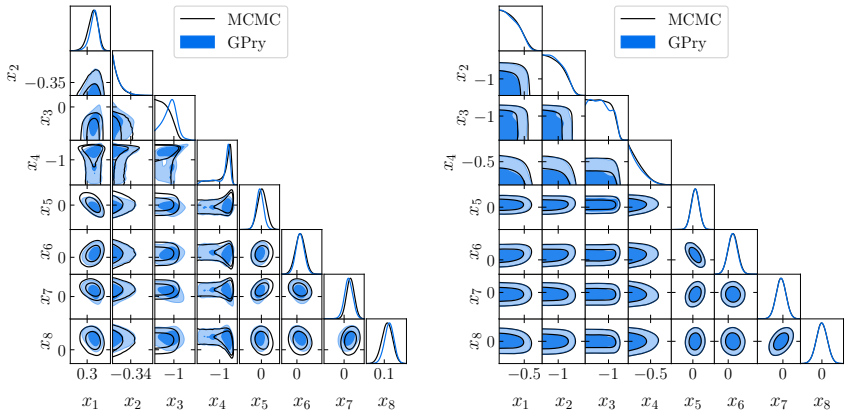


Figure 9. Triangle plots of the 8-dimensional log-Gaussian example at convergence according to `CorrectCounter` with default settings. *Left*: premature convergence (928 posterior evaluations, $D_{\mathrm{KL}}^{\mathrm{sym}} = 1.49$). The mode has been found but not been explored completely. *Right*: correct classification as converged (736 posterior evaluations, $D_{\mathrm{KL}}^{\mathrm{sym}} = 0.08$). Even though our target $D_{\mathrm{KL}}^{\mathrm{sym}}$ of 0.05 has not been reached the contours are still recovered correctly.

6.2.2 Curved degeneracies

We also investigated whether more general curved degeneracies with different length-scales in the different parameter dimensions could be modeled correctly. We use three examples.

1. Example one is a "banana"-shaped curved degeneracy, a slightly modified version of a benchmark found in [88], which is based upon an eight-order polynomial in the exponent and exhibits a long tail in the $x_1 \approx 4x_0^4$ direction. The log-likelihood of this

JCAP10(2023)021

distribution is

$$\log \mathcal{L}(x_0, x_1) = -(10 \cdot (0.45 - x_0))^2/4 - (20 \cdot (x_1/4 - x_0^4))^2 \,. \tag{6.4}$$

Figure 10(a) shows how `GPry` performs at sampling this distribution. The posterior shape is correctly recovered (at around $\sim$ 40 posterior evaluations) and shows good match with MCMC.

2. Example two has a fourth-order polynomial in the exponent, but in this case the parameters are tuned in order to exhibit an extremely sharp cutoff away from the degeneracy direction and an extremely long tail along the degeneracy. This particularly pathological case is the Rosenbrock function, commonly used to test minimization algorithms. It is described by

$$\log \mathcal{L}(x_0, x_1) = -\frac{1}{2}\left[(a - x_0)^2 + b(x_1 - x_0^2)^2\right] , \tag{6.5}$$

where we set the parameters to their typical values of $a = 1$ and $b = 100$. It has a long, narrow, parabolic "ridge" along which the maximum lies. Since the parabolic degeneracy direction changes throughout, this is a good test for the robustness of `GPry` for distributions which do not show a clear axis of correlation or symmetry. We impose a uniform prior between $[-4, 4]$ for both x_0 and x_1. The results for this posterior are displayed in figure 10(b), which shows that even such a pathological posterior function can be accurately described by the `GPry` code, while still requiring a reasonably small number of posterior evaluations (~ 60).

3. The third example is a sharp ring-like posterior. The log-likelihood of this distribution is given by

$$\log \mathcal{L}(x_0, x_1) = -\frac{1}{2}\left[\frac{\left(\sqrt{x_0^2 + x_1^2} - \mu\right)^2}{\sigma} + \log(2\pi\sigma^2)\right] , \tag{6.6}$$

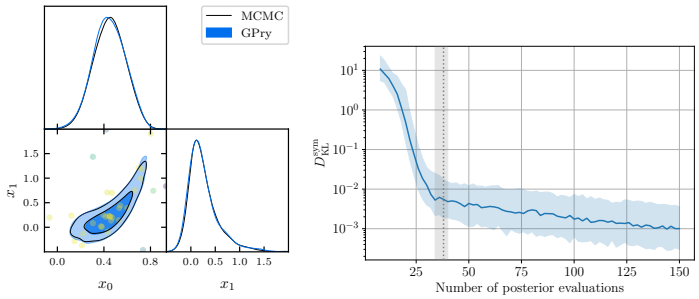
with $\mu = 1$ and $\sigma = 0.05$. This produces a ring-shaped posterior distribution with the two very different scales μ (the location of the ring) and σ (the width of the ring). Furthermore the maximum of this function is the ridge of the ring, making it especially hard to capture the full mode and sample the distribution correctly. Nevertheless our algorithm efficiently captures this mode within ~ 75 posterior evaluations and agrees well with MCMC.

We note that for all of these non-Gaussian examples more posterior evaluations are required for convergence compared to the multivariate Gaussian examples with the same dimensionalities. This is because the surrogate model requires more training samples to correctly capture the non-trivial shape and the extended tails.

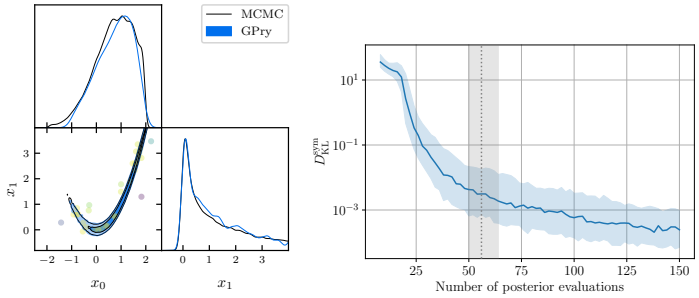
6.2.3 Multi-modal posteriors

We also want to check the robustness of the `GPry` tool against mild multi-modality. For this, we make use of a modified Himmelblau function (which is commonly used in minimization studies). The log-posterior is defined as

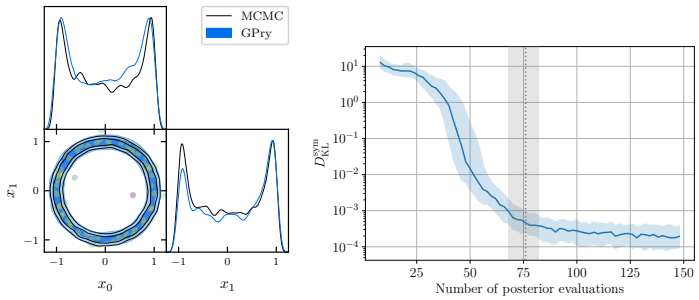
$$\log \mathcal{L}(x_0, x_1) = -\frac{1}{2}\left[a \cdot (x_0^2 - x_1 - 11)^2 + (x_0 + x_1^2 - 7)^2\right] \tag{6.7}$$



(a) "Banana"-shaped degeneracy.



(b) Rosenbrock likelihood.



(c) Gaussian ring.

Figure 10. Performance tests for the non-Gaussian likelihoods with curved degeneracies presented in section 6.2.2. For each case *left*: 2d and 1d posterior distributions for typical converged runs (40, 62 and 68 posterior evaluations, respectively); *right*: convergence against number of accepted steps, where the blue and grey bands are defined as in figure 6. Even though these distributions display very non-Gaussian behaviours, their shape is correctly recovered without needing a large number of samples.

JCAP10(2023)021

where the term in the brackets corresponds to the Himmelblau function for $a = 1$. We include this scaling factor a in the first term in order to create a "mild" multi-modal posterior ($a = 0.1$) with relatively connected modes which we compare to the full Himmelblau function ($a = 1$).

We show the results of sampling this distribution in figures 11 to 13. We observe that many runs do not correctly capture the modes. In general, we can distinguish three modes of failure of the `GPry` algorithm, nicely demonstrated in these examples.

1. The algorithm can find and sample all modes, but not weigh them correctly in relation to each other. This is clearly visible in figure 11, where all modes are reliably sampled, but the 1D posterior reveals the incorrect weighting.

2. The `CorrectCounter` criterion may falsely claim convergence and stop the sampling when some of the modes have been well explored, while further sampling might have revealed modes that have not been mapped. This is shown in figure 12, where we compare the convergence to the true distribution (through the $D_{\mathrm{KL}}^{\mathrm{sym}}$) when the `CorrectCounter` criterion has claimed convergence, with that of the runs at a larger number of samples (150 in this case). We observe that if the sampling had continued further, they would have been able to better map the underlying modes. See also figure 13 for two examples of these first two failure modes for the $a = 1$ case.

3. The SVM could characterize a whole region as irrelevant due to a very deep intermediate valley even though a mode is present there. In that case, no amount of additional sampling would reveal the hidden mode. This failure mode does not occur for the $a = 0.1$ or $a = 1$ cases as the valleys are not deep enough there to be characterized as irrelevant.

As such, we would like to stress that this package was designed with a focus on uni-modal distributions and that there is no guarantee that *in general* all modes are captured or weighed correctly by the algorithm. Deeper investigations into multi-modal GP algorithms are left for future work. Note that for this distribution we used the PolyChord nested sampler [9, 10] for generating our reference contours and MC samples of the GP surrogate as it — unlike MCMC — reliably finds and explores all modes.

6.2.4 Performance for non-Gaussian and multi-modal distributions

In section 6.2 we have demonstrated that non-Gaussian distributions need a larger number of training samples in order to converge when compared to Gaussian distributions with equal dimensionality (in the particular examples presented in this section, the ratio of required samples seems to be approximately 5). This need for a larger number of posterior evaluations for convergence is also true for traditional algorithms. We could perform a similar analysis to the one presented in figure 5 to check whether the comparison with MCMC and `PolyChord` generalizes to non-Gaussian cases. Due to the wide variety of possible non-Gaussian shapes, the required number of samples and the corresponding overhead will depend dramatically on the distribution at hand. We therefore refrain from performing such an analysis at this point, and leave it for future work, where a range of more realistic non-Gaussian distributions would be tested, instead of the particularly pathological cases discussed here.

JCAP10(2023)021

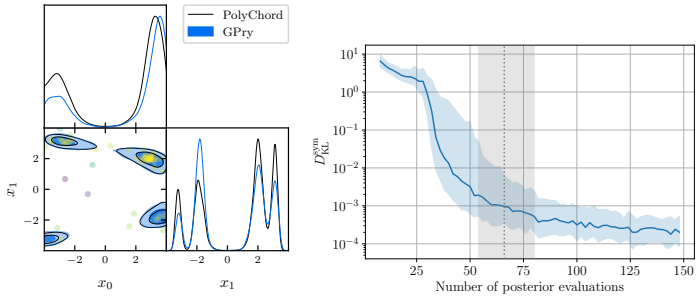


Figure 11. 2d and 1d posterior distributions of a typical, converged runs of the "mild" Himmelblau function (left) at convergence (58 posterior evaluations) and convergence against number of accepted steps (right), where the blue and grey bands are defined as in figure 6. The function has four modes which are all sampled but not weighed correctly by `GPry`. `GPry` on average needs few ($\lesssim 75$) samples to claim convergence.

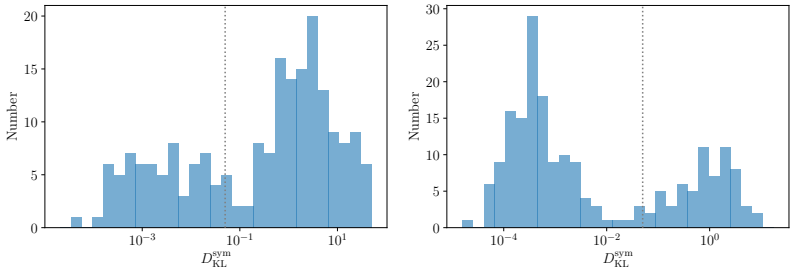


Figure 12. *Left*: distribution of KL divergences at convergence according to `CorrectCounter` of the standard Himmelblau function. In many cases convergence is declared while not all of the four modes of the function have been explored, leading to large values of $D_{\mathrm{KL}}^{\mathrm{sym}}$. *Right*: distribution of KL divergences for the same Himmelblau function at a budgeted, large number of samples (in this case 150). The distribution shows that sampling beyond reported convergence of the `CorrectCounter` criterion would aid in improving the interpolation. Nonetheless, there still remain two modes: one at low values of $D_{\mathrm{KL}}^{\mathrm{sym}}$ (around $D_{\mathrm{KL}}^{\mathrm{sym}} = 10^{-3}$) where all modes have been found and one at high values (around $D_{\mathrm{KL}}^{\mathrm{sym}} = 1$) where some of the modes were not explored. Examples of this behaviour are shown in figure 13.

6.3 Cosmology

We also test the `GPry` tool in the context of cosmological applications, such as the inference of the posterior for Planck CMB anisotropy measurements (using the nuisance-marginalised Planck Lite likelihood of [89, 90] in the context of the 6-dimensional ΛCDM model). We performed 75 separate runs of the `GPry` tool, converging on average within only around 500 evaluations of the underlying theory code.[14] The convergence history as well as the final

[14]Here we happen to be using `CLASS` [91], but since the `GPry` tool is fully interfaced with `Cobaya` [86], other theory codes can be used as well.

JCAP10(2023)021

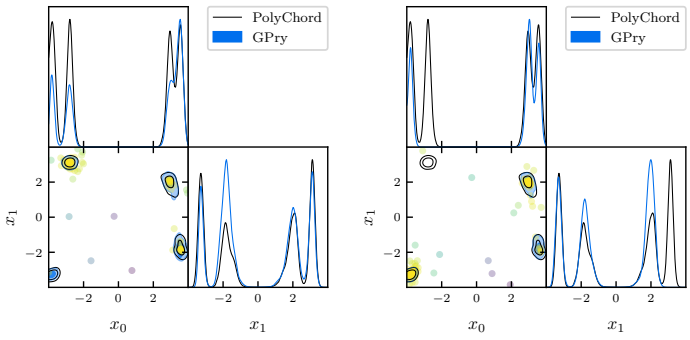


Figure 13. Exemplary 2d and 1d posterior distributions of the full Himmelblau function ($a = 1$). *Left*: contours of the algorithm finding all modes and converging at 102 posterior evaluations (although the 1D posteriors are not weighed correctly). *Right*: example of the algorithm missing a mode completely and falsely claiming convergence. This problem arises when the posterior distribution to map has several disconnected modes. If one of the modes is missed completely early in the sampling procedure the GP surrogate and hence the acquisition procedure may deem this region irrelevant and not sample there. This behaviour is especially severe when the SVM classifies the region which contains the additional mode(s) as infinite.

KL-divergence upon termination through the convergence criterion are shown in figure 14. An exemplary case (close to the median in terms of required number of samples) is also shown in figure 14, where we can see that the constraints are very well aligned with those of the true posterior.

We note that the full Planck likelihood (including nuisance parameters) can also be modeled with `GPry`, but in this case the high dimensionality of the parameter space (27 dimensions in our case) makes the proposal of new points to start the acquisition optimization from (see line 12 of algorithm 1) rather difficult. If one uses the bestfit and covariance matrix of the Planck chains to propose these points instead, the acquisition function can be well optimized and the run does correctly map the posterior. We leave investigations of reaching convergence for the full Planck likelihood without any kind of a priori information (such as covariance matrix or bestfit) for future work.

Another illustrative example is that of the combined Big Bang Nucleosynthesis (BBN) and Baryon Acoustic Oscillatons (BAO) measurements. We combine low redshift BAO from 6dFGS galaxies, the DR7 main galaxy sample, DR12 luminous red galaxies (together low-z), as well as high redshift BAO from DR16 quasars, DR16 Lyman-α based BAO, and their cross-correlations (together high-z), in order to constrain the Hubble constant and the matter composition of the Universe. The BBN likelihood we adopt is the same as in [92]. For this case we vary the number of effective neutrinos $N_{\rm eff}$, corresponding to the addition of dark radiation to the ΛCDM model. This results in three possible data likelihood combinations, depending on whether we combine with the BBN data the low redshift galaxy based BAO likelihoods ("low-z"), with the higher redshift Lyman-α and quasar BAO likelihoods ("high-z"), or we use both redshift samples ("combined"). For every combination, we sample the four-dimensional posterior using both MCMC and `GPry`. In figure 15 we show the resulting triangle plot which is in excellent agreement, demonstrating again the flexibility of GPry even

JCAP10(2023)021

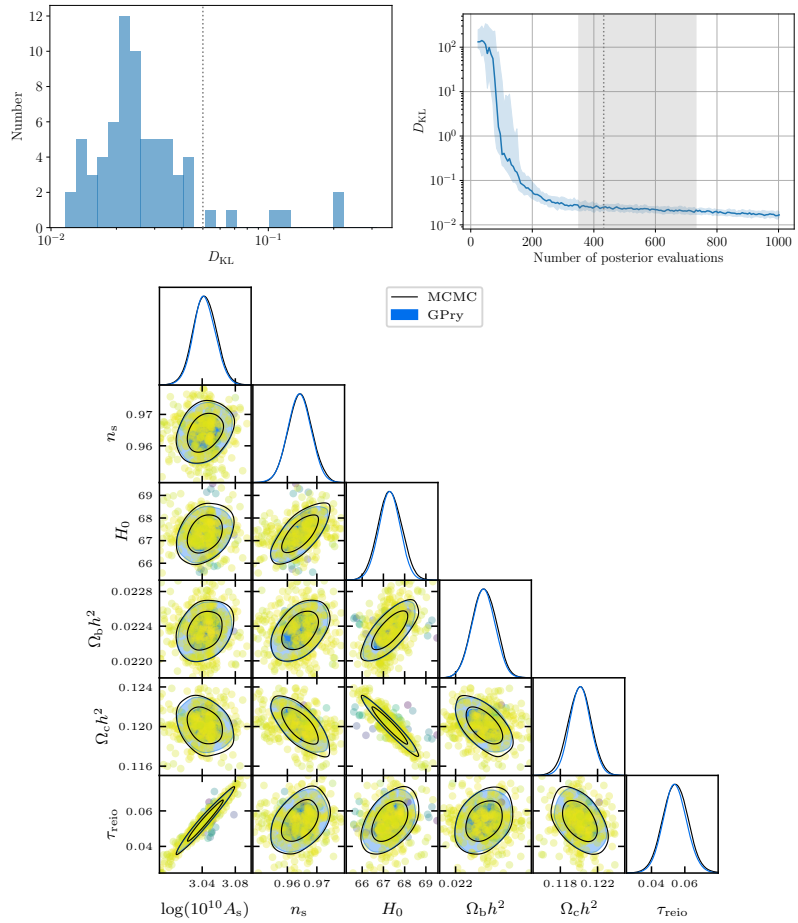


Figure 14. Constraints and convergence statistics in a ΛCDM model from Planck 2018 (TT, TE, EE, lensing) using the nuisance-marginalised Planck Lite likelihood. The given constraints could be obtained sampling only around ~ 500 (in this case 420) evaluations of the underlying theory code and likelihood. *Top Left*: distribution of KL divergences at convergence according to `CorrectCounter` with default settings. *Top Right*: KL divergences against number of accepted steps, as defined in figure 6. For both top plots, these distributions refer to 75 independent runs with identical training settings. *Bottom*: 1D posteriors and (68.3%, 95.4%) contours of the 2D posteriors for one representative run at convergence. The dots show the training samples in the order in which they were acquired, where darker samples were added early and yellow ones late.

when the underlying model or the used data sets are varied. The contours can be recovered with only around ~ 100 posterior evaluations (as opposed to the $\sim 10^4$ points used for the MCMC chains).

JCAP10(2023)021

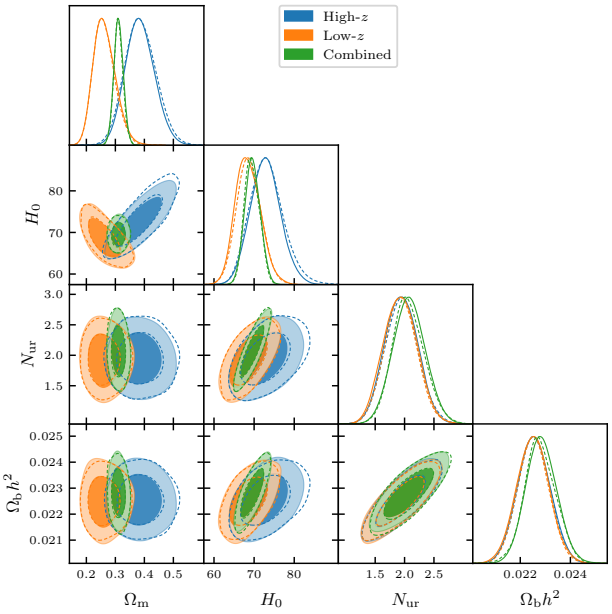


Figure 15. Triangle plot showing the marginalised constraints of the four-dimensional likelihood of BBN+BAO measurements for high-z, low-z and combined likelihoods. `GPry` is able to recover all contours correctly with only very few (124, 108, 80) posterior evaluations. The contours that we recover are in excellent agreement with the constraints from MCMC.

7 Conclusions

In this paper we presented the `GPry` algorithm and Python package implementation. As shown with both synthetic and cosmological likelihoods our algorithm requires vastly less posterior evaluations for generating a fair Monte Carlo sample for Bayesian Inference than current state-of-the-art MCMC and nested samplers. We report up to multiple orders of magnitude improvements in the number of posterior evaluations required, as well as in wall-clock computation time savings, making this algorithm very promising for slow likelihood codes. This not only speeds up inference significantly but also reduces its carbon footprint. Furthermore, we open a new window of possibilities by enabling inference from extremely slow likelihoods ($\gtrsim$ minutes per evaluation), which otherwise would be impossible to sample, since traditional samplers might take months to converge. In addition, since our algorithm does not rely on specialized hardware (such as GPUs) or any kind of pre-training, it can be used as a drop-in replacement for traditional Monte Carlo samplers. Particularly in the case of cosmological applications it benefits from an interface with `Cobaya`.

Despite the algorithm's impressive performance, there is still ample room for improvement both in terms of speed and robustness. In a future series of papers we plan to explore

JCAP10(2023)021

four main avenue: (i) the overhead of constructing the GP surrogate model and the acquisition procedure could be further minimized by using clever numerical techniques, allowing `GPry` to outcompete traditional MC samplers even for fast likelihoods. (ii) As discussed in section 6.2.3 `GPry` currently is optimized for unimodal posterior distributions; it would be desirable to increase the robustness towards strongly multi-modal posteriors by generating the starting points for the acquisition optimization in a special way, and using clustering algorithms to track different modes separately. (iii) For likelihood distributions with significant stochastic or numerical noise, it would be beneficial to automatically adapt the noise term in equation (2.3) without requiring prior knowledge. (iv) For high dimensionalities the current methods of proposing additional points for restarting hyperparameter optimization and sample acquisition are still relatively naive. Similarly, the overhead of the underlying operations performed on the GP increases strongly with the number of acquired samples. Both of these hurdles can be overcome with novel approaches, potentially unlocking even the regime of high-dimensional likelihoods for further optimization with `GPry`.

The `GPry` algorithm and python package presented in this work enables parameter inference in cosmology without high computational and environmental costs. This opens up new possibilities for Bayesian inference on costly likelihood functions which have been computationally unfeasible before. With many avenues of optimization of the code-base and algorithm still left to explore, `GPry` will only continue to improve in efficiency and accuracy.

Acknowledgments

We thank Julien Lesgourgues, Antony Lewis, Andrew Liddle and Marcos Pellejero Ibáñez for useful discussions. This project was initiated when all authors were working at the TTK institute of RWTH Aachen University. We also acknowledge the use of the JARA computing cluster of the RWTH Aachen University under project `jara0184`. N.S. acknowledges support from the Maria de Maetzu fellowship grant: CEX2019-000918-M, financiado por MCIN/AEI/10.13039/501100011033. J.E. acknowledges support by the ROMFORSK grant project no. 302640. The authors thank the referee for the many helpful comments and suggestions that helped improve the quality of this manuscript. J.T. acknowledges support from the STARS@UNIPD2021 project GWCross.

A Posterior scale in higher dimensions

When considering a problem with a larger number of dimensions, there are a few aspects of the problem that require special care. It is a well-known fact that for a 1-dimensional Gaussian the region defined by one standard deviation around the mean contains $\approx 68\%$ of the total probability mass. The generalisation to higher dimensionality is non-trivial: for multivariate Gaussians, considering distances defined in units of the covariance matrix (*Mahalanobis distance*), the region defined by a unit away from the mean contains a smaller and smaller fraction of the total probability mass as dimensionality goes up. This is, of course, nothing more than the *curse of dimensionality*, and it will be present in most of the inference problems that we target in this study.[15]

In our context of modelling a probability density function, this is reflected in the dynamic range of log-probability that needs to be carefully modelled, meaning that given some

[15] It affects distributions with significant tails, and most well-behaved distributions show tails, including those in the exponential family (which includes multivariate Gaussians).

JCAP10(2023)021

confidence limit (CL) up to which we want our model to be especially precise, the difference between the log-posterior corresponding to that CL and the maximum log-posterior will depend on dimensionality. This dynamic range will show up at three different steps of the algorithm, explicitly in the treatment of infinities and extreme values in section 3.3 and the convergence criterion in section 4.3, and implicitly in the choice of the acquisition hyperparameter in section 4.1.2. Taking into account this dimensionality scaling in the ways explained below has proven to dramatically improve the performance of our algorithm.

In order to give a rough order-of-magnitude estimate for this log-posterior scaling, we can turn towards a multivariate Gaussian distribution of the same dimensionality. Treated as a random variable itself, a multivariate Gaussian log-probability is proportional to the sum of N_d independent standard 1-dimensional Gaussian random variables (up to a linear covariance-diagonalizing transformation). Thus the value of the Gaussian log-posterior when multiplied by -2 follows a χ^2 distribution with N_d degrees of freedom. Defining $\Delta\chi^2 = 2[\max(\log p) - \log p]$ we find $\Delta\chi^2 \sim {\chi^2}_{N_d}$.

We can use this to compute the posterior range corresponding to different CLs defined by the posterior mass ϵ that they leave out, using the χ^2 cumulative distribution function F_{N_d} (where N_d is the number of degrees of freedom):

$$1 - \varepsilon = F_{N_d}(\Delta\chi^2)\,. \tag{A.1}$$

When referring to CLs in higher dimensions, we can alternatively name them as their 1D equivalent normal Gaussian extent ($1-\epsilon = 0.683$ for 1-σ, $1-\epsilon = 0.954$ for 2-σ, etc.). As such, in the following when we refer to a n–σ contour in an arbitrary dimensionality within this paper, we explicitly refer to the CL corresponding to that number of standard deviations in a $1D$ Gaussian. Explicitly, since $F_1(x) = \mathrm{erf}(\sqrt{x/2})$, and given that the value of a χ^2_1 random variable represents the squared number of standard deviations away from the mean in the corresponding Gaussian, we can simply write $1 - \varepsilon = \mathrm{erf}(n/\sqrt{2})$ for a given n-σ CL.

With this, we can get the expected scaling in N_d dimensions corresponding to a n-σ probability mass as

$$[\Delta\chi^2](n) = F^{-1}_{N_d}\left[\mathrm{erf}(n/\sqrt{2})\right]\,. \tag{A.2}$$

As an example, the 2-σ ($1 - \epsilon = 0.954$) contour corresponds to a range $[\Delta\chi^2](2) = 9.72$ in 4 dimensions and $[\Delta\chi^2](2) = 15.79$ in 8 dimensions.

In section 3.3 we have used this result to derive the threshold value T for the SVM (the criterion for deciding if a sample has a sufficient log-posterior to be added to the GP) by imposing $T = [\Delta\chi^2](n_T)/2$ which is the scaling of the log-posterior for $n_T = 20$ ($\epsilon \approx 5.5 \cdot 10^{-89}$), and has been found to work well for most practical applications (including all examples in this work). This prescription ensures dimensional consistency: choices of T as absolute values do not work well across different dimensions, causing the SVM to be too permissive in low dimensions (does not capture extreme values efficiently) and too stringent in high dimensions (points with significant log-posterior are excluded).

We have also used this result to scale the tolerance of the convergence criterion in section 4.3. In particular, since the absolute threshold ϵ_{abs} is compared against differences in absolute values of $\log p$, we are scaling it as these differences do for a fixed difference in credibility, in particular that of the first σ credible (hyper)volume: $\epsilon_{\mathrm{abs}} = 0.01[\Delta\chi^2](1)$.

Regarding the dimensionality scaling of the learning hyperparameter, it is not trivial to find an analytic prescription to write ζ as a function of $[\Delta\chi^2](n)$. As discussed in section 4.1.2 we have derived an experimental scaling $\zeta = N_d^{-0.85}$, which corresponds to the value of ζ

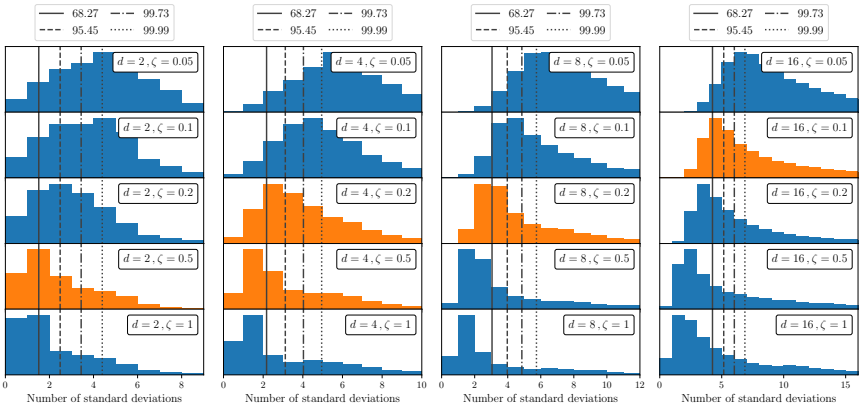


Figure 16. Histograms of aggregated Mahalanobis distances of the points in the training sets of the realizations used in figure 2, for different dimensionalities (columns) and values of ζ (rows). The optimal ζ's from the experimental relation $\zeta = N_d^{-0.85}$ are highlighted in orange/clear (for $d = 4$, the two closest values are both highlighted). A remarkable result is that efficiency at converging (the criterion imposed to get the optimal ζ's) is maximised when the distribution of training points are centered around the same CL (68%) in all dimensionalities, likely imposed indirectly by using the dimensionally-consistent KL divergence to assess convergence when selecting optimal ζ's.

that leads to convergence in the smallest number of posterior evaluations, as demonstrated in figure 2. We can check a posteriori how these optimal dimensional-dependent ζ's relate to the CL's at these dimensionalities. To do that, we compute the Mahalanobis distances of all points in the training sets of all the realizations used for figure 2, and create histograms of these distances for each dimensionality and ζ in figure 16. In this figure, we highlight the cases that converged most efficiently in orange/clear, as assessed by the dimensionally-consistent Kullback-Leibler divergence. We observe that convergence is achieved more efficiently when ζ is such that the distribution of training points is centered around the same CL (68%) in all dimensionalities. This underlines the idea that the dimensionally-dependent CL's should set the relevant scales in the surrogate model for optimal efficiency, and is possibly in fact a consequence of having used a dimensionally-consistent method (the Kullback-Leibler divergence) to assess convergence.

B KL divergence

A natural way of assessing how well a given distribution can approximate another reference distribution is the Kullback-Leibler (KL) divergence. The KL divergence of the continuous probability distribution P from the distribution Q, with respective probability density functions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, is defined as [93]

$$D_{\mathrm{KL}}(P||Q) = \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x}\,. \tag{B.1}$$

The KL divergence as defined above more strongly weighs disagreements between the two probability distributions where $p(\boldsymbol{x})$ is large. Since we want the approximation to be equally

JCAP10(2023)021

accurate in all regions where either distribution is large, we use a symmetrized version of the divergence (often called Jeffreys divergence). It is defined as

$$D_{\mathrm{KL}}^{\mathrm{sym}}(P,Q) = \frac{1}{2}\left(D_{\mathrm{KL}}(P||Q) + D_{\mathrm{KL}}(Q||P)\right) . \tag{B.2}$$

A smaller value means that the two posteriors are in better agreement, and one typically wants $D_{\mathrm{KL}}^{\mathrm{sym}}(P||Q) \ll 1$ for good agreement. The dimensionality consistency of the KL divergence guarantees that a given value for the divergence characterizes similar differences across dimensionalities.

To compute the KL divergence explicitly, one can use the fact that the points in a Monte Carlo sample of P are distributed as $p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$. One can thus approximate the integral as a sum of the quantity $\log p(\boldsymbol{x}_i) - \log q(\boldsymbol{x}_i)$ over all points in the MC sample (multiplied by their respective weights/multiplicities).

We can use the KL divergence to assess the convergence towards the true distribution of a GP surrogate model, if a sample from the true distribution can be obtained with the usual MC methods (e.g. in the test cases presented in section 6). In that case, $\log p(\boldsymbol{x}_i)$ would be the true log-posterior at point i in the MC sample, and $\log q(\boldsymbol{x}_i)$ would be the emulated log-posterior from `GPry` at that same point. In practical applications where an MC sample of the true posterior is not possible to obtain, the KL divergence can be used in a similar fashion to define a convergence criterion by comparing GP surrogate models at consecutive iterations of the `GPry` algorithm, summing over an MC sample of the GP surrogate model at a particular step (see section 4.3).

In order to save a significant amount of memory, when using the KL divergence for the purpose of a convergence criterion, instead of integrating a full MCMC, we only store the information from the mean and the covariance matrix. This is equivalent to approximating the underlying distributions as multivariate Gaussian distributions (with mean $\mathbb{m}$ and covariance $\boldsymbol{C}$). While this is a bad description for the distribution itself, it is often the case that when the multivariate Gaussian approximation of a distribution agrees to a high level of precision with that of another distribution, so do the underlying distributions. Under this approximation the KL divergence is simply given by

$$D_{\mathrm{KL}}(P||Q) \approx \frac{1}{2}\left(\mathrm{tr}\left(\boldsymbol{C}_Q^{-1}\boldsymbol{C}_P\right) - d + (\mathbb{m}_Q - \mathbb{m}_P)^T\boldsymbol{C}_Q^{-1}(\mathbb{m}_Q - \mathbb{m}_P) + \log\left(\frac{\det\boldsymbol{C}_Q}{\det\boldsymbol{C}_P}\right)\right) . \tag{B.3}$$

Whether using the MC-summed or the Gaussian approximation for the KL divergence, using it to naively define a convergence criterion, can be problematic, since running a full Monte Carlo sample at every acquired point, or at every iteration, would dominate the overhead of the algorithm. To reduce this computational cost, we take a number of decisions: before deciding whether to re-run the Monte Carlo sample, we reweigh the previous one and compute the KL divergence between it and the previous estimate. We then re-use the reweighed one if the KL divergence between original and reweighed is small enough. We also relax the convergence criterion of the Monte Carlo algorithm early in the sampling procedure as convergence there is rather unlikely, so we do not need a high-quality estimation of the mean and covariance at that point.

Convergence is then determined by defining a threshold c value such that the algorithm stops when $D_{\mathrm{KL}}^{\mathrm{sym}} < c$ during n iterations, suggesting that the interpolation of the posterior distribution has stabilised. We set $n = 2$ as the default.

JCAP10(2023)021

Even with these improvements this method still produces considerable computational overhead, mainly due to the fact that running a Monte Carlo chain needs a large number of samples from the GP, especially as the number of dimensions increases.

References

[1] A. Lewis and S. Bridle, *Cosmological parameters from CMB and other data: A Monte Carlo approach*, *Phys. Rev. D* **66** (2002) 103511 [astro-ph/0205436] [INSPIRE].

[2] A. Lewis, *Efficient sampling of fast and slow cosmological parameters*, *Phys. Rev. D* **87** (2013) 103529 [arXiv:1304.4473] [INSPIRE].

[3] D. Foreman-Mackey, D.W. Hogg, D. Lang and J. Goodman, *emcee: The MCMC Hammer*, *Publ. Astron. Soc. Pac.* **125** (2013) 306 [arXiv:1202.3665] [INSPIRE].

[4] J. Akeret, S. Seehars, A. Amara, A. Refregier and A. Csillaghy, *CosmoHammer: Cosmological parameter estimation with the MCMC Hammer*, *Astron. Comput.* **2** (2013) 27.

[5] J. Skilling, *Nested Sampling*, *AIP Conf. Proc.* **735** (2004) 395.

[6] F. Feroz and M.P. Hobson, *Multimodal nested sampling: an efficient and robust alternative to MCMC methods for astronomical data analysis*, *Mon. Not. Roy. Astron. Soc.* **384** (2008) 449 [arXiv:0704.3704] [INSPIRE].

[7] F. Feroz, M.P. Hobson and M. Bridges, *MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics*, *Mon. Not. Roy. Astron. Soc.* **398** (2009) 1601 [arXiv:0809.3437] [INSPIRE].

[8] F. Feroz, M.P. Hobson, E. Cameron and A.N. Pettitt, *Importance Nested Sampling and the MultiNest Algorithm*, *Open J. Astrophys.* **2** (2019) 10 [arXiv:1306.2144] [INSPIRE].

[9] W.J. Handley, M.P. Hobson and A.N. Lasenby, *PolyChord: nested sampling for cosmology*, *Mon. Not. Roy. Astron. Soc.* **450** (2015) L61 [arXiv:1502.01856] [INSPIRE].

[10] W.J. Handley, M.P. Hobson and A.N. Lasenby, *polychord: next-generation nested sampling*, *Mon. Not. Roy. Astron. Soc.* **453** (2015) 4385 [arXiv:1506.00171] [INSPIRE].

[11] E. Higson, W. Handley, M. Hobson and A. Lasenby, *Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation*, *Stat. Comput.* **29** (2018) 891.

[12] J.S. Speagle, *dynesty: a dynamic nested sampling package for estimating Bayesian posteriors and evidences*, *Mon. Not. Roy. Astron. Soc.* **493** (2020) 3132 [arXiv:1904.02180] [INSPIRE].

[13] E.D. Feigelson, R.S. de Souza, E.E.O. Ishida and G.J. Babu, *21st Century Statistical and Computational Challenges in Astrophysics*, *Ann. Rev. Stat. App.* **8** (2021) 493 [arXiv:2005.13025] [INSPIRE].

[14] R. Alves Batista et al., *EuCAPT White Paper: Opportunities and Challenges for Theoretical Astroparticle Physics in the Next Decade*, arXiv:2110.10074 [INSPIRE].

[15] A.R.H. Stevens, S. Bellstedt, P.J. Elahi and M.T. Murphy, *The imperative to reduce carbon emissions in astronomy*, *Nature Astron.* **4** (2020) 843 [arXiv:1912.05834] [INSPIRE].

[16] S. Portegies Zwart, *The Ecological Impact of High-performance Computing in Astrophysics*, *Nature Astron.* **4** (2020) 819 [arXiv:2009.11295] [INSPIRE].

[17] M. Kaplinghat, L. Knox and C. Skordis, *Rapid calculation of theoretical CMB angular power spectra*, *Astrophys. J.* **578** (2002) 665 [astro-ph/0203413] [INSPIRE].

[18] R. Jimenez, L. Verde, H. Peiris and A. Kosowsky, *Fast cosmological parameter estimation from microwave background temperature and polarization power spectra*, *Phys. Rev. D* **70** (2004) 023005 [astro-ph/0404237] [INSPIRE].

[19] T. Auld, M. Bridges, M.P. Hobson and S.F. Gull, *Fast cosmological parameter estimation using neural networks*, *Mon. Not. Roy. Astron. Soc.* **376** (2007) L11 [astro-ph/0608174] [INSPIRE].

JCAP10(2023)021

[20] T. Auld, M. Bridges and M.P. Hobson, *CosmoNet: Fast cosmological parameter estimation in non-flat models using neural networks*, *Mon. Not. Roy. Astron. Soc.* **387** (2008) 1575 [astro-ph/0703445] [INSPIRE].

[21] J. Albers, C. Fidler, J. Lesgourgues, N. Schöneberg and J. Torrado, *CosmicNet. Part I. Physics-driven implementation of neural networks within Einstein-Boltzmann Solvers*, *JCAP* **09** (2019) 028 [arXiv:1907.05764] [INSPIRE].

[22] A. Manrique-Yus and E. Sellentin, *Euclid-era cosmology for everyone: neural net assisted MCMC sampling for the joint* 3×2 *likelihood*, *Mon. Not. Roy. Astron. Soc.* **491** (2020) 2655 [arXiv:1907.05881] [INSPIRE].

[23] A. Mootoovaloo, A.F. Heavens, A.H. Jaffe and F. Leclercq, *Parameter Inference for Weak Lensing using Gaussian Processes and MOPED*, *Mon. Not. Roy. Astron. Soc.* **497** (2020) 2213 [arXiv:2005.06551] [INSPIRE].

[24] A. Nygaard, E.B. Holm, S. Hannestad and T. Tram, *CONNECT: a neural network based framework for emulating cosmological observables and cosmological parameter inference*, *JCAP* **05** (2023) 025 [arXiv:2205.15726] [INSPIRE].

[25] J. Donald-McCann, F. Beutler, K. Koyama and M. Karamanis, *matryoshka: halo model emulator for the galaxy power spectrum*, *Mon. Not. Roy. Astron. Soc.* **511** (2022) 3768 [arXiv:2109.15236] [INSPIRE].

[26] J. Donald-McCann, K. Koyama and F. Beutler, *matryoshka II: accelerating effective field theory analyses of the galaxy power spectrum*, *Mon. Not. Roy. Astron. Soc.* **518** (2022) 3106 [arXiv:2202.07557] [INSPIRE].

[27] M. Bonici, L. Biggio, C. Carbone and L. Guzzo, *Fast emulation of two-point angular statistics for photometric galaxy surveys*, arXiv:2206.14208 [INSPIRE].

[28] A. Mootoovaloo, A.H. Jaffe, A.F. Heavens and F. Leclercq, *Kernel-based emulator for the* 3*D matter power spectrum from CLASS*, *Astron. Comput.* **38** (2022) 100508 [arXiv:2105.02256] [INSPIRE].

[29] S. Günther et al., *CosmicNet II: emulating extended cosmologies with efficient and accurate neural networks*, *JCAP* **11** (2022) 035 [arXiv:2207.05707] [INSPIRE].

[30] A. Spurio-Mancini, D. Piras, J. Alsing, B. Joachimi and M.P. Hobson, *CosmoPower: emulating cosmological power spectra for accelerated Bayesian inference from next-generation surveys*, *Mon. Not. Roy. Astron. Soc.* **511** (2022) 1771 [arXiv:2106.03846] [INSPIRE].

[31] C.-H. To, E. Rozo, E. Krause, H.-Y. Wu, R.H. Wechsler and A.N. Salcedo, *LINNA: Likelihood Inference Neural Network Accelerator*, *JCAP* **01** (2023) 016 [arXiv:2203.05583] [INSPIRE].

[32] S. Khan and R. Green, *Gravitational-wave surrogate models powered by artificial neural networks*, *Phys. Rev. D* **103** (2021) 064015 [arXiv:2008.12932] [INSPIRE].

[33] M. Chianese, A. Coogan, P. Hofma, S. Otten and C. Weniger, *Differentiable Strong Lensing: Uniting Gravity and Neural Nets through Differentiable Probabilistic Programming*, *Mon. Not. Roy. Astron. Soc.* **496** (2020) 381 [arXiv:1910.06157] [INSPIRE].

[34] F. Lanusse, R. Mandelbaum, S. Ravanbakhsh, C.-L. Li, P. Freeman and B. Póczos, *Deep generative models for galaxy image simulations*, *Mon. Not. Roy. Astron. Soc.* **504** (2021) 5543.

[35] K.K. Rogers, H.V. Peiris, A. Pontzen, S. Bird, L. Verde and A. Font-Ribera, *Bayesian emulator optimisation for cosmology: application to the Lyman-α forest*, *JCAP* **02** (2019) 031 [arXiv:1812.04631] [INSPIRE].

[36] T. McClintock et al., *The Aemulus Project. Part II. Emulating the Halo Mass Function*, *Astrophys. J.* **872** (2019) 53 [arXiv:1804.05866] [INSPIRE].

[37] M.-F. Ho, S. Bird and C.R. Shelton, *Multifidelity emulation for the matter power spectrum using Gaussian processes*, *Mon. Not. Roy. Astron. Soc.* **509** (2021) 2551 [arXiv:2105.01081] [INSPIRE].

JCAP10(2023)021

[38] C.J. Moore and J.R. Gair, *Novel Method for Incorporating Model Uncertainties into Gravitational Wave Parameter Estimates*, *Phys. Rev. Lett.* **113** (2014) 251101 [arXiv:1412.3657] [INSPIRE].

[39] C. Chen, Y. Li, F. Villaescusa-Navarro, S. Ho and A. Pullen, *Learning the Evolution of the Universe in N-body Simulations*, in proceedings of the *34th Conference on Neural Information Processing Systems*, online conference, Canada, 6–12 December 2020, arXiv:2012.05472 [INSPIRE].

[40] S. Bird, K.K. Rogers, H.V. Peiris, L. Verde, A. Font-Ribera and A. Pontzen, *An Emulator for the Lyman-α Forest*, *JCAP* **02** (2019) 050 [arXiv:1812.04654] [INSPIRE].

[41] K. Cranmer, J. Brehmer and G. Louppe, *The frontier of simulation-based inference*, *Proc. Nat. Acad. Sci.* **117** (2020) 30055 [arXiv:1911.01429] [INSPIRE].

[42] J.-M. Lueckmann, J. Boelts, D.S. Greenberg, P.J. Gonçalves and J.H. Macke, *Benchmarking Simulation-Based Inference*, arXiv:2101.04653.

[43] A. Delaunoy et al., *Lightning-Fast Gravitational Wave Parameter Inference through Neural Amortization*, arXiv:2010.12931 [INSPIRE].

[44] J. Alsing, T. Charnock, S. Feeney and B. Wandelt, *Fast likelihood-free cosmology with neural density estimators and active learning*, *Mon. Not. Roy. Astron. Soc.* **488** (2019) 4440 [arXiv:1903.00007] [INSPIRE].

[45] B.K. Miller, A. Cole, G. Louppe and C. Weniger, *Simulation-efficient marginal posterior estimation with swyft: stop wasting your precious time*, arXiv:2011.13951 [INSPIRE].

[46] J. Hermans, N. Banik, C. Weniger, G. Bertone and G. Louppe, *Towards constraining warm dark matter with stellar streams through neural simulation-based inference*, *Mon. Not. Roy. Astron. Soc.* **507** (2021) 1999 [arXiv:2011.14923] [INSPIRE].

[47] F. Gerardi, S.M. Feeney and J. Alsing, *Unbiased likelihood-free inference of the Hubble constant from light standard sirens*, *Phys. Rev. D* **104** (2021) 083531 [arXiv:2104.02728] [INSPIRE].

[48] D. Huppenkothen and M. Bachetti, *Accurate X-ray timing in the presence of systematic biases with simulation-based inference*, *Mon. Not. Roy. Astron. Soc.* **511** (2022) 5689 [arXiv:2104.03278] [INSPIRE].

[49] A. Rouhiainen, U. Giri and M. Münchmeyer, *Normalizing flows for random fields in cosmology*, arXiv:2105.12024 [INSPIRE].

[50] K. Zhang, J.S. Bloom, B.S. Gaudi, F. Lanusse, C. Lam and J.R. Lu, *Real-time Likelihood-free Inference of Roman Binary Microlensing Events with Amortized Neural Posterior Estimation*, *Astron. J.* **161** (2021) 262.

[51] C.-H. Hahn et al., *SIMBIG: A Forward Modeling Approach To Analyzing Galaxy Clustering*, arXiv:2211.00723 [INSPIRE].

[52] M. Reza, Y. Zhang, B. Nord, J. Poh, A. Ciprijanovic and L. Strigari, *Estimating Cosmological Constraints from Galaxy Cluster Abundance using Simulation-Based Inference*, in proceedings of the *39th International Conference on Machine Learning Conference*, Baltimore, MD, U.S.A., 17–23 July 2022, arXiv:2208.00134 [INSPIRE].

[53] S.S. Boruah, T. Eifler, V. Miranda and P.M. Sai Krishanth, *Accelerating cosmological inference with Gaussian processes and neural networks — an application to LSST Y1 weak lensing and galaxy clustering*, *Mon. Not. Roy. Astron. Soc.* **518** (2022) 4818 [arXiv:2203.06124] [INSPIRE].

[54] K.H. Scheutwinkel, W. Handley and E. de Lera Acedo, *Bayesian evidence-driven likelihood selection for sky-averaged 21 cm signal extraction*, *Publ. Astron. Soc. Austral.* **40** (2023) e016 [arXiv:2204.04491] [INSPIRE].

[55] D. Grandón and E. Sellentin, *Bayesian error propagation for neural-net based parameter inference*, *Open J. Astrophys.* **5** (2022) 12 [arXiv:2205.11587] [INSPIRE].

JCAP10(2023)021

[56] P. Lemos et al., *Robust simulation-based inference in cosmology with Bayesian neural networks*, *Mach. Learn. Sci. Tech.* **4** (2023) 01LT01 [arXiv:2207.08435] [INSPIRE].

[57] C.E. Rasmussen and C.K.I. Williams, *Gaussian processes for machine learning*, in *Adaptive Computation and Machine Learning*, MIT Press, Cambridge, MA, U.S.A. (2006).

[58] K.P. Murphy, *Machine Learning — A Probabilistic Perspective*, MIT Press, Cambridge, MA, U.S.A. (2012).

[59] C. Cortes and V. Vapnik, *Support-vector networks*, *Mach. Learn.* **20** (1995) 273.

[60] T. Gunter, M. Osborne, R. Garnett, P. Hennig and S. Roberts, *Sampling for Inference in Probabilistic Models with Fast Bayesian Quadrature*, in proceedings of the *27th International Conference on Neural Information Processing Systems (NIPS'14)*, Montréal, QC, Canada, 8–13 December 2014, Curran Associates, Inc. (2014), pp. 2789–2797 http://papers.nips.cc/paper/5483-sampling-for-inference-in-probabilistic-models-with-fast-bayesian-quadrature.pdf.

[61] M. Osborne, R. Garnett, Z. Ghahramani, D.K. Duvenaud, S.J. Roberts and C. Rasmussen, *Active Learning of Model Evidence Using Bayesian Quadrature*, in proceedings of the *25th International Conference on Neural Information Processing Systems (NIPS'12)*, Lake Tahoe, Nevada, U.S.A., 3–6 December 2012, *Advances in Neural Information Processing Systems* **25**, F. Pereira, C.J.C. Burges, L. Bottou and K.Q. Weinberger eds., Curran Associates, Inc. (2012), pp. 46–54 https://proceedings.neurips.cc/paper/2012/file/6364d3f0f495b6ab9dcf8d3b5c6e0b01-Paper.pdf.

[62] K. Kandasamy, J. Schneider and B. Póczos, *Bayesian active learning for posterior estimation*, in proceedings of the *24th International Joint Conference on Artificial Intelligence (IJCAI'15)*, Buenos Aires, Argentina, 25–31 July 2015, pp. 3605–3611.

[63] H. Wang and J. Li, *Adaptive Gaussian Process Approximation for Bayesian Inference with Expensive Likelihood Functions*, *Neural Comput.* **30** (2018) 3072 [arXiv:1703.09930].

[64] H. Chai and R. Garnett, *Improving Quadrature for Constrained Integrands*, arXiv:1802.04782.

[65] M. Pellejero-Ibañez, R.E. Angulo, G. Aricó, M. Zennaro, S. Contreras and J. Stücker, *Cosmological parameter estimation via iterative emulation of likelihoods*, *Mon. Not. Roy. Astron. Soc.* **499** (2020) 5257 [arXiv:1912.08806] [INSPIRE].

[66] K.K. Rogers, H. Peiris, A. Pontzen, S. Bird, L. Verde and A. Font-Ribera, *Bayesian emulator optimisation for cosmology: application to the Lyman-α forest*, *JCAP* **02** (2019) 031 [arXiv:1812.04631] [INSPIRE].

[67] L. Acerbi, *Variational Bayesian Monte Carlo*, in proceedings of the *32nd International Conference on Neural Information Processing Systems*, Montréal, QC, Canada, 3–8 December 2018, *Advances in Neural Information Processing Systems* **31**, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett eds., Curran Associates, Inc. (2018), pp. 8223–8233 https://proceedings.neurips.cc/paper_files/paper/2018/file/747c1bcceb6109a4ef936bc70cfe67de-Paper.pdf.

[68] L. Acerbi, *Variational Bayesian Monte Carlo with Noisy Likelihoods*, in proceedings of the *34th International Conference on Neural Information Processing Systems (NIPS'20)*, Vancouver, BC, Canada, 6–12 December 2020, *Advances in Neural Information Processing Systems* **33**, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan and H. Lin eds., Curran Associates, Inc. (2020), pp. 8211–8222 [arXiv:2006.08655].

[69] B. Huggins, C. Li, M. Tobaben, M.J. Aarnos and L. Acerbi, *PyVBMC: Efficient Bayesian inference in Python*, arXiv:2303.09519 [DOI:10.21105/joss.05428].

[70] C. Li, G. Clarté and L. Acerbi, *Fast post-process Bayesian inference with Sparse Variational Bayesian Monte Carlo*, arXiv:2303.05263.

[71] A. Saha, K. Bharath and S. Kurtek, *A Geometric Variational Approach to Bayesian Inference*, arXiv:1707.09714.

JCAP10(2023)021

[72] P. Frank, R. Leike and T.A. Enßlin, *Geometric variational inference*, arXiv:2105.10470 [DOI:10.3390/e23070853].

[73] C.A. Micchelli, Y. Xu and H. Zhang, *Universal Kernels*, *J. Mach. Learn. Res.* **7** (2006) 2651.

[74] M. Kupperman, *Probabilities of Hypotheses and Information-Statistics in Sampling from Exponential-Class Populations*, *Ann. Math. Stat.* **29** (1958) 571.

[75] C. Zhu, R.H. Byrd, P. Lu and J. Nocedal, *Algorithm 778: L-BFGS-B*, *ACM Trans. Math. Software* **23** (1997) 550.

[76] Y. Sui, V. Zhuang, J. Burdick and Y. Yue, *Stagewise Safe Bayesian Optimization with Gaussian Processes*, in proceedings of the *35th International Conference on Machine Learning*, Stockholmsmässan, Stockholm, Sweden, 10–15 July 2018, *Proceedings of Machine Learning Research* **80**, J. Dy and A. Krause eds., PMLR (2018), pp. 4781–4789 http://proceedings.mlr.press/v80/sui18a.html.

[77] F. Berkenkamp, A. Krause and A.P. Schoellig, *Bayesian Optimization with Safety Constraints: Safe and Automatic Parameter Tuning in Robotics*, arXiv:1602.04450.

[78] L. Acerbi, *An Exploration of Acquisition and Mean Functions in Variational Bayesian Monte Carlo*, in proceedings of the *1st Symposium on Advances in Approximate Bayesian Inference*, Montréal, QC, Canada, 2 December 2018, F. Ruiz, C. Zhang, D. Liang and T. Bui eds., *Proceedings of Machine Learning Research* **96**, PMLR (2019), pp. 1–1 https://proceedings.mlr.press/v96/acerbi19a.html.

[79] F.L. Fernández, L. Martino, V. Elvira, D. Delgado and J. López-Santiago, *Adaptive Quadrature Schemes for Bayesian Inference via Active Learning*, *IEEE Access* **8** (2020) 208462.

[80] K. Kandasamy, J. Schneider and B. Póczos, *Bayesian active learning for posterior estimation*, in proceedings of the *24th International Joint Conference on Artificial Intelligence (IJCAI'15)*, Buenos Aires, Argentina, 25–31 July 2015, pp. 3605–3611.

[81] T. Desautels, A. Krause and J.W. Burdick, *Parallelizing Exploration-Exploitation Tradeoffs in Gaussian Process Bandit Optimization*, *J. Mach. Learn. Res.* **15** (2014) 4053.

[82] C. Chevalier and D. Ginsbourger, *Fast Computation of the Multi-Points Expected Improvement with Applications in Batch Selection*, in *Learning and Intelligent Optimization*, proceedings of the *7th International Conference (LION 7)*, Catania, Italy, 7–11 January 2013, *Lecture Notes in Computer Science* **7997**, Springer (2013), pp. 59–69 [DOI:10.1007/978-3-642-44973-4_7].

[83] D. Ginsbourger, R. Le Riche and L. Carraro, *A Multi-points Criterion for Deterministic Parallel Global Optimization based on Gaussian Processes*, hal-00260579 (2008).

[84] D. Ginsbourger, R.L. Riche and L. Carraro, *Kriging Is Well-Suited to Parallelize Optimization*, in *Computational Intelligence in Expensive Optimization Problems*, Springer (2010), pp. 131–162 [DOI:10.1007/978-3-642-10701-6_6].

[85] J. González, Z. Dai, P. Hennig and N. Lawrence, *Batch Bayesian Optimization via Local Penalization*, in proceedings of the *19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, Cadiz, Spain, 7–11 May 2016, *Proceedings of Machine Learning Research* **51**, PMLR (2016), pp. 648–657 https://proceedings.mlr.press/v51/gonzalez16a.html.

[86] J. Torrado and A. Lewis, *Cobaya: Code for Bayesian Analysis of hierarchical physical models*, *JCAP* **05** (2021) 057 [arXiv:2005.05290] [INSPIRE].

[87] J. Torrado, N. Schöneberg and J.E. Gammal, *Parallelized Acquisition for Active Learning using Monte Carlo Sampling*, arXiv:2305.19267 [INSPIRE].

[88] E. Cameron and A. Pettitt, *Recursive Pathways to Marginal Likelihood Estimation with Prior-Sensitivity Analysis*, *Statist. Sci.* **29** (2014) 397.

[89] PLANCK collaboration, *Planck 2018 results. Part V. CMB power spectra and likelihoods*, *Astron. Astrophys.* **641** (2020) A5 [arXiv:1907.12875] [INSPIRE].

JCAP10(2023)021

[90] Planck collaboration, *Planck 2018 results. Part VIII. Gravitational lensing*, *Astron. Astrophys.* **641** (2020) A8 [arXiv:1807.06210] [inSPIRE].

[91] D. Blas, J. Lesgourgues and T. Tram, *The Cosmic Linear Anisotropy Solving System (CLASS). Part II. Approximation schemes*, *JCAP* **07** (2011) 034 [arXiv:1104.2933] [inSPIRE].

[92] N. Schöneberg, J. Lesgourgues and D.C. Hooper, *The BAO+BBN take on the Hubble tension*, *JCAP* **10** (2019) 029 [arXiv:1907.11594] [inSPIRE].

[93] S. Kullback and R.A. Leibler, *On Information and Sufficiency*, *Ann. Math. Stat.* **22** (1951) 79 [inSPIRE].

JCAP10(2023)021

Paper II

Parallelized Acquisition for Active Learning using Monte Carlo Sampling

Parallelized Acquisition for Active Learning using Monte Carlo Sampling

Jesús Torrado
Dipartimento di Fisica e Astronomia "G. Galilei"
Università degli Studi di Padova
Via Marzolo 8, I-35131 Padova, Italy
jesus.torrado@pd.infn.it

Nils Schöneberg
Institut de Ciències del Cosmos
Universitat de Barcelona
Martí i Franquès 1, Barcelona E08028, Spain
nils.science@gmail.com

Jonas El Gammal
Department of Mathematics and Physics
University of Stavanger
NO-4036 Stavanger, Norway
jonas.e.elgammal@uis.no

Abstract

Bayesian inference remains one of the most important tool-kits for any scientist, but increasingly expensive likelihood functions are required for ever-more complex experiments, raising the cost of generating a Monte Carlo sample of the posterior. Recent attention has been directed towards the use of emulators of the posterior based on Gaussian Process (GP) regression combined with active sampling to achieve comparable precision with far fewer costly likelihood evaluations. Key to this approach is the batched acquisition of proposals, so that the true posterior can be evaluated in parallel. This is usually achieved via sequential maximisation of the highly multimodal acquisition function. Unfortunately, this approach parallelizes poorly and is prone to getting stuck in local maxima. Our approach addresses this issue by generating nearly-optimal batches of candidates using an almost-embarrassingly parallel Nested Sampler on the mean prediction of the GP. The resulting nearly-sorted Monte Carlo sample is used to generate a batch of candidates ranked according to their sequentially conditioned acquisition function values at little cost. The final sample can also be used for inferring marginal quantities. Our proposed implementation (NORA) demonstrates comparable accuracy to sequential conditioned acquisition optimization and efficient parallelization in various synthetic and cosmological inference problems.

1 Introduction

One of the fundamental tools of science is the comparison of observations with theory. In many Bayesian inference pipelines, this involves inferring the parameters of a model (or models themselves) given some observed or generated data. This is often realised directly using Bayes theorem: Given some model parameters $\boldsymbol{x} \in \mathbb{R}^d$ and data $\mathcal{D}$, the conditioned probability $p(\boldsymbol{x}|\mathcal{D})$ (the so-called *posterior*) is given by

$$p(\boldsymbol{x}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{x})p(\boldsymbol{x})}{p(\mathcal{D})} \,. \tag{1}$$

where $p(\mathcal{D}|\boldsymbol{x}) \equiv L(\boldsymbol{x})$ is called the *likelihood*, $p(\boldsymbol{x}) \equiv \pi(\boldsymbol{x})$ the *prior*, and $p(\boldsymbol{D}) \equiv E$ the *evidence*. We are dropping the explicit dependence on $\mathcal{D}$ as it is fixed for a given inference problem. Traditionally

Preprint. Under review.

the posterior distribution is sampled with Monte Carlo (MC) samplers such as Markov Chain Monte Carlo (MCMC) or Nested Sampling (NS). Unfortunately though, $L(\boldsymbol{x})$ is often an expensive to evaluate black box function, either because calculating observables from the theoretical model involves expensive computations, because the amount of data is large, or both. This makes sampling in such circumstances unfeasible with MC samplers, since they typically require $\mathcal{O}(10^3 - 10^6)$ posterior evaluations for dimensionalities up to $\mathcal{O}(10)$.

There exist multiple approaches to accelerating such inference problems using machine-learning: enhanced pre-conditioning to accelerate traditional MC methods [1–5], simulation-based, implicit-likelihood inference algorithms [6–12], and emulators of underlying physical quantities (for Cosmological applications, see [13–18]). In this work we are going to focus on emulating the likelihood as a function of its parameters, using a Gaussian Process (for previous approaches see [19–25] using GP, and [26–28] enhancing the GP with a variational approximation). Any such emulation (such as that based on a GP) will typically require a set of samples of the function at various parameter points. In order to maximize the amount of information about the behavior of the function captured by these samples, often times an active sampling approach is used: New samples are proposed, based on the current best emulation, where the greatest probability of the estimated improvement of the future emulation is located [29]. This is typically measured by an acquisition function. In this work we will tackle the question of how the active sampling algorithm can be performed in a highly parallel fashion while producing optimal or near-optimal batches of new proposed sampling locations.

In order to acquire a nearly optimal batch of proposed sampling locations for the active learning algorithm in a highly parallel fashion, naive maximization of the acquisition function is not sufficient. This is not only due to the multi-modal nature of a typical acquisition function – making it easy for the optimizer to get stuck in local maxima, especially in moderately high dimensionality – but also due to the inherent lack of parallelization of standard optimization routines. This is caused by the sequential nature of the maximization algorithm, and more importantly requiring the result of a given maximization in order to compute the conditional acquisition function, which will be used for a subsequent maximization.

Our implementation combines the solution to both of these problems in an efficient way: First, by making use of a MC sampling algorithm it is possible to acquire samples of growing function value in a parallel fashion that is much more likely to find the global maximum. Second, through the usage of a ranked pool (see Section 3) we are also able to create a batch of multiple proposed sampling locations simultaneously. Both of these solutions combine to give us a highly efficient algorithm to acquire multiple near-optimal active learning sampling positions.

In Section 3 we describe the general methodology employed in our algorithm. In Section 4 we show the scaling with MPI processes as well as the acquisition histories for a number of toy examples, and we conclude in Section 5. We also show further examples in the context of cosmological inference in appendix E.

2 Theoretical background

2.1 Gaussian Processes

In this section we briefly summarize the main notation and theory. For a review, see [30]. A Gaussian process (GP) is based on a probabilistic model of a function value at any point x, which follows a conditioned Gaussian with a mean μ and a standard deviation σ. Any two sampling locations x and x' are correlated in a multivariate Gaussian way with a correlation function given by $k(x, x')$, the kernel. The choice of the kernel and its hyperparameters encodes assumptions about the behavior of the underlying function (such as differentiability) into the GP. Given a set of sampling locations X_1,...,X_N and corresponding function values y_1,...,y_N the conditioned mean of the GP gives an emulation of the underlying function, while its conditional standard deviation describes the uncertainty in the emulation. A common choice for the kernel function and the one that we will use throughout this paper is the radial basis function (RBF) kernel given in one dimension by

$$k(x, x') = C \cdot \exp\left(\frac{(x - x')^2}{2l^2}\right) \tag{2}$$

where we will call C the *output scale* and l the length scale of the kernel. In multiple dimensions ($\boldsymbol{x} \in \mathbb{R}^d$) we construct the kernel function as a product of RBF kernels, each acting on one dimension

and have different length scales l_i:

$$k(\boldsymbol{x}, \boldsymbol{x}') = C \cdot \exp\left(\sum_{i=1}^{d} \frac{(x_i - x_i')^2}{2l_i^2}\right) \tag{3}$$

By optimizing the marginal log-likelihood of the hyperparameters $\theta = \{C, \boldsymbol{l}\}$ one can fit the GP to a set of sampled points. A good choice of such samples, such as through an acquisition procedure for obtaining new locations is fundamental to the final performance of the GP emulation.

2.2 Acquisition procedure

The second part, the acquisition of samples with which to train the GP relies on maximizing a so called *acquisition function* which, given the current GP, is a measure of the assumed information gained by sampling at any given location. We will denote the already sampled point as the training set in this context. As we want more precision towards the top of the mode for the final inference steps we encode this by choosing the acquisition function

$$a(\mu, \sigma|\boldsymbol{x}) = 2\zeta(\mu(\boldsymbol{x}) - p_{\max}) + \log(\sigma(\boldsymbol{x}))\ . \tag{4}$$

where ζ is an empirically determined dimensional regularization factor,[1] and $p_{\max}$ is the current maximum log-posterior value of the training set. We introduce this current maximum, as in most realistic cases the posterior distribution is not necessarily normalized and hence the scale of the peak not known.

In order to make use of the massive parallelization allowed for by current scientific computing systems, we require not a single optimal point, but a set of simultaneously-optimal sampling locations (batch acquisition). This would in principle require maximizing a joint acquisition function (as a function of multiple locations), which is a high-dimensional multi-modal problem. However, this can be approximated in a simpler way by sequentially acquiring a batch of points, each conditioned to the previous ones using the Kriging believer method [31–36]. In that method one optimizes the acquisition function, conditions the GP on the emulated mean μ at the previous maximum (which is a comparatively cheap operation)[2] and recomputes the acquisition function using this conditioned GP. The true posterior can then be evaluated in parallel at these locations. Throughout this paper, we call this procedure *sequential optimization*.

2.3 Nested sampling

Nested sampling (NS) [37–44] is a family of Monte Carlo sampling algorithms and, simultaneously, an integrator for probability density functions (or positive functions in general). It is based on the idea that the marginal likelihood computation can be substituted by a one-dimensional integration:

$$\int L(\boldsymbol{x})\pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_0^1 \mathfrak{L}(X)\mathrm{d}X\ , \tag{5}$$

where $\mathfrak{L}(X)$ is defined as the inverse of the cumulant prior mass containing only likelihood values greater than a given threshold λ:

$$X(\lambda) = \int_{L(\boldsymbol{x})>\lambda} \pi(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\ . \tag{6}$$

The function $\mathfrak{L}(X)$ is then sampled in increasing order by narrowing (nested) regions that contain only posterior values greater than this threshold. This is performed by tracking a set of *live* points, and sequentially discarding the one with the lowest likelihood value and substituting it for a newly-sampled one. The discarded point is weighed correspondingly to the estimated posterior volume contained within the prior shell defined between the likelihood value of the discarded point and the one of the next lowest-likelihood live point. Due to the nature of NS as an integration, Monte Carlo

[1] We use $\zeta = d^{-0.85}$, which has been shown in [25] to provide a good balance between exploration and exploitation in a variety of dimensionalities.

[2] This is because only the kernel matrix changes in this step, while the hyperparameters do not need to be refitted. Indeed, the highest cost of this operation is solely a single kernel matrix decomposition and inversion required for future predictions on this conditioned GP.

samples from NS produce a better representation of the dynamic range of the distribution than other MC samplers. A review of possible implementations and application to physical sciences can be found in [45].

In this paper, we will use the publicly available PolyChord code [41, 42], in particular the PolyChordLite python wrapper available at `https://github.com/PolyChord/PolyChordLite`. The advantage of using this implementation: the code is well known to allow for massively parallel exploration of the desired function (see [42] for the weak and strong scaling), due to the use of slice sampling to sample from constrained likelihood contours it scales mildly with dimensionality, and due to its cluster identification algorithm it also very good at identifying global maxima even when multiple local maxima are present (see e.g. Rastrigin example in [42]).

3 Method

3.1 Monte Carlo sampling

The basis of our method is substituting the sequential optimisation of the acquisition function using Kriging believer by the exploitation of a Monte Carlo sample of the mean of the GP. Individual samples are ranked according to their acquisition function, as explained below, in a way that reproduces a conditioned ranking similar to what would be obtained via sequential optimization.

Since the target of the acquisition procedure is the optimisation of the aquisition function, it might seem most logical to generate the MC sample directly from it. Nevertheless, there are a number of convincing arguments in favor of sampling on the mean of the GP instead:

Speed: Predicting the GP mean and standard deviation at multiple points simultaneously is much faster due to the possible use of vectorized matrix multiplication routines, but such vectorization is hard to exploit during optimization. While the prediction of the mean is a matrix multiplication of size $(N_{\rm new}, N_{\rm train}) \times N_{\rm train}$, the evaluation of the standard deviation requires at least the matrix multiplication of size $(N_{\rm train}, N_{\rm train}) \times (N_{\rm train}, N_{\rm new})$.[3] As such, it is often times cheaper to first predict only the mean during the sampling (sequential) and then evaluate the standard deviation. These are then used for the acquisition function computation in a single vectorized call.

Simplicity: The acquisition function is often very multi-modal and rather difficult to sample while the mean for a typical well-behaved likelihood is comparatively simple. This reduces the runtime of the MC sampler (sometimes quite drastically) for a given convergence criterion of the MC sample.

Regions of Interest: While there almost surely exist regions far away from the mode with large standard deviations and corresponding acquisition function values, it is not always a good idea to actually sample these. This is because the actual posterior mode defines a region of interest where the accuracy of the GP is desired to be high, while other regions are not necessarily important to sample. This becomes especially interesting in moderately-high dimensionality where the volume contained by the mode becomes an ever smaller fraction of the total prior volume. However, we stress that the nested sampling employed in this work typically explores all regions relevant to the acquisition function in our examples – This region of interest is thus not an entirely strict notion.

Reusability: Since the nested sampling run is performed on the mean of the GP, this is effectively giving us a sample of the emulated posterior at this step, useful for inferring marginal quantities (such as credible intervals, means, variances, marginal distributions, etc.).

The use of NS in particular is advantageous with respect to Markov-chain Monte Carlo methods in this particular case: it naturally balances *exploration* and *exploitation*, since it samples the full dynamic range of the target distribution, including its tails, where low-value-but-high-variance optimal locations dwell; it is also almost-embarrasingly parallel up a number of processes similar to the number of *live* points tracked during sampling, and, depending on implementation, has a mild divergence with dimensionality (true for PolyChord).

[3] There is also a trace of a matrix product, requiring additional $N_{\rm train} \cdot N_{\rm new}$ operations, but this is always subdominant in runtime.

After all samples have been drawn, we compute the acquisition function at these locations simultaneously and use the result to create a batch of new active sampling location proposals.[4]

3.2 Ranked acquisition pool

Instead of the sequential optimisation approach discussed in Section 2.2, we develop an algorithm to rank the MC samples according to their acquisition function value conditioned to the rest of the candidates, using Kriging believer, until an optimal batch of candidates is found. We call this approach a ranked acquisition pool (RAP). To rank a set of points, we start with the sample with the highest unconditioned acquisition as our accepted starting point. From there, we condition the GP to the already accepted samples, and rank all other points according to their acquisition function value conditioned to those accepted samples (i.e. compute the acquisition using the uncertainty of the GP conditioned to the accepted samples). We include an empty slot at the bottom of the pool for temporary sorting. Any sample in that slot will be eventually discarded. Importantly, for any acquisition function monotonic in the GP uncertainty the conditioning can only lower the acquisition value of a point.

We separate the algorithm of proposing a new sample into three main steps, and make use of Figure 1 to show examples for each (description in italics at the end of a step).

1. Initial rejection: A sample is only added if its unconditioned acquisition function is larger than the lowest conditioned acquisition function. *The sample d is rejected from the acquisition pool since its unconditioned acquisition function is smaller than those of samples a, b, c already present in the pool.*
2. Insertion and conditioning: If a sample is not rejected, it is initially inserted at the rank corresponding to its unconditioned acquisition function. If it isn't inserted at the top, it has to subsequently be conditioned to all the points above it (which typically decreases its acquisition function). If it is now lower than the next rank, it is inserted and re-conditioned there. This process is repeated until it is higher than the next rank (goes to step 3), or at the bottom of the pool and thus rejected. *Sample e is proposed to the pool, and in its unconditioned state ranks in the second position. However, after conditioning it to the first point, it performs worse than sample b and is pushed one rank down. It is then conditioned to the two points above it. This time, it performs better than sample c and thus is inserted into its current position. Since its current position is the last position of the pool no resorting is necessary.*
3. Resorting: If a sample has been inserted at any rank but the lowest, all the other ranks below are now conditioned to the wrong samples, and need to be re-conditioned and correspondingly re-ranked. This happens in an iterative fashion, where all samples in the current pool compete for the next highest position under the inserted sample (using the same conditioned GP), and the highest conditional acquisition sample is inserted there. Then the process repeats until all the slots have been filled. *The element f is added to the pool. Its unconditioned acquisition function places it at the top, and it does not need to be conditioned. This invalidates all other ranks, necessitating a full re-sorting of the pool. Next, all of (a,b,e) compete for the second slot by computing the acquisition function value conditioned to the first rank (here sample b wins). Samples a and e now compete for the third slot by computing their acquisition value when conditioned to ranks 1 and 2 simultaneously (sample e wins).*

In order to speed up especially the computations of the conditional acquisition function, the ranked pool works with a cached model of the GP regressor instances, in order to quickly compute acquisition function values conditioned to a certain rank (and those above it). A technical description of the algorithm can be found in Algorithm 1.

By giving up maximization in favour of sampling, our candidates are not the true optima of information gain, but they will be close enough to them. It is more important to get a batch of near-optimal candidates at the same time than getting just a few perfect ones.

[4]The authors of [24] also perform an MC of the mean GP, but do not take care of conditioning when selecting optimal candidates, as we do in the next section.

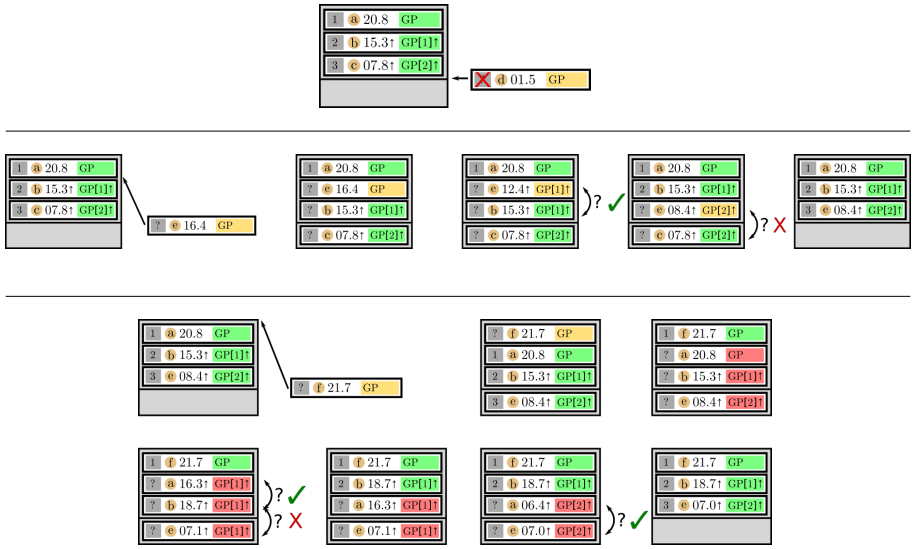


Figure 1: Three insertion cases in a ranked pool of size 3, with one empty slot below it. Each rectangular box represents a single sample. The number in the grey box represents current known rank of the sample (? means un-ranked), the letter in the orange circles is an identifier of the points, the number next to it is the current (conditioned) acquisition function, and the green/orange/red box at the end shows if a point is conditioned to a given rank and those above it (number in angular brackets) and the status of the acquisiton value (green=up-to-date, orange=newly inserted and possibly in need of conditioning, red=invalidated by insertion at higher rank). The three cases are described in the main text.

Algorithm 1 The ranked pool updating routine in pythonic pseudo-code.

Known: Samples $X_1 ... X_N$ with stored conditional acquisitions $a[0] ... a[N]$
Require: New sample X, a(X)

▷ Rejection check

1: $i \leftarrow N$
2: **if** a(X) > a[N] **then**
3: reject(X)
4: **end if**

▷ Finding the correct insertion position

5: **while** i>0 **do**
6: c = $a(X)|(i-1), ..., 1$
7: **if** c > a[i-1] **then**
8: $i \leftarrow i-1$
9: **else**
10: insert(X,i)
11: **end if**
12: **end while**

▷ Resorting + rebuilding the cache:

13: **while** i < N-1 **do**
14: **for** j **in range**(i+1,N) **do**
15: compute $c_j = a(X_j)|i, ..., 1$ ▷ Update conditioned acquisitions
16: **end for**
17: m = argmax[$c_{i+1}, ..., c_n$]
18: swap(X_{i+1},X_m)
19: $a[i+1] \leftarrow c_m$
20: **end while**

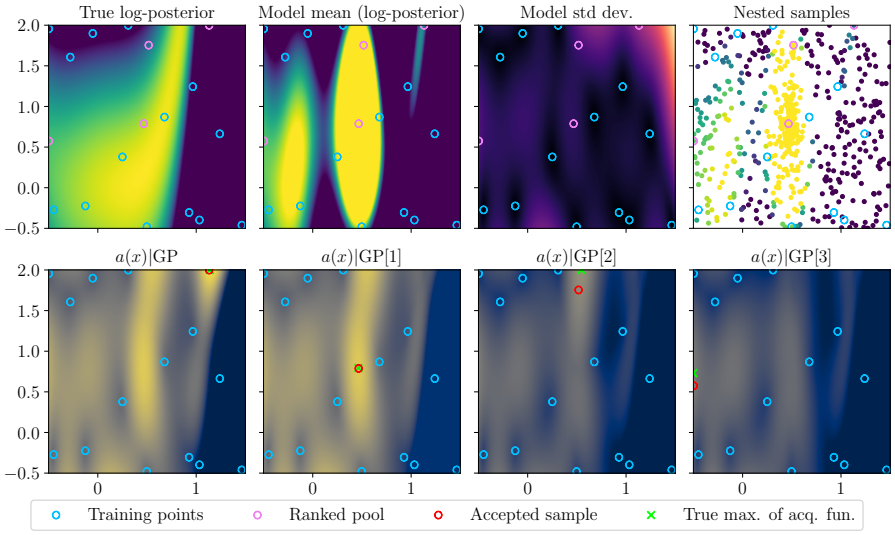


Figure 2: Acquisition procedure with a ranked pool of size $N = 4$. The top row shows from left to right: The true function to be emulated, the current GP mean prediction, it's standard deviation, the nested samples (dead points) from PolyChord. The bottom row shows the acquisition function for the unconditioned GP on the left, and for the conditioned GPs in the three right panels (each conditioned to all samples added to its left). Blue circles are current training samples, pink circles are samples that have been accepted into the ranked pool (top), and red circles are each respective optimal sample for the conditioned GP (bottom). Note that this example is very early in the active sampling so the mode has not been well mapped. Nevertheless it is visible that even with very few samples the locations of the nested samples still cover the regions of high acquisition function well.

4 Results

The combination of the nested sampling approach with the ranked acquisition pool is implemented as NORA (**N**ested sampling **O**ptimization for **R**anked **A**cquistion), based on the GP treatment from [25, 46] (as well as useful functionality from [47, 48]). A demonstration of the acquisition procedure in NORA can be found in fig. 2.

We tested NORA on a number of synthetic likelihoods to demonstrate both the accuracy and the highly parallel nature of our approach. The likelihoods for accuracy tests include a curved degeneracy, a ring, and the multi-modal Himmelblau function. Further discussion of these synthetic examples can be found in appendix D, while real-world applications to cosmological data can be found in appendix E.

The curved degeneracy (see also [25, 49]) has a tight ridge in the $x_2 \approx 4x_1^4$ direction, and its log-likelihood is

$$\log L(x_1, x_2) = -(10 \cdot (0.45 - x_1))^2/4 - (20 \cdot (x_2/4 - x_1^4))^2 \,. \tag{7}$$

The log-likelihood for the ring example is instead

$$\log L(x_1, x_2) = -\frac{1}{2}\left[\frac{(\sqrt{x_1^2 + x_2^2} - \mu)^2}{\sigma} + \log(2\pi\sigma^2)\right] \,, \tag{8}$$

where $\mu = 1$ and $\sigma = 0.05$ in our example. We show in Figure 3 that in both of these cases the accuracy and efficiency is very comparable to the sequential method (while much more parallelizable, see below). Both reach about the same level of agreement between emulation and the true function

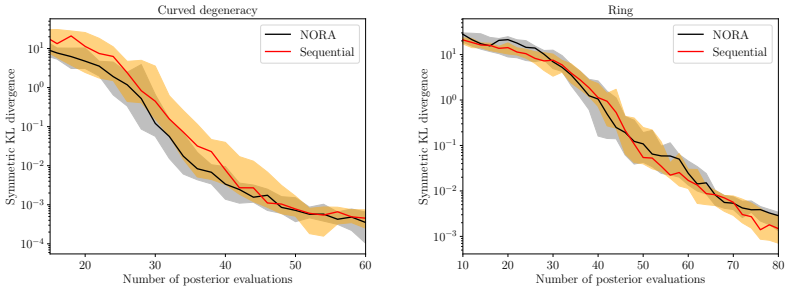


Figure 3: Comparison of the efficiency and accuracy of the acquisition procedure between the naive sequential optimization approach and the NORA approach. We show the agreement between the emulated and the true posterior (specified by the symmetric KL divergence) as a function of the number of samples (posterior evaluations). The solid line is the median, and the shaded region is the 25% to the 75% quantiles of 20 realizations. In this case NORA shows similar performance to the sequential algorithm.

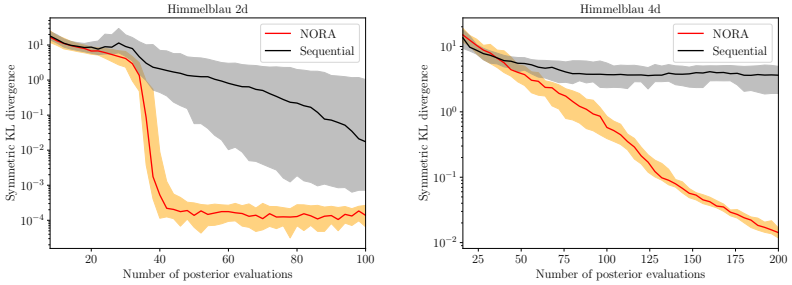


Figure 4: Same as Figure 3 for the Himmelblau function (left) and a four-dimensional extension with 4 modes in two of the dimensions (right). In the other two dimensions it is flat. In these multi-modal cases the NORA algorithm is far more efficient than the sequential sampling algorithm.

(as captured by their symmetric KL divergence, which is further explain in appendix C). We also investigate a multi-modal example like the Himmelblau function with log-likelihood

$$\log L(x_1, x_2) = -\frac{1}{2}\left[100 \cdot (x_1^2 - x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2\right] \tag{9}$$

We furthermore construct a four-dimensional version of this function which retains the four maxima in two dimensions but is constant along the other two dimensions. This combines the multimodality with the problem of correctly mapping and exploring the flat dimensions. We show the results in Figure 4. Since in this case the nested sampling has a far higher chance of quickly discovering a mode of the function far from the already known ones, the NORA approach is much more efficient than the Sequential optimization approach in this case (we show examples of explicit modeling for 100 posterior evaluations in Figure 5).

In order to assess that gains in modelling do not come at the cost of overhead in the acquisition step, we have performed a number of tests in Gaussian likelihoods at different dimensionalities. The comparison with the costs of acquisition with sequential optimization and NORA, as well as the scaling with parallelization is shown in Table 1. We see that the overhead of NORA is comparable to that of sequential optimization for the same number of MPI processes. However, sequential optimization will only profit from parallelization up to the number of restarts of the optimizer while nested sampling will parallelize virtually infinitely (up to the large number of live points).

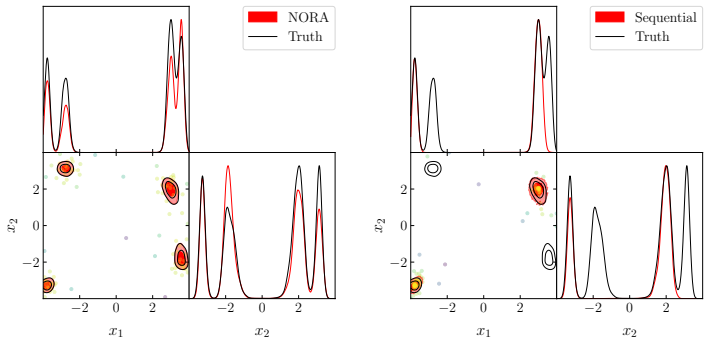


Figure 5: Example of a failure case of the naive sequential optimization compared to the same cases treated with NORA, which due to its nested sampling correctly identifies all modes. The sampling locations are shown with colour denoting how late they were sampled (yellower=later). It is clearly visible that the sequential optimization is sampling more aggressively towards the top of the mode and showing less explorative behaviour. Both are allowed 100 posterior evaluations.

	$d = 2$		$d = 4$		$d = 8$	
SeqOpt	[2]	[2.4, **2.55**, 2.7]	[4]	[11.2, **12.1**, 12.7]	[8]	[177, **183**, 191]
NORA	[2]	[3.67, **3.84**, 4.34]		—		—
NORA	[4]	[1.43, **1.54**, 1.75]	[4]	[9.3, **9.9**, 10.6]		—
NORA	[8]	[0.96, **1.05**, 1.13]	[8]	[6.29 **6.58** 6.90]	[8]	[147, **160**, 220]
NORA	[16]	[0.30, **0.33**, 0.35]	[16]	[4.52, **4.77**, 5.16]	[16]	[122, **129**, 171]

Table 1: Comparison of wall-clock runtimes for the acquisition step between NORA and sequential optimization (dubbed "SeqOpt", with $5 \cdot d$ restarts of the optimizer). We add in angular brackets the number of MPI processes. In each dimensionality we show the [25, **50**, 75] percent quantiles. We run 50 runs in 2- and 4 dimensions, and 20 runs in 8 dimensions, with respective truth evaluation budgets 20, 60 and 400. Convergence in terms of symmetric KL divergence is similar in all cases and of magnitude $\mathcal{O}(0.01)$. We additionally allow multi-threading (useful e.g. for BLAS [50] matrix operations) for each MPI process up to a total of 32 cores.

5 Conclusion

Sequential optimization for active learning is facing a variety of challenges, such as difficult parallelization, and a lack of robustness to getting stuck in local maxima, thus requiring many restarts of the optimizer in high dimensions to properly explore the target inference space. To overcome these challenges we propose a new algorithm, called NORA, that substitutes the sequential optimization of the acquisition function by combining Monte Carlo exploration of the GP's mean using Nested Sampling, and ranking of the Monte Carlo samples according to their conditional acquisition function values, to generate a nearly optimal batch of sampling locations. These two steps can be performed in a nearly perfectly-parallelizable way, and the same Monte Carlo sample can be reused in consecutive iterations for lowering computational costs.

We apply NORA to a number of synthetic Bayesian inference problems to assess its performance, and compare it to a reasonably good implementation of sequential optimisation of the acquisition function.

We find that NORA and sequential optimization perform equally well at comparable computational costs for simple unimodal likelihoods for $d < 10$, and for highly non-Gaussian likelihoods in small dimensionalities. NORA greatly outperforms sequential optimization for multi-modal likelihoods, due to the more exploratory approach to acquisition, despite producing less precise acquisition batches than sequential optimization.

The limitations of the NORA algorithm are similar to those of other approaches to Bayesian inference based on surrogate GP models: Their strong divergence with dimensionality due to the increasingly large number of training points needed for good posterior modelling, and the $\mathcal{O}(n^3)$ scaling when fitting of the hyperparameters of the GP (see also [51]). Furthermore, one particular shortcoming of NORA compared to sequential optimization is that due to its less aggressive acquisition it will converge later in simple problems, e.g. Gaussian likelihoods. Our methodology also does not address the problem of stochastic likelihood evaluations (see [27]).

Acknowledgments and Disclosure of Funding

J. T. acknowledges support from the STARS@UNIPD2021 project *GWCross*. N. S. acknowledges support from the Maria de Maetzu fellowship grant: CEX2019-000918-M, financiado por MCIN/AEI/10.13039/501100011033. J. E. acknowledges support by the ROMFORSK grant project no. 302640.

A Description of the surrogate model

In this section we describe details of the GP surrogate model (see also [25] for a detailed description).

A.1 Choice of kernel function

On top of the choice of the kernel function itself, as defined in Equation (3), some knowledge of the target function is also incorporated in the priors for the hyperparameters. Our assumption is that the length scales should be of an order of magnitude close to that of the posterior modes, while the latter would be of an order of magnitude not much smaller than that of the prior ranges for the parameters of the posterior. We express this belief by setting the prior of the length scales to being uniform between 0.01 and 1 in units of the prior length in each direction. This condition assumes that the size of the mode is larger than about $1/100$th of the prior width in each dimension, which we find reasonably permissive. The prior of the output scale C is chosen to be very broad and allows for values between 0.001 and 10000. The $d+1$ free hyperparameters $\theta \equiv \{C, l_i\}$ are then chosen such that they maximize

$$-\log p(\boldsymbol{y}|\boldsymbol{X},\theta) = \frac{1}{2}\boldsymbol{y}^T(\boldsymbol{k}(\boldsymbol{X},\boldsymbol{X}) + \sigma_n^2\boldsymbol{I})^{-1}\boldsymbol{y} + \frac{1}{2}\log|\boldsymbol{k}(\boldsymbol{X},\boldsymbol{X}) + \sigma_n^2\boldsymbol{I}| - \frac{N_s}{2}\log 2\pi\,. \quad (10)$$

where σ_n is a small noise parameter that typically improves numerical stability of the matrix inversion.

A.2 Parameter space transformations

To ensure numerical stability we use a number of transformations during the modelling with the GP:

Firstly, we sample the log-posterior distribution to reduce the scale of the function that the GP interpolates. Furthermore, the characteristic length scale of isotropic kernels tends to be larger when sampling the log-posterior, which implies that the GP surrogate generalizes better to distant parts of the function, making the GP more predictive.

In addition, at every iteration of the algorithm, we *internally* re-scale the modeled function using the mean and standard deviation of the current samples set as

$$\log\widetilde{p}(\boldsymbol{X}) = \frac{\log p(\boldsymbol{X}) - \overline{\boldsymbol{y}}}{s_{\boldsymbol{y}}}\,, \quad (11)$$

where $\overline{\boldsymbol{y}}$ and $s_{\boldsymbol{y}}$ are the sample mean and standard deviation respectively. This re-scaling acts like a non-zero mean function, causing the GP to return to the mean value far away from sampling locations. This in turn encourages exploration when most samples are close to the mode and exploitation when most samples have low posterior values.

As for the space of parameters $\boldsymbol{x}$, we transform the samples such that the prior boundary becomes a unit-length hypercube. This usually leads to comparable correlation length scales of the GP across dimensions, which increases the effectiveness of the limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS-B) constrained optimizer [52], used to optimize the GP hyperparameters.

B Some details of the algorithm

Active-sampling Bayesian inference algorithms based on surrogate models in the literature [19–28] usually follow a fixed procedure: after an initial batch of training samples is either provided or drawn from the prior, the algorithm iterates on a cycle of (1) optimising an acquisition function to obtain candidates for evaluation of the true posterior, (2) evaluation of the true posterior at the proposed locations, (3) refitting of the surrogate model, and (4) convergence checks. In this study we do not concern ourselves with the initial proposal (in our case sampled from the prior) or the convergence checks (in most examples we have fixed budgets of how many true posterior points are sampled), since the focus of this study is on the acquisition step.

As discussed in the main text, our acquisition procedure has two steps: first the mean GP is explored using nested sampling, and, second, the resulting MC samples are ranked according to their conditioned acquisition function value. In this short appendix, we discuss some particularities of these procedures.

B.1 Scaling of Nested Sampling precision parameters

The two fundamental parameters of a nested sampler are the number of *live points*, and the fraction of the total posterior mass (evidence) contained in the final set of live points. Additional parameters depend on the particular implementation of NS. In POLYCHORD, our sampler of choice, the aforementioned parameters are called respectively `nlive` and `precision_criterion`. In addition, and among others, POLYCHORD has two more important parameters: the length of the slice-sampling chains (`num_repeats`), and the size of the initial prior sample from which the live points are extracted (`nprior`). It is a natural choice that in the early stages of learning, where limited precision when optimising the acquisition function is enough, this would translate in our procedure into lower precision settings for POLYCHORD. To reflect this, we scale the number of live points to be proportional to the number of points in the training set (by a factor of 3 by default), with a cap equal to the default precision criterion of POLYCHORD, which is 25 times the dimensionality of the problem. On the other hand, we have found that the accuracy of our algorithm benefits from more accuracy than the default for the length of slice chains (`num_repeats`, 5 times the dimensionality instead of 2), whereas the evidence fraction contained in the live points (`precision_criterion`) can be relaxed with respect to the default by a factor of 5, since we are not interested in an accurate calculation of the model evidence.

B.2 Byproducts of Nested Sampling

The nested sampling step produces both a MC sample and a calculation of the evidence of the model. The first one is a useful by-product, which can be used for inference once the run has converged, or to implement a global convergence criterion, such as one based on the calculation of KL divergences between iterations. The value of the evidence is also a useful output, in particular to define a further convergence criterion, but it needs to be taken into account that the resulting NS uncertainty does not include the uncertainty due to the probabilistic nature of the GP, or the uncertainty over the choice of hyperparamenters values, as Bayesian Quadrature approaches do.

B.2.1 Parallelization

Nested samplers parallelize effectively up to the number of live points (`nlive`), since parallel evaluation of the target function increases the chance that at every iteration an acceptable sample will be found at the cost of a single evaluation. Since this number is usually a few tens of times the dimensionality, this step of our algorithm will effectively parallelize linearly with the number of simultanous processes. POLYCHORD does not do vectorized evaluation of the target function, i.e. the target function is always called with a single argument. Hence for this step we prefer to invest CPU cores into separate MPI processes, as opposed to multiple threads.

The ranking step of the algorithm when running NORA in parallel occurs in two steps: first the MC sample is split in as many equal parts as running processes, for evaluation of the GP standard deviation and the acquisition function value, and the individual ranking of each subset into ranked pools with as many points as the desired Kriging believer steps; and later all the ranked pools are combined an re-ranked in a single process. The first of these two steps can be effectively parallelized, but the second one is not parallelizable by definition, and may at most benefit for multi-threading. In most situations, unless the size of the training set is very large, the first step is costlier and thus a larger number of MPI processes is more beneficial than a larger number of threads per process.

Finally, the evaluation step occurs always in parallel when MPI processes are available, but its parallelization is limited by the number of Kriging believer steps we have decided to take. This number must be kept in check because the quality of the batch of proposals decreases when conditioning on increasingly bad information, making our model larger and more computationally expensive. We have found that a number of Kriging believer steps equal to the dimensionality is a good choice in most cases. Highly multimodal posterior can benefit from larger number of Kriging beliver steps, since their acquisition functions have more local maxima, but it would not be wise to go beyond a few times the number of dimensions. Thus, the evaluation step benefits from the number of MPI processes in a limited way, and may be faster if more cores are left available for multi-threading, thus accelerating the evaluation of the true posterior.

The difference between the acquisition step benefiting from a large number of MPI processes, and the evaluation step potentially benefiting more from a large number of threads, makes the choice of the

ratio of MPI processes to threads per process dependent on the speed of the evaluation step and the overhead costs, which scale with dimensionality: fast true posteriors in high dimensionality call for larger number of MPI threads, and very slow posterior with an implementation that benefits from multi-threading would call for a larger amount of threads and a smaller amount of MPI processes. In the future, we will look at substituting PolyChord by a nested sampler that can perform vectorized calls to the target function, in order to make multi-threading an overall better choice, beyond the small necessary MPI parallelization for Kriging believer.

C Kullback-Leibler Divergences

We define the Kullback-Leibler (KL) divergence of the continuous probability distribution P with respect to Q with probability density functions $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ as

$$D_{\mathrm{KL}}(P||Q) = \int p(\boldsymbol{x}) \log\left(\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}\right) d\boldsymbol{x} \,. \tag{12}$$

The KL divergence as defined above more strongly weighs disagreements between the two probability distributions where $p(\boldsymbol{x})$ is large. Since we want the approximation to be equally accurate in all regions where either distribution is large, we use a symmetrized version of the divergence (often called Jeffreys divergence). It is defined as

$$D_{\mathrm{KL}}^{\mathrm{sym}}(P,Q) = \frac{1}{2}\left(D_{\mathrm{KL}}(P||Q) + D_{\mathrm{KL}}(Q||P)\right) \,. \tag{13}$$

A smaller value means that the two posteriors are in better agreement, and one typically wants $D_{\mathrm{KL}}^{\mathrm{sym}}(P||Q) \ll 1$ for good agreement. The dimensionality consistency of the KL divergence guarantees that a given value for the divergence characterizes similar differences across dimensionalities.

To compute the KL divergence explicitly, one can use the fact that the points in a Monte Carlo sample of P are distributed as $p(\boldsymbol{x})d\boldsymbol{x}$. One can thus approximate the integral as a sum of the quantity $\log p(\boldsymbol{x}_i) - \log q(\boldsymbol{x}_i)$ over all points in the MC sample (multiplied by their respective weights/multiplicities). This can be done by evaluating either the real model or the GP emulated posterior for the given points.

There also exists a Gaussian approximation for the KL divergence which is particularly useful when computing the true log-posteriors at each point of the MC sample is computationally undesirable (such as the cosmological examples below). It is defined as

$$D_{\mathrm{KL}}(P||Q) \approx \frac{1}{2}\left(\mathrm{tr}\left(\boldsymbol{C}_Q^{-1}\boldsymbol{C}_P\right) - d + (\mathfrak{m}_Q - \mathfrak{m}_P)^T \boldsymbol{C}_Q^{-1}(\mathfrak{m}_Q - \mathfrak{m}_P) + \log\left(\frac{\det \boldsymbol{C}_Q}{\det \boldsymbol{C}_P}\right)\right) \,. \tag{14}$$

with $\boldsymbol{C}_Q$ and $\boldsymbol{C}_P$ being the respective covariance matrices of the two probability distributions, while $\mathfrak{m}_Q$ and $\mathfrak{m}_P$ are the respective means. While the approximation of the individual distribution as multivariate Gaussian is certainly incorrect in non-Gaussian cases, it is typically the case that a good agreement of the Gaussian KL signals a good compatibility of the true KL as well. We always compute the true symmetric KL unless explicitly stated otherwise.

D Test functions

Here we comment further on the test functions presented in Section 4 of the main text.

Figure 6 and Figure 7 show exemplary corner plots of the examples used in Section 4. In all three multi-modal cases presented in that section, NORA correctly recovers the contours.

In Section 4 we also presented a study of the parallelization of the overhead costs of NORA. In this context, in Figure 8 we show comparisons in convergence between NORA and sequential optimization for Gaussians drawn with random correlations in 2, 4 and 8 dimensions. For these very easy-to-model functions NORA converges as fast as sequential optimization. The slightly slower convergence in $d = 8$ is likely due to the somewhat more exploratory behaviour of NORA compared to sequential optimization.

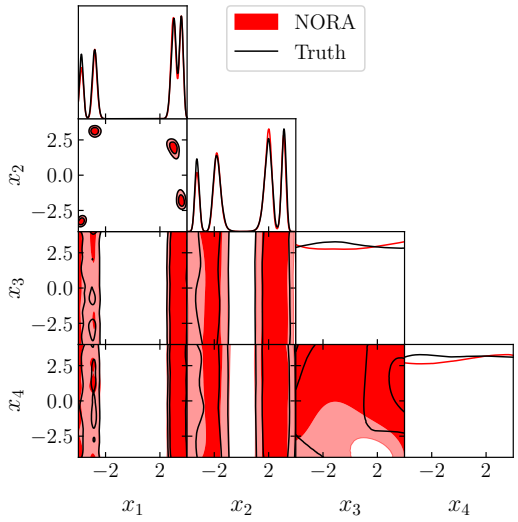


Figure 6: Example of inference on the 4d Himmelblau example. The four modes are in the x_1-x_2 direction while the other two directions are flat. The example shows NORA sampling with a budget of 200 posterior evaluations. Both contours are in good agreement with each other.

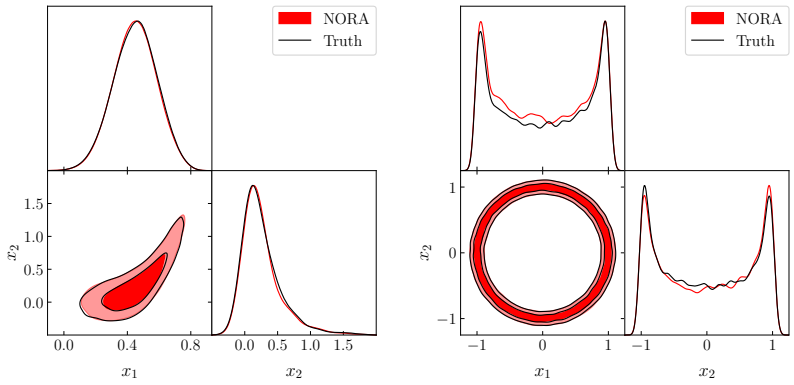


Figure 7: Example of inference on the curved degeneracy example (left) and the ring (right). NORA correctly recovers both contours. Both runs have been performed with a budget of 80 posterior evaluations.

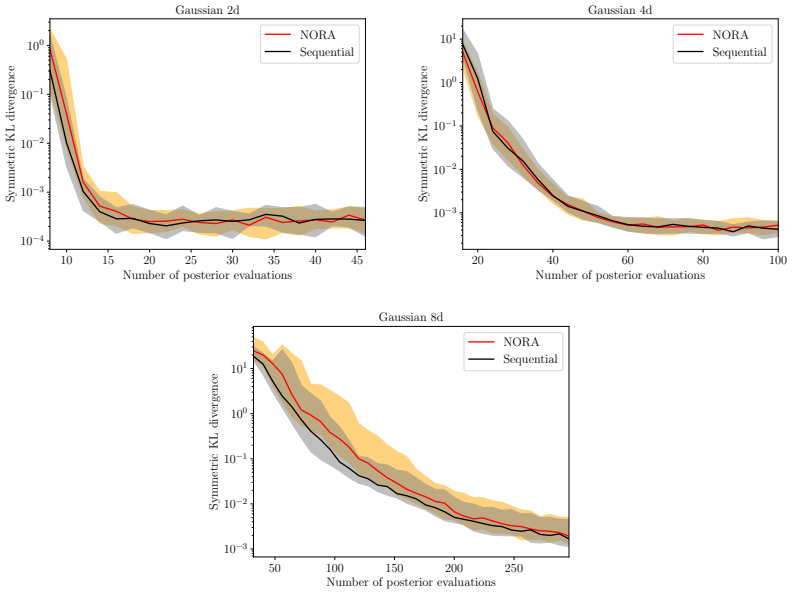


Figure 8: Comparing convergence of NORA vs. Sequential optimization for randomly drawn Gaussians in 2, 4, 8 dimensions. NORA and sequential optimization perform nearly equally well for these easy-to-model likelihoods.

E Cosmological examples

In order to test the applicability and robustness of the NORA implementation to real-world examples, we also apply it to a number of inference runs commonly used in cosmology. In particular, we use the Planck 2018 temperature, polarization, and lensing data (using the nuisance-marginalized 'lite' version, as described in [53, 54]), and consider either a model of a curved universe (ΛCDM + Ω_k) or a model with sinusoidal variations of the primordial power spectrum, similar to [55, Sec 7.1.1]. For the ΛCDM baseline model in both cases we adopt the common 6 cosmological parameters $\{\ln(10^{10}A_s), n_s, H_0, \Omega_b h^2, \Omega_{\rm cdm} h^2, \tau_{\rm reio}\}$ and adopt a single massive neutrino with mass 0.06eV (see [56] for a more detailed description of this baseline model).

We show in Figure 9 the results for a model of a curved universe, an extension of the ΛCDM model described above with an additional seventh parameter Ω_k representing the energy density-equivalent of the curvature. It presents a particularly strong degeneracy between the curvature parameter Ω_k and the Hubble constant H_0. We observe that the contours are in good agreement ($D_{\rm KL}^{\rm sym,Gaussian} = 0.08$ using the Gaussian approximation of Equation (14)).

Next we try to fit a sinusoidal oscillation with three parameters (amplitude, wavelength and phase) to the primordial power spectrum of Planck 2018, fixing the parameters of the ΛCDM model. This is a low-dimensional problem, but with a highly multi-modal behavior in the frequency and phase of the oscillation, since we are effectively fitting experimental noise. The result can be seen in Figure 10: most of the distribution is well recovered, despite its complexity.

F Reproducibility

The NORA implementation and all scripts required to reproduce the tests will be released after review of this manuscript.

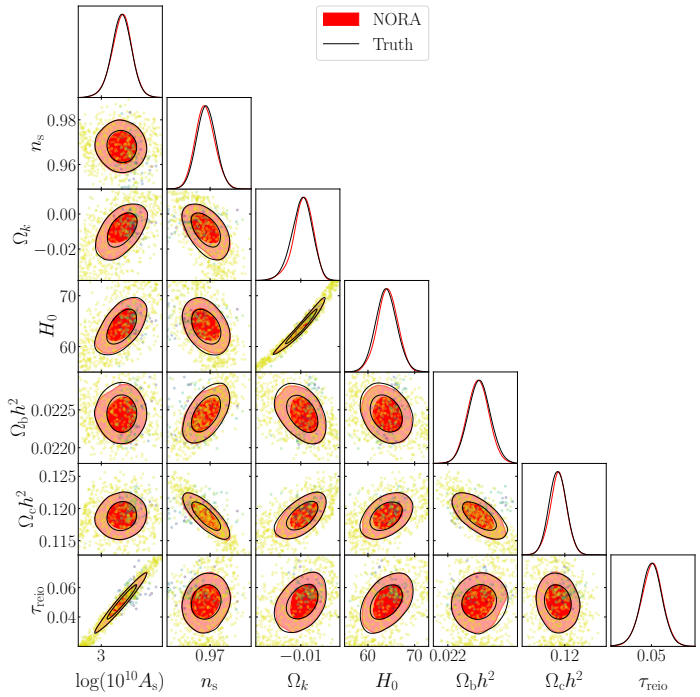


Figure 9: Inference of the cosmological parameters of the Planck 2018 likelihood (Planck lite) with curvature Ω_k sampled in addition. NORA correctly recovers the contours with only 903 evaluations of the likelihood function.

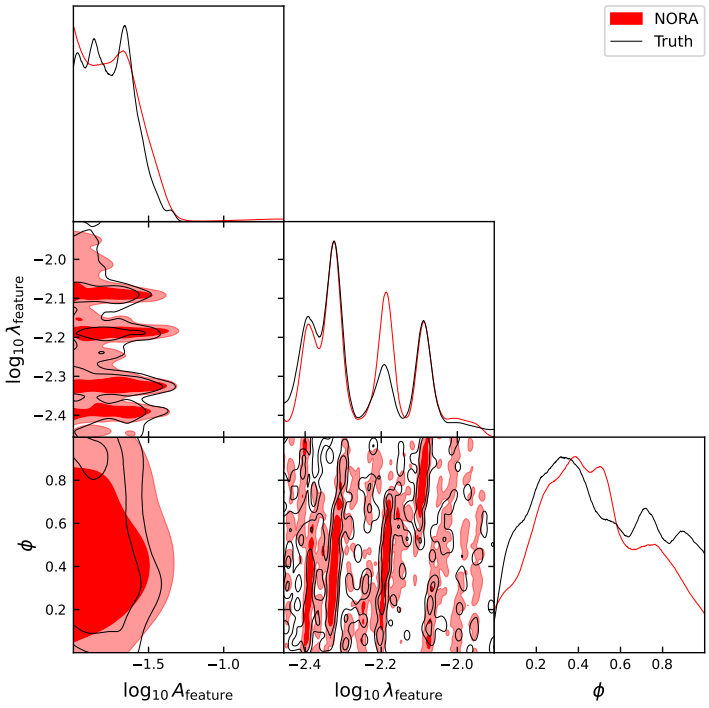


Figure 10: A three dimensional model of an oscillation in the primordial power spectrum of the Planck 2018 CMB sky constrained by NORA, and the reference Nested Sampling results with `PolyChord`. NORA accurately captures the constraints with a budget of 2000 evaluations, performed in parallel batches of 6 evaluations.

References

[1] Philip Graff, Farhan Feroz, Michael P. Hobson, and Anthony Lasenby. BAMBI: blind accelerated multimodal bayesian inference. *Monthly Notices of the Royal Astronomical Society*, jan 2012. doi: 10.1111/j.1365-2966.2011.20288.x. URL `https://doi.org/10.1111%2Fj.1365-2966.2011.20288.x`.

[2] Adam Moss. Accelerated Bayesian inference using deep learning. *Mon. Not. Roy. Astron. Soc.*, 496(1):328–338, 2020. doi: 10.1093/mnras/staa1469.

[3] Hector J. Hortua, Riccardo Volpi, Dimitri Marinelli, and Luigi Malago. Accelerating MCMC algorithms through Bayesian Deep Networks. In *34th Conference on Neural Information Processing Systems*, 11 2020.

[4] Michael J. Williams, John Veitch, and Chris Messenger. Nested sampling with normalizing flows for gravitational-wave inference. *Phys. Rev. D*, 103(10):103006, 2021. doi: 10.1103/PhysRevD.103.103006.

[5] Minas Karamanis, David Nabergoj, Florian Beutler, John A. Peacock, and Uros Seljak. pocoMC: A Python package for accelerated Bayesian inference in astronomy and cosmology. *J. Open Source Softw.*, 7(79):4634, 2022. doi: 10.21105/joss.04634.

[6] Florent Leclercq. Bayesian optimization for likelihood-free cosmological inference. *Phys. Rev. D*, 98(6):063511, 2018. doi: 10.1103/PhysRevD.98.063511.

[7] Justin Alsing, Tom Charnock, Stephen Feeney, and Benjamin Wandelt. Fast likelihood-free cosmology with neural density estimators and active learning. *Mon. Not. Roy. Astron. Soc.*, 488 (3):4440–4458, 2019. doi: 10.1093/mnras/stz1960.

[8] Benjamin Kurt Miller, Alex Cole, Gilles Louppe, and Christoph Weniger. Simulation-efficient marginal posterior estimation with swyft: stop wasting your precious time. 11 2020.

[9] T. Lucas Makinen, Tom Charnock, Justin Alsing, and Benjamin D. Wandelt. Lossless, scalable implicit likelihood inference for cosmological fields. *JCAP*, 11(11):049, 2021. doi: 10.1088/1475-7516/2021/11/049. [Erratum: JCAP 04, E02 (2023)].

[10] Biwei Dai and Uros Seljak. Translation and rotation equivariant normalizing flow (TRENF) for optimal cosmological analysis. *Mon. Not. Roy. Astron. Soc.*, 516(2):2363–2373, 2022. doi: 10.1093/mnras/stac2010.

[11] Justine Zeghal, Francois Lanusse, Alexandre Boucaud, Benjamin Remy, and Eric Aubourg. Neural Posterior Estimation with Differentiable Simulator. In *Machine Learning for Astrophysics*, page 52, July 2022. doi: 10.48550/arXiv.2207.05636.

[12] Pablo Lemos, Miles Cranmer, Muntazir Abidi, ChangHoon Hahn, Michael Eickenberg, Elena Massara, David Yallup, and Shirley Ho. Robust simulation-based inference in cosmology with Bayesian neural networks. *Mach. Learn. Sci. Tech.*, 4(1):01LT01, 2023. doi: 10.1088/2632-2153/acbb53.

[13] Andrea Manrique-Yus and Elena Sellentin. Euclid-era cosmology for everyone: neural net assisted MCMC sampling for the joint 3×2 likelihood. *Mon. Not. Roy. Astron. Soc.*, 491(2): 2655–2663, 2020. doi: 10.1093/mnras/stz3059.

[14] Arrykrishna Mootoovaloo, Alan F. Heavens, Andrew H. Jaffe, and Florent Leclercq. Parameter Inference for Weak Lensing using Gaussian Processes and MOPED. *Mon. Not. Roy. Astron. Soc.*, 497(2):2213–2226, 2020. doi: 10.1093/mnras/staa2102.

[15] Alessio Spurio Mancini, Davide Piras, Justin Alsing, Benjamin Joachimi, and Michael P. Hobson. CosmoPower: emulating cosmological power spectra for accelerated Bayesian inference from next-generation surveys. *Mon. Not. Roy. Astron. Soc.*, 511(2):1771–1788, 2022. doi: 10.1093/mnras/stac064.

[16] Chun-Hao To, Eduardo Rozo, Elisabeth Krause, Hao-Yi Wu, Risa H. Wechsler, and Andrés N. Salcedo. LINNA: Likelihood Inference Neural Network Accelerator. *JCAP*, 01:016, 2023. doi: 10.1088/1475-7516/2023/01/016.

[17] Andreas Nygaard, Emil Brinch Holm, Steen Hannestad, and Thomas Tram. CONNECT: a neural network based framework for emulating cosmological observables and cosmological parameter inference. *JCAP*, 05:025, 2023. doi: 10.1088/1475-7516/2023/05/025.

[18] Sven Günther, Julien Lesgourgues, Georgios Samaras, Nils Schöneberg, Florian Stadtmann, Christian Fidler, and Jesús Torrado. CosmicNet II: emulating extended cosmologies with efficient and accurate neural networks. *JCAP*, 11:035, 2022. doi: 10.1088/1475-7516/2022/11/035.

[19] Michael Osborne, Roman Garnett, Zoubin Ghahramani, David K Duvenaud, Stephen J Roberts, and Carl Rasmussen. Active learning of model evidence using bayesian quadrature. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/6364d3f0f495b6ab9dcf8d3b5c6e0b01-Paper.pdf.

[20] Tom Gunter, Michael A Osborne, Roman Garnett, Philipp Hennig, and Stephen J Roberts. Sampling for inference in probabilistic models with fast bayesian quadrature. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/e94f63f579e05cb49c05c2d050ead9c0-Paper.pdf.

[21] K. Kandasamy, J. Schneider, and B. Póczos. Bayesian active learning for posterior estimation. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 3605–3611, 2015. ISBN 978-1-577-35738-4.

[22] Henry Chai and Roman Garnett. Improving Quadrature for Constrained Integrands. *arXiv e-prints*, art. arXiv:1802.04782, February 2018. doi: 10.48550/arXiv.1802.04782.

[23] Hongqiao Wang and Jinglai Li. Adaptive Gaussian Process Approximation for Bayesian Inference with Expensive Likelihood Functions. *Neural Computation*, 30(11):3072–3094, 11 2018. ISSN 0899-7667. doi: 10.1162/neco_a_01127. URL https://doi.org/10.1162/neco_a_01127.

[24] Marcos Pellejero-Ibañez, Raul E. Angulo, Giovanni Aricó, Matteo Zennaro, Sergio Contreras, and Jens Stücker. Cosmological parameter estimation via iterative emulation of likelihoods. *Mon. Not. Roy. Astron. Soc.*, 499(4):5257–5268, 2020. doi: 10.1093/mnras/staa3075.

[25] Jonas El Gammal, Nils Schöneberg, Jesús Torrado, and Christian Fidler. Fast and robust Bayesian Inference using Gaussian Processes with GPry. 11 2022.

[26] Luigi Acerbi. Variational bayesian monte carlo. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/747c1bcceb6109a4ef936bc70cfe67de-Paper.pdf.

[27] Luigi Acerbi. Variational bayesian monte carlo with noisy likelihoods. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 8211–8222. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/5d40954183d62a82257835477ccad3d2-Paper.pdf.

[28] Bobby Huggins, Chengkun Li, Marlon Tobaben, Mikko J. Aarnos, and Luigi Acerbi. Pyvbmc: Efficient bayesian inference in python, 2023.

[29] Zoubin Ghahramani and Carl Rasmussen. Bayesian monte carlo. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002. URL https://proceedings.neurips.cc/paper_files/paper/2002/file/24917db15c4e37e421866448c9ab23d8-Paper.pdf.

[30] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Mass. [u.a.], 2006. ISBN 0-262-18253-X and 978-0-262-18253-9.

[31] Thomas Desautels, Andreas Krause, and Joel W. Burdick. Parallelizing exploration-exploitation tradeoffs in gaussian process bandit optimization. *Journal of Machine Learning Research*, 15 (119):4053–4103, 2014. URL http://jmlr.org/papers/v15/desautels14a.html.

[32] Clément Chevalier and David Ginsbourger. Fast computation of the multi-points expected improvement with applications in batch selection. In Giuseppe Nicosia and Panos Pardalos, editors, *Learning and Intelligent Optimization*, pages 59–69, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-44973-4.

[33] J. González, Z. Dai, P. Hennig, and N. Lawrence. Batch bayesian optimization via local penalization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51 of *JMLR Workshop and Conference Proceedings*, pages 648–657, May 2016. URL http://jmlr.org/proceedings/papers/v51/gonzalez16a.pdf.

[34] David Ginsbourger, Rodolphe Le Riche, and Laurent Carraro. A Multi-points Criterion for Deterministic Parallel Global Optimization based on Gaussian Processes. Technical report, March 2008. URL https://hal.archives-ouvertes.fr/hal-00260579.

[35] David Ginsbourger, Rodolphe Le Riche, and Laurent Carraro. *Kriging Is Well-Suited to Parallelize Optimization*, volume 2, pages 131–162. 01 2010. doi: 10.1007/978-3-642-10701-6_6.

[36] R. J. Barnes and A. G. Watson. Efficient updating of kriging estimates and variances. *Mathematical Geology*, 24(1):129–133, Jan 1992. ISSN 1573-8868. doi: 10.1007/BF00890091.

[37] John Skilling. Nested sampling. *AIP Conference Proceedings*, 735(1):395–405, 2004. doi: 10.1063/1.1835238.

[38] F. Feroz and M. P. Hobson. Multimodal nested sampling: an efficient and robust alternative to markov chain monte carlo methods for astronomical data analyses. *Monthly Notices of the Royal Astronomical Society*, 384(2):449–463, Jan 2008. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2007.12353.x.

[39] F. Feroz, M. P. Hobson, and M. Bridges. Multinest: an efficient and robust bayesian inference tool for cosmology and particle physics. *Monthly Notices of the Royal Astronomical Society*, 398(4):1601–1614, Oct 2009. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2009.14548.x.

[40] Farhan Feroz, Michael P. Hobson, Ewan Cameron, and Anthony N. Pettitt. Importance nested sampling and the multinest algorithm. *The Open Journal of Astrophysics*, 2(1), Nov 2019. ISSN 2565-6120. doi: 10.21105/astro.1306.2144.

[41] W. J. Handley, M. P. Hobson, and A. N. Lasenby. PolyChord: nested sampling for cosmology. *Mon. Not. Roy. Astron. Soc.*, 450(1):L61–L65, 2015. doi: 10.1093/mnrasl/slv047.

[42] W. J. Handley, M. P. Hobson, and A. N. Lasenby. polychord: next-generation nested sampling. *Mon. Not. Roy. Astron. Soc.*, 453(4):4385–4399, 2015. doi: 10.1093/mnras/stv1911.

[43] Edward Higson, Will Handley, Michael Hobson, and Anthony Lasenby. Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation. *Statistics and Computing*, 29(5):891–913, dec 2018. doi: 10.1007/s11222-018-9844-0. URL https://doi.org/10.1007%2Fs11222-018-9844-0.

[44] Joshua S Speagle. dynesty: a dynamic nested sampling package for estimating bayesian posteriors and evidences. *Monthly Notices of the Royal Astronomical Society*, 493(3):3132–3158, feb 2020. doi: 10.1093/mnras/staa278. URL https://doi.org/10.1093%2Fmnras%2Fstaa278.

[45] Greg Ashton et al. Nested sampling for physical scientists. *Nature*, 2, 2022. doi: 10.1038/s43586-022-00121-x.

[46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[47] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

[48] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020. doi: 10.1038/s41586-020-2649-2. URL `https://doi.org/10.1038/s41586-020-2649-2`.

[49] Ewan Cameron and Anthony Pettitt. Recursive pathways to marginal likelihood estimation with prior-sensitivity analysis. *Statistical Science*, 29(3):397–419, 2014. ISSN 08834237, 21688745. URL `http://www.jstor.org/stable/43288518`.

[50] L Susan Blackford, Antoine Petitet, Roldan Pozo, Karin Remington, R Clint Whaley, James Demmel, Jack Dongarra, Iain Duff, Sven Hammarling, Greg Henry, et al. An updated set of basic linear algebra subprograms (blas). *ACM Transactions on Mathematical Software*, 28(2): 135–151, 2002.

[51] Sivaram Ambikasaran, Daniel Foreman-Mackey, Leslie Greengard, David W. Hogg, and Michael O'Neil. Fast Direct Methods for Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38:252, June 2015. doi: 10.1109/TPAMI.2015.2448083.

[52] Ciyou Zhu, Richard H. Byrd, Peihuang Lu, and Jorge Nocedal. Algorithm 778: L-bfgs-b: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, dec 1997. ISSN 0098-3500. doi: 10.1145/279232.279236.

[53] N. Aghanim et al. Planck 2018 results. V. CMB power spectra and likelihoods. 2019.

[54] N. Aghanim et al. Planck 2018 results. VIII. Gravitational lensing. 2018.

[55] Y. Akrami et al. Planck 2018 results. X. Constraints on inflation. *Astron. Astrophys.*, 641:A10, 2020. doi: 10.1051/0004-6361/201833887.

[56] N. Aghanim et al. Planck 2018 results. VI. Cosmological parameters. *Astron. Astrophys.*, 641: A6, 2020. doi: 10.1051/0004-6361/201833910. [Erratum: Astron.Astrophys. 652, C4 (2021)].

Paper III

Accelerating LISA inference with Gaussian processes

Accelerating LISA inference with Gaussian processes

Jonas El Gammal [1,2,*] Riccardo Buscicchio [3,4,5] Germano Nardini [1] and Jesús Torrado [6,7,8]

[1] *Department of Mathematics and Physics, University of Stavanger, NO-4036 Stavanger, Norway*
[2] *Como Lake Center for Astrophysics, Department of Science and High Technology, University of Insubria, via Valleggio 11, I-22100, Como, Italy*
[3] *Dipartimento di Fisica "G. Occhialini", Università degli Studi di Milano-Bicocca, Piazza della Scienza 3, 20126 Milano, Italy*
[4] *INFN, Sezione di Milano-Bicocca, Piazza della Scienza 3, 20126 Milano, Italy*
[5] *Institute for Gravitational Wave Astronomy & School of Physics and Astronomy, University of Birmingham, Birmingham, B15 2TT, UK*
[6] *Dipartimento di Fisica e Astronomia "G. Galilei", Università degli Studi di Padova, via Marzolo 8, I–35131 Padova, Italy*
[7] *INFN, Sezione di Padova, via Marzolo 8, I–35131 Padova, Italy*
[8] *Instituto de Estructura de la Materia, CSIC, Serrano 121, 28006 Madrid, Spain*

(Dated: March 31, 2025)

Source inference for deterministic gravitational waves is a computationally demanding task in LISA. In a novel approach, we investigate the capability of Gaussian Processes to learn the posterior surface in order to reconstruct individual signal posteriors. We use `GPry`, which automates this reconstruction through active learning, using a very small number of likelihood evaluations, without the need for pretraining. We benchmark `GPry` against the cutting-edge nested sampler `nessai`, by injecting individually three signals on LISA noisy data simulated with `Balrog`: a white dwarf binary (DWD), a stellar-mass black hole binary (stBHB), and a super-massive black hole binary (SMBHB). We find that `GPry` needs $\mathcal{O}(10^{-2})$ fewer likelihood evaluations to achieve an inference accuracy comparable to `nessai`, with Jensen-Shannon divergence $D_{\rm JS} \lesssim 0.01$ for the DWD, and $D_{\rm JS} \lesssim 0.05$ for the SMBHB. Lower accuracy is found for the less Gaussian posterior of the stBHB: $D_{\rm JS} \lesssim 0.2$. Despite the overhead costs of `GPry`, we obtain a speed-up of $\mathcal{O}(10^2)$ for the slowest cases of stBHB and SMBHB. In conclusion, active-learning Gaussian process frameworks show great potential for rapid LISA parameter inference, especially for costly likelihoods, enabling suppression of computational costs without the trade-off of approximations in the calculations.

I. INTRODUCTION

In the last decade, the direct detection of gravitational waves (GWs) has transformed from a remarkable, singular accomplishment into a routine procedure. Currently, the LIGO-Virgo-KAGRA collaboration has observed approximately a hundred systems emitting GWs in the 10 – 1000 Hz frequency range [1]. Moreover, pulsar timing array experiments are possibly on the verge of gathering enough statistics to announce the first direct detection of GWs in the nHz range [2–5]. Gravitational waves in the mHz frequency range remain unobserved. LISA, with construction commissioned now and launch scheduled in a decade, is set to delve into this uncharted territory [6].

LISA poses data analysis challenges that are radically different from those of the other GW experiments, as it is a signal-dominated one, expected to observe a multitude of Galactic binaries, supermassive BHBs, EMRIs, and stellar-mass BHBs constantly populating the datastreams [6]. A primordial stochastic gravitational wave background (SGWB) as loud as the astrophysical sources might also be present [7]. To further complicate things, the zoology of LISA signals does not admit a common detection and reconstruction strategy: within the experiment's lifetime, some sources are monochromatic, others slowly drift, and others move fast outside the LISA frequency sensitivity. Analyzing the data in time chunks makes the identification of the long-duration sources more difficult, while keeping the whole datastream a priori makes the likelihood evaluations too heavy. On the other hand, splitting the data into frequency intervals is suitable for monochromatic sources [8], less so for those that are largely chirping. Despite their difficulty, these challenges must be solved to achieve the groundbreaking science promised by LISA [7, 9, 10].

Concerning the likelihood evaluation cost, several improvements are conceivable (see e.g., [11] for a review): speeding up waveform evaluation (e.g., through hardware acceleration or approximation schemes) [12–14], bypassing the likelihood evaluation (e.g., using simulation-based inference methods (SBI) [15–22]), or building surrogates of the likelihood function itself [23, 24]. In this paper, we focus on the latter, while agnostically retaining the waveform content and the signal Bayesian model intact. Thus, no approximation is made to either the Fourier transforms of the modeled signals or the likelihood computation.

Instead, we adopt the machine learning framework implemented in `GPry` [25, 26]. Within it, we interpolate the posterior with a Gaussian process [27], trained on a small number of evaluations that are sequentially proposed in an optimal way to minimize their number [28, 29]. As we will see, this approach can produce accurate inference with $\mathcal{O}(10^{-2})$ fewer likelihood evaluations than traditional Monte Carlo approaches. This translates into a speed-up of inference by a factor of 100 in the regime in

* jonas.e.elgammal@uis.no

which the overhead of `GPry` is subdominant, i.e., when the likelihood evaluation time is over a few seconds, and the dimensionality of the problem is $\mathcal{O}(10)$. The output is a surrogate model for the posterior that can be sampled with a Monte Carlo algorithm at virtually zero cost.

Within the general context of machine-learning accelerated inference of GW sources, the likelihood-based, *active-learning* approach taken by `GPry` differs from the likelihood-free, *amortized* approaches such as SBI in a number of ways: a) amortized approaches are much faster at the point of inference, in exchange for some costly pre-training, whereas the more expensive active-learning frameworks can be run with no upfront costs for variations of data or waveform modelling; b) likelihood-based approaches do not necessitate the simulated data to contain stochastic noise, and possess a direct way to evaluate goodness-of-fit. Both approaches are complementary and can thus coexist in the LISA parameter inference pipeline. To estimate the benefits for LISA of a machine learning framework similar to the one just described, we use `GPry` to perform parameter inference on some benchmark LISA signals simulated through `Balrog`, and compare the speed and accuracy of the results to those obtained with the state-of-the-art nested sampler `nessai`.

The paper is organized as follows: in Sec. II A we describe the target sources of our study: a supermassive black-hole binary (SMBHB), a stellar mass black hole binary (stBHB) and a Galactic double white dwarf (DWD) system; in Sec. II B we compare the waveform modeling available in literature; in Sec. II C we briefly describe the inference scheme, for individual source parameter estimation in the three source scenarios previously mentioned; in Sec. III A we present previous approaches for exact or approximate inference, and how they can be used as a starting point for our pipeline; in Sec. III B we briefly introduce how our algorithm models a posterior using a Gaussian process interpolator; in Sec. III C we detail how we perform, evaluate and compare our inference runs; in Sec. IV we present our results for the three source types above, and compare `GPry` and `nessai` on performance and accuracy; in Sec. V we draw conclusions and outline possible future developments.

II. SOURCES

A. Source types

We explore three different source classes, roughly categorized by their signal spectral content. Double white dwarfs (DWDs) are observed by LISA during their early inspiral, emitting quasi-monochromatic GWs largely detectable within the Galactic neighborhood [30–32]. Each signal persists in the LISA datastream for the entire mission, Doppler modulated by the satellite-constellation orbital motion within a very narrow frequency band ($\Delta f/f \leq 10^{-4}$). For DWDs emitting above approximately 2 mHz, the GW-driven orbital tightening reaches a frequency evolution $\dot{f} \gtrsim 10^{-15}\,\mathrm{Hz}^2$ which LISA can measure over the nominal mission duration $T_{\mathrm{LISA}} = 4\,\mathrm{yr}$. As many as 10^7 DWD sources are expected to emit in the LISA band, with up to 1% individually detectable. They are unambiguously the most numerous deterministic sources expected for LISA, and their collective brightness makes its datastream strongly signal dominated below a few mHz. The brightest of these sources are identifiable after a few months [33], once a sufficient phase coherence emerges from the noisy datastream. In most of the available literature, a phenomenological parametrization of their signal is preferred over a physically-motivated one. Waveform models accurately taking into account the LISA response [34, 35] are typically fast, often leveraging frequency domain representation and heterodyning. It is uncertain whether such advantages will be retained in more realistic data analysis setups (see, e.g., [36], for recent developments on a frequency domain treatment of gaps).

stBHBs are the second class of sources we consider. They are expected to populate the whole LISA spectrum, the largest majority slowly drifting in frequency within the LISA mission duration. Only a handful of them will exit the band on its upper end in less than a year and eventually merge in the ground-based detector frequency band (10 Hz to 1 kHz) [37, 38]. Their waveforms are comparatively more complex than the DWD ones, with a parameter space equipped to describe eccentric, precessing, unequal-mass binaries [39, 40]. The in-band persistence of DWD and stBHB signals allows for a coherent integration of the data over millions of cycles during the nominal mission duration, making their detection heavily phase dominated.

Finally, we consider SMBHBs as the third category of GW sources: they are the most massive binaries expected to emit GWs in the LISA band. In the lifetime of the mission, SMBHBs will be detected as transient signals reaching signal-to-noise ratios (SNRs) as large as $\sim 10^3$, therefore being the loudest individual sources among the LISA ones. SMBHBs rapid evolution towards merger in band makes them prototypical to excite GW higher multipole modes, spin-precession [41]. However, orbit circularization [42] prevents from measuring large orbital eccentricities. State-of-the-art waveform models are phenomenological ones, calibrated against numerical relativity simulations with mass-ratios up to 1:18 [43]. Their computational efficiency is granted by decades of waveform developments for ground-based detectors, though the signal brightness questions the level of accuracy required to achieve unbiased parameter estimation [44]. The broadband nature of SMBHBs waveforms makes them the most expensive to compute in frequency domain (only second to extreme mass ratio inspirals), with up to 10^5 datapoints required for the lightest, most distant sources merging at about 10 mHz. Even though time-domain truncation may reduce the number of frequencies to evaluate the waveform at, advanced global inference schemes (e.g. Gibbs-like sampling or SBI tech-

niques) may require the usage of conditional data with full-resolution frequency series.

B. Signal model

We model LISA data d as the linear superposition of noise n and signal s. Observations are collected through time-delay-interferometric variables, synthetic time series constructed from suitable delayed combinations of single-link inter-spacecraft laser phase measurements [45]. For simplicity, we assume the three LISA satellites orbiting in an equilateral triangular configuration with constant armlength of 2.5×10^9 m. Under such an approximation, the three interferometric variables, often referred to in literature as X, Y, Z, are linearly combined into the A, E, T variables, such that the respective noises are uncorrelated.

We model the GW strain emitted from a distant DWD as a quasi-monochromatic signal. Its two polarizations are described by

$$h_+(t;\boldsymbol{\theta}) = A(1+\cos^2\iota)\cos(2\pi f_{\rm GW}(t)/{\rm Hz} - \phi)\,, \quad (1)$$

$$h_\times(t;\boldsymbol{\theta}) = -2A(\cos\iota)\sin(2\pi f_{\rm GW}(t)/{\rm Hz} - \phi)\,, \quad (2)$$

where A denotes the GW amplitude, ι represents the source inclination with respect to the line-of-sight, $f_{\rm GW}(t)$ is the instantaneous GW frequency measured in the solar system barycenter frame, and ϕ is the binary orbital phase at the time t_0 at which LISA observations start. The amplitude A can be expressed as

$$A = \frac{2(G\mathcal{M}_c)^{5/3}}{c^4 d_L}(\pi f)^{2/3}, \quad (3)$$

while, to leading order, $f_{\rm GW}(t)$ reads

$$f_{\rm GW}(t) = f + \dot{f}(t - t_0)\,, \quad (4)$$

with f and $\dot{f}$ being the orbital frequency and its (solar system barycenter frame) time derivative at time t_0, respectively. In Eq. (3), d_L denotes the source luminosity distance whose redshift is z, and $\mathcal{M}_c$ denotes the chirp mass

$$\mathcal{M}_c = \frac{(m_1 m_2)^{3/5}}{(m_1+m_2)^{1/5}}\,, \quad (5)$$

for a binary system of two component masses m_1 and m_2. Eq. (5) is frame-invariant, however we consider only solar system barycenter frame quantities hereafter.

The LISA detector response introduces an additional dependence upon the source position in the sky. We parametrize it by the source Ecliptic latitude b and longitude λ, and an overall polarization angle ψ. For inference purposes, we also reparameterize ϕ, ψ with two circular initial phases $\phi_L = \phi + \psi$ and $\phi_R = \phi - \psi$, respectively.

We assume the DWD source orbital evolution to be GW driven when injecting it into LISA data. Therefore, the injected f and $\dot{f}$ must satisfy the constraint

$$\dot{f} = \frac{96}{5}\frac{(G\mathcal{M}_c)^{5/3}}{\pi c^5}(\pi f)^{11/3}\,. \quad (6)$$

However, we infer f and $\dot{f}$ as free independent parameters. Due to the signal being narrowband and at a much lower frequency than the LISA data sampling rate ($f_s = 0.2$ Hz), we speed up likelihood evaluations through heterodyning, filtering, and downsampling, resulting in a few hundred of datapoints per waveform evaluation.

Concerning stBHBs, we model their GW signal by following [46] where the waveform is computed through an adiabatic inspiral post-Newtonian expansion. As stBHBs drift much faster than DWDs in the LISA frequency band, their waveform exhibits (mild) sensitivity to the component masses m_1, m_2 and dimensionless spins χ_1, χ_2. For simplicity, we consider aligned-spin systems in circular orbits only, leaving the investigation of eccentric, precessing ones for future work. We reduce inference correlations with a convenient physical parameterization through the binary chirp mass $\mathcal{M}_c$, reduced mass ratio $\delta\mu = (m_1 - m_2)/(m_1 + m_2)$, and component dimensionless spin magnitudes $\chi_{1,2}$; its initial orbital frequency f_0, and left- and right-handed phases ϕ_L, ϕ_R. The extrinsic parameters are decomposed as follows: the source position and inclination are parameterized by the square root of two circular amplitudes $A_{L,R} = (1 \pm \cos\iota)/\sqrt{2 d_L}$, the sin-ecliptic latitude $\sin\beta$, and longitude λ. TDIs are constructed through a rigid adiabatic approximation [47]. In previous work [46], waveforms were evaluated only at a few hundreds of points, employing Clenshaw-Curtis quadrature to approximate the likelihood in Eq. (7). This was made possible by analyses of noiseless data, whose smoothness allows for such an integration scheme. In turn, in this work we focus on noisy data, and hence we use the full GW frequency content, resulting in around 10^4 data points per waveform.

Finally, SMBHBs signals are described through phenomenological, numerical-relativity calibrated waveforms, as implemented in `IMRPhenomXHM` [48]. This waveform family smoothly captures the inspiral-merger-ringdown structure of a binary merger signal in frequency domain, accounting for higher-modes emission. Despite being extremely fast, thanks to decades-long optimization for current and future ground-based detectors [49], the LISA frequency resolution makes the waveform array typically long: in this study we consider a system emitting up to 3.5 mHz, reaching its merger $\tau_m = 4.138 \times 10^6$s after the start of the mission. We do not consider any time-domain truncation scheme, and model the signal at the highest frequency resolution available. The signal is parameterized by the binary chirp mass, its reduced mass ratio, the component dimensionless spin magnitudes (assumed aligned with respect to the angular momentum), the time-to-merger τ_m, the luminosity distance d_L, the sine-ecliptic latitude $\sin\beta$ and ecliptic longitude λ, the cosine inclination $\cos\iota$, the initial orbital phase $\phi_0^{\rm orb}$, and the polarization angle ψ. All quantities are defined in the

solar system barycenter frame, and non-conserved ones are defined at a reference frequency $f_{\rm ref} = 10^{-4}\,$Hz.

Throughout this work, we assume perfectly known, Gaussian, instrumental noise [50], superimposed on likewise perfectly known, Gaussian, confusion noise, whose level is modeled as in [51] as a function of $T_{\rm LISA}$. In the larger context of global fit pipelines, this is equivalent to performing inference on the three chosen sources after all resolvable ones have been identified and perfectly subtracted from the data. We further simplify the two noise models assuming both zero mean and perfectly stationary, thus reducing their entire description to simple power spectral densities [50].

C. Likelihood

Given the assumptions detailed in Sec. II B, the likelihood of observed data $d_k = A, E, T$ in frequency domain reads

$$\log \mathcal{L}(d|\boldsymbol{\theta}) = -\sum_k \frac{\langle d_k - s_k(\boldsymbol{\theta}) \mid d_k - s_k(\boldsymbol{\theta})\rangle_k}{2} + \text{const.} \quad (7)$$

where d_k denotes the superposition of noises realizations and each injected signal as described in Table I, II, and III, respectively. Thanks to the stationarity of noise in each datastream and uncorrelatedness across them, the inner product is simply given by

$$\langle x \mid y\rangle_k = 4\text{Re}\int_0^{+\infty} \mathrm{d}f \frac{\tilde{x}(f)\tilde{y}^\dagger(f)}{S_{n,k}(f)}\,. \quad (8)$$

Finally, $s_k(f;\boldsymbol{\theta})$ denotes a proposed GW signal with parameters $\boldsymbol{\theta}$, as observed in the k-th datastream, and $S_{n,k}$ the noise power spectral density in the same datastream. We characterize the overall source brightness with the SNR, defined as

$$\text{SNR}^2 = \sum_{k=A,E,T} \langle s_k(f;\boldsymbol{\theta}) \mid s_k(f;\boldsymbol{\theta})\rangle_k\,. \quad (9)$$

In this study, we present two approaches to obtain posterior samples for each inference, according to

$$p(\boldsymbol{\theta}|d) \propto \mathcal{L}(d|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\,, \quad (10)$$

where $\pi(\boldsymbol{\theta})$ denotes the prior assumption $\boldsymbol{\theta}$. In Sec. III A we detail the construction of priors for each source category, which we assume to be uniform over the prescribed ranges.

III. INFERENCE

A. Setting priors

Pre-constraining the parameter space of the inference problem down to a region around the location of the bulk probability mass, henceforth referred to as "mode", makes surrogate-posterior approaches such as `GPry` significantly faster and more robust. This can usually be achieved with methods that avoid the evaluation of the expensive posterior, via e.g. approximations in the likelihood, template matching with an approximate waveform or machine-learning forward modeling [12, 14, 52]. These methods can produce rough estimates of the location and span of the posterior mode at a very low computational cost.

The DWDs live in a narrow frequency band and can be initially constrained using frequentist triggers with a sliding-window method that scans the frequency domain. Additionally, by using an optimizer, it is possible to obtain a maximum likelihood estimate (MLE), and an estimate of the Fisher information matrix. In combination, these methods allow the setting of priors that sufficiently encapsulate the mode of the posterior distribution, as done in [8]. Since the DWD are mostly a test case for our study, we set conservative priors by hand, encapsulating $\sim 10\sigma$ for each unbounded parameter.

or stBHBs our approach to pre-constraining the parameter space is the one introduced in [53] and successfully applied to LISA data in [54]. This method employs a semi-coherent search combined with Particle Swarm Optimization (PSO) to efficiently scan the large parameter space involved. The semi-coherent approach divides the data into frequency-domain segments, analyzing each individually, and then combining the results. This technique balances sensitivity and computational efficiency by widening the posterior distribution over the parameter space, thus helping to locate the posterior bulk.

The path traced by the particles in the PSO can then be used to find regions in the parameter space with high posterior density values. For our stBHB analysis, we use a subset of 5000 samples from the PSO paths, obtained from 256 data segments, and evaluate the posterior in Eq. (10) at these locations. We then restrict the prior to the smallest hyper-rectangle containing PSO posterior samples within a 10σ confidence region from the peak, assuming a multivariate Gaussian distribution as the posterior distribution (for a detailed discussion, see App. A of [25]). In addition, we use a small set of these samples close to the top of the mode as an initial training set for `GPry`. Together, the shrunken prior and the initial training set eliminate the need to explore the parameter space and let `GPry` focus on mapping the mode, thus combining the strengths of both approaches: a fast initial exploration of the parameter space by the PSO followed by `GPry` which maps the mode with very few evaluations of the relatively slow-to-evaluate posterior distribution. The PSO search takes $\mathcal{O}(10\,\text{min})$, adding only very little overhead to our pipeline. We perform the initial PSO on noiseless data to introduce an additional bias beyond the one arising from their segmentation.

For the less explored case of SMBHBs, we assume that a similar PSO approach, a neural-network or a frequentist one can be used to approximate the mean and covariance

of the posterior mode (see, e.g., [52, 55]).

Due to the simpler structure of the posterior, we can set a larger uniform prior covering $> 10\sigma$ in each dimension. As a proxy for a search, we generate Monte Carlo samples of the noiseless posterior distribution and use its mean and covariance to draw a set of 35 samples from a multivariate Gaussian distribution. `GPry` is initialized with these samples which are close to the top of the mode but biased. If multimodalities are present, as is expected in the sky location for low latency searches [56–58], `GPry` would be initialized with points from all modes.

B. Gaussian process posterior interpolation

The inference algorithm employed in this study uses a Gaussian process regressor (GPR) to create an approximate model of the posterior density function, using a small set of evaluations performed at optimal locations. This approximation is then used as a *surrogate* model from which we can draw, at very low computational cost, Monte Carlo (MC) samples that very closely resemble samples from the true posterior. Contrary to amortized machine learning-based approaches such as SBI, our surrogate model is built sequentially at runtime (an approach known as *active learning*), and does not rely on previous training. `GPry`'s approach is more closely related to variational inference (see, e.g., [59]), with the difference that `GPry` does not need derivatives of the posterior. As we will see, the necessary number of evaluations of the GW signal likelihood is at least $\mathcal{O}(10^{-2})$ smaller than those needed by `nessai` (which is already more efficient than traditional Nested Sampling implementations).

We use `GPry` [25, 26, 60] to construct such a surrogate model. In this subsection, we adopt the notation most commonly used in the context of Gaussian processes where $\mathbf{x}$ refers to a vector in the sampling space (equivalent to θ above) and y is the value of the target function. `GPry` iteratively proposes points $\mathbf{x}$ in parameter space at locations where the expected gain in information about the posterior is maximized. With them, at every iteration, it builds an approximation of the posterior log-density function $\log p(\mathbf{x}|\mathcal{D})$ under some data $\mathcal{D}$ as the mean of a Gaussian process conditioned on the current set of *training* samples $(\mathbf{X}, \mathbf{y})$, where $\mathbf{X} = \{\mathbf{x}^{(i=1,\dots)}\}$ and $\mathbf{y} = \{\log p(\mathbf{x}^{(i=1,\dots)})\}$:

$$\log p(\mathbf{x}|\mathcal{D}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')|\mathbf{X}, \mathbf{y})\,. \tag{11}$$

Here $k(\mathbf{x}, \mathbf{x}')$ represents the covariance function, for which `GPry` uses a d-dimensional inverse-squared Radial Basis Function (RBF) kernel allowing for a different length-scale in each dimension of the sampled parameter space:

$$k(\mathbf{x}, \mathbf{x}') = C^2 \prod_{i=1}^{d} \exp\left(\frac{(x_i - x_i')^2}{2l_i^2}\right), \tag{12}$$

with C and l representing, respectively, the output and length scales of the Gaussian process. The null mean of the Gaussian process prior in Eq. (11) applies to a transformed set of $\mathbf{y}$ log-posterior values so that they have null mean and unit standard deviation. Hereon we drop the explicit dependence on the training data $\mathbf{X}, \mathbf{y}$.

The mean of the conditioned Gaussian process, with which we approximate the log-posterior density, is computed as

$$\mu(\mathbf{x}_*) = \mathbf{k}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{y}\,, \tag{13}$$

where $(\mathbf{k}_*)_i = k(\mathbf{x}^{(i)}, \mathbf{x}_*)$, $(\mathbf{K})_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$, and σ_n^2 is an estimate of the numerical uncertainty of log-posterior values. The standard deviation of the conditioned Gaussian process, used in the acquisition function defined below, is $\sigma(\mathbf{x}) = \sqrt{\mathrm{diag}(\Sigma(\mathbf{x}))}$, where

$$\Sigma(\mathbf{x}_*) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^T (\mathbf{K} + \sigma_n^2 \mathbf{I})^{-1} \mathbf{k}_*. \tag{14}$$

The hyperparameters of the kernel, i.e., its output and length scales, hereon denoted collectively as $\mathbf{\Lambda}$, are determined by maximizing their marginalized likelihood [27]. The mean and standard deviation of a Gaussian process conditioned on a set of training samples can be seen in the upper row of Fig. 1.

Optimizing the kernel hyperparameters $\mathbf{\Lambda}$ eventually dominates the overhead of the algorithm, as it requires multiple kernel matrix inversions that scale as $\mathcal{O}(N^3)$, with N being the number of training samples. In order to mitigate this, we only perform a full re-fit of the hyperparameters at every few iterations of the algorithm (see Appendix B). In general, overhead costs start making `GPry` an impractical approach for dimensionalities larger than a few tens, depending on the cost of the likelihood. The number of training samples, which drives the overhead costs, needed for accurate posterior reconstruction depends on the dimensionality of the problem. In exchange for this overhead, `GPry` reduces the number of posterior evaluations required with respect to traditional samplers by a factor of $\mathcal{O}(10^2)$. Therefore, `GPry`'s advantage in performance increases for low dimensions and large costs per posterior evaluation. An approximate rule of thumb is that `GPry` is faster for dimensionalities lower than a few tens when the posterior evaluation time is $\mathcal{O}(1\,\mathrm{s})$ or higher.

As a further refinement of the surrogate model, we multiply the GPR by a Support Vector Machine (SVM) classifier, of comparatively negligible computational cost, trained both on the evaluations used for the GPR, and those rejected because their log-posterior density is either negative infinity or very low with respect to the best training point. This SVM is used to partition the parameter space into regions in which the true log-posterior is expected to return a finite, or negative infinity value; the latter is used as an exclusion region where future candidates are automatically rejected without evaluating their true log-posterior.

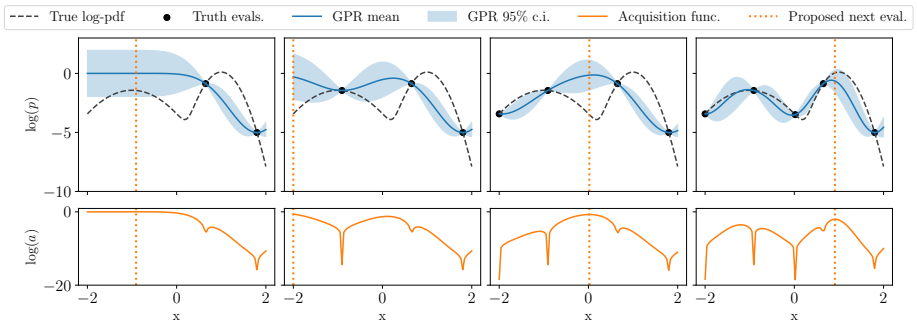


FIG. 1. Simplified illustration of the `GPry` algorithm on a 1-dimensional Gaussian mixture test function. Each column, corresponding to consecutive iterations, shows on the top the true target log-pdf (dashed), the current set of evaluations (black points), and the current GPR model mean from Eq. (13) (blue, solid) and 95% confidence interval defined by Eq. (14) (blue, shaded); the bottom panel shows the current acquisition function values from Eq. (15), whose maximum (dotted orange) will be proposed for evaluation for the next iteration. Not illustrated are more complex aspects of the algorithm, such as the batch proposal of points [25], and the procedure to obtain approximate maxima of the acquisition function [26].

To enable active sampling, we introduce an acquisition function, denoted as $a(\mathbf{x})$, which guides the sampling process by quantifying the expected utility of sampling the true posterior at each point in the parameter space:

$$a(\mathbf{x}) = \exp\left(2\zeta \cdot \mu\left(\mathbf{x}\right)\right)\left(\sigma(\mathbf{x}) - \sigma_n\right) , \qquad (15)$$

ζ is a scaling factor that balances exploration and exploitation. The learning efficiency is maximized when this scaling factor is made dimensionality-dependent, increasingly encouraging exploration for larger dimensionalities: $\zeta = d^{-c}$, with $c > 0$ [25]. In this paper, we empirically set this scaling factor to $\zeta = d^{-0.65}$, promoting exploitation slightly more than the value derived in [25] for Gaussian distributions. The effect of evaluating sequentially at the optimum of the acquisition function can be seen in Fig. 1.

The acquisition function is optimized through the NORA active sampling strategy described in [26]: we draw MC samples from the mean $\mu(\mathbf{x})$ of the GPR using a Nested Sampler (NS), in our study `PolyChord` [61, 62]. The acquisition function is then evaluated at the resulting NS samples, and its value is used to produce a pool of candidate points. This pool is ranked using the Kriging believer [63] prescription so that the n-th point is assigned a conditioned acquisition function value assuming a true posterior evaluation at the $n-1$ points above it. The optimal batch size is approximately equal to d [25], making the `GPry` algorithm efficiently parallelizable up to d processes using MPI.

The use of a NS at the acquisition step, that explores the full surrogate posterior (as opposed to directly maximizing the acquisition function), makes it easier for `GPry` to map a multimodal posterior, such as those expected in the sky localization parameters for low-latency signals, as demonstrated in [26]. This ability can be further boosted by making the acquisition function more exploratory (lower scaling factor ζ in Eq. (15)), and the exploration of the posterior more thorough (larger number of *live points* of the NS).

At the end of every iteration, convergence is checked and considered reached as soon as one of two criteria is fulfilled at least twice consecutively: the value of the likelihood at the new proposed sampling locations is close enough to their GPR-predicted value (see [25] for clarification), or the Gaussian-approximated Kullback-Leibler divergence[1] between consecutive NORA NS runs is small enough.

After convergence of the Bayesian optimization loop has been reached, MCMC samples of the surrogate model are generated. This typically only takes a few seconds since the evaluation of the surrogate model is very fast at $\mathcal{O}(10^{-5}\,\mathrm{s})$. To do so, we use `Cobaya`'s implementation of the MCMC sampler of `CosmoMC` [64, 65].[2] A flow chart of the algorithm is shown in Fig. 2.

[1] I.e., the Kullback-Leibler divergence, see Eq. (17), when distributions are approximated as multivariate Gaussians defined by their respective empirical means and covariance matrices.

[2] Notice that any other sampler, including those that require gradients, could be used without changing our conclusions.

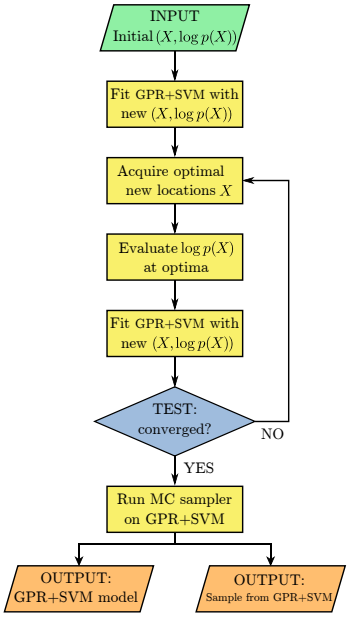


FIG. 2. Simplified flow chart of the `GPry` algorithm. Looking at Fig. 1, the GPR at the top of its first column presents the initial stage, where the GPR has been fit to an initial set of two samples. The main loop (*"acquire→evaluate→fit"*) corresponds sequentially to finding the location of the maximum of the acquisition function in the bottom row (dotted vertical line), evaluating the log-posterior there, and fitting the GPR to obtain the new model at the top of the following column.

C. Inference strategy and methodology for validating the results

`GPry` adopts default values for the parameters controlling the some of the aspects of the algorithm mentioned above, based on test runs on typical scenarios [25, 26]. The peculiarities of the problem at hand motivate changing some of these defaults in each case, as detailed in Table IV in Appendix B, and summarized below:

- For all three sources, especially for the stBHB and SMBHB, the log-likelihood presents significant numerical noise with respect to small changes in the waveform parameters. If not correctly accounted for, `GPry` interprets this sizable noise contribution as physically meaningful, which may lead to overfitting. We alleviate this problem by choosing large values of the expected noise scale σ_n in Eq. (14). Away from the mode, for very low likelihood values, the numerical noise dominates, so we raise the SVM classifier cutoff to exclude low-valued regions from the GPR.
- For the sources with the slowest likelihood, the stBHB and especially the SMBHB, it makes sense to increase the overhead of the algorithm in exchange for reducing the number of necessary true posterior evaluations for convergence. Hence, we increase the frequency and the number of restarts for the GPR hyperparameters optimization. Similarly, we update the set of NS samples from mean GPR more often, and, for the SMBHB, reduce the number of Kriging steps.
- Since the set of initial points for the stBHB and SMBHB is very informative (see Sec. III A), it is advantageous to define a *trust region* around the current training set restricting the area where new evaluations are proposed. This region is the minimal hyper-rectangle containing training samples with posterior density above some cutoff with respect to the best one.

For each source type, we consider a high-SNR source signal as a noiseless LISA data stream, then inject it in multiple simulated noise realizations The three easier noiseless inference problems are used for consistency checks (e.g., robustness with respect to initialization) and are not presented below.

In order to benchmark `GPry`'s performance, both in terms of computational cost and inference accuracy, we pair every `GPry` run in each noise realization with a similar run with the machine-learning-enhanced nested sampler `nessai`, which has proven to be an efficient and reliable sampler in the context of GW data analysis [66–68].

We perform two tests on the two sets of runs. The first focuses on the accuracy of the full pipeline, from signal and noise generation to MC sampling. In literature, this test is often referred to as a *pp*-plot [69, 70]. For a given sampler choice and source category, we perform N inference runs on independent noise realizations, and compute the empirical quantiles $\{q_i\}_i^N$ corresponding to the injected parameters for the inferred posterior. In the limit $N \to \infty$ the cumulative distribution function of quantiles across runs is theoretically expected to approach that of the uniform distribution over the unit interval. Deviations from the asymptotic distribution due to finite N can be estimated numerically, and confidence intervals constructed accordingly. We present results of this test across source categories in Figs. 3a, 4a and 5a, respectively.

In the second test, we focus instead on a direct comparison between posteriors obtained through inference with `nessai` and `GPry` in paired runs on the same noise realizations. To do so, we evaluate the Jensen-Shannon (JS) divergence D_{JS} between each `nessai` posterior distribution P and the `GPry` surrogate model P_{GP} over the

parameter space [71]

$$D_{\text{JS}}(P||P_{\text{GP}}) = \frac{1}{2}\left(D_{\text{KL}}(P||M) + D_{\text{KL}}(P_{\text{GP}}||M)\right) , \quad (16)$$

where $M = \frac{1}{2}(P + P_{\text{GP}})$ is the mixture distribution of P and P_{GP}. The Kullback-Leibler (KL) divergence between two continuous probability distributions P, M with densities $p(x)$, $m(x)$ is defined as

$$D_{\text{KL}}(P||M) = \int p(x) \log\left(\frac{p(x)}{m(x)}\right) \mathrm{d}x \ . \quad (17)$$

In practice, we compute the KL divergence as a Monte Carlo sum of the samples from `GPry` and `nessai`. In this paper, we use natural logarithms for the divergence calculations. The JS-divergence $D_{\text{JS}}(P||Q)$ approaches zero if and only if P and Q describe the same distribution and is upper-bounded by $\log 2$. For inference purposes, values of $D_{\text{JS}} \lesssim 0.05$ would make `GPry` as accurate as traditional samplers, whereas values up to $D_{\text{JS}} = 0.1$ could be considered precise enough, given the large computational trade-off. We show the distribution of D_{JS} for different source categories in Figs. 3b, 4b and 5b, respectively.

Following the formulas in Appendix A, for the dimensionality of our problems $D_{\text{JS}} = 0.05$ ($D_{\text{JS}} = 0.1$) would translate into a mean deviation in each parameter of $\approx 0.08\sigma$ ($\approx 0.11\sigma$) if assuming similar covariances, or alternatively a misestimation of the error of $\sim 15\%$ ($\sim 25\%$) if assuming similar means.

IV. RESULTS

A. Double white dwarf system

For a single injection of a DWD system, the waveform and subsequent likelihood computations are fast ($\sim 10^{-3}$ s). Hence, we do not expect significant savings in wall clock computation time between `GPry` and `nessai`. We therefore use it to test the `GPry` algorithm and gain some insight on its reliability.

Parameter	**Symbol**	**Value**
Ecliptic longitude	λ	2.0 rad
Ecliptic sine-latitude	$\sin\beta$	0.479
Amplitude	A	$2 \cdot 10^{-23}$
Frequency	f	0.00377 Hz
Frequency derivative	$\dot{f}$	2×10^{-18} Hz^2
Cosine-inclination	$\cos\iota$	0.4
Left phase	ϕ_L	1.3 rad
Right phase	ϕ_R	1.5 rad
SNR		23.64

TABLE I. Injected values for the sampled parameters of the DWD system and total source SNR.

The injected parameters for the benchmark source are shown in Table I. We draw 200 noise realizations according to our model in Sec. II B. For this source all parameters are constrained and the posterior distribution exhibits a single, localized nearly-Gaussian mode. For each noise realization, we perform separate inference runs with `GPry` and `nessai`, with the `nessai` runs performed with 2000 live points. We then generate a PP plot comparing the performance of both algorithms (see Fig. 3a), and find a similar accuracy for the reconstruction. Furthermore, we compute the JS divergence, D_{JS}, between `GPry` and `nessai` for each noise realization, and show its histogram in Fig. 3b. This comparison shows that both samplers are in excellent agreement for all but one noise realization. In Fig. 8 we show a corner plot overlaying the posterior contours obtained by `GPry` and `nessai` for the realization corresponding to the median D_{JS}. There we can observe the clear agreement between the two approaches, achieved with 1/300 fewer likelihood evaluations by `GPry` compared to `nessai`. The locations of the `GPry` evaluations can be seen in the upper triangle of Fig. 8.

In Fig. 9 we show a corner plot and the posterior contours obtained by `GPry` and `nessai` corresponding to the highest $D_{\text{JS}} = 0.12$. Although the mode has been found by `GPry` in this example, it remains underexplored. Tightening the convergence criterion would eliminate this problem in exchange for higher computational costs, but maintaining the two-orders-of-magnitude difference in the number of likelihood evaluations with respect to `nessai`. Only one of the 200 runs performed shows this behavior with $D_{\text{JS}} > 0.05$ which leads us to conclude that the precision and accuracy of `GPry` is sufficient in this context.

B. Stellar origin binary black holes

For one injection of a stBHB system, the cost of a single evaluation of the inference pipeline (waveform and likelihood calculations) is $\sim 10^{-1}$ s, which is significantly higher than for DWDs. Thus, the savings here are potentially higher, which makes `GPry` a worthy approach.

The benchmark source's injected parameters are shown in Table II. Unfortunately, as the semi-coherent search presented in Sec. III A does not provide us with a reliable estimate of the phases, sampling these proves to be difficult with `GPry`. This is further complicated by the periodic nature of these parameters. We therefore fix the values to the injected ones. Contrary to the DWD case, the resulting posterior is highly non-Gaussian: it is heavy-tailed and has a large curving degeneracy. As discussed in [25], exploring the full posterior in a reasonable amount of time poses a challenge to `GPry`.

We generate 100 noise realizations and perform inference runs for each of them with `GPry` and `nessai`, with the `nessai` runs performed with 2000 live points. We then generate both a PP plot (see Fig. 4a), and compute the JS divergence for each pair of runs, whose histogram is shown in Fig. 4b. From the PP plot, it is clear that `GPry` performs worse than `nessai`, even if both samplers

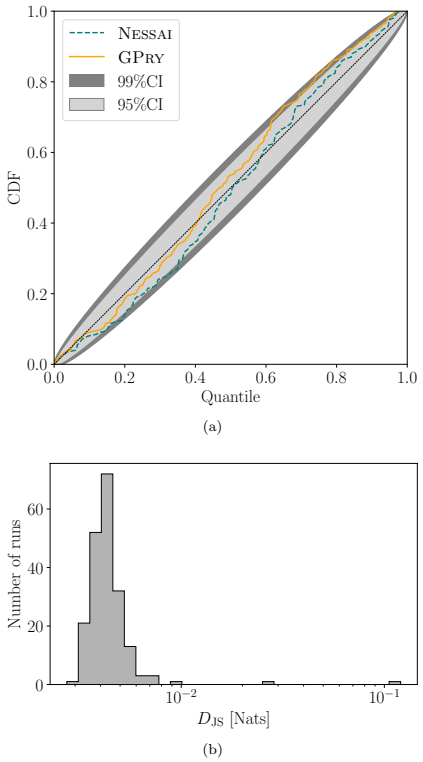


FIG. 3. PP plot **(a)** and Jensen-Shannon divergence **(b)** for 200 DWD runs with different noise realizations. `nessai` and `GPry` show comparable accuracy in the former, consistent at 99% confidence (dark gray shaded area) with the theoretical prediction (dotted black line) across all runs, and at 95% confidence (light gray shaded area) for the largest majority of them. Relatively to `nessai`, `GPry` reconstructs the posterior shape reliably with only one run exceeding the target of $D_{\mathrm{JS}} = 0.05$.

show reasonably good performance. This is reflected in the higher JS divergences between `nessai` and `GPry` (see Fig. 4b), localized mostly in the $[0.1, 0.25]$ interval, with a few $D_{\mathrm{JS}} \gtrsim 0.3$ outliers.

The effect of the $D_{\mathrm{JS}} \sim 0.2$ divergence is illustrated in Fig. 10, which shows the result of the median D_{JS} run with `GPry` and `nessai`. As we can see, although the resulting mode for `GPry` is localized correctly towards the injected value, it fails to explore a fraction of the posterior corresponding to the large-values tail of the $(\mathcal{M}_c, \delta\mu)$ degeneracy. The handful of cases with higher D_{JS} (up to 0.6) present the same sort of effect, and small ($< 1\sigma$) biases for other parameters.

Parameter	**Symbol**	**Value**
Redshifted chirp mass	$\mathcal{M}_c$	$48.618\,\mathrm{M}_\odot$
Reduced mass-ratio	$\delta\mu$	0.5
Ecliptic longitude	λ	0.19 rad
Ecliptic sine-latitude	$\sin\beta$	0.82 rad
Initial orbital frequency	f_0	1.87 mHz
Left phase (fixed)	ϕ_L	0.97 rad
Right phase (fixed)	ϕ_R	1.76 rad
Left square-root amplitude	$\sqrt{A_L}$	$12.57 \cdot 10^{-5}$
Right square-root amplitude	$\sqrt{A_R}$	$1.13 \cdot 10^{-5}$
Dimensionless spin	χ_1	0.223
Dimensionless spin	χ_2	0.262
SNR		16.79

TABLE II. Injected values for the parameters of the stBHB system, and total source SNR. All parameters are sampled except for the phases, for which the method described in Sec. III A failed to provide reliable estimates. The detector-frame individual masses are $m_1 = 99.55\,\mathrm{M}_\odot$ and $m_2 = 33.18\,\mathrm{M}_\odot$.

Possible mitigation strategies include fine-tuning of the `GPry` hyperparameters, to increase the chance that it fits this particular problem better (e.g. that it does not converge prematurely), as well as the use of alternative parameterizations whose posterior would not present these strong non-Gaussian features. We leave this endeavor for future work. It must be remarked that this difficulty also affects `nessai`, whose precision we had to increase in order to map this posterior correctly, so that $\mathcal{O}(10^2)$ more evaluations are needed than for the other two test sources (see Fig. 6a).

As explained in Sec. III A, we are seeding the stBHB runs discussed in this section with high-likelihood points from a noiseless Semi-Coherent PSO run. Since a noise realization introduces a bias in the inferred source parameters with respect to the noiseless case, we have investigated whether this under performance may be related to the use of a biased initial set of samples in the `GPry` runs. To do this, we have performed 100 additional paired runs with a noiseless injection, but found the same under-exploration effect with a similar magnitude.

We find that our approach reliably recovers the expected central values, and therefore could be used for source subtraction or fast, preliminary analysis; it is however suboptimal for full statistical source characterization. There is ongoing development of `GPry` addressing more robust inference in highly non-Gaussian distributions such as this stBHB posterior.

C. Supermassive binary black-hole

We now focus on the injection of a single SMBHB in noisy LISA data. Here, the cost per evaluation of the inference pipeline is larger than a few seconds and there-

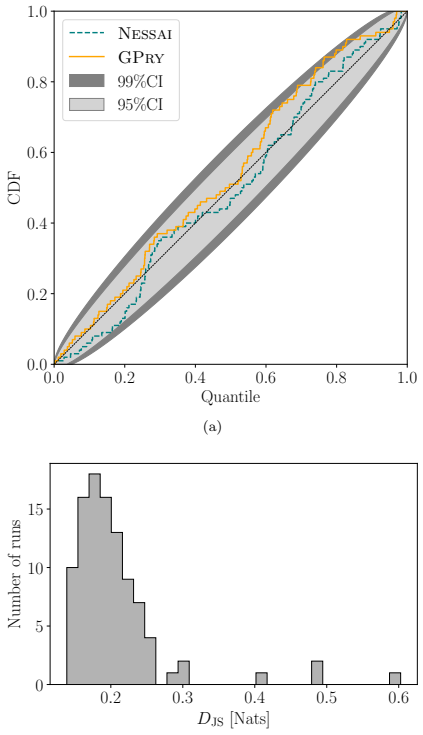


FIG. 4. PP plot **(a)** and Jensen-Shannon divergence **(b)** for 100 stBHB runs with different noise realizations. While `nessai` and `GPry` show comparable accuracy in the former, consistent at 99% CL (dark shaded region) with the theoretical prediction (black dotted line), the distribution of D_{JS} that is entirely above the target value of 0.05 indicates insufficient characterization of the posterior mode. Indeed, Fig. 10 shows that `GPry` underestimates the tails, especially in the $\mathcal{M}_c, \delta\mu$ direction which leads to the large discrepancy.

.

fore `GPry` shows great potential: it could turn days- or weeks-long inference runs with `nessai` into hours-long ones.

The injected parameters for the benchmark source are shown in Table III. For this high signal-to-noise case, the posterior is nearly Gaussian.

In this case, we generate 100 noise realizations (of which one was discarded due to an HPC error) and perform inference runs with `GPry` and `nessai`. The `nessai` runs were performed with 500 live points, instead of 2000

Parameter	**Symbol**	**Value**
Redshifted chirp mass	$\mathcal{M}_c$	$6.5744 \times 10^6\,\mathrm{M}_\odot$
Reduced mass-ratio	$\delta\mu$	0.12864
Luminosity distance	d_L	18.7 Gpc
Ecliptic longitude	λ	2.15 rad
Ecliptic sine-latitude	$\sin\beta$	-0.34 rad
cosine-inclination	$\cos\iota$	0.86 rad
Orbital phase	ϕ_0^{orb}	5.86 rad
Polarization	ψ	-0.136 rad
Time to merger	τ_m	4.138×10^6 s (47.89 days)
Dimensionless spin	χ_1	0.9874
Dimensionless spin	χ_2	0.9876
SNR		1944.8

TABLE III. Injected values for the sampled parameters of the SMBHB system and total source SNR. The detector-frame individual masses are $m_1 = 8.61 \times 10^6\,\mathrm{M}_\odot$ and $m_2 = 6.65 \times 10^6\,\mathrm{M}_\odot$. The reference frequency is $f_{\mathrm{ref}} = 10^{-4}$ Hz.

as for the other sources, for reasons of limited computational capacity. The resulting PP plot can be seen in Fig. 5a, and the relative JS divergence for each pair of runs in Fig. 5b. Both `GPry` and `nessai` show very good performance in the PP plot, and agree very well, with a median $D_{\mathrm{JS}} \approx 0.05$ and no run with $D_{\mathrm{JS}} \geq 0.1$. The runs with the median and highest D_{JS} are shown in Figs. 11 and 12, respectively. Therein we contrast the respective `GPry` runs with two higher-resolution (2000 live points) `nessai` runs performed to show finer contours for comparison.

For the SMBHB runs `GPry` needs $n < 10^3$ evaluations, which amounts to $\approx 30\%$ of the total computation time when the learning overhead is accounted for. In contrast, `nessai`, despite being run with a significantly low resolution for reasons of time, performs $n \sim 10^5$ evaluations.

D. Number of posterior evaluations and speedup

`GPry`'s main advantage compared to more traditional samplers is a drastic reduction in the number of posterior evaluations needed for inference, as shown in Fig. 6a. It is clear that `GPry` consistently performs $\mathcal{O}(10^2) - \mathcal{O}(10^3)$ fewer evaluations than `nessai` to converge to the posterior mode. This, however, comes at the price of the relatively large amount of time required for the acquisition of new optimal sampling locations and fitting the GPR hyperparameters. The size of this overhead depends mainly on the dimensionality of the sampling space and, to a lesser extent, on the Gaussianity of the posterior. The dimensionality scaling of the overhead can be clearly observed in Fig. 7 in Appendix B. Ultimately, the potential speedup with respect to an alternative sampler depends on a combination of a slow-enough posterior and a reasonable dimensionality.

In Fig. 6b we show a comparison between the distribution of wall-clock times of the `GPry` runs in this paper, and an optimistic (assuming no overhead) estimate

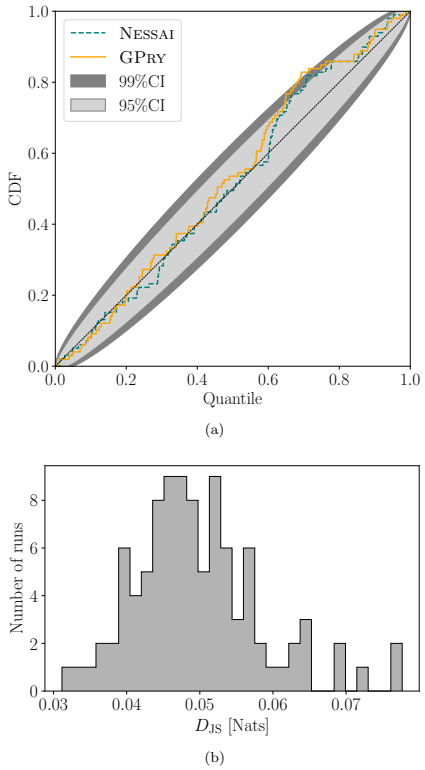


FIG. 5. PP plot **(a)** and Jensen-Shannon divergence **(b)** for 99 SMBHB noisy runs. The former shows that both `GPry` and `nessai` show comparable accuracy, consistent at 99% confidence (dark gray shaded area) with the theoretical prediction (black dotted line). The distribution of D_{JS} clusters around our target value of 0.05. This might partially be caused by `nessai` running at low resolution, but also by `GPry` occasionally under-exploring the tails of the posterior.

for the paired `nessai` runs. The quoted clock times are obtained by multiplying the number of their likelihood evaluations by their evaluation on the same hardware as for the `GPry` runs.[3] As we can see there, in the case of the DWD source `GPry` does not outperform `nessai`, needing roughly twice the time on average, due to the very short computation time of the DWD likelihood. However, for the stBHBs and SMBHBs, where likelihood computations are more expensive, the speed up is highly significant, reducing the time for inference from $\sim 10^6$ core seconds (around 11 days) to $\sim 10^4$ core seconds (around 3 hours). Of course, both of these numbers can be reduced through parallel processing but the advantage would still be evident.

When taking the reliability and accuracy of the inference into account, it is clear that the biggest potential for speed up is currently in the inference of SMBHBs.

[3] The `nessai` runs needed to be performed on a different platform due to limitations in our computing budget.

V. CONCLUSIONS

We demonstrated that active sampling methods with Gaussian processes have the ability to produce accurate inference on individual injections of three different GW sources expected in the LISA band, DWDs, stBHBs and SMBHBs, employing $\mathcal{O}(10^{-2})$ fewer evaluations of the GW signal likelihood than a state-of-the-art nested sampler, and with a significant speedup, going up to a $\mathcal{O}(10^{-2})$ wall-clock time reduction for likelihood evaluation times approaching $\mathcal{O}(1\,\mathrm{s})$ and above. They do so with some, but little preconditioning, that can be provided by frequentist searches or other faster but less accurate approximate inference schemes. Crucially, no expensive pretraining is required with these methods as would be in amortized approaches.

Using `GPry` as an active learning framework, we found the advantages with respect to traditional Monte Carlo samplers to be problem-dependent:

- Inference for DWDs can be provided quickly and robustly, but the fast-to-evaluate DWD waveforms mean that the overhead of acquiring samples and fitting the Gaussian process outweigh the time saved by reducing the number of sampling steps. This in turn means that we report no savings in terms of wall-clock time. However, in the presence of gaps in the data, as expected in LISA, the computational cost of the likelihood will go up. In this case our approach could be competitive.

- Inferring the parameters for stBHBs was possible with considerable time savings of 2 orders of magnitude. However, this comes at the cost of underestimating the tails of the distribution, though we retain the ability to reliably recover the central values. There is ongoing development of `GPry` aimed at addressing this shortcoming.

- The best combination of speed-up and accuracy was achieved for the SMBHBs, whose likelihood is very slow to evaluate at $\mathcal{O}(1\,\mathrm{s})$. We report a speed-up of two orders of magnitude compared to nested sampling while retaining a comparable accuracy. This reduces the computational cost of the inference from $\sim 10^6$ core-seconds (~ 11 core-days) to merely $\sim 10^4$ core-seconds (~ 3 core-hours).

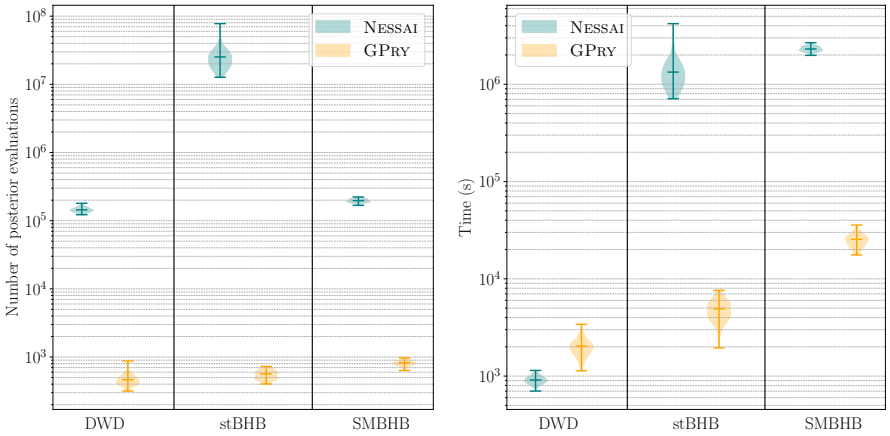


FIG. 6. Violin plots comparing `GPry` (orange) and `nessai` (teal) on each of the three test sources on noisy LISA data, according to **(a)** the total number of posterior evaluations needed, and **(b)** the hypothetical wall-clock time required for inference in a single-core setup (see text for a precise definition of these time estimates). The violins show the distribution the number of posterior samples and times, the minimum, median, and maximum values are marked with horizontal bars.

The integration of `GPry` (or a different active learning approach) into the LISA Global Fit pipeline would allow the characterization of expensive-likelihood signals (such as ones with strong time dependence) with low latency, by spawning it on their conditional likelihoods at any point. Even in cases in which `GPry` would be outperformed at inference time by amortized approaches (such as simulation-based inference), or if the calculation of waveforms or likelihoods are significantly accelerated, `GPry` opens the door to explore new physics (e.g., modified GR at emission or propagation) or characterize exotic signals, for neither of which a pre-trained emulator may be available or cost-effective.

`GPry` can also be a very powerful tool for prototyping waveforms, theoretical models, and the inference pipeline with mock data. It requires no pretraining, and no noise to be present in the data and accounted for in the likelihood. This enables quick forecasting and testing without the need for dedicated computing infrastructure to be in place.

The resulting surrogate posterior can be stored as kB-sized object, a size much smaller than the data necessary to reproduce the inference problem, and can be upsampled at very low computational cost. As an analytic function, it can be easily used as a prior in subsequent searches or constraints.

In the future, we aim to go beyond the results of this paper in parallel with the ongoing development of `GPry`, improving its accuracy in highly non-Gaussian and highly multimodal cases (e.g., extreme mass ratio inspirals), and its performance in larger dimensionalities such as inference problems with multiple sources.

ACKNOWLEDGMENTS

The authors would like to thank A. Klein, D. Bandopadhyay, C. Moore, and all the `Balrog` developers for useful discussions and insightful comments. JE and GN acknowledge support from the ROMFORSK grant project no. 302640. JE acknowledges support by the Spoke 3 (INAF) of the Italian Center for SuperComputing (ICSC), funded by the European Union - NextGenerationEU program, under the grant agreement N. C53C22000350006 (acronym Fab-HPCc). RB acknowledges support from the ICSC National Research Center funded by NextGenerationEU, and the Italian Space Agency grant *Phase A activity for LISA mission*, Agreement n.2017-29-H.0. JT acknowledges financial support from the Supporting TAlent in ReSearch@University of Padova (STARS@UNIPD) for the project "Constraining Cosmology and Astrophysics with Gravitational Waves, Cosmic Microwave Background and Large-Scale Structure cross-correlations", and from a Ramón y Cajal contract by the Spanish Ministry for Science, Innovation and Universities with Ref. RYC2023-045660-I. Computa-

tional resources were provided by University of Birmingham BlueBEAR High Performance Computing facility, by CINECA through EuroHPC Benchmark access call grant EUHPC-B03-24, by the Google Cloud Research Credits program with the award GCP19980904, and by the CloudVeneto initiative of the Università di Padova and the INFN – Sezione di Padova.

Software: We acknowledge usage of `Mathematica` [72] and of the following `Python` [73] packages for modeling, analysis, post-processing, and production of results throughout: `nessai` [74], `matplotlib` [75], `numpy` [76], `scipy` [77], `scikit-learn` [78], `Cobaya` [79] and `corner` [80].

[1] R. Abbott *et al.* (KAGRA, VIRGO, LIGO Scientific), Phys. Rev. X **13**, 041039 (2023), arXiv:2111.03606 [gr-qc].
[2] A. Afzal *et al.* (NANOGrav), Astrophys. J. Lett. **951**, L11 (2023), [Erratum: Astrophys.J.Lett. 971, L27 (2024), Erratum: Astrophys.J. 971, L27 (2024)], arXiv:2306.16219 [astro-ph.HE].
[3] J. Antoniadis *et al.* (EPTA, InPTA:), Astron. Astrophys. **678**, A50 (2023), arXiv:2306.16214 [astro-ph.HE].
[4] D. J. Reardon *et al.*, Astrophys. J. Lett. **951**, L6 (2023), arXiv:2306.16215 [astro-ph.HE].
[5] H. Xu *et al.*, Res. Astron. Astrophys. **23**, 075024 (2023), arXiv:2306.16216 [astro-ph.HE].
[6] P. Amaro-Seoane *et al.* (LISA), arXiv e-prints (2017), arXiv:1702.00786 [astro-ph.IM].
[7] P. Auclair *et al.* (LISA Cosmology Working Group), Living Rev. Rel. **26**, 5 (2023), arXiv:2204.05434 [astro-ph.CO].
[8] S. H. Strub, L. Ferraioli, C. Schmelzbach, S. C. Stähler, and D. Giardini, Phys. Rev. D **106**, 062003 (2022), arXiv:2204.04467 [astro-ph.IM].
[9] P. A. Seoane *et al.* (LISA), Living Rev. Rel. **26**, 2 (2023), arXiv:2203.06016 [gr-qc].
[10] N. Afshordi *et al.* (LISA Consortium Waveform Working Group), arXiv e-prints (2023), arXiv:2311.01300 [gr-qc].
[11] E. Cuoco *et al.*, Mach. Learn. Sci. Tech. **2**, 011002 (2021), arXiv:2005.03745 [astro-ph.HE].
[12] P. Canizares, S. E. Field, J. Gair, V. Raymond, R. Smith, and M. Tiglio, Phys. Rev. Lett. **114**, 071104 (2015), arXiv:1404.6284 [gr-qc].
[13] R. Smith, S. E. Field, K. Blackburn, C.-J. Haster, M. Pürrer, V. Raymond, and P. Schmidt, Phys. Rev. D **94**, 044031 (2016), arXiv:1604.08253 [gr-qc].
[14] S. E. Field, C. R. Galley, J. S. Hesthaven, J. Kaye, and M. Tiglio, Phys. Rev. X **4**, 031006 (2014), arXiv:1308.3565 [gr-qc].
[15] M. Dax, S. R. Green, J. Gair, J. H. Macke, A. Buonanno, and B. Schölkopf, Phys. Rev. Lett. **127**, 241103 (2021), arXiv:2106.12594 [gr-qc].
[16] U. Bhardwaj, J. Alvey, B. K. Miller, S. Nissanke, and C. Weniger, Phys. Rev. D **108**, 042004 (2023), arXiv:2304.02035 [gr-qc].
[17] M. Andrés-Carcasona, M. Martinez, and L. M. Mir, Mon. Not. Roy. Astron. Soc. **527**, 2887 (2023), arXiv:2309.04303 [gr-qc].
[18] D. Chatterjee *et al.*, Mach. Learn. Sci. Tech. **5**, 045030 (2024), arXiv:2407.19048 [gr-qc].
[19] M. Dax, S. R. Green, J. Gair, N. Gupte, M. Pürrer, V. Raymond, J. Wildberger, J. H. Macke, A. Buonanno, and B. Schölkopf, Nature **639**, 49 (2025), arXiv:2407.09602 [gr-qc].
[20] V. Raymond, S. Al-Shammari, and A. Göttel, arXiv e-prints (2024), arXiv:2406.03935 [gr-qc].
[21] I. M. Vílchez and C. F. Sopuerta, arXiv e-prints (2024), arXiv:2406.00565 [gr-qc].
[22] H. Sun, H. Wang, and J. He, "Accelerating bayesian sampling for massive black hole binaries with prior constraints from conditional variational autoencoder," (2025), arXiv:2502.09266 [astro-ph.IM].
[23] H. Gabbard, C. Messenger, I. S. Heng, F. Tonolini, and R. Murray-Smith, Nature Phys. **18**, 112 (2022), arXiv:1909.06296 [astro-ph.IM].
[24] M. Dax, S. R. Green, J. Gair, M. Pürrer, J. Wildberger, J. H. Macke, A. Buonanno, and B. Schölkopf, Phys. Rev. Lett. **130**, 171403 (2023), arXiv:2210.05686 [gr-qc].
[25] J. EL Gammal, N. Schöneberg, J. Torrado, and C. Fidler, JCAP **10**, 021 (2023), arXiv:2211.02045 [astro-ph.CO].
[26] J. Torrado, N. Schöneberg, and J. El Gammal, "Parallelized acquisition for active learning using monte carlo sampling," (2023), arXiv:2305.19267 [stat.ML].
[27] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, Adaptive computation and machine learning (MIT Press, Cambridge, Mass., 2006) pp. XVIII, 248 S.
[28] M. Osborne, R. Garnett, Z. Ghahramani, D. K. Duvenaud, S. J. Roberts, and C. Rasmussen, in *Advances in Neural Information Processing Systems*, Vol. 25, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Curran Associates, Inc., 2012) pp. 46–54.
[29] T. Gunter, M. Osborne, R. Garnett, P. Hennig, and S. Roberts, in *Advances in Neural Information Processing Systems 27* (Curran Associates, Inc., 2014) pp. 2789–2797.
[30] M. Georgousi, N. Karnesis, V. Korol, M. Pieroni, and N. Stergioulas, MNRAS **519**, 2552 (2023), arXiv:2204.07349 [astro-ph.GA].
[31] V. Korol, V. Belokurov, C. J. Moore, and S. Toonen, MNRAS **502**, L55 (2021), arXiv:2010.05918 [astro-ph.GA].
[32] M. Colpi, K. Danzmann, M. Hewitson, K. Holley-Bockelmann, and et al., arXiv e-prints , arXiv:2402.07571 (2024), arXiv:2402.07571 [astro-ph.CO].
[33] E. Finch, G. Bartolucci, D. Chucherko, B. G. Patterson, and et al., MNRAS **522**, 5358 (2023), arXiv:2210.10812 [astro-ph.SR].
[34] C. Cutler, Phys. Rev. D **57**, 7089 (1998), arXiv:gr-qc/9703068 [gr-qc].
[35] M. Katz, "mikekatz04/gbgpu: First official public release!" (2022).
[36] O. Burke, S. Marsat, J. R. Gair, and M. L. Katz, arXiv e-prints , arXiv:2502.17426 (2025), arXiv:2502.17426 [gr-qc].

[37] R. Buscicchio, J. Torrado, C. Caprini, G. Nardini, N. Karnesis, M. Pieroni, and A. Sesana, JCAP **01**, 084 (2025), arXiv:2410.18171 [astro-ph.HE].
[38] A. Klein, G. Pratten, R. Buscicchio, P. Schmidt, and et al., arXiv e-prints , arXiv:2204.03423 (2022), arXiv:2204.03423 [astro-ph.HE].
[39] A. Klein, arXiv e-prints , arXiv:2106.10291 (2021), arXiv:2106.10291 [gr-qc].
[40] G. Morras, G. Pratten, and P. Schmidt, arXiv e-prints , arXiv:2502.03929 (2025), arXiv:2502.03929 [gr-qc].
[41] G. Pratten, P. Schmidt, H. Middleton, and A. Vecchio, Phys. Rev. D **108**, 124045 (2023), arXiv:2307.13026 [gr-qc].
[42] P. C. Peters and J. Mathews, Phys. Rev. **131**, 435 (1963).
[43] C. García-Quirós, M. Colleoni, S. Husa, H. Estellés, and et al., Phys. Rev. D **102**, 064002 (2020), arXiv:2001.10914 [gr-qc].
[44] LISA Consortium Waveform Working Group, N. Afshordi, S. Akçay, P. Amaro Seoane, and et al., arXiv e-prints , arXiv:2311.01300 (2023), arXiv:2311.01300 [gr-qc].
[45] M. Tinto and S. V. Dhurandhar, Living Reviews in Relativity **8**, 4 (2005).
[46] R. Buscicchio, A. Klein, E. Roebber, C. J. Moore, and et al., Phys. Rev. D **104**, 044065 (2021), arXiv:2106.05259 [astro-ph.HE].
[47] L. J. Rubbo, N. J. Cornish, and O. Poujade, Phys. Rev. D **69**, 082003 (2004), arXiv:gr-qc/0311069 [gr-qc].
[48] G. Pratten, C. García-Quirós, M. Colleoni, A. Ramos-Buades, and et al., Phys. Rev. D **103**, 104056 (2021), arXiv:2004.06503 [gr-qc].
[49] M. Pürrer and C.-J. Haster, Physical Review Research **2**, 023151 (2020), arXiv:1912.10055 [gr-qc].
[50] LISA Science Study Team, *LISA Science Requirements Document*, Tech. Rep. 1.0 (ESA, 2018).
[51] N. Karnesis, S. Babak, M. Pieroni, N. Cornish, and T. Littenberg, Phys. Rev. D **104**, 043019 (2021), arXiv:2103.14598 [astro-ph.IM].
[52] H. Sun, H. Wang, and J. He, arXiv e-prints , arXiv:2502.09266 (2025), arXiv:2502.09266 [astro-ph.IM].
[53] D. Bandopadhyay and C. J. Moore, Physical Review D **108** (2023), 10.1103/physrevd.108.084014.
[54] D. Bandopadhyay and C. J. Moore, Physical Review D **110** (2024), 10.1103/physrevd.110.103026.
[55] G. Cabourn Davies, I. Harry, M. J. Williams, D. Bandopadhyay, and et al., Phys. Rev. D **111**, 043045 (2025), arXiv:2411.07020 [hep-ex].
[56] G. Pratten, A. Klein, C. J. Moore, H. Middleton, and et al., Phys. Rev. D **107**, 123026 (2023), arXiv:2212.02572 [gr-qc].
[57] L. Piro, M. Colpi, J. Aird, A. Mangiagli, and et al., MNRAS **521**, 2577 (2023), arXiv:2211.13759 [astro-ph.HE].
[58] A. Mangiagli, C. Caprini, M. Volonteri, S. Marsat, and et al., Phys. Rev. D **106**, 103017 (2022), arXiv:2207.10678 [astro-ph.HE].
[59] M. Vallisneri, M. Crisostomi, A. D. Johnson, and P. M. Meyers, arXiv e-prints (2024), arXiv:2405.08857 [gr-qc].
[60] J. El Gammal, N. Schöneberg, J. Torrado, and C. Fidler, "GPry: Bayesian inference of expensive likelihoods with Gaussian processes," Astrophysics Source Code Library, record ascl:2212.006 (2022), ascl:2212.006.
[61] W. J. Handley, M. P. Hobson, and A. N. Lasenby, Mon. Not. Roy. Astron. Soc. **450**, L61 (2015), arXiv:1502.01856 [astro-ph.CO].
[62] W. J. Handley, M. P. Hobson, and A. N. Lasenby, Mon. Not. Roy. Astron. Soc. **453**, 4384 (2015), arXiv:1506.00171 [astro-ph.IM].
[63] D. Ginsbourger, R. Le Riche, and L. Carraro, "Kriging is well-suited to parallelize optimization," (Springer, 2010) pp. 131–162.
[64] A. Lewis and S. Bridle, Phys. Rev. **D66**, 103511 (2002), arXiv:astro-ph/0205436 [astro-ph].
[65] A. Lewis, Phys. Rev. **D87**, 103529 (2013), arXiv:1304.4473 [astro-ph.CO].
[66] M. J. Williams, "nessai: Nested sampling with artificial intelligence," (2021).
[67] M. J. Williams, J. Veitch, and C. Messenger, Phys. Rev. D **103**, 103006 (2021), arXiv:2102.11056 [gr-qc].
[68] M. J. Williams, J. Veitch, and C. Messenger, Mach. Learn. Sci. Tech. **4**, 035011 (2023), arXiv:2302.08526 [astro-ph.IM].
[69] S. R. Cook, A. Gelman, and D. B. Rubin, Journal of Computational and Graphical Statistics **15**, 675 (2006).
[70] M. B. Wilk and R. Gnanadesikan, Biometrika **55**, 1 (1968).
[71] J. Lin, IEEE Transactions on Information Theory **37**, 145 (1991).
[72] Wolfram Research Inc., "Mathematica," (2022).
[73] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual* (CreateSpace, Scotts Valley, CA, 2009).
[74] M. J. Williams, J. Veitch, and C. Messenger, Phys. Rev. D **103**, 103006 (2021), arXiv:2102.11056 [gr-qc].
[75] J. D. Hunter, Computing in Science and Engineering **9**, 90 (2007).
[76] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, and et al., Nature **585**, 357 (2020), arXiv:2006.10256 [cs.MS].
[77] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, and et al., Nature Methods **17**, 261 (2020), arXiv:1907.10121 [cs.MS].
[78] F. Pedregosa *et al.*, Journal of Machine Learning Research **12**, 2825 (2011).
[79] J. Torrado and A. Lewis, JCAP **05**, 057 (2021), arXiv:2005.05290 [astro-ph.IM].
[80] D. Foreman-Mackey, The Journal of Open Source Software **1**, 24 (2016).

Appendix A: Approximate Jensen-Shannon divergence between multivariate Gaussians

The KL divergence D_{KL}, defined in Eq. (17), has an analytical representation when computed between two d-dimensional multivariate Gaussians, $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma_1})$ and $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma_2})$, with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and covariance matrices $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$. Conversely, an analytical representation for the JS divergence, defined in Eq. (16), does not exist in this case. However, we can find an approximate expression if the mixture distribution M (with mean $\boldsymbol{\mu}_M = \frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$) is a multivariate normal distribution with covariance $\boldsymbol{\Sigma}_M = \frac{1}{2}(\sigma_1^2 + \sigma_2^2)\mathbb{I}$. The approximation holds if the multivariate Gaussians are sufficiently similar, i.e. $|\Delta\boldsymbol{\mu}| \equiv |\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2| \ll |\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2|$ and $\sigma_1 \approx \sigma_2$. In this case, the JS divergence reads

$$D_{\text{JS}}[\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)||\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)] \approx \frac{1}{4}\Delta\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}\Delta\boldsymbol{\mu} + \frac{1}{2}\log\left(\frac{|\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}}\right) - \frac{d}{2}\log 2\ . \tag{A1}$$

For $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \sigma^2\mathbb{I}$ this simplifies to

$$D_{\text{JS}} \approx \frac{(\Delta\boldsymbol{\mu})^2}{8\sigma^2}\ , \tag{A2}$$

and for $|\Delta\boldsymbol{\mu}| = 0$, $\boldsymbol{\Sigma}_1 = \sigma_1^2\mathbb{I}$, $\boldsymbol{\Sigma}_2 = \sigma_2^2\mathbb{I}$ to

$$D_{\text{JS}} \approx \frac{d}{2}\log\frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2}\ . \tag{A3}$$

Therefore, when only the mean of the distribution is misestimated, D_{JS} is a function of the distance $\Delta\boldsymbol{\mu}$ in units of σ (see Eq. (A2)). This is independent of the number of dimensions if only one parameter is misestimated whereas it is proportional to d if the misestimation occurs for multiple parameters. On the other hand, if the mean is properly estimated but the spread of the distribution is not, then the result is proportional to d whenever, on average, the spread is wrong by the same amount (see Eq. (A3)). We find that in less than 11 dimensions (the highest number of inferred parameters that we consider in this paper), the approximation holds up to $D_{\text{JS}} \approx 0.1$.

Appendix B: Hyperparameters and overhead of `GPry`

Table IV shows the values of the `GPry` settings adapted to this study, as motivated in Sec. III C. An in-depth explanation of the meaning of each setting can be found in `GPry`'s documentation[4]. In Fig. 7 we show a breakdown of the computation costs of the `GPry` runs into posterior evaluation time and overhead from the two computationally expensive steps of the Bayesian optimization loop.

Setting	Description	DWD	stBHB	SMBHB
`noise_level`	Expected level of numerical noise	0.1	4	2
`inf_threshold`	Cutoff in log-posterior of the SVM classifier	20σ	10σ	30σ
`fit_full_every`	Number of iterations between GPR hyperparameter optimizations	2	2	1
`n_restarts_optimizer`	Number of restarts per GPR hyperparameter optimization	$2d$	d	$4d$
`mc_every`	Number of iterations between NS runs to generate proposals	5	3	3
`n_points_per_acq`	Number of Kriging steps determining the proposal batch size	7	9	8
`trust_region_nstd`	Cutoff in log-posterior for defining the trust region	—	3σ	3σ
`trust_region_factor`	Enlargement factor of the trust region	—	2.5	2.5

TABLE IV. Non-default settings for `GPry`, as discussed in Sec. III C. In the parameter values, a number followed by d is multiplied by the dimensionality of the problem, whereas one followed by σ represents the difference between the log-posterior of the cutoff and that of the best training sample as the equivalent number of 1-dimensional standard deviations, i.e. 2σ represents the log-posterior difference from the top of the distribution of a d-dimensional Gaussian that leaves 95% of the mass above it.

[4] https://gpry.readthedocs.io

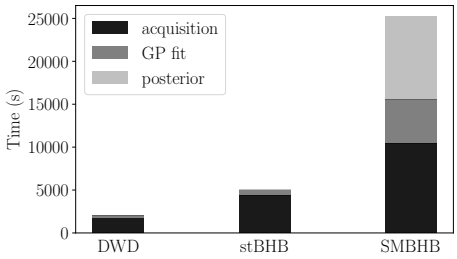


FIG. 7. Graph showing the dominant part of the overhead (acquisition and hyperparameter fits of the GPR) vs the total time spent on evaluating the log-posterior. Sub-dominant or non-necessary contributions to `GPry`'s overhead have been omitted such as determining convergence, checkpointing and the generation of the final MC sample, which typically add up to a few seconds. For the relatively fast to evaluate posteriors of the DWDs and stBHBs the runtime is dominated by `GPry`'s overhead which – due to the much lower number of posterior evaluations – still leads to a speedup over `nessai` in the case of the stBHBs and SMBHBs. For the slow SMBHB posterior, despite only roughly 1/3 of the time being spent on posterior evaluations, the large reduction in their number with respect to `nessai` still leads to a significant speedup (see Fig. 6b).

Appendix C: Corner plots

In this appendix, we show some corner plots for the sources studied in Sec. IV. In the upper part of each plot we furthermore show the sampling locations of `GPry`, omitting the samples that are far away from the mode (typically a few percent). The contours for `GPry` are obtained by sampling the surrogate model with an MCMC.

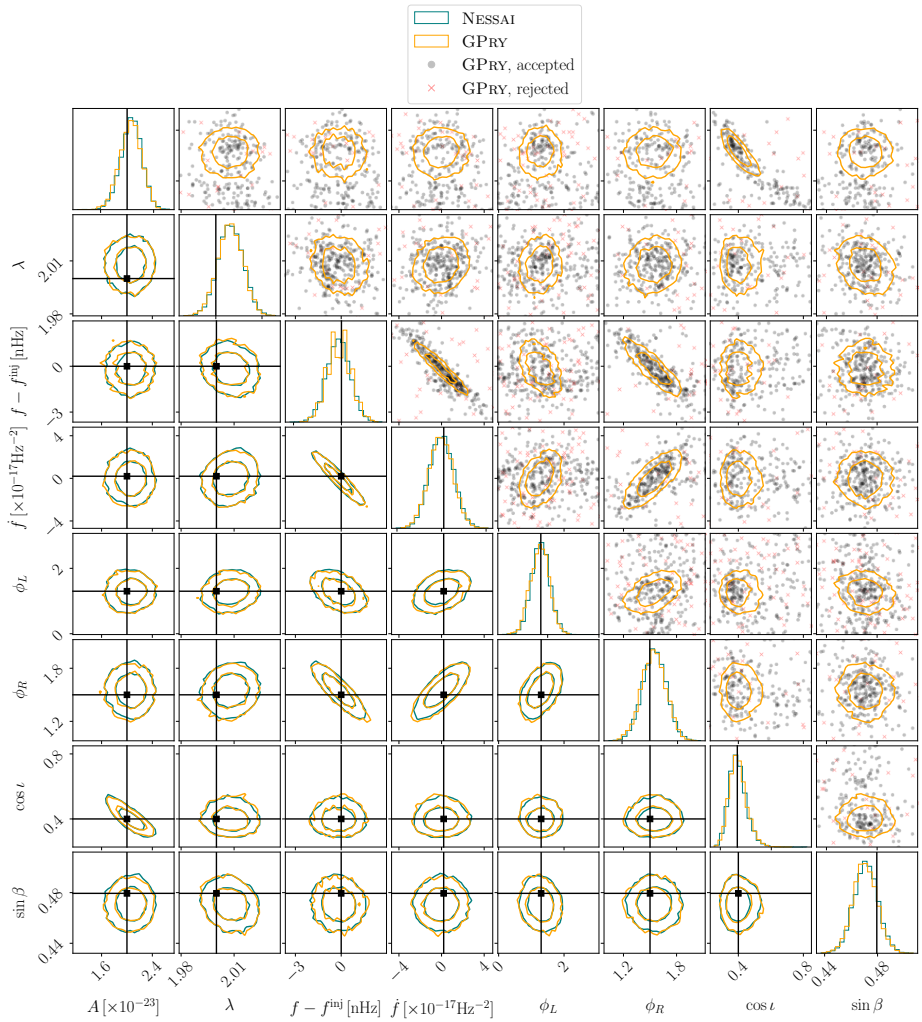


FIG. 8. Corner plot comparing `nessai` to `GPry` inference on the DWD source with the parameters specified in Table I for the run with the median JS divergence $D_{\rm JS} = 0.0043$ (see Fig. 3b for the distribution). The number of likelihood evaluations for `GPry` was ≈ 500 (shown in the upper triangle, missing a few percent that would fall outside the ranges of the plot), and for `nessai` it was ≈ 136500. The 2d contour levels show the 68% and 95% CL constraints. On the upper triangular we show the locations where `GPry` has evaluated the true posterior distribution. The gray dots represent accepted samples (samples that are used to train the GPR), while the red crosses are rejected (used to train the SVM classifier).

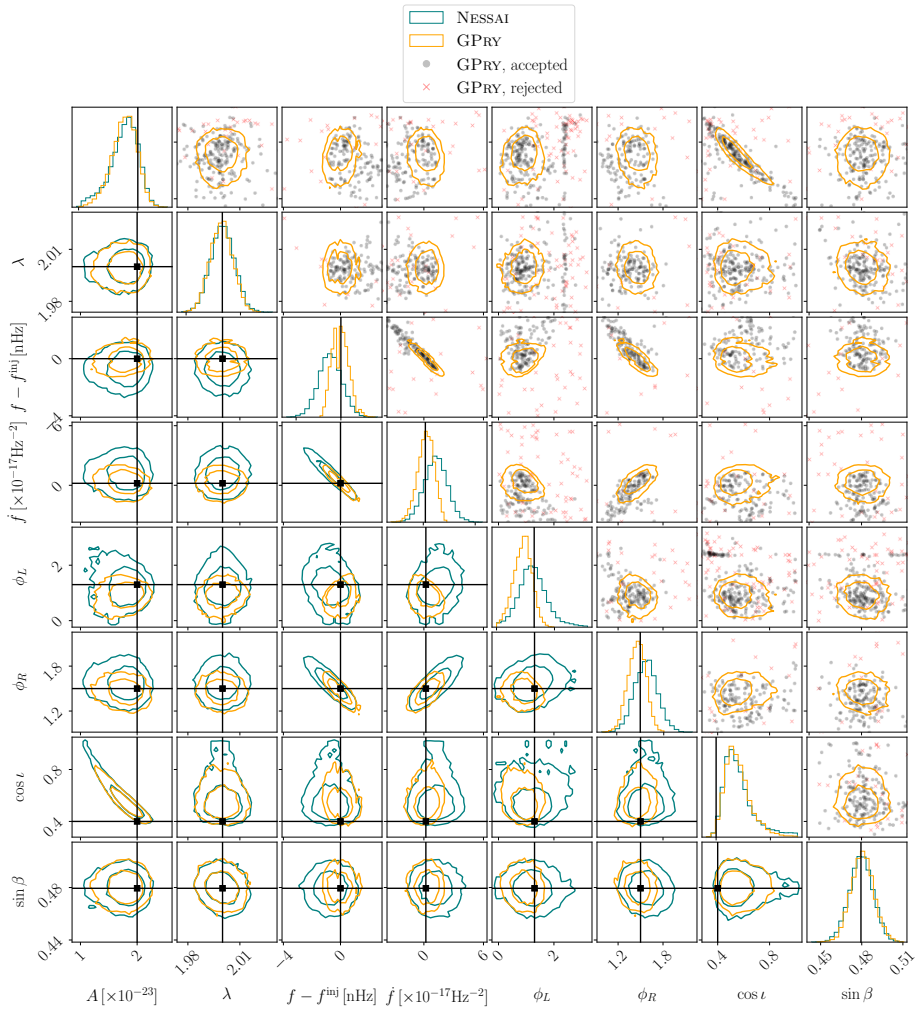


FIG. 9. Same as Fig. 8, comparing `nessai` to `GPry` inference on the DWD source with the parameters specified in Table I for the run with the highest JS divergence $D_{\rm JS} = 0.12$ (see Fig. 3b for the distribution). The number of likelihood evaluations for `GPry` was ≈ 350 (shown in the upper triangle), and for `nessai` it was ≈ 146000.

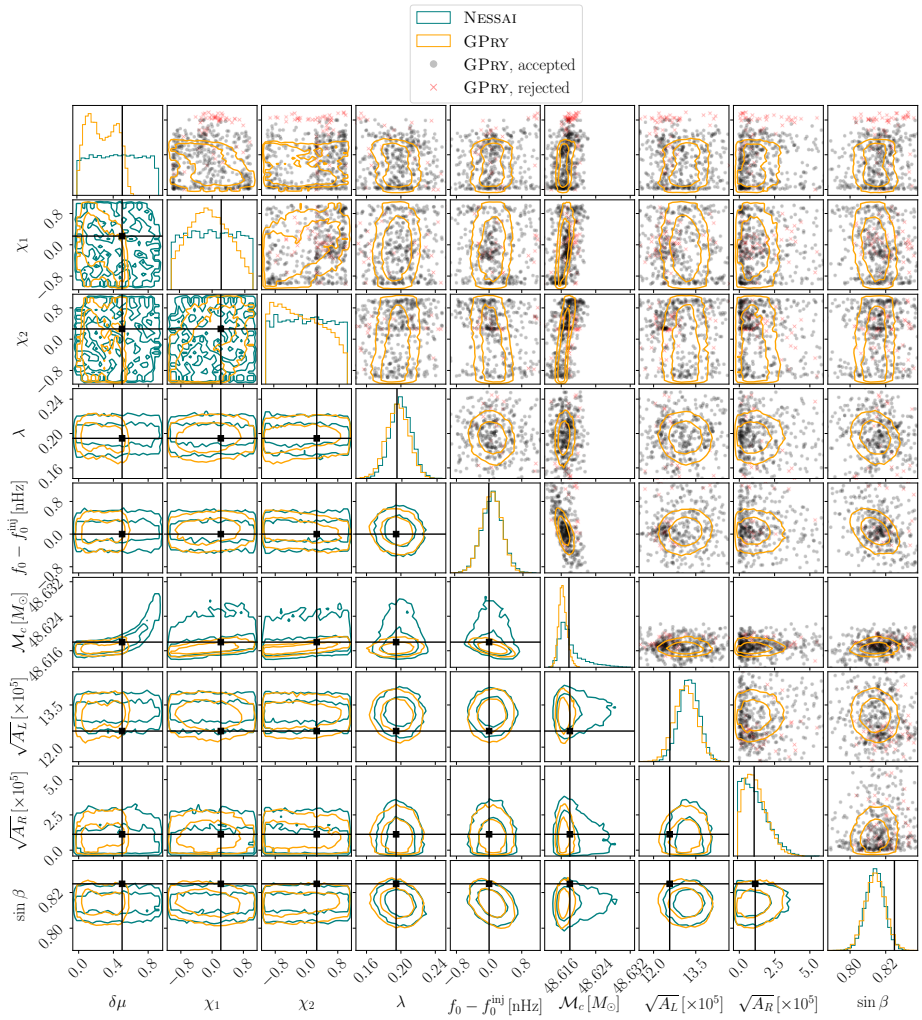


FIG. 10. Same as Fig. 8, comparing `nessai` to `GPry` inference on the stBHB source with the parameters specified in Table II for the run with the median JS divergence $D_{\rm JS} = 0.19$ (see Fig. 4b for the distribution). The number of likelihood evaluations for `GPry` was ≈ 450, and for `nessai` it was ≈ 23484500.

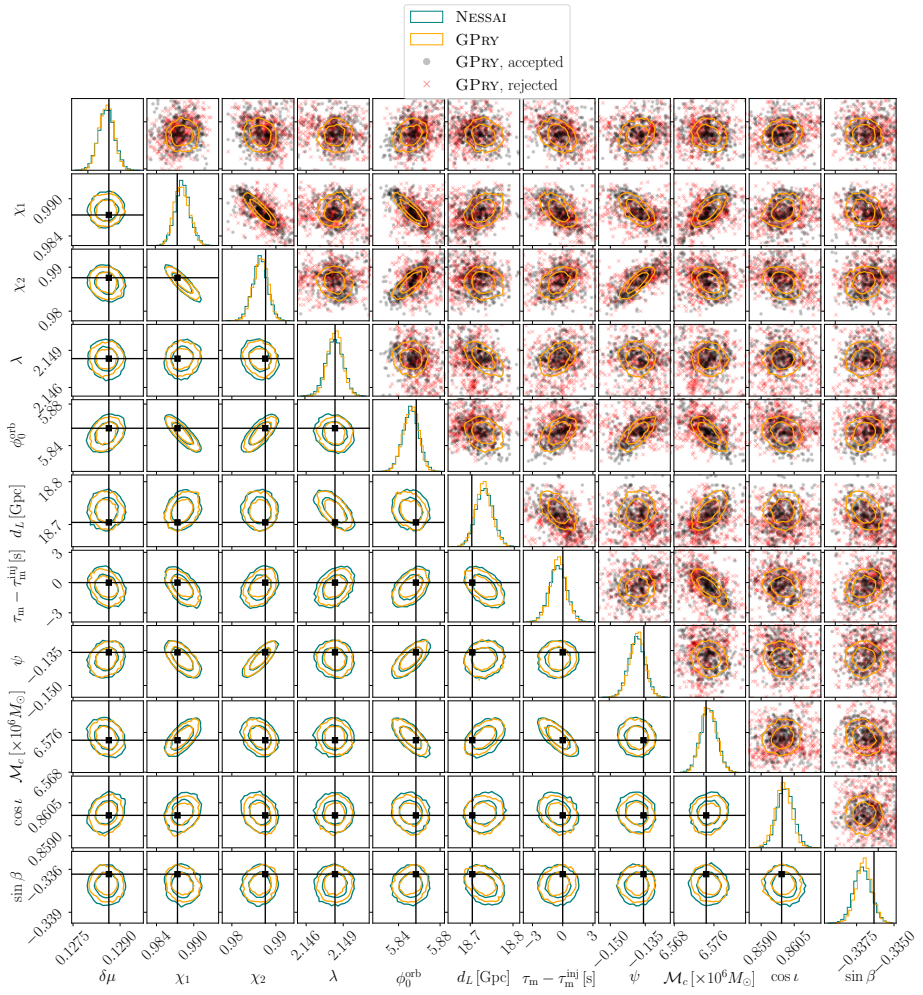


FIG. 11. Same as Fig. 8, comparing `nessai` to `GPry` inference on the SMBHB source with the parameters specified in Table III for the run with the median JS divergence $D_{\rm JS} = 0.048$ (see Fig. 5b for the distribution). The number of likelihood evaluations for `GPry` was ≈ 850, and for `nessai` it was ≈ 207000 with 500 live points (used for the $D_{\rm JS}$ calculation and the PP plot), and ≈ 567000 for the high resolution run (2000 live points) whose contours are shown.

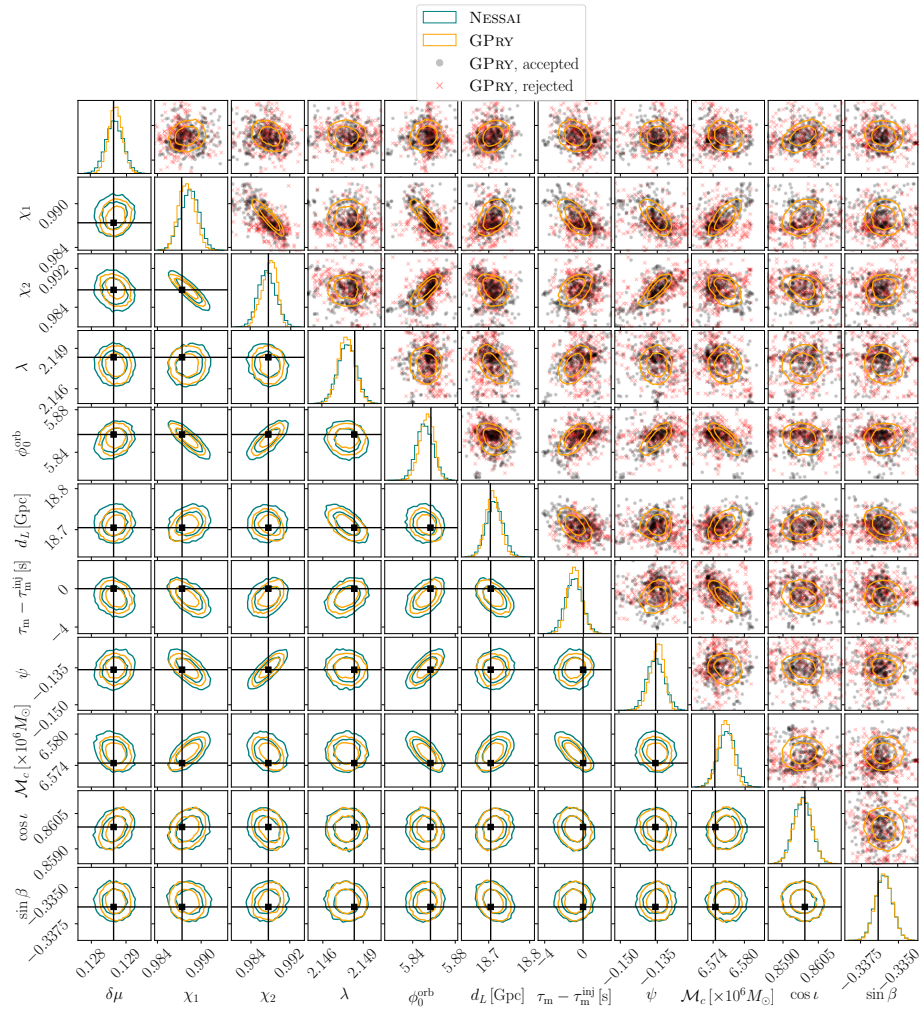


FIG. 12. Same as Fig. 8, comparing `nessai` to `GPry` inference on the SMBHB source with the parameters specified in Table III for the run with the highest JS divergence $D_{\rm JS} = 0.078$ (see Fig. 5b for the distribution). The number of likelihood evaluations for `GPry` was ≈ 650, and for `nessai` it was ≈ 198500 at low resolution (500 live points), ≈ 364000 at high resolution (2000 live points), whose contours are shown.

Paper IV

Reconstructing Primordial Curvature Perturbations via Scalar-Induced Gravitational Waves with LISA

CERN-TH-2024-217

Reconstructing Primordial Curvature Perturbations via Scalar-Induced Gravitational Waves with LISA

Jonas El Gammal[1],[a] **Aya Ghaleb**,[b] **Gabriele Franciolini**[2],[c] **Theodoros Papanikolaou**,[def] **Marco Peloso**,[gh] **Gabriele Perna**[3],[gh] **Mauro Pieroni**,[c] **Angelo Ricciardone**,[ij] **Robert Rosati**[4],[k] **Gianmassimo Tasinato**,[blm]

Matteo Braglia,[n] **Jacopo Fumagalli**,[o] **Jun'ya Kume**,[ghr] **Enrico Morgante**,[pq] **Germano Nardini**,[a] **Davide Racco**,[st] **Sébastien Renaux-Petel**,[u] **Hardi Veermäe**,[v] **Denis Werth**,[u] **and Ivonne Zavala**[b]

(For the LISA Cosmology Working Group)

[a] *Department of Mathematics and Physics, University of Stavanger, NO-4036 Stavanger, Norway*
[b] *Department of Physics, Faculty of Science and Engineering, Swansea University, Singleton Park, SA2 8PP, Swansea, United Kingdom*
[c] *CERN, Theoretical Physics Department, Esplanade des Particules 1, Geneva 1211, Switzerland*
[d] *Scuola Superiore Meridionale, Largo San Marcellino 10, 80138 Napoli, Italy*
[e] *Istituto Nazionale di Fisica Nucleare (INFN), Sezione di Napoli, Via Cinthia 21, 80126 Napoli, Italy*
[f] *National Observatory of Athens, Lofos Nymfon, 11852 Athens, Greece*
[g] *Dipartimento di Fisica e Astronomia "G. Galilei", Università degli Studi di Padova, via Marzolo 8, I-35131 Padova, Italy*
[h] *INFN, Sezione di Padova, via Marzolo 8, I-35131 Padova, Italy*
[i] *Dipartimento di Fisica "Enrico Fermi", Università di Pisa, Largo Bruno Pontecorvo 3, Pisa I-56127, Italy*
[j] *INFN, Sezione di Pisa, Largo Bruno Pontecorvo 3, Pisa I-56127, Italy*
[k] *NASA Marshall Space Flight Center, Huntsville, AL 35812, USA*
[l] *Dipartimento di Fisica e Astronomia, Università di Bologna*
[m] *INFN, Sezione di Bologna, I.S. FLAG, viale B. Pichat 6/2, 40127 Bologna, Italy*

[1] Corresponding author: jonas.el.gammal@rwth-aachen.de
[2] Project coordinator: gabriele.franciolini@cern.ch
[3] Corresponding author: gabriele.perna@phd.unipd.it
[4] Project coordinator: robert.j.rosati@nasa.gov, NASA Postdoctoral Program Fellow

[n] *Center for Cosmology and Particle Physics, New York University, 726 Broadway, New York, NY 10003, USA*

[o] *Departement de Física Quàntica i Astrofisica and Institut de Ciències del Cosmos (ICC), Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain*

[p] *Dipartimento di Fisica, Università di Trieste, Strada Costiera 11, I-34151 Trieste, Italy*

[q] *INFN, Sezione di Trieste, Via Valerio 2, 34127 Trieste, Italy*

[r] *Research Center for the Early Universe (RESCEU), Graduate School of Science, The University of Tokyo, Hongo 7-3-1 Bunkyo-ku, Tokyo 113-0033, Japan*

[s] *Institut für Theoretische Physik, ETH Zürich,Wolfgang-Pauli-Str. 27, 8093 Zürich, Switzerland*

[t] *Physik-Institut, Universität Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland*

[u] *Institut d'Astrophysique de Paris, UMR 7095 du CNRS et de Sorbonne Université, 98 bis bd Arago, 75014 Paris, France*

[v] *Keemilise ja bioloogilise füüsika instituut, Rävala pst. 10, 10143 Tallinn, Estonia*

Abstract: Many early universe scenarios predict an enhancement of scalar perturbations at scales currently unconstrained by cosmological probes. These perturbations source gravitational waves (GWs) at second order in perturbation theory, leading to a scalar-induced gravitational wave (SIGW) background. The LISA detector, sensitive to mHz GWs, will be able to constrain curvature perturbations in a new window corresponding to scales $k \in [10^{10}, 10^{14}]\,\mathrm{Mpc}^{-1}$, difficult to probe otherwise. In this work, we forecast the capabilities of LISA to constrain the source of SIGWs using different approaches: *i)* agnostic, where the spectrum of curvature perturbations is binned in frequency space; *ii)* template-based, modeling the curvature power spectrum based on motivated classes of models; *iii)* ab initio, starting from first-principles model of inflation featuring an ultra-slow roll phase. We compare the strengths and weaknesses of each approach. We also discuss the impact on the SIGW spectrum of non-standard thermal histories affecting the kernels of SIGW emission and non-Gaussianity in the statistics of the curvature perturbations. Finally, we propose simple tests to assess whether the signal is compatible with the SIGW hypothesis. The pipeline used is built into the `SIGWAY` code.

Contents

1 Introduction

The Laser Interferometer Space Antenna (LISA) [1] represents a groundbreaking gravitational wave (GW) observatory aimed to probe and impact our understanding of fundamental physics, astronomy, and cosmology [2–4]. With the first-ever direct probe of the stochastic gravitational-wave background (SGWB) in the millihertz frequency range, LISA provides the opportunity to unveil processes that occurred in the first stages of the Universe, including inflation [5, 6]. Probing a primordial SGWB at millihertz frequencies corresponds to exploring comoving scales that lie between those accessible to ground-based GW interferometers and those probed by pulsar-timing arrays, cosmic microwave background (CMB), or large-scale structure surveys. These scales correspond to comoving wavenumbers in the range $k \sim [10^{10}, 10^{14}]\,\mathrm{Mpc}^{-1}$. Several cosmological models (see e.g. refs. [7–10] for reviews) predict a measurable SGWB at these scales, often without any other complementary distinctive signature, placing LISA in a unique position to test these scenarios.

Inflationary models exhibiting amplified scalar fluctuations are one of the candidates for sourcing an SGWB in the LISA frequency band. Enhanced scalar fluctuations generate scalar-induced gravitational waves (SIGWs) at second order in perturbations [11–20], resulting in a potentially large SGWB. Amplified scalar fluctuations naturally arise in single-field inflationary scenarios with features in the potential like those leading to ultra-slow-roll (USR) phases, multi-field setups, and mechanisms such as preheating or early matter-dominated eras. Intriguingly, the same perturbations that seed the SIGWs can also trigger the formation of primordial black holes (PBHs) [21–25]. SIGWs in the millihertz frequency band arise in correspondence with the asteroidal-mass window for PBHs, a viable candidate for addressing the dark matter puzzle [26–28]. By detecting or setting upper bounds on SIGWs, LISA would not only shed light on the inflationary epoch but also on dark matter and non-astrophysical black hole formation channels. This makes SIGWs a high-gain, well-motivated target for LISA.

SIGWs are also powerful tools to investigate non-Gaussianity (NG) in the early universe since their production is highly sensitive to the statistical properties of scalar curvature fluctuations [29–38]. NG is typically characterized by parameters such as $f_{\rm NL}$, which quantifies NG at the bispectrum level (three-point correlation function), and $\tau_{\rm NL}$, which appears in the trispectrum (four-point correlation function). Earlier analysis have focused on contributions to the SIGW involving $f_{\rm NL}$, [39–50]. More recent studies have extended this analysis to higher-order NG terms [45, 51–53]. Recently, a Fisher forecast analysis for LISA about NG has been performed in [53]. The tensor power spectrum of SIGWs is directly related to the four-point correlation function of the curvature fluctuations. Such a correlation function has both connected and disconnected contributions. While the latter contributes only to the Gaussian SIGW power spectrum, the former is directly linked to the trispectrum through the $\tau_{\rm NL}$ parameter [49], which is the key observable to constrain NG from SIGWs.

The detection and characterization of the primordial SGWB is one of the most challenging objectives of the LISA mission [54]. LISA is a signal-dominated detector, where a multitude of transient or quasi-monochromatic events overlap in time and frequency with the stochastic superposition of all unresolved astrophysical events and, potentially also with a significant primordial SGWB. Additionally, the stationary component of the instrumental noise can mimic a SGWB to some extent. Completing the LISA science program for the SGWB therefore requires:

i) Determining whether a primordial SGWB is present in the data.

ii) Reconstructing the SGWB frequency shape and, if possible, its statistical properties.

iii) Setting upper limits on cosmological sources of SGWB not supported by the data.

iv) Constraining the parameters of the most likely SGWB source candidates.

ESA and NASA plan to address these tasks through the so-called "global fit", a data analysis procedure where modules fitting each class of sources (galactic binaries, supermassive black hole binaries, SGWB, etc.) iterate until convergence [54]. Recently, successful prototype global fit analyses became available in the literature [55–57], tested on the *Sangria* LISA Data Challenge (LDC) dataset [58], which contains no primordial SGWB. It is still an open question how the global fit should support a primordial SGWB search, and how the SGWB properties should be represented in the detection catalogs that the space agencies will publish. A recent study [59] has attempted to perform an SGWB search directly on the global fit residual.

In this work, we aim to bridge these gaps by providing elements of the global fit SGWB module useful for the tasks *i) - iv)* in the presence of a SGWB due to SIGWs. To develop and test our rationale, we work in the limit that all resolvable events have been precisely reconstructed[1], leaving us with data containing the stationary component of the noise, the

[1] Although this optimistic working hypothesis may seem unrealistic, it is the correct one to use in a global fit module. All current implementations of the global fit are based on a blocked Gibbs sampling scheme, where each source type is sampled independently, assuming a perfect subtraction of the other source types.

SIGW background, and the foregrounds from the unresolved galactic and extragalactic binaries.

As we focus on SIGW sources, we perform an analysis starting from the properties of the source curvature power spectra $\mathcal{P}_\zeta(k)$ of the source, instead of the GW energy density $\Omega_{\rm GW}(f)$ generated by these power spectra.

Concerning *i)* and *ii)*, we prototype a model-agnostic method that reconstructs the power spectrum $\mathcal{P}_\zeta(k)$ by binning it in frequency space. This approach allows for maximal flexibility in capturing unknown SIGW features. It is however not as agnostic as other generic SGWB searches [60–66] since it requires, by construction, a SIGW source, i.e. a SGWB that can be derived as the proper convolution of a generic $\mathcal{P}_\zeta(f)$. Due to this additional information, the method is expected to be more sensitive to SIGW signals than other fully-agnostic approaches. It can be particularly useful for placing upper bounds on the SIGW amplitude if no signal is detected in the LISA data, or act as a key ingredient for SIGW model selection if a signal is present.

On the other hand, if the computational resources available to the mission allow running the global fit for every SGWB template, several modules for the template-based SIGW reconstruction have been conceived since the first iterations of the global fit.[2] The advantage of such a possibility is clear: the more signal characterization is included in the search, the higher the sensitivity to that signal. This process also reduces the risk of SGWB misreconstructions that the global fit might absorb into the parameter estimation of other sources. To address the points *i)* and *ii)* above within this framework, we collect several well-motivated inflationary models with known $\mathcal{P}_\zeta(k)$ predictions, we design template classes that effectively parameterize these $\mathcal{P}_\zeta(k)$, and we prototype the SIGW template-based searches for them.

Accurately performing *i)* and *ii)* enables LISA to identify the most favored SIGW models and then proceed with tasks *iii)* and *iv)*. Accordingly, we implement a prototype data analysis pipeline, choosing the USR inflationary setup as a representative example. In particular, we develop a fast numerical algorithm determining $\mathcal{P}_\zeta(k)$ once the inflationary model parameters are known. Thanks to its speed, the algorithm allows for rapid likelihood evaluations in the fundamental-parameter space, enabling direct inference on the USR model parameters from GW data.

As a proof of concept, we further perform inference on $\mathcal{P}_\zeta(k)$ in cosmological scenarios where standard assumptions on Gaussianity of curvature perturbations and on the standard thermal history of the Universe are relaxed. We evaluate the SIGW energy density sourced by non-Gaussian contributions parametrized by $\tau_{\rm NL}$, resorting to the local ansatz emerging from a perturbative expansion of scalar fluctuations. This contribution is known to modify

By periodically alternating which source type is being sampled over, imperfect source subtraction and source type confusion are properly included and fully modeled in the resulting posterior. When working after the global fit, with one sample of the residual as we assume in this work, these possible degeneracies are not fully modeled. Additionally, if the global fit has experienced a convergence failure (the MCMC is still "burning in"), unmodeled source power may still be in the data and lead to false detections of an SGWB. Ref. [59] studies how this convergence failure affects stochastic background recovery in the available prototype global fit residuals.

[2] See ref. [67] for other template-based reconstructions suitable for inflationary models.

its frequency profile compared to the Gaussian counterpart [41, 42, 46, 49, 52, 53]. Since for models of inflation with local type NG, where the curvature perturbation is dominated by one degree of freedom, $\tau_{\rm NL}$ and $f_{\rm NL}$ are related, we take advantage of such a relation to forecast the ability of LISA of probing NG, focusing the analysis on $f_{\rm NL}$ and discussing its implications for $\tau_{\rm NL}$. It is worth mentioning that, as recently argued by [68], the effects of NG on the SIGW background may not always be accurately captured by an expansion around a Gaussian field. Properly accounting for the full impact of intrinsic non-linearities may significantly suppress or enhance the spectrum compared to the predictions based on the local ansatz. Achieving this would require the development of fully non-perturbative approaches to compute the SIGW spectrum, which are beyond the scope of this work. Finally, we perform some diagnostic tests to assess whether the reconstructed signal is consistent with the SIGW hypothesis. Such tests could help to rule out a scalar-induced origin as a viable explanation for some SGWB spectral shapes.

The core of our numerical analysis is implemented in the `SIGWAY` code.[3] This stand-alone `Python` code addresses tasks *i) - iv)* for SIGW signals by offering the following functionalities:

- A fast vectorized numerical integrator for computing the SGWB resulting from any spectrum of primordial curvature fluctuations of modes reentering the Hubble radius during radiation domination or a phase of early matter domination.
- An integration algorithm for computing the SGWB assuming a binned spectrum of $\mathcal{P}_\zeta(k)$ for agnostic reconstructions of $\mathcal{P}_\zeta$.
- Solvers for the background- and perturbation equations of motion for the inflaton in a single-field scenario that can be called by the SIGW integrator starting from the inflaton Lagrangian.
- Capabilities for computing the SGWB including non-Gaussian contributions for a lognormal shape of $\mathcal{P}_\zeta$.
- Functionality for pairing to the `SGWBinner` pipeline [61, 62] for computing the LISA likelihood, and performing inference on the parameters governing the primordial curvature fluctuations.

The paper is organized as follows. In Sec. 2 we review some representative models predicting enhanced power spectra of curvature perturbations. In Sec. 3 we identify functional forms that describe the shapes of the aforementioned spectra in terms of effective parameters. In Sec. 4 we describe the analytic and numerical tools that we implement in the `SIGWAY` code. The functionality of the `SGWBinner` code that is relevant for performing inference is briefly described in Sec. 5. Sec. 6 illustrates, for representative benchmark signals, how the elements built in this work help tackle the key tasks *i) – iv)*, including testing whether the SIGW hypothesis is compatible with the putative signal in Sec. 7. Finally, Sec. 8 presents our main conclusions, while App. A and App. B discuss technicalities and subtleties regarding the proposed SIGW signal reconstruction and interpretation.

[3] https://github.com/jonaselgammal/SIGWAY

Notation. We indicate with k the comoving wavenumber while with f the associated frequency $f = k/2\pi$. For scales relevant to LISA, we translate wavenumbers in units of Hz using $c = \mathrm{Hz}\,\mathrm{Mpc}/(1.023 \times 10^{14})$. For presentation purposes, we differentiate between frequencies and momenta arbitrarily denoting them with units of Hz and s^{-1}, respectively. As usually done in the literature, we report the GW spectral energy density $\Omega_{\rm GW}$ multiplied by the rescaled Hubble rate $h = H_0/(100\,\mathrm{km/s/Mpc})$ squared. Finally, we indicate vectors with bold symbols (e.g. $\vec{x} \equiv \boldsymbol{x}$) while their magnitude with the lower-case letter.

2 Early universe models leading to enhanced curvature spectra

In this section, we summarise the main classes of models predicting enhanced curvature power spectra at small scales, which can lead to potential GW signatures in LISA.

2.1 Single field inflation

In the simplest models of inflation, a single scalar field known as the inflaton moves gradually down its potential under the influence of Hubble friction, resulting in a slow-roll (SR) phase. The generated fluctuations are nearly scale-invariant, Gaussian, and adiabatic: they freeze on super-Hubble scales, producing a universe that is statistically homogeneous and isotropic [69–73].

In certain models, however, the inflaton potential in the Einstein frame can contain features such as a flat region or a mild bump that causes the field velocity to decrease rapidly in a brief USR[4] phase [74–79] followed by another SR or a constant-roll phase [80, 81]. The shape of the enhanced spectral features is determined by the amplification of the curvature perturbations during the USR phase, as well as by the specifics of the transition into the USR era. If the field accelerates significantly during that transt epoch, it can cause sizeable spectral modulation, and even be the dominant source of amplification. The resulting perturbations deviate significantly from scale invariance, exhibiting the strongest amplification for wavelengths exiting the horizon around the USR era.

Generically, models of this kind can be subdivided into four categories: *i)* Quasi-inflection points and plateaus [44, 79, 82–112]; *ii)* Upward [113, 114] or downward [114–117] steps; *iii)* Models in which the inflaton rolls through a global minimum/double-well potentials [118–123]; and *iv)* Potentials with stacked features/oscillating potentials [124–130]. It was also suggested that models going beyond a non-minimally coupled inflaton, e.g. within modified gravity theories, can introduce features in terms other than the inflaton potential or the non-minimal coupling [131–147]. See e.g. [148] for a model-building review. Even though the literature on these models is quite vast, many scenarios predict a curvature power spectrum that is enhanced at small scales with similar properties. In particular, most models, especially those in category *i)*, produce a single peak in the power spectrum that is approximately captured by a broken power law. However, models with sharp features

[4] We consider a slightly broader definition of USR, which is often characterized by $\eta_H \equiv -\ddot{H}/(2H\dot{H}) = 3$ (or $\eta \equiv \dot{\epsilon}/(H\epsilon) = -6$) and can be realized on a flat plateau. However, models considered in the literature often exhibit a small bump instead, which results in $\eta_H > 3$ due to the curvature of the potential. We will include such deviations in our definition of USR.

can produce spectral oscillations at and after the peak in the power spectrum. Such enhancement mechanisms are mainly in the categories *ii)* and *iii)*. Non-standard potentials in category *iv)* can also deviate strongly from this picture.

We finish this section with a note on the theoretical consistency of these enhancing mechanisms. There has been significant debate about the potential impact of loop corrections, induced by the enhanced modes during an USR phase, on long-wavelength scales, with some even challenging the validity of perturbative computations in these scenarios. This discussion is also relevant for the possible interpretation of a stochastic signal as originating from SIGW. The question of whether these corrections can become sufficiently large to undermine the predictive power of inflationary models related to PBH and SIGW has been also explored in Res. [149–159], while other studies have questioned the very existence of these corrections and proposed an argument against their existence [160–164].

2.2 Multi-field inflation

Hybrid/multi-field inflation. Given the high number of degrees of freedom within multi-field inflationary setups, we can separate and control more efficiently two stages responsible respectively for the generation of nearly scale-invariant primordial curvature perturbations on large CMB scales, and for enhanced curvature perturbations on small scales which lead to the production of SIGWs. Hybrid/multi-field models of inflation tend to generate slowly-growing lognormal-like peaks in the curvature power spectrum [165–170] while strong deviations from a geodesic trajectory in field space may lead to sharp peaks and features such as spectral oscillations in $\mathcal{P}_\zeta$ [171–177].

Curvaton models Within curvaton scenarios [178], one can realize setups with the curvaton field being characterized by a steep blue spectrum either due to interactions with the inflaton or other degrees of freedom during inflation [179, 180] or due to a non-trivial kinetic term [181, 182]. In particular, within axion-like curvaton setups, the curvaton field is identified with the phase of a complex field whose modulus decreases rapidly during inflation [183–186]. We should also highlight that in any curvaton model the curvature perturbations on small scales originate from non-adiabatic curvaton field fluctuations during inflation, leading to a non-Gaussian probability distribution function for the primordial curvature perturbations [187–190] with important consequences at the level of the SIGW signal [42, 53, 191, 192] (but see also [193]).

Axion-gauge field coupling Enhanced scalar [194] and tensor [194, 195] modes can be produced by gauge fields amplified by their pseudo-scalar $\phi F\tilde{F}$ coupling with a rolling inflaton or spectator [196, 197] axion during inflation. The gauge field amplitude is exponentially sensitive to the axion velocity, thus providing naturally blue signals. These enhanced curvature modes can lead to PBH and SIGW [40, 198–200]. The precise shape of these signals is sensitive to the axion evolution, which is significantly impacted by the backreaction of the amplified gauge fields, which is recently being explored via lattice simulations [201–204].

2.3 Other classes of models

Preheating. During the preheating phase following inflation, as the inflaton field undergoes coherent oscillations around the minimum of its potential, a striking phenomenon emerges: the resonant amplification of quantum inflaton field fluctuations, which drives particle production [205, 206]. These enhanced quantum fluctuations are accompanied by a resonant amplification of the scalar metric fluctuations (usually quoted as metric preheating [207–210]), or, in other terms, with enhanced curvature perturbations, responsible for the generation of SIGWs [211] and potentially for PBH formation [212–214] (see however [215] for an assessment on the role of non-linearities and anharmonicities). Most studies have focused on multi-field inflationary setups since in such scenarios the enhancement of entropic (isocurvature) fluctuations can give rise to the enhancement of the adiabatic/curvature fluctuations in the broad resonance regime [216–221]. This leads to a notable amplification of the primordial curvature power spectrum, deviating from the standard scale-invariant behavior at small scales [222–224]. Interestingly, recent works also suggest that a parametric amplification of the curvature perturbations can occur even in the narrow regime in the case of single field inflation [212–215, 225]. We should also note that these gravitational waves are typically peaked at MHz or GHz frequencies far above those of LISA, although some scenarios do allow for a peak in LISA's range [226].

Matter Bouncing Scenarios. In non-singular matter bouncing cosmological models [227], the matter contracting phase inevitably amplifies super-horizon curvature perturbations. This enhancement can lead to an enhanced primordial curvature power spectrum on small scales compared to the ones probed by CMB. As these perturbations cross the cosmological horizon, either during the contracting phase [228–230] or the expanding Hot Big Bang phase [231, 232], they can lead to the abundant production of SIGWs.

Early PBH domination. At distances much larger than the mean PBH separation length, a population of PBHs can be viewed as an effective pressureless fluid. One can then treat this PBH fluid within the context of cosmological perturbation theory showing that the PBH energy fluctuations are isocurvature in nature [233, 234] and can convert to adiabatic curvature perturbations in an early matter-dominated era driven by light PBHs ($m_{\rm PBH} < 10^9$g) occurring before BBN. Interestingly enough, these PBH-induced curvature perturbations can source abundant SIGWs detectable by GW observatories [233–235]. Notably, these PBH associated SIGWs [236, 237] can serve as a novel portal to probe primordial non-Gaussianities (NGs) [238, 239] at small scales ($k > {\rm Mpc}^{-1}$) as well the underlying gravity theory [240–242] and Hawking evaporation [243–250].

2.4 A simple benchmark scenario: single field USR

In this work, we consider one of the simplest realizations of the single-field scenarios discussed in the previous section. This class of models is described by the inflaton potential in the Einstein frame, $V(\phi)$. The corresponding action can be written as

$$\mathcal{S} = \int {\rm d}^4x\sqrt{-g}\left(\frac{1}{2}M_{\rm P}^2 R - \frac{1}{2}(\partial^\mu\phi)^2 - V(\phi)\right), \tag{2.1}$$

where R is the Ricci scalar and $M_{\rm P}$ is the reduced Planck mass. Assuming a flat FLRW background geometry $\mathrm{d}s^2 = -\mathrm{d}t^2 + a^2\mathrm{d}x_i^2$, where a is the scale factor, the background evolution is governed by the Friedmann equation (dots indicate time derivatives)

$$3M_{\rm P}^2H^2 = \dot{\phi}^2/2 + V(\phi)\,, \tag{2.2}$$

with $H = \dot{a}/a$, and the Klein-Gordon equation (a prime denotes a derivative with respect to the field)

$$\ddot{\phi} + 3H\dot{\phi} + V'(\phi) = 0\,. \tag{2.3}$$

In order to produce an enhancement of perturbations at LISA scales, and at the same time comply with CMB bounds at large scales, the inflationary potential should feature a shallower region or an inflection point, which breaks the SR evolution exponentially decelerating the field velocity.

In the class of models considered here, the dynamics can be understood in relatively simple terms. In SR, the inflaton evolves with a negligible acceleration, and the SR solution gives $\dot{\phi} = -V'/(3H)$. As the inflaton begins to approach the inflection point, the SR conditions are violated primarily due to a rapid change in the second SR parameter. Having almost reached the local maximum ϕ_*, the inflaton will spend $\mathcal{O}(10)$ e-folds crossing it and its evolution is thus dictated by $\ddot{\phi} + 3H\dot{\phi} + \eta_V(\phi_*)H^2(\phi - \phi_*) \simeq 0$, where $\eta_V \equiv M_{\rm P}^2V''/V$ denotes the second potential SR parameter. The two solutions of this equation describe two phases: First, a USR-like phase in which the inflaton rapidly decelerates, which leads to an amplification of the power spectrum. Second, a subsequent constant roll or a SR phase that is dual to the initial USR-like phase [81, 251].

The linear superposition of these solutions describes a smooth transition between these epochs. The second SR parameter $\eta_V(\phi_*)$ determines the spectral slope after the peak $n_s - 1 = 3(1 - \sqrt{1 - (4/3)\eta_V(\phi_*)})$. Thus, as exact USR ($\eta_V(\phi_*) = 0$) would produce a scale-invariant spectrum at scales *above* the spectral peak, the violation of scale invariance in the UV is directly related to the deviation from an exact USR.

As an example, we consider the potential given by the rational function proposed in [82] (see also [85, 252])

$$V(\phi) = \frac{\lambda}{12}\phi^2(vM_{\rm P})^2\left(6 - 4b_l\frac{\phi}{vM_{\rm P}} + 3\frac{\phi^2}{(vM_{\rm P})^2}\right)\left(1 + b\frac{\phi^2}{(vM_{\rm P})^2}\right)^{-2}\,. \tag{2.4}$$

The presence of an inflection point is enforced by setting

$$b = (1 + b_f)\left[1 - \frac{b_l^2}{3} + \frac{b_l^2}{3}\left(\frac{9}{2b_l^2} - 1\right)^{2/3}\right]\,, \tag{2.5}$$

where we included a tuning parameter b_f allowing for deviation from perfect inflection points (with $b_f > 0$ the inflection point becomes a shallow minimum). The field ϕ appearing in the action (2.1) is canonically normalized, and minimally coupled to gravity. This is a proxy for more realistic models in which the inflaton field has a quartic potential and couples non-minimally to gravity via a $\xi R\phi^2$ term, see e.g. [83, 84, 99, 253]. After moving

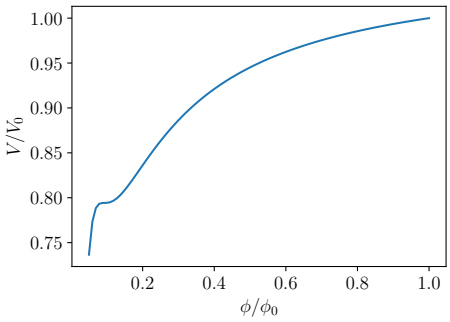


Figure 1. Example of a single field model (2.4), with our benchmark parameters (2.6), leading to an early SR phase consistent with CMB data as well as an USR phase which leads to an enhancement of perturbations within the LISA frequency range.

to the Einstein frame, the factor in the denominator appears, which flattens the potential at large field values. In this case, one would further need to canonically normalize the field, and possibly add logarithmic corrections to the coefficients of the monomials in Eq. (2.4). We do not discuss the origin of such a potential, as our goal is solely to provide a simple representative model to work with.

We define our benchmark potential by choosing the following parameters (close to the ones used in [252])

$$\begin{aligned} \lambda &= 1.4731 \times 10^{-6}, \qquad v = 0.19688, \\ b_l &= 0.71223, \qquad b_f = 1.87 \times 10^{-5}, \end{aligned} \tag{2.6}$$

leading to good agreement with CMB (within 3σ of current Planck 2018 data [254]) and at the same time to a peak of the curvature spectrum: $\mathcal{P}_\zeta(k_{\rm peak}) \simeq 10^{-3}$ at LISA scales. In Eq. (2.6) we report 5 significant digits because of the required tuning of the USR potential [252]. In Sec. 3.3 we describe how to compute the spectrum of curvature perturbations in detail. The benchmark potential is depicted in Fig. 1, where we have arbitrarily normalized the axes using the initial values $\phi_0 = 3M_{\rm P}$ and $V_0 = 2.3 \cdot 10^{-10} M_{\rm P}^4$, which are set well before the SR phase that governs the CMB scales.

3 Modeling the curvature power spectrum

Throughout this paper, we describe the metric as a small perturbation of the FLRW metric in the longitudinal (conformal Newtonian) gauge

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu = -a^2(1+2\Phi)\mathrm{d}\eta^2 + a^2\left[(1-2\Phi)\delta_{ij} + \frac{1}{2}h_{ij}\right]\mathrm{d}x^i \mathrm{d}x^j, \tag{3.1}$$

where η is the conformal time, and we neglect vector perturbations and anisotropic stress (and so we can identify the two scalar Bardeen potentials, $\Phi = \Psi$).

At linear order in perturbation theory, the time and momentum dependence of the scalar potential Φ can be factored out by introducing the transfer function $T(\eta, k)$: the

Bardeen's potential is related to the gauge-invariant comoving curvature perturbation $\zeta(\boldsymbol{k})$ by

$$\Phi(\eta, \boldsymbol{k}) = \frac{3+3w}{5+3w} T(\eta, k)\zeta(\boldsymbol{k}) \tag{3.2}$$

where w is the equation of state characterizing the fluid dominating the energy density in the universe. In a radiation-dominated universe, $w = 1/3$, and the prefactor becomes $2/3$. As long as a mode k is super-Hubble ($k\eta < 1$), its evolution is frozen and the transfer function tends to 1. For a homogeneous and isotropic universe, the two-point function of curvature perturbations reads

$$\langle \zeta(\boldsymbol{k}_1)\zeta(\boldsymbol{k}_2)\rangle \equiv (2\pi)^3\delta(\boldsymbol{k}_1+\boldsymbol{k}_2)P_\zeta(k_1) \equiv (2\pi)^3\delta(\boldsymbol{k}_1+\boldsymbol{k}_2)\frac{2\pi^2}{k_1^3}\mathcal{P}_\zeta(k_1), \tag{3.3}$$

where we introduced the dimensionful power spectrum $P_\zeta(k)$, and the dimensionless one $\mathcal{P}_\zeta(k)$. We adopt different methodologies to model the primordial curvature power spectrum $\mathcal{P}_\zeta(k)$ generated in various inflationary models, by:

- Developing a model-independent parametrization of the spectrum, enabling an agnostic reconstruction approach;
- Constructing analytical templates for $\mathcal{P}_\zeta(k)$. These templates are motivated by specific scenarios and are consequently less flexible than the previous case. However, given the lower number of parameters controlling the templates as well as their simple analytical descriptions, they alleviate the computational challenges posed by models which would require expensive computation of the curvature power spectrum;
- Establishing a robust pipeline for the computation of primordial curvature power spectra $\mathcal{P}_\zeta(k)$ for the benchmark model described in Sec. 2.

3.1 Model independent parameterization: binned spectrum approach

In this section, we discuss a parameterization of the power spectrum $\mathcal{P}_\zeta$ in terms of a *binned spectrum approach*, which is useful in developing a template-independent procedure for computing the SIGW. We represent the curvature power spectrum as a sum of discrete components defined within small momentum intervals. Specifically, we decompose $\mathcal{P}_\zeta$ as a sum of $N-1$ top hat functions as

$$\mathcal{P}_\zeta(p) = \sum_{i=1}^{N-1} A_i \Theta(p-p_i)\Theta(p_{i+1}-p)\,. \tag{3.4}$$

The coefficients A_i – which can be approximated as constants – represent the amplitude of the spectrum within the i-th bin. The latter is defined within the momentum boundaries p_i and p_{i+1}, using the Heaviside step functions $\Theta(p)$.[5] Any curvature spectrum $\mathcal{P}_\zeta$ can then be associated with a specific vector A_i of the coefficients appearing in Eq. (3.4).

[5] One could extend this approach to include a tilt of the spectrum at each bin and enforcing continuity for adjacent bins, at the cost of doubling the parameters controlling the power in each bin. We leave this, and other possible extensions for future work.

Therefore, our computation does not require an *a priori* choice of a particular template for the momentum dependence of $\mathcal{P}_\zeta(p)$.

The SIGW depends quadratically on $\mathcal{P}_\zeta$ using convolution integrals. Hence, we can expect that the ansatz (3.4) allows us to express the SIGW in terms of the double sum of (model-dependent) vector components A_i, contracted over a general, model-independent matrix kernel. In Sec. 4.4 we elaborate a simple procedure aimed at performing this task.

3.2 Analytical templates for curvature spectra

We divide the list of templates into classes depending on their properties. We can define "smooth shapes" and shapes with features.

3.2.1 Smooth templates

We divide this class of templates into two cases:

Lognormal (LN). A typical class of spectral peaks can be characterized by a LN-shape

$$\mathcal{P}_\zeta^{\rm LN}(k) = \frac{A_s}{\sqrt{2\pi\Delta^2}}\,\exp\left[-\frac{1}{2\Delta^2}\log^2(k/k_*)\right]\,. \tag{3.5}$$

Such spectra appear e.g. in a subset of hybrid inflation and curvaton models, as well as from axion-gauge field coupling, see the discussion in Sec. 2. This template can describe scenarios in which the peak is typically narrow and symmetric in log space. Interestingly, this template allows for an analytic derivation of the GW power spectrum in the broad/narrow peak approximations $\Delta \gg 1$ or $\Delta \ll 1$ (see Ref. [255] and improvements in [256]). We will consider the following benchmark scenario choosing

$$\log_{10} A_s = -2.50, \quad \log_{10}\Delta = \log_{10}(0.5) \approx -0.301, \quad \log_{10}\left(k_*/{\rm s}^{-1}\right) = -2.00\,. \tag{3.6}$$

Broken power law (BPL). Another broad class of spectra is encountered, for instance, in single field inflation and curvaton models and can be described by a BPL

$$\mathcal{P}_\zeta^{\rm BPL}(k) = A_s\frac{(\alpha+\beta)^\gamma}{\left(\beta\,(k/k_*)^{-\alpha/\gamma} + \alpha\,(k/k_*)^{\beta/\gamma}\right)^\gamma}, \tag{3.7}$$

where $\alpha, \beta > 0$ describe respectively the growth and decay of the spectrum around the peak, while γ is an $\mathcal{O}(1)$ model dependent parameter. The normalization is such that $\mathcal{P}_\zeta^{\rm BPL}(k) = A_s$ at the peak. This template provides a very close approximation to the shape of $\mathcal{P}_\zeta$ obtained from single-field USR scenarios. There, one typically finds $\alpha \simeq 5 - |1 - 2\eta_H| \approx 4$ [80, 81, 257], where η_H is the second Hubble SR parameter (see footnote 4) *before* the USR phase, which is typically close to SR, that is, $\eta_H \ll 1$. For this template, it is possible to find an analytic GW spectrum with $\gamma = 1$, see [258]. We will use the following benchmark with parameters

$$\begin{aligned} \log_{10} A_s = -2.71, \qquad \log_{10}(k_*/{\rm s}^{-1}) = -1.58, \\ \alpha = 3.11, \qquad \beta = 0.221, \qquad \gamma = 1.25. \end{aligned} \tag{3.8}$$

which provides a curvature power spectrum whose peak is within LISA's sensitivity and that fits the spectrum produced in the ab initio USR scenario described below (see Sec. 3.3).

3.2.2 Templates with oscillations

Oscillations in the primordial power spectrum have been extensively studied to seek for deviations from the standard SR inflationary paradigm mainly at scales relevant for CMB or large-scale structure (see e.g. [34, 38, 259, 260]). These oscillations, denoted as primordial features, typically arise due to a sudden transition during inflation [261], occurring over a short time-scale of the order of one e-fold. SIGW provides an opportunity to probe such primordial features at small-scales ($\ll$ Mpc) [173, 175, 262–264]. Specifically, if the mechanism responsible for the enhancement in $\mathcal{P}_\zeta$ is active when modes are sufficiently deep inside the horizon, the resulting spectrum exhibits oscillations of order one. Due to their large amplitude, these oscillations could potentially leave their imprints and be detected in the SGWB. This phenomenon can occur in both single-field and multi-field inflationary models. As benchmarks for these large amplitude oscillations, we will consider a small-scale feature induced by a genuine multi-field mechanism and, secondly, weaker oscillations arising from a rapid transition between SR and USR phases in single-field inflation.

Turns in multi-field inflation. In multi-field inflationary setups, a common phenomenon is the presence of turns in the field space, being equivalent to the inflationary trajectory deviating from a geodesic in field space. This bending is quantified through the parameter $\eta_\perp$, measuring the acceleration of the inflationary trajectory perpendicular to its direction [265, 266] or equivalently the deviation of the trajectory from a geodesic in the target field space. One then can show that sharp and strong (large $\eta_\perp$) turns can lead to the following curvature power spectrum, modulated by order-one rapid oscillations[6] [171–173]

$$\mathcal{P}_\zeta^{\rm ST}(\kappa) = \mathcal{P}_\zeta^{\rm env}(\kappa)\left[1+(\kappa-1)\cos\left(2e^{-\frac{\delta}{2}}\eta_\perp\kappa\right)+\sqrt{(2-\kappa)\kappa}\,\sin\left(2e^{-\frac{\delta}{2}}\eta_\perp\kappa\right)\right]\Theta(2-\kappa)\,, \tag{3.9}$$

and the envelope

$$\mathcal{P}_\zeta^{\rm env}(\kappa) = A_s\exp\left(-2\eta_\perp\delta\right)\exp\left[2\sqrt{(2-\kappa)\kappa}\,\eta_\perp\delta\right]/\left[4(2-\kappa)\kappa\right]\,, \tag{3.10}$$

where A_s denotes the amplitude of the power spectrum in the absence of transient instability and $\kappa \equiv k/k_*$ with k_* being associated with the maximally enhanced scale, deep inside the cosmological Hubble sphere at the time of the sharp turn. The parameter δ is the duration in e-folds of the turn. $\delta \gtrsim \log\eta_\perp$ stands for broad turns and $\delta \lesssim \log\eta_\perp$ stands for sharp turns. Finally, $\Theta(x)$ denotes the Heaviside theta function.

We define a parameterization in which the oscillations are switched off via the parameter $F \in [0,1]$, continuously interpolating between (3.9) when $F=1$ and (3.10) when $F=0$

$$\mathcal{P}_\zeta(k) = F\mathcal{P}_\zeta^{\rm ST}(k) + (1-F)\mathcal{P}_\zeta^{\rm env}(k). \tag{3.11}$$

We consider the benchmark scenario whose parameters are given by

$$\begin{gathered}\log_{10}A_s = -1.5, \quad \log_{10}\left(k_*/s^{-1}\right) = -1.5,\\ \delta = 0.5, \quad \eta_\perp = 14, \quad F = 1.\end{gathered} \tag{3.12}$$

[6] Equation (3.9) assumes the entropic field to be massless during the turn. Generalized expressions for other mass choices can be found in [173], where the qualitative features remain analogous.

Rapid transitions between SR and USR phases. In most USR/inflection-point scenarios, the transition from the initial SR phase to the USR-like phase depends on the properties of the potential. In the presence of sufficiently sharp transitions between these phases the spectrum of curvature perturbations deviates from a BPL profile due to oscillatory features.

A fully analytic spectrum can be obtained when the initial SR to USR transition is instantaneous. In that case, the peak can be described as (Eq. (3.8) in [81])

$$\mathcal{P}_\zeta(k) = \frac{\kappa^2}{4\pi}\Bigg| -\frac{\Gamma\left(1+\nu_{\rm II}\right)}{\zeta_1}\left(\kappa/2\right)^{-\nu_{\rm II}+\frac{1}{2}} J_{\nu_{\rm II}}\left(\kappa\right) H_{\nu_{\rm I}}\left(\kappa\right) \\ + \frac{\Gamma\left(\nu_{\rm II}\right)}{\zeta_2}\left(\kappa/2\right)^{-\nu_{\rm II}+\frac{3}{2}}\left[J_{\nu_{\rm II}}\left(\kappa\right) H_{\nu_{\rm I}-1}\left(\kappa\right) + J_{\nu_{\rm II}+1}\left(\kappa\right) H_{\nu_{\rm I}}\left(\kappa\right)\right]\Bigg|^2, \tag{3.13}$$

where, $\kappa = k/k_*$, $H_\nu(x)$ and $J_\nu(x)$ denote the Hankel- and Bessel functions of the first kind, respectively, k_* is the scale of the mode that exits the Hubble sphere during the SR to USR transition, and $\nu_{\rm I}$ and $\nu_{\rm II}$ are related to the spectral slopes in the attractor phases before and after the transition as $n_s = 2(2-\nu)$. The first line does not contribute to the peak and can be omitted in case there is no sensitivity to spectral features away from the peak. The parameters ζ_1, ζ_2 control the amplitude at the IR (or CMB) scales and the peak, respectively.

This spectrum resembles a broken power law with a modulation around the peak. This modulation has a period of $2k_*$ and is damped as $1/k$. It is generated in an instantaneous SR to USR transition and is typically suppressed or removed when the transition is non-instantaneous [81, 267, 268]. In this way, such oscillations carry information about the evolution of the inflaton during the transition. Moreover, spectral oscillations can be greatly enhanced in some cases. For instance, when the inflaton rolls through a deep minimum before entering the USR phase [118–123].

We can test for the sensitivity of LISA to resolve these oscillations. To this aim, we only consider the second line of the spectrum (3.13), which describes the peak, and consider a generalized form[7]

$$\mathcal{P}_\zeta^{\rm osc}(k) = F\mathcal{P}_\zeta^{\rm osc,B}(k) + (1-F)\mathcal{P}_\zeta^{\rm osc,BPL}(k), \tag{3.14}$$

where $F \in [0,1]$ and with

$$\mathcal{P}_\zeta^{\rm osc,B}(k) = \pi^2\kappa^{5-2\nu_{II}} A_s \left|J_{\nu_{II}}(2\kappa)H^{(1)}_{\nu_I-1}(2\kappa) + J_{\nu_{II}+1}(2\kappa)H^{(1)}_{\nu_I}(2\kappa)\right|^2, \tag{3.15}$$

$$\mathcal{P}_\zeta^{\rm osc,BPL}(k) = A_s\left[\left(\kappa^{7/2-\nu_I}\frac{(\nu_I+\nu_{II})\Gamma(\nu_I-1)}{\Gamma(\nu_{II}+2)}\right)^{-\gamma} + \left(\kappa^{3/2-\nu_{II}}\right)^{-\gamma}\right]^{-2/\gamma}. \tag{3.16}$$

The rewriting in terms of the two contributions in Eqs. (3.15) and (3.16) is equivalent to (3.13) when imposing $F = 1$, but it is done to separate the smooth BPL contribution from

[7] For notational simplicity, the template uses a different normalization and scaling of the argument than (3.13).

the one including the oscillations. This way, changing F smoothly transitions between an oscillating ($F = 1$) and a non-oscillating $F = 0$ power spectra.

The shape of the BPL template is recovered using (3.7) with $\nu_I = (7 - \alpha)/2$ and $\nu_{II} = (3 + \beta)/2$ when $\alpha \leq 5$. To obtain an exact match with (3.7) the normalization and the momenta must also be rescaled so that $\mathcal{P}_\zeta^{\rm BPL}(k) = n\mathcal{P}_\zeta^{\rm osc,BPL}(bk)$, where

$$n \equiv b^\beta (1 + \beta/\alpha)^\gamma \,, \qquad b \equiv \left[\frac{4(\alpha/\beta)^\gamma \Gamma\left((\beta+7)/2\right)^2}{(-\alpha+\beta+10)^2 \Gamma\left((5-\alpha)/2\right)^2} \right]^{\frac{1}{\alpha+\beta}} \tag{3.17}$$

or, equivalently, as $A_s \to nA_s$ and $k_* \to k_*/b$. The benchmark scenario corresponds to

$$\begin{aligned} \log_{10} A_s = -2.58 \,, \quad \log_{10}(k_*/\mathrm{s}^{-1}) = -2.02 \\ \nu_I = 1.95 \,, \quad \nu_{II} = 1.61 \,, \quad \gamma = 1.67 \,, \quad F = 1. \end{aligned} \tag{3.18}$$

We fix the parameters to match the BPL template in the limit $F = 0$.

3.3 Computation of $\mathcal{P}_\zeta$ in single field USR scenarios

The benchmark model we introduced in Sec. 2.4 is based on a single-field model of inflation featuring a phase of USR. We now describe in detail how to compute the spectrum of curvature perturbations using linear perturbation theory.

As a warm-up, let us define the system of equations in terms of dimensionless variables rescaled to the corresponding relevant quantities. This typically improves the numerical stability of a code computing the spectrum of curvature fluctuations. We define

$$x \equiv \phi/M_{\rm P}, \qquad U(x) \equiv V(\phi)/V_0, \tag{3.19}$$

where we introduce the suffix 0 to indicate the quantities evaluated at the initial conditions of the background evolution. We can also define the dimensionless time and Hubble rate using

$$\tau \equiv tV_0^{1/2}/M_{\rm P}, \qquad h \equiv HM_{\rm P}/V_0^{1/2}. \tag{3.20}$$

We can now change the evolution variable from time to the number of e-folds N, defined as $\mathrm{d}N \equiv H\mathrm{d}t = h\mathrm{d}\tau$, setting $N = 0$ at τ_0. In this way, the background equations of motion (2.2) and (2.3) become

$$y' + 3\left[1 - \frac{y^2}{6}\right] y + \frac{U_{,x}}{h^2} = 0, \tag{3.21a}$$

$$x' - y = 0, \tag{3.21b}$$

$$h' + \frac{(x')^2}{2} h = 0, \tag{3.21c}$$

where prime denotes derivatives with respect to N, while $U_{,x} \equiv \mathrm{d}U(x)/\mathrm{d}x$. The initial conditions for the inflaton and the Hubble parameters can be found assuming that initially, SR is satisfied, which for a given initial x_0 leads to

$$y_0 = -\frac{3}{U_{0,x}} \left(\sqrt{1 + \frac{2}{3} U_{0,x}^2} - 1 \right), \qquad h_0 \equiv \frac{1}{\sqrt{6}} \left(\sqrt{1 + \frac{2}{3} U_{0,x}^2} - 1 \right)^{1/2}. \tag{3.22}$$

Finally, in the code, we keep track of the equation of state during inflation $w_{\rm inf} \equiv p_\phi/\rho_\phi$, which can be written in our notation as $w = [(x')^2 - 3]/3$. As inflation stops when the equation of state becomes larger than $w > -1/3$, we stop the evolution when $(x')^2 = 2$.

The resulting inflationary background can be described by the evolution of the Hubble rate H. This is dictated by dynamical equations relating H to the SR parameters, which are defined as

$$\epsilon_H \equiv -\frac{\dot{H}}{H^2} = \frac{1}{2M_{\rm P}^2}\left(\frac{d\phi}{dN}\right)^2 \equiv \frac{y^2}{2}, \tag{3.23}$$

$$\eta_H \equiv -\frac{\ddot{H}}{2H\dot{H}} = \epsilon_H - \frac{1}{2}\frac{d\log\epsilon_H}{dN} = \frac{y^2}{2} - \frac{y'}{y}, \tag{3.24}$$

where we introduced $y = x'$. Notice that our definition of η_H differs from another definition often used in the literature, which is expressed as $\dot{\epsilon}_H/(\epsilon_H H)$. During the USR phase, this definition equals approximately $-2\eta_H$ neglecting ϵ_H corrections.

As long as the SR approximation is valid, i.e. $\epsilon_H \ll 1$ and $\eta_H \ll 1$, the power spectrum of curvature perturbations (see e.g. [269]) is given by

$$\mathcal{P}_\zeta(k) = \frac{H_k^2}{8\pi^2 M_{\rm P}^2 \epsilon_{H,k}} = \frac{V_0}{M_{\rm P}^4}\left(\frac{h_k}{2\pi y_k}\right)^2, \tag{3.25}$$

where the suffix k indicates that these quantities are evaluated at Hubble crossing $k = aH$. Consequently, one can also show that the scalar spectral index $n_s - 1 \equiv \mathrm{d}\ln\mathcal{P}_\zeta/\mathrm{d}\ln k$ and the tensor-to-scalar ratio $r \equiv \mathcal{P}_h/\mathcal{P}_\zeta$ are given by the well-known expressions

$$n_s \approx 1 - 4\epsilon_H + 2\eta_H, \tag{3.26}$$

$$r \approx 16\epsilon_H, \tag{3.27}$$

at leading order in the SR parameters.

The approximations above can not be used in the context of USR scenarios, as the large deceleration of the inflaton velocity causes $\eta_H \sim \mathcal{O}(1)$ ($\eta_H = 3$ in the limit of exponential deceleration). We go beyond the SR approximation by solving the Mukhanov-Sasaki (MS) equation [270, 271]. Introducing momentarily the conformal time η such that $\mathrm{d}\eta \equiv \mathrm{d}t/a$, we can define the MS variable $v \equiv a\delta\phi$ in terms of the inflaton perturbations in the spatially flat gauge, which satisfies the EoM

$$\frac{\partial^2 v}{\partial\eta^2} + \left(k^2 - \frac{1}{z}\frac{\partial^2 z}{\partial\eta^2}\right)v = 0, \tag{3.28}$$

where $z \equiv a^2(\partial\phi/\partial\eta)/(\partial a/\partial\eta)$. We can relate the curvature perturbation to v at linear order in perturbation theory using $\zeta = v/z = H\delta\phi/(\partial\phi/\partial t) = \delta x/y$. Assuming initial adiabatic vacuum, each mode k is fixed in the asymptotic past at $\eta_{\rm in} \ll -1/k$ as

$$v(\eta) = e^{-ik(\eta-\eta_{\rm in})}/\sqrt{2k}, \tag{3.29}$$

which translates into boundary conditions for Eq. (3.28) of the form

$$v_{\rm in} = \frac{1}{\sqrt{2k}}, \qquad \frac{\partial v_{\rm in}}{\partial\eta} = -i\sqrt{\frac{k}{2}}. \tag{3.30}$$

One can improve the stability of the numerical computation by defining a rescaled variable $\widetilde{\delta\phi} \equiv a_{\rm in} e^{ik(\eta-\eta_{\rm in})}\delta\phi\sqrt{2k}$ and the corresponding dimensionless quantity $\widetilde{\delta x} \equiv \widetilde{\delta\phi}/M_{\rm P}$. The prefactor absorbs the sub-Hubble oscillations, simplifying the time evolution of the sub-Hubble phase. Next, we introduce the dimensionless momentum $\kappa \equiv kM_{\rm P}/V_0^{1/2}$. Finally, moving to the number of e-folds as the time variable, the MS equation becomes

$$\frac{{\rm d}^2\widetilde{\delta x}}{{\rm d}N^2} + \left(3 - \frac{1}{2}y^2 - \frac{2i\kappa}{e^N h}\right)\frac{{\rm d}\widetilde{\delta x}}{{\rm d}N} + \left(\frac{U_{,xx}+2U_{,x}y}{h^2} + 3y^2 - \frac{1}{2}y^4 - \frac{2i\kappa}{e^N h}\right)\widetilde{\delta x} = 0, \tag{3.31}$$

to be solved with initial conditions $\widetilde{\delta\phi}_{\rm in} = 1$, and $\widetilde{\delta\phi}'_{\rm in} = -1$, while the curvature power spectrum is extracted using

$$\mathcal{P}_\zeta(k) = \frac{1}{4\pi^2}\frac{V_0}{M_{\rm P}^4}\left(\frac{\kappa}{e^{N_{\rm in}}}\right)^2 \frac{|\widetilde{\delta x}|^2}{y^2}. \tag{3.32}$$

In Fig. 2 we show the background evolution of $x(N)$, $y(N)$ and $h(N)$ obtained in the benchmark scenario (2.6). For convenience, we define $N_{\rm end}$ as the number of e-folds at the end of inflation. In the same plot (top panels), we report the evolution of the Hubble SR parameters and the curvature power spectrum. For simplicity, we identify the momentum with the corresponding time of Hubble crossing $k = a(N)H(N)$. On top, we also indicate the momentum in units of $1/s$. The CMB reference scale crosses the Hubble sphere $N - N_{\rm end} = -58$ e-folds before the end of inflation, which is indicated with a dashed vertical line. In the top panel, the curvature power spectrum is shown both using the SR approximation (3.25) (blue line), which however fails to reproduce the spectrum around the peak where the USR phase takes place. We show with an orange line the full spectrum computed solving Eq. (3.31). At $k = 0.05/{\rm Mpc}$ where the SR expressions (3.25)– (3.27) are valid, we find $\mathcal{P}_\zeta = 2.12 \times 10^{-9}$, $n_s = 0.952$, and $r = 0.00726$. These numbers are compatible with the latest observational bounds [254, 272]

$$\mathcal{P}_\zeta = (2.10 \pm 0.03) \times 10^{-9}, \quad n_s = 0.9649 \pm 0.0042, \quad r < 0.036 \tag{3.33}$$

at the scale $k_{\rm ref} \equiv 0.05/{\rm Mpc}$, reported here at 68% CL for A_s and n_s, and at 95% for r. The low value of n_s is a rather common feature of models featuring an USR phase not sufficiently far from the region of the potential controlling the CMB scales [83, 84, 99, 253], as it is the case if one considers enhancements in the LISA band. This benchmark USR scenario leads to a power spectrum within the LISA band which can be fitted with a BPL template with parameters (3.8).

Let us mention here that it was recently suggested that USR dynamics, in the extremal case of large spectral enhancement with $\mathcal{P}_\zeta \sim \mathcal{O}(10^{-2})$ leading to PBH formation, may violate perturbativity and induce loop corrections that could also affect much longer modes associated with the scales observed through the CMB [149]. While the existence and magnitude of this effect for soft modes is still under debate [128, 150–156, 158–164, 273–282], recent analyses show that for realistic transitions into and out of USR perturbativity is retained (see e.g. for analytical [152, 155, 159] and lattice results [283]). In this work, we restrict to adopting linear perturbation theory to compute the $\mathcal{P}_\zeta$ at the LISA scales and leave further refinements including loop corrections to the spectrum of curvature perturbations for future work.

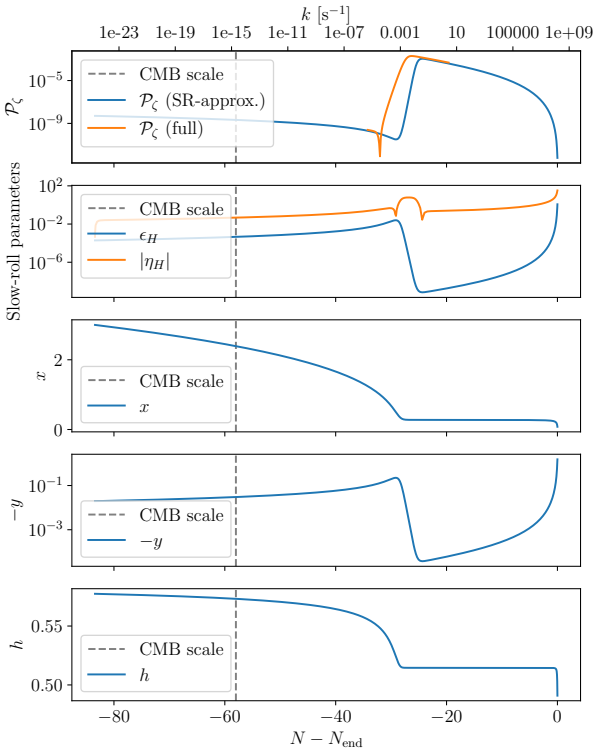


Figure 2. From top to bottom: curvature power spectrum $\mathcal{P}_\zeta$; Hubble SR parameters ϵ_H and η_H; dimensionless inflaton field x; dimensionless inflaton field velocity y; rescaled Hubble parameter h. In the second panel, η_H is initially small and negative and it transitions to positive values close to $\eta_H \sim 3$ during the USR phase. All quantities are plotted as a function of the number of e-folds to the end of inflation. The vertical dashed lines indicate the epoch when the CMB pivot scale $k_{\rm ref} = 0.05/\text{Mpc}$ crossed the Hubble sphere.

Non-Gaussianities. As a final note, let us comment on the NGs expected in this benchmark scenario. In this class of USR models, the peak in $\mathcal{P}_\zeta$ is generated during a brief phase of USR inflation, which transitions into constant-roll (i.e. constant η_H) inflation afterwards [81, 284, 285]. In these scenarios, NGs are controlled by the details of the transition between USR and the subsequent phase, which is related to the UV tilt of the spectrum [93, 286–288]. We can expect the non-linear curvature perturbation to take the form

$$\zeta = -\frac{2}{\beta}\log\left(1 - \frac{\beta}{2}\zeta_{\rm G}\right) = \zeta_{\rm G} + \frac{3}{5}f_{\rm NL}\zeta_{\rm G}^2 + \dots, \tag{3.34}$$

with $f_{\rm NL} = 5\beta/12 \simeq 0.092$ in our benchmark scenario (3.8). Note, however, that this only takes into account the local generation of non-Gaussianity on super-Hubble scales by considering Gaussian inflaton fluctuations, while it has recently been shown using simulations that nonlinear interactions can also generate a large amount of non-Gaussianity intrinsic to the inflaton [283]. Finally, let us mention that such small NG for perturbations of the typical amplitude considered here would lead to negligible contributions to the SIGW spectrum (see more discussion in Sec. 4.5), even beyond standard perturbation theory [68].

4 Computation of the scalar-induced GW background

Primordial scalar perturbations are frozen during their super-Hubble evolution; only when re-entering the Hubble radius after inflation ends do they start evolving in time. Moreover, even if scalar, vector and tensor modes are independent at the first order in perturbation theory, they couple when going to higher orders in fluctuations. For example, scalar modes source tensor modes and thus produce GWs.[8].

In the metric defined in Eq. (3.1), we ignore tensor perturbations generated at first order $h_{ij}^{(1)}$ and consider scalar perturbations that act as a source of second order tensor modes at $h_{ij}^{(2)}$. The exact evolution of $h_{ij}^{(2)}$ can be obtained from the spatial part of the Einstein equations after applying the projection tensor $\mathcal{T}_{ij}^{lm}$, which selects the transverse-traceless component. In the absence of anisotropic stress, at second order (note that we drop the superscript indicating the order in perturbation theory from now on) one obtains [15, 16, 18, 19]

$$h_{ij}''(\eta,\boldsymbol{x}) + 2\mathcal{H}h_{ij}'(\eta,\boldsymbol{x}) - \nabla^2 h_{ij}(\eta,\boldsymbol{x}) = -4\mathcal{T}_{ij}{}^{lm}\mathcal{S}_{lm}(\eta,\boldsymbol{x}), \tag{4.1}$$

where $'$ is the derivative with respect to conformal time η, $\mathcal{H} = a'/a$ denotes the conformal Hubble parameter, and $\mathcal{S}_{ij}$ is the source term

$$\mathcal{S}_{ij}(\eta,\boldsymbol{x}) = 4\Phi\partial_i\partial_j\Phi + 2\partial_i\Phi\partial_j\Phi - \frac{4}{3(1+w)}\partial_i\left(\frac{\Phi'}{\mathcal{H}} + \Phi\right)\partial_j\left(\frac{\Phi'}{\mathcal{H}} + \Phi\right). \tag{4.2}$$

In the last equation w is the equation-of-state parameter of the Universe, and the scalar perturbation $\Phi(\eta,\boldsymbol{x})$ can be related to the gauge-invariant comoving curvature perturbation ζ. Solving Eq. (4.1) in Fourier space, as we review in Sec. 4.1, we obtain the (time-averaged) dimensionless power spectrum $\mathcal{P}_h(\eta_f,k)$ of GWs at a time η_f after the end of their production. The fractional energy density of GWs per logarithmic interval in frequency is given by

$$\Omega_{\rm GW}(\eta_f,k) \equiv \frac{\rho_{\rm GW}(\eta_f,k)}{\rho_{\rm c}(\eta_f)} = \frac{1}{24}\left(\frac{k}{\mathcal{H}(\eta)}\right)^2 \overline{\mathcal{P}_h(\eta_f,k)}\,, \tag{4.3}$$

with

$$\rho_{\rm GW}(\eta) = \int \mathrm{d}\ln k\, \rho_{\rm GW}(\eta,k)\,, \tag{4.4}$$

[8]The production of scalar modes from primordial tensor modes is discussed in [289, 290] Moreover, very recently, SIGWs sourced by scalar-tensor perturbations have also been analyzed [291, 292].

the bar denoting an oscillation average and ρ_c is the critical energy density. GWs redshift as relativistic species and the current abundance of the GWB can be obtained by accounting for entropy injections in the standard thermal history of the Universe:

$$\Omega_{\rm GW}(k)h^2 = \Omega_{r,0}h^2 \left(\frac{g_*(\eta_f)}{g_*^0}\right)\left(\frac{g_{*\rm S}^0}{g_{*,\rm S}(\eta_f)}\right)^{4/3} \Omega_{\rm GW}(\eta_f, k)\,. \tag{4.5}$$

Here g_* ($g_{*,\rm S}$) are the relativistic degrees of freedom in energy (entropy), $\Omega_{r,0}h^2 = 4.2\cdot 10^{-5}$ is the current radiation density if the neutrino were massless [293]. Assuming standard model degrees of freedom, one finds that [294]

$$c_g(\eta_f) \equiv \left(\frac{g_*(\eta_f)}{g_*^0}\right)\left(\frac{g_{*\rm S}^0}{g_{*,\rm S}(\eta_f)}\right)^{4/3} \simeq 0.39, \tag{4.6}$$

for η_f of relevance for LISA.

4.1 Source of GWs at second order in the scalar perturbations

The solution to Eq. (4.1) can be easily found in Fourier space, where the equation for the GW amplitude h for each polarization state s becomes

$$h_s''(\boldsymbol{k},\eta) + 2\mathcal{H}h_s'(\boldsymbol{k},\eta) + k^2 h_s(\boldsymbol{k},\eta) = 4\mathcal{S}_s(\boldsymbol{k},\eta) \tag{4.7}$$

and where $\mathcal{S}_s(\boldsymbol{k},\eta)$ encloses the Fourier transform of the (projected) source given by

$$\mathcal{S}_s(\boldsymbol{k},\eta) = \int \frac{d^3\boldsymbol{p}}{(2\pi)^3} Q_s(\boldsymbol{k},\boldsymbol{p}) f(|\boldsymbol{k}-\boldsymbol{p}|,p,\eta)\zeta(\boldsymbol{p})\zeta(\boldsymbol{k}-\boldsymbol{p})\,. \tag{4.8}$$

In the latter equation, we introduced

$$\begin{aligned} f(|\boldsymbol{k}-\boldsymbol{p}|,p,\eta) =& \frac{3(1+w)}{(5+3w)^2}\left[2(5+3w)T(|\boldsymbol{k}-\boldsymbol{p}|\eta)T(p\eta) + \frac{4}{\mathcal{H}^2}T'(|\boldsymbol{k}-\boldsymbol{p}|\eta)T'(p\eta)\right. \\ &\left. + \frac{4}{\mathcal{H}}\left(T(|\boldsymbol{k}-\boldsymbol{p}|\eta)T'(p\eta) + T'(|\boldsymbol{k}-\boldsymbol{p}|\eta)T(p\eta)\right)\right] \end{aligned} \tag{4.9}$$

and we introduced the curvature perturbation transfer function $T(k\eta)$, and the spherical coordinates (p,θ,ϕ) for the internal momentum $\boldsymbol{p}$ and aligned the axes $(\hat{x},\hat{y},\hat{z})$ with $(e_+(\boldsymbol{k}), e_\times(\boldsymbol{k}), \boldsymbol{k})$ (with $e_+, e_\times$ being the polarisation tensors for the GW) so that

$$Q_s(\boldsymbol{k},\boldsymbol{p}) = e_s^{ij}(\hat{\boldsymbol{k}})p_i p_j = \frac{p^2}{\sqrt{2}}\sin^2\theta \times \begin{cases} \cos 2\phi, & s = + \\ \sin 2\phi, & s = \times \end{cases}\,. \tag{4.10}$$

A solution to the Fourier transform of the inhomogeneous equation of motion for $h_{ij}(\eta,k)$, Eq. (4.1), can be obtained using the Green's function $G_{\boldsymbol{k}}(\eta,\bar{\eta})$, that solves the homogeneous equation

$$G_{\boldsymbol{k}}''(\eta,\bar{\eta}) + \left(k^2 - \frac{a''}{a}\right)G_{\boldsymbol{k}}(\eta,\bar{\eta}) = \delta\left(\eta-\bar{\eta}\right), \tag{4.11}$$

with the boundary conditions $\lim_{\eta\to\bar{\eta}} G_{\boldsymbol{k}}(\eta,\bar{\eta}) = 0$ and $\lim_{\eta\to\bar{\eta}} G'_{\boldsymbol{k}}(\eta,\bar{\eta}) = 1$. The Green's function depends on $k = |\boldsymbol{k}|$ by isotropy and can be expressed analytically in terms of Bessel functions as

$$k\, G_{\boldsymbol{k}}(\eta,\bar{\eta}) = \sqrt{x\bar{x}}\,[y_\nu(x)j_\nu(\bar{x}) - j_\nu(x)y_\nu(\bar{x})]\,, \tag{4.12}$$

where $x = k\eta, \bar{x} = k\bar{\eta}$, and $\nu = 3(1-w)/[2(1+3w)]$ and j_ν and y_ν are respectively the spherical Bessel function of the first and second kind. For example, during RD, $kG_k(\eta,\bar{\eta}) = \sin(x-\bar{x})\Theta(\eta-\bar{\eta})$, where Θ is the Heaviside function.

The amplitude of the tensor modes can then be written as

$$\begin{aligned} h_s(\boldsymbol{k},\eta) =& 4\int_{\eta_i}^{\eta} d\bar{\eta}\, G_{\boldsymbol{k}}(\eta,\bar{\eta})\frac{a(\bar{\eta})}{a(\eta)}\mathcal{S}_s(\boldsymbol{k},\bar{\eta}) \\ =& 4\int_{\eta_i}^{\eta} d\bar{\eta}\, G_{\boldsymbol{k}}(\eta,\bar{\eta})\frac{a(\bar{\eta})}{a(\eta)}\int\frac{d^3\boldsymbol{p}}{(2\pi)^3}Q_s(\boldsymbol{k},\boldsymbol{p})f(|\boldsymbol{k}-\boldsymbol{p}|,p,\bar{\eta})\zeta(\boldsymbol{p})\zeta(\boldsymbol{k}-\boldsymbol{p})\,, \end{aligned} \tag{4.13}$$

with s indicating the polarisation and η_i the initial emission time. The GW two-point function, needed to obtain the energy density of GWs as a function of the scalar power spectrum, is defined in terms of the GW power spectrum

$$\langle h^r(\eta,\boldsymbol{k}_1)h^s(\eta,\boldsymbol{k}_2)\rangle \equiv (2\pi)^3\delta(\boldsymbol{k}_1+\boldsymbol{k}_2)\,\delta^{rs}\,\frac{2\pi^2}{k_1^3}\mathcal{P}_h(k_1)\,. \tag{4.14}$$

After substituting Eq. (4.12) into Eq. (4.13), the final expression for the second-order induced tensor power spectrum reads [295, 296]

$$\begin{aligned} \langle h^{s_1}(\eta,\boldsymbol{k}_1)h^{s_2}(\eta,\boldsymbol{k}_2)\rangle \;=\; & 16\int\frac{d^3\boldsymbol{p}_1}{(2\pi)^3}\frac{d^3\boldsymbol{p}_2}{(2\pi)^3}Q_{s_1}(\boldsymbol{k}_1,\boldsymbol{p}_1)Q_{s_2}(\boldsymbol{k}_2,\boldsymbol{p}_2) \\ & \times I(|\boldsymbol{k}_1-\boldsymbol{p}_1|,p_1,\eta_1)I(|\boldsymbol{k}_2-\boldsymbol{p}_2|,p_2,\eta_2)\delta^{(3)}(\boldsymbol{k}_1+\boldsymbol{k}_2)\langle\zeta_{\boldsymbol{p}_1}\zeta_{\boldsymbol{k}-\boldsymbol{p}_1}\zeta_{\boldsymbol{p}_2}\zeta_{\boldsymbol{k}-\boldsymbol{p}_2}\rangle\,. \end{aligned} \tag{4.15}$$

In the Gaussian case, where only the disconnected part of the trispectrum survives (hence the "d" in the following equation, we discuss in Sec. 4.5 the impact of primordial NG), one obtains

$$\overline{\mathcal{P}_{h,\mathrm{d}}(\eta,k)} = 4\int_0^\infty \mathrm{d}v\int_{|1-v|}^{1+v}\mathrm{d}u\left[\frac{4v^2-(1+v^2-u^2)^2}{4uv}\right]^2\overline{I^2(u,v,k,\eta)}\mathcal{P}_\zeta(kv)\mathcal{P}_\zeta(ku), \tag{4.16}$$

where we introduced the dimensionless variables

$$v \equiv \frac{p}{k}, \qquad u \equiv \frac{|\boldsymbol{k}-\boldsymbol{p}|}{k}\,. \tag{4.17}$$

The overline stands for an oscillation average, and the kernel function $I(u,v,\eta)$ is defined in terms of Green's function as

$$I(|\boldsymbol{k}-\boldsymbol{p}|,p,\eta) = \int_{\eta_i}^{\eta} d\bar{\eta}\, G_{\boldsymbol{k}}(\eta,\bar{\eta})\frac{a(\bar{\eta})}{a(\eta)}f(|\boldsymbol{k}-\boldsymbol{p}|,p,\bar{\eta})\,. \tag{4.18}$$

Since the integration domain of (4.16) is rectangular, for computational purposes, it is convenient to rotate the coordinates into

$$t \equiv u+v-1, \qquad s \equiv u-v \tag{4.19}$$

and the SIGW spectrum $\overline{\mathcal{P}_h(\eta,k)}$ can be then rewritten as

$$\overline{\mathcal{P}_h(\eta,k)} = 4\int_0^\infty \mathrm{d}t \int_0^1 \mathrm{d}s \left[\frac{t(2+t)(1-s^2)}{(1-s+t)(1+s+t)}\right]^2 \overline{I^2(t,s,k,\eta)} \times \mathcal{P}_\zeta\left(k\frac{t+s+1}{2}\right)\mathcal{P}_\zeta\left(k\frac{t-s+1}{2}\right), \quad (4.20)$$

where the integration in s is restricted to positive values due to the integrand being an even function of s. The integration kernel $I(t,s,\eta)$ contains information about the time evolution of the source during emission, as well as the propagation of the emitted GWs after emission, and thus depends on the thermal history when the relevant modes re-entered the Hubble radius. Let us consider different assumptions on the thermal history in the following sections.

The generation of tensor modes at second-order in perturbation theory raises concerns about the potential gauge dependence of results commonly calculated in the Newtonian gauge. Unlike first-order tensor modes, which are gauge invariant, second-order tensor modes are gauge dependent [297]. During the phase when the source is still active, the result is expected to remain gauge-dependent, as one cannot directly identify the tensor mode with the freely propagating GW. However, when the source becomes inactive after the GWs are produced, it effectively decouples from the GWs. This happens for example during the radiation-dominated era of the Universe and in the other cases considered in this draft. Therefore, in the late-time limit well inside the cosmological horizon, tensor mode behaves as linear metric perturbations and the initial gauge dependence no longer affects the final result [298–301], ensuring that the spectra computed in this work are unaffected by this issue.

4.2 Radiation-dominated era

If the emission takes place in a RD universe, the kernel function in the deep sub-horizon regime $k\eta \to \infty$ takes the form [49, 295, 296]

$$\overline{I^2_{\mathrm{RD}}(t,s,k\eta\to\infty)} = \frac{1}{2(k\eta)^2} I_A^2(u,v)\left[I_B^2(u,v)+I_C^2(u,v)\right] \quad (4.21)$$

where

$$\begin{aligned} I_A(u,v) &= \frac{3(u^2+v^2-3)}{4u^3v^3}\,, \\ I_B(u,v) &= -4uv + (u^2+v^2-3)\log\left|\frac{3-(u+v)^2}{3-(u-v)^2}\right|\,, \\ I_C(u,v) &= \pi(u^2+v^2-3)\Theta(u+v-\sqrt{3})\,, \end{aligned} \quad (4.22)$$

u and v have been introduced above and again $\Theta(x)$ is the Heaviside function. Notice that the unphysical divergence obtained in the limit $|\boldsymbol{k}-\boldsymbol{p}|+p=\sqrt{3}k$, which is also retained in the spectrum produced by monochromatic scalar perturbations [18]. The factor of $\sqrt{3}$ originates from the (inverse) sound speed in RD appearing in the transfer function, and in the limit $|\boldsymbol{k}-\boldsymbol{p}|+p=\sqrt{3}k$ the contributions from some of the source terms add up

constructively and build up logarithmically over time [295]. Assuming RD up to $\eta \to \infty$ leads to this logarithmic divergence, which is regularized in the integral for $\mathcal{P}_h$ if $\mathcal{P}_\zeta(k)$ is smooth enough (e.g. has a nonzero width), or if the emission time is finite. We do not introduce this regulator here, as it is numerically irrelevant for spectra with finite width [67].

4.3 Transition from an early matter-dominated to the radiation-dominated era

We further consider the alternative thermal history wherein an early matter-dominated (eMD) era may precede the RD era. We follow the prescription of [302], assuming a sudden transition at a conformal time $\eta = \eta_{\rm R}$, where the subscript R indicates the reheating time [296]. In this scenario, the dominant contribution comes from GWs induced during the RD era by the scalar perturbations that have entered the horizon during an eMD era.

Interestingly, in this case, one observes a resonantly enhanced production of GWs. In particular, during the transition, the time derivative of the Bardeen potential Φ, which is the source of the SIGW signal, goes very quickly from $\Phi' = 0$ (since in an eMD era $\Phi =$ constant) to $\Phi' \neq 0$ (see [234, 299] for more details), leading to an enhanced secondary tensor mode production sourced mainly by the $\mathcal{H}^2\Phi'^2$ term in Eq. (4.2). In a more physical scenario, the transition happens more gradually [303–305].

For the kernel function $I(t, s, k, \eta)$, one finds two dominant contributions at the onset of the late RD era, i.e. at $\eta = \eta_{\rm R}$, when most of the GWs are expected to be produced [299, 306]. The first contribution to $I(t, s, k, \eta)$ is given by $k_{\rm max}/k \sim 1$, at $t_0 = \sqrt{3} - 1$ [302]

$$\overline{I^2_{\rm lRD,res}}(t_0, s, k, \eta_{\rm R}) \simeq Y \frac{9\left(-5 + s^2 + 2t + t^2\right)^4 x_{\rm R}^8}{2^{17}5^4(1 - s + t)^2(1 + s + t)^2} {\rm Ci}^2\left(y\right), \tag{4.23}$$

where $x_{\rm R} = k\eta_{\rm R}$. The variable y is defined as $y \equiv \frac{|t+1-c_s^{-1}|x_{\rm R}}{2c_s^{-1}} = \frac{|t+1-\sqrt{3}|x_{\rm R}}{2\sqrt{3}}$, and Y is a fudge factor to absorb the uncertainty in the integration boundary, set here to be 2.3 as in [302]. At $t_0 = \sqrt{3} - 1$, the logarithmic singularity of the function Ci is reached, giving rise to the peak in the spectrum.

The second contribution to $I(t, s, k, \eta)$ comes from the wave-numbers satisfying $k_{\rm max}/k \gg 1$, hence the integrations are dominated by the large t region $u \sim v \sim t \gg 1$. Therefore, the dependence on s is lost. Setting $s = 0$, the kernel function reads

$$\overline{I^2_{\rm lRD,LV}}(t \gg 1, s, k, \eta_{\rm R}) \simeq \frac{9t^4 x_{\rm R}^8}{2^{17}5^4}\left[4{\rm Ci}^2(x_{\rm R}/2) + \left(\pi - 2{\rm Si}(x_{\rm R}/2)\right)^2\right]. \tag{4.24}$$

The integration region is

$$0 \leq s \leq 1 \quad \text{and} \quad 0 \leq t \leq -s + 2\frac{k_{\rm max}}{k} - 1 \qquad \text{for} \quad k \leq k_{\rm max}, \tag{4.25}$$

and the result obtained from this integration region is then doubled to account for $s \to -s$. The two contributions are computed separately and added to give

$$\Omega_{\rm GW}^{\rm eMDRD} \simeq \Omega_{\rm GW}^{\rm (LV)} + \Omega_{\rm GW}^{\rm (res)}. \tag{4.26}$$

4.4 Computation of SIGW with the binned spectrum approach

In this section, we discuss in more detail the procedure sketched in Sec. 3.1 for computing the SIGW in terms of a template-free approach to the curvature power spectrum. We express the power spectrum as a sum over momentum bins, as in Eq. (3.4). The specific profile for the power spectrum $\mathcal{P}_\zeta$ is then associated with a vector of coefficients A_i. By plugging Eq. (3.4) into Eqs. (4.3) and (4.20), we recast the SIGW density into the sum

$$\Omega_{\rm GW}(k) = \sum_{i,j}^{N-1} \Omega_{\rm GW}^{(i,j)}(k)\, A_i A_j \tag{4.27}$$

performed over the momentum bins. The kernel for this sum is the matrix

$$\begin{aligned}\Omega_{\rm GW}^{(i,j)}(k) = \frac{1}{12}\left(\frac{k}{aH}\right)^2 \int_0^\infty dt \int_0^1 ds \left[\frac{t(2+t)(s^2-1)}{(1-s+t)(1+s+t)}\right]^2 \overline{I^2(t,s,k,\eta)} \\ \times \Theta\left(k\, v(s,t) - p_i\right) \Theta\left(p_{i+1} - k\, v(s,t)\right) \Theta\left(k\, u(s,t) - p_j\right) \Theta\left(p_{j+1} - k\, u(s,t)\right) ,\end{aligned} \tag{4.28}$$

where p_i, p_j are the boundaries of the momentum bins entering in Eq. (3.4).

Importantly, we stress that the matrix $\Omega_{\rm GW}^{(i,j)}$ of Eq. (4.27) is *independent* of the specific scalar spectrum considered, and the information on $\mathcal{P}_\zeta$ is stored only in the coefficients A_i appearing quadratically in Eq. (4.27). This implies that the computation of Eq. (4.28) depends only on the kernel function, and its entries can be computed once for all for any given cosmology: for example, we can use one of the kernels discussed in Sec. 4.2 or Sec. 4.3. The nested integrals appearing in Eq. (4.28) should then be performed a single time for each kernel. Once $\Omega_{\rm GW}^{(i,j)}$ is determined, it can be used to swiftly compute the resulting $\Omega_{\rm GW}$ for *any* $\mathcal{P}_\zeta$, by means of the contractions in Eq. (4.27). In fact, with this method, we reduce the problem of computing $\Omega_{\rm GW}$ to perform the simple sum of Eq. (4.27).

This approach is useful in scenarios where the underlying shape of the curvature spectrum is not accurately known, for example, due to the presence of peaks or breaks, whose position depends on the underlying physics we wish to probe. In fact, the method allows us to scan over different sets of A_i components, swiftly computing the SIGW frequency profile, which can then be compared with data. Other approaches for reconstructing the properties of the underlying $\mathcal{P}_\zeta$ from SIGW data can be found for example in [307–309].

We tabulate the matrix (4.27) assuming a varying number of bins N in the range of relevance for LISA, which is $k \in [1.26 \times 10^{-4}, 6.28]$/s both for the internal (i,j) indices, as well as external momentum k. This range is chosen to match the one used by the `SGWBinner` code adopted to perform the LISA forecasts. See the discussion in Sec. 5.

4.5 Non-Gaussian imprints on the SIGW spectrum

From the solution of the SIGW, Eq. (4.13), one can relate the tensor power spectrum to the four-point correlation function of the curvature perturbation, see Eq. (4.15). As anticipated above, the latter can be decomposed into disconnected and connected contributions, where the connected part vanishes when primordial fluctuations are drawn from a Gaussian distribution. The disconnected contribution gives rise to Eq. (4.16) that can

be solely expressed in terms of the scalar power spectrum $\mathcal{P}_\zeta(k)$. However, the connected contribution depends on the primordial trispectrum $\langle\zeta_{\boldsymbol{k}_1}\zeta_{\boldsymbol{k}_2}\zeta_{\boldsymbol{k}_3}\zeta_{\boldsymbol{k}_4}\rangle'_{\rm c} = T_\zeta(\boldsymbol{k}_1,\boldsymbol{k}_2,\boldsymbol{k}_3,\boldsymbol{k}_4)$, whose corresponding tensor power spectrum reads [49][9]

$$\begin{aligned}\overline{\mathcal{P}_{h,\rm c}} = \frac{1}{4\pi}\int_0^\infty {\rm d}v_1\int_{|1-v_1|}^{1+v_1}{\rm d}u_1\int_0^\infty {\rm d}v_2\int_{|1-v_2|}^{1+v_2}{\rm d}u_2\int_0^{2\pi}{\rm d}\psi \\ \times\frac{\cos(2\psi)}{(u_1v_1u_2v_2)^{5/4}}[4v_1^2-(1+v_1^2-u_1^2)^2][4v_2^2-(1+v_2^2-u_2^2)^2] \\ \times\overline{I(u_1,v_1,k,\eta)I(u_2,v_2,k,\eta)}\mathcal{T}_\zeta(u_1,v_1,u_2,v_2,\psi)\,,\end{aligned} \tag{4.29}$$

and we use the variables u_i and v_i defined above. The dimensionless trispectrum function $\mathcal{T}_\zeta$ is defined as

$$\mathcal{T}_\zeta(\boldsymbol{k}_1,\boldsymbol{k}_2,\boldsymbol{k}_3,\boldsymbol{k}_4) = \frac{(k_1k_2k_3k_4)^{9/4}}{(2\pi)^6}T_\zeta(\boldsymbol{k}_1,\boldsymbol{k}_2,\boldsymbol{k}_3,\boldsymbol{k}_4)\,, \tag{4.30}$$

and is evaluated at $\boldsymbol{k}_1=\boldsymbol{p}_1, \boldsymbol{k}_2=\boldsymbol{k}-\boldsymbol{p}_1$, $\boldsymbol{k}_3=-\boldsymbol{p}_2$, $\boldsymbol{k}_4=-\boldsymbol{k}+\boldsymbol{p}_2$, with $\psi=\phi_1-\phi_2$ the difference between the azimuthal angles of $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$ with respect to $\boldsymbol{k}$. The integration kernel for emission during radiation domination is given by

$$\begin{aligned}\overline{I(u_1,v_1,k,\eta)I(u_2,v_2,k,\eta)} = \frac{1}{2(k\eta)^2}I_A(u_1,v_1)I_A(u_2,v_2) \\ \times\left[I_B(u_1,v_1)I_B(u_2,v_2)+I_C(u_1,v_1)I_C(u_2,v_2)\right]\,,\end{aligned} \tag{4.31}$$

in terms of $I_{A,B,C}$ defined in Eq. (4.22) and with $x=k\eta$. As for the disconnected contribution, it is numerically convenient to change the integration variables from (u_i,v_i) to (t_i,s_i), with

$$\int_0^\infty {\rm d}v_i\int_{|1-v_i|}^{1+v_i}{\rm d}u_i(\ldots) = \frac{1}{2}\int_0^\infty {\rm d}t_i\int_{-1}^{1}{\rm d}s_i(\ldots)\,, \tag{4.32}$$

where we did not assume symmetry between positive and negative s, to retain full generality in this case.

On general grounds, the properties of the trispectrum and the symmetries of the kernel Eq. (4.31) enable one to split the connected contribution (4.29) into three inequivalent channels [49]. Hence, any trispectrum function of the unordered set $\{\boldsymbol{k}_1,\boldsymbol{k}_2,\boldsymbol{k}_3,\boldsymbol{k}_4\}$, can be written as

$$T_\zeta\left[\{\boldsymbol{k}_1,\boldsymbol{k}_2,\boldsymbol{k}_3,\boldsymbol{k}_4\}\right] = \tilde{T}_\zeta\left[(\boldsymbol{k}_1,\boldsymbol{k}_2,\boldsymbol{k}_3,\boldsymbol{k}_4)\right] + \ 23 \text{ perm.}\,, \tag{4.33}$$

where the individual contributions $\tilde{T}_\zeta$ are not in general invariant under permutations of their arguments. Moreover, it can be conveniently written as

$$T_\zeta = (T_{\mathfrak{s}}+T_{\mathfrak{t}}+T_{\mathfrak{u}}) + \ 7 \text{ perm.}\,, \text{ with } \begin{cases} T_{\mathfrak{s}} = \tilde{T}_\zeta\left[(\boldsymbol{k}_1,\boldsymbol{k}_2;\boldsymbol{k}_3,\boldsymbol{k}_4)\right] \\ T_{\mathfrak{t}} = \tilde{T}_\zeta\left[(\boldsymbol{k}_1,\boldsymbol{k}_3;\boldsymbol{k}_2,\boldsymbol{k}_4)\right] \\ T_{\mathfrak{u}} = \tilde{T}_\zeta\left[(\boldsymbol{k}_1,\boldsymbol{k}_4;\boldsymbol{k}_2,\boldsymbol{k}_3)\right] \end{cases}, \tag{4.34}$$

[9] In [49], which uses different conventions, their Eq. (2.6) should be divided by 4, as well as subsequent results. This has been corrected in [50] in a study of SIGWs including parity violation.

where the channels $\mathfrak{s}, \mathfrak{t}$ and $\mathfrak{u}$ correspond to the three unordered pairs $\{\{\boldsymbol{k}_1, \boldsymbol{k}_2\}, \{\boldsymbol{k}_3, \boldsymbol{k}_4\}\}$, $\{\{\boldsymbol{k}_1, \boldsymbol{k}_3\}, \{\boldsymbol{k}_2, \boldsymbol{k}_4\}\}$ and $\{\{\boldsymbol{k}_1, \boldsymbol{k}_4\}, \{\boldsymbol{k}_2, \boldsymbol{k}_3\}\}$ together with their "exchanged momenta" $\mathfrak{s} = |\boldsymbol{k}_1 + \boldsymbol{k}_2|$, $\mathfrak{t} = |\boldsymbol{k}_1 + \boldsymbol{k}_3|$ and $\mathfrak{u} = |\boldsymbol{k}_1 + \boldsymbol{k}_4|$ respectively, and where the seven permutations in (4.34) preserves the exchanged momentum of each channel. Furthermore one can show that the trispectrum-induced GW spectrum (4.29) can be written in terms of only three fundamental contributions corresponding to the seeds $T_{\mathfrak{s}}$, $T_{\mathfrak{t}}$ and $T_{\mathfrak{u}}$:

$$\overline{\mathcal{P}_{h,\mathrm{c}}} = 8(\overline{\mathcal{P}^{\mathfrak{s}}_{h,\mathrm{c}}} + \overline{\mathcal{P}^{\mathfrak{t}}_{h,\mathrm{c}}} + \overline{\mathcal{P}^{\mathfrak{u}}_{h,\mathrm{c}}}) \,, \tag{4.35}$$

with $\overline{\mathcal{P}^{\mathfrak{s}}_{h,\mathrm{c}}}$ simply corresponding to $\overline{\mathcal{P}_{h,\mathrm{c}}}$ with T_ζ replaced by $T_{\mathfrak{s}}$, etc.

Computing from first principles the trispectrum generated in models relevant for GW astronomy is a difficult task, as these scenarios often involve a strong breaking of scale invariance as well as enhanced fluctuations that can jeopardize perturbative computations, see e.g. [114, 128, 130, 173, 283, 310–313]. In the following, we assume that the trispectrum is of the local τ_{NL} type:

$$T^{\mathrm{loc}}_\zeta = \tau_{\mathrm{NL}} \left[(P_\zeta(\mathfrak{s}) P_\zeta(k_1) P_\zeta(k_3) + \text{ 3 perm.}) + \ (\mathfrak{s} \leftrightarrow \mathfrak{t}) + (\mathfrak{s} \leftrightarrow \mathfrak{u}) \right] . \tag{4.36}$$

We give more details on the computation of the spectrum in the presence of local NGs in App. A.4.

Local NGs, typical of multi-field models, generically arise from the non-linear evolution of cosmological fluctuations on super-Hubble scales (see e.g. [36] for a review). Besides the τ_{NL} type, which emerges microscopically from the exchange of scalar particles through cubic interactions, the local trispectrum also acquires in general a g_{NL} component, coming from contact quartic interactions. However, its momentum dependence is such that it does not contribute to the GW spectrum, and hence we can disregard it for our purpose. We stress that our choice, Eq. (4.36), is a first methodological step motivated by simplicity, in particular because the momentum dependence of the trispectrum is fully characterized by the one of the power spectrum $P_\zeta(k)$, which is not the case in general (see [49] for a study of the impact of various trispectrum shapes on the GW spectrum). Let us also highlight a conceptual aspect. Several works in the literature consider a local ansatz in which the real-space curvature perturbation $\zeta(\boldsymbol{x})$ is expanded in powers of a Gaussian variable $\zeta_{\mathrm{G}}(\boldsymbol{x})$ as

$$\zeta = \zeta_{\mathrm{G}} + \frac{3}{5} f_{\mathrm{NL}} \zeta^2_{\mathrm{G}} + \dots \,, \tag{4.37}$$

and compare the corresponding GW spectrum with the one obtained by keeping only the first term, fully characterized by the power spectrum $P_{\zeta_{\mathrm{G}}}$, see e.g. [41, 42, 45, 46, 48, 52, 53, 277, 309, 314–317]. At leading order, such an expansion does lead to the trispectrum (4.36) with P_ζ replaced by $P_{\zeta_{\mathrm{G}}}$, and $\tau_{\mathrm{NL}} = (6 f_{\mathrm{NL}}/5)^2$. However, the nonlinear terms in Eq. (4.37) also imply that the curvature power spectrum does not coincide with $P_{\zeta_{\mathrm{G}}}$. Instead, keeping only the quadratic term shown in Eq. (4.37) for definiteness, one finds the power spectrum

$$P_\zeta(k) = P_{\zeta_{\mathrm{G}}}(k) + \frac{1}{2} \left(\frac{6}{5} f_{\mathrm{NL}} \right)^2 \int \frac{\mathrm{d}^3 \boldsymbol{p}}{(2\pi)^3} P_{\zeta_{\mathrm{G}}}(p) P_{\zeta_{\mathrm{G}}}(|\boldsymbol{k} - \boldsymbol{p}|) \,. \tag{4.38}$$

Hence, as described in [49], in this approach, one considers on similar grounds the impact of primordial NG on the SIGW, through the trispectrum, and the difference between a putative P_{ζ_G} to which we have no access, and the power spectrum of ζ which is anyway the only observable quantity. Again, we emphasize that the effects of non-linearities on the SIGW spectrum may not always be fully captured by the local ansatz, Eq. (4.37). As a result, the predicted amplitude of the resulting SGWB could differ significantly [68], potentially being suppressed or enhanced by several orders of magnitude.

In our analysis, whose results are shown in Sec. 6.5, we find it conceptually clearer to take as a benchmark the disconnected prediction from the purely Gaussian theory (4.16) with a given power spectrum $\mathcal{P}_\zeta(k)$, which will take to be of the log-normal form (3.5), and to compare it with the addition of the non-Gaussian, connected, contribution, Eq. (4.29) with trispectrum (4.36). Note that for the latter, the $\mathfrak{s}$-channel contribution vanishes as the corresponding $\mathcal{T}_\zeta$ in Eq. (4.29) does not depend on the azimuthal angle ψ. We are thus left with the two $\mathfrak{t}$ and $\mathfrak{u}$ contributions in Eq. (4.35). Overall, we stress that the parameter to be constrained from observations is $\tau_{\rm NL}$, which measures the non-Gaussian contribution to the SIGW spectrum coming from the trispectrum of curvature perturbations. On scales relevant to LISA, there is *a priori* no constraint on $\tau_{\rm NL}$ except that it is positive in known concrete realizations of inflation. Its size is also *a priori* arbitrary, although from a theoretical perspective, perturbative control during inflation typically implies $\tau_{\rm NL}\mathcal{P}_\zeta < 1$.

5 Mock signal reconstructions with the `SGWBinner` and `SIGWAY` codes

This section outlines the analysis method employed in this work. Before presenting the LISA data model adopted in our analysis (Sec. 5.1) and functionalities of the code (Sec. 5.2), let us briefly illustrate the measurement of GWs with LISA.

The observatory will consist of three satellites ($\alpha = 1, 2, 3$) that orbit at the vertices of an approximately equilateral triangle with sides about 2.5 million kilometers long. Each satellite contains two Test Masses (TMs), whose positions are constantly monitored, and two lasers emitting toward the other satellites. By monitoring the fractional Doppler frequency shifts of photons traveling along the arms between satellites, LISA measures the relative displacements of the TMs. The path connecting two satellites is typically dubbed "link" and the single link measurement can be denoted as $\eta_{\alpha\beta}(t)$, where the laser emitted from the satellite β at time $t - L_{\alpha\beta}/c$ is recorded at time t in the satellite α. These measurements are, however, dominated by laser frequency noise, which is expected to be several orders of magnitude greater than the required sensitivity [1]. To suppress this noise contribution, LISA will employ a post-processing technique called Time-Delay Interferometry (TDI) [318–326]. In practice, TDI can be understood as the operation of 3×6 matrix on the six single link measurements $\eta_{\alpha\beta}(t)$ [64, 327] that returns the three TDI channels where the laser frequency noise is strongly suppressed.

As in the previous studies using the `SGWBinner` code [61, 62, 67, 328, 329], in this work we assume for simplicity i) equal and static arm lengths and ii) equality of noise at each

link. While, in reality, these hypotheses will not be perfectly satisfied[10], it has been shown that the signal reconstruction is almost unaffected by unequal (but static) arm length and unequal noise amplitudes [327, 332]. Under the equal and static arm length assumption, the so-called first-generation TDI variables suffice to achieve laser noise cancellation.[11] In the $\{\mathrm{X}, \mathrm{Y}, \mathrm{Z}\}$ basis, they are expressed as

$$\mathrm{X} \equiv (1 - D_{13}D_{31})(\eta_{12} + D_{12}\eta_{21}) + (D_{12}D_{21} - 1)(\eta_{13} + D_{13}\eta_{31}) \,, \tag{5.1}$$

with Y and Z being cyclic permutations of X. Here $D_{\alpha\beta}$ is the delay operator acting on any time-dependent function $x(t)$ as $D_{\alpha\beta}\, x(t) = x(t - L_{\alpha\beta})$ and we take $L_{\alpha\beta} = L = 2.5 \times 10^9$ m. For SGWB signal searches, it is convenient to combine the XYZ variables to obtain the so-called AET basis [334, 335], defined as

$$\mathrm{A} \equiv \frac{\mathrm{Z} - \mathrm{X}}{\sqrt{2}}, \qquad \mathrm{E} \equiv \frac{\mathrm{X} - 2\mathrm{Y} + \mathrm{Z}}{\sqrt{6}}, \qquad \mathrm{T} \equiv \frac{\mathrm{X} + \mathrm{Y} + \mathrm{Z}}{\sqrt{3}} \,, \tag{5.2}$$

which, in the limit of equal arms and equal noises, can be shown to have vanishing cross-correlations and simplify the likelihood computation. Moreover, due to its symmetric structure, the T channel strongly suppresses GW signals at small frequencies compared to instrumental noise. For this reason, the T channel can be treated as a quasi-null channel that is mostly sensitive to instrumental noise.[12]

5.1 Data streams from LISA TDI channels

We represent the three time-domain data streams as $d_i(t)$, where i runs over the channels of the TDI basis. These quantities are real-valued functions defined on the interval $[-\tau/2, \tau/2]$ with τ being the duration of a data segment. The Fourier transforms of these data streams are then given by

$$\tilde{d}_i\,(f) = \int_{-\tau/2}^{\tau/2} \mathrm{d}t\; \mathrm{e}^{2\pi i f t} d_i\,(t) \;. \tag{5.3}$$

Our central assumption is that all transients including loud deterministic signals and glitches in the noise are subtracted from the time stream through some appropriate methods within the LISA global fit scheme [55–57, 337–339].[13] That is, as adopted in previous studies [61, 62, 67, 328, 329, 332], the data considered in our analysis only contain the stochastic contributions to the noise, $\tilde{n}_i^{\nu}$, and the stochastic signal $\tilde{s}_i^{\sigma}$ due to the unresolved binary signals and, possibly, the SGWB:

$$\tilde{d}_i(f) = \sum_{\nu} \tilde{n}_i^{\nu}(f) + \sum_{\sigma} \tilde{s}_i^{\sigma}(f) \,, \tag{5.4}$$

[10] With realistic orbits, LISA will not be perfectly equilateral and arm-lengths vary at the percent level [330] (see also Appendix A of [331]).

[11] To account for non-static arm lengths and the associated Doppler shifts, the second-generation TDI variables [323, 324, 326, 333] would be required.

[12] The T channel does not remain a null channel in general with unequal and flexing arms and differing noise levels in the different spacecraft, although other quasi-null channels are available [336].

[13] See Ref. [340] for the application of simulation-based inference to the SGWB search performed by LISA in the presence of transient signals, which goes beyond the framework of purely stochastic analysis.

where ν and σ run over different noise and signal components, respectively. In the following, we will assume that all these components are stationary and obey Gaussian statistics with zero mean and variance given by

$$\langle \tilde{n}_i^{\nu}(f)\tilde{n}_j^{\nu *}(f')\rangle = \frac{1}{2}\delta\left(f-f'\right) P_{N,ij}^{\nu}(f)\,, \qquad \langle \tilde{s}_i^{\sigma}(f)\tilde{s}_j^{\sigma *}(f')\rangle = \frac{1}{2}\delta\left(f-f'\right) P_{S,ij}^{\sigma}(f)\,, \quad (5.5)$$

where we define the one-side power-spectral density (PSD) (for $i=j$) and cross-spectral density (CSD) (for $i\neq j$) of noise and signal components as $P_{N,ij}^{\nu}(f)$ and $P_{S,ij}^{\sigma}(f)$, respectively. Assuming all these components to be uncorrelated with one another, we obtain

$$\begin{aligned}\langle \tilde{d}_i(f)\tilde{d}_j^{*}(f')\rangle &= \frac{1}{2}\delta\left(f-f'\right)\left[\sum_{\nu} P_{N,ij}^{\nu}(f) + \sum_{\sigma} P_{S,ij}^{\sigma}(f)\right] \\ &\equiv \frac{1}{2}\delta\left(f-f'\right)\left[P_{N,ij}(f) + P_{S,ij}(f)\right]\,,\end{aligned} \quad (5.6)$$

where $P_{N,ij}(f)$ and $P_{S,ij}(f)$ are the total noise and signal PSDs and CSDs. By denoting the response functions for isotropic SGWB signals as $\mathcal{R}_{ij}(f)$ (see Refs. [62, 341] for expressions for this quantity), the SGWB (in either strain $S_h^{\sigma}(f)$ or abundance $\Omega_{\rm GW}^{\sigma}(f)$) projected onto the data PSDs and CSDs can be expressed as

$$P_{S,ij}(f) = \mathcal{R}_{ij}(f)\sum_{\sigma} S_h^{\sigma}(f) \;= \mathcal{R}_{ij}(f)\frac{3H_0^2}{4\pi^2 f^3}\sum_{\sigma} h^2\Omega_{\rm GW}^{\sigma}(f)\,, \quad (5.7)$$

where H_0 is the present Hubble constant and h is the normalized one as $H_0/h \simeq 3.24\times10^{-18}$ 1/s. Note again that under the assumptions stated above in this section, one finds that $R_{ij}(f)$ is diagonal in the AET basis.

It is common practice to quantify the predicted primordial SGWB signal in terms of $h^2\Omega_{\rm GW}(f)$; therefore, for later convenience, we define

$$P_{N,ij}^{\Omega}(f) = \frac{4\pi^2 f^3}{3H_0^2} P_{N,ij}(f)\,. \quad (5.8)$$

In the following, we provide more detailed descriptions of the noise PSDs in the AET basis and of the astrophysical foregrounds which are included in $h^2\Omega_{\rm GW}^{\sigma}(f)$.

5.1.1 Instrumental noise

Our current knowledge of the LISA noise is based on the LISA Pathfinder [342] and laboratory tests. As a first approximation, the stochastic component of the noise in each TDI channel can be grouped into two effective components: "Optical Metrology System" (OMS) noise and TM noise. The former accounts for noise in the readout frequency, such as laser shot noise, while the latter models the noise sources causing accelerations of the TMs, e.g., by environmental disturbances. Introducing the transfer functions for these two noise sources $\mathcal{T}_{ij,\alpha\beta}^{\nu}(f)$ (for details, see e.g. Refs. [62, 327, 333, 343]), which project those contributions onto the TDI channels, the total noise PSDs and CSDs can be expressed as

$$P_{N,ij}(f) = \sum_{\nu} P_{N,ij}^{\nu}(f) = \sum_{\alpha\beta}\left[\mathcal{T}_{ij,\alpha\beta}^{\rm TM}(f) S_{\alpha\beta}^{\rm TM}(f) + \mathcal{T}_{ij,\alpha\beta}^{\rm OMS}(f) S_{\alpha\beta}^{\rm OMS}(f)\right]. \quad (5.9)$$

As customary in the literature, we assume stationary, Gaussian, and uncorrelated noises at each link with identical spectral shapes given by

$$S_{\alpha\beta}^{\rm TM}(f) = 7.7\times10^{-46}\times A_{\alpha\beta}^2\left(\frac{f_c}{f}\right)^2\left[1+\left(\frac{0.4{\rm mHz}}{f}\right)^2\right]\left[1+\left(\frac{f}{8{\rm mHz}}\right)^4\right]\times{\rm s}\,, \quad (5.10)$$

$$S_{\alpha\beta}^{\rm OMS}(f) = 1.6\times10^{-43}\times P_{\alpha\beta}^2\left(\frac{f}{f_c}\right)^2\left[1+\left(\frac{2{\rm mHz}}{f}\right)^4\right]\times{\rm s}\,, \quad (5.11)$$

where $A_{\alpha\beta}$ and $P_{\alpha\beta}$ represent the amplitudes of the TM and OMS noises in the different links. Moreover, we have introduced $f_c \equiv (2\pi L/c)^{-1} \simeq 19{\rm mHz}$ representing the characteristic frequency of the detector. As mentioned above, we assume the noise amplitudes for all links to be identical, i.e. $A_{\alpha\beta} = A_{\rm noise}$ and $P_{\alpha\beta} = P_{\rm noise}$, and, following the ESA mission specifications [54], with fiducial values $A_{\rm noise} = 3$ and $P_{\rm noise} = 15$. In this case, the noise spectra reduce to $S_{\alpha\beta}^{\rm TM}(f) = S^{\rm TM}(f,A)$ and $S_{\alpha\beta}^{\rm OMS}(f) = S^{\rm OMS}(f,P)$. With $\mathcal{T}_{ij,\alpha\beta}^{\nu}(f)$ in the equal arm length limit, the PSDs in the AET basis read

$$\begin{aligned} P_{N,{\rm AA}}(f) &= P_{N,{\rm EE}}(f) \\ &= 8\sin^2 x\left\{4\left[1+\cos x+\cos^2 x\right]S^{\rm TM}(f,A_{\rm noise}) + [2+\cos x]\,S^{\rm OMS}(f,P_{\rm noise})\right\}\,, \end{aligned} \quad (5.12)$$

and

$$P_{N,{\rm TT}}(f) = 16\sin^2 x\left\{2\left[1-\cos x\right]^2 S^{\rm TM}(f,A_{\rm noise}) + [1-\cos x]\,S^{\rm OMS}(f,P_{\rm noise})\right\}\,, \quad (5.13)$$

where we have defined $x \equiv f/f_c$. Here the CSDs vanish, i.e. $P_{N,ij}(f) = 0$ $(i \neq j)$, so that the noise covariance matrix is diagonal.

5.1.2 Astrophysical foregrounds

Numerous weak and unresolvable signals from astrophysical sources will superimpose incoherently generating astrophysical SGWB [4, 344–350]. There are at least two guaranteed components in the LISA band. Below a few millihertz, the dominant contribution will come from Compact Galactic Binaries (CGBs) mostly composed of Double White Dwarfs (DWDs) [351, 352]. At higher frequencies, another contribution is expected from all the extragalactic compact objects including Stellar Origin Binary Black Holes (SOBBHs) and binary neutron stars (BNS) [353]. In the remainder of this section, we provide the templates for these foreground components implemented in the `SGWBinner` code that was recently used in Refs. [67, 328, 329, 332].

Galactic foreground. This component represents the contribution from the unresolved sub-threshold mergers of CGBs that remain after the removal of loud signals from the population of CGBs in the galactic disk [354]. Due to the angular dependence of the response functions and LISA yearly orbit, this component exhibits an annual modulation. While, in principle, this characteristic can help distinguish the galactic component from other stationary contributions, e.g., by accounting for variations in each segment [331,

355, 356], we average over anisotropies, which leads to suboptimal (but conservative[14]) foreground extraction. Similarly, because this foreground is formed by the superposition of many unresolvable sources, it is expected to have Gaussian statistics. Recently, Refs. [59, 361, 362] have called into question whether the populations entering the foreground are sufficient for the central limit theorem to apply at all frequencies, and imply a Gaussian description of the foreground may be biased. The non-stationarity of the foreground also in principle induces some non-Gaussianity when time-averaging.

Nevertheless, we use the empirical model from Ref. [363], which describes the sky-averaged and Gaussian contribution by

$$h^2\Omega_{\rm GW}^{\rm Gal}(f) = \frac{1}{2}\left(\frac{f}{1\,{\rm Hz}}\right)^{2/3} e^{-(f/f_1)^\alpha}\left[1+\tanh\frac{f_{\rm knee}-f}{f_2}\right] h^2\Omega_{\rm Gal}\,, \tag{5.14}$$

where the value of f_1 and $f_{\rm knee}$ depends on the total observation time $T_{\rm obs}$ as

$$\begin{aligned}\log_{10}(f_1/{\rm Hz}) &= a_1\log_{10}(T_{\rm obs}/{\rm year}) + b_1\,,\\ \log_{10}(f_{\rm knee}/{\rm Hz}) &= a_k\log_{10}(T_{\rm obs}/{\rm year}) + b_k\,.\end{aligned} \tag{5.15}$$

The exponential factor $e^{-(f/f_1)^\alpha}$ accounts for the loss of stochasticity at higher frequency [363], while the last tanh term models the expected complete subtraction of CGBs signal at frequencies $f > f_{\rm knee}$. In order to keep the notation compact, we define $\log_{10}(h^2\Omega_{\rm Gal}) \equiv \alpha_{\rm Gal}$. From Ref. [363], we set the fiducial values $a_1 = -0.15$, $b_1 = -2.72$, $a_k = -0.37$, $b_k = -2.49$, $\alpha = 1.56$, $f_2 = 6.7\times10^{-4}$Hz and $\alpha_{\rm Gal} = -7.84$.

Extragalactic foreground. The extragalactic foreground, arising from the incoherent superposition of all extragalactic compact object mergers, includes potential contributions from SOBBHs, BNSs, EMRIs, and DWDs in their inspiral phase. In this work, we focus on only the SOBBH+BNS contribution, leaving any potential EMRI and DWD contribution to future work. Recent studies suggest that extreme mass-ratio inspirals can largely contribute to the foreground but only in somewhat extreme population synthesis scenarios [347]. Extragalactic DWDs may also be more abundant than previously estimated, with a relevant impact on the extragalactic foreground [349, 364] which ongoing analyses are verifying [365]. In the lack of a firmer understanding, we assume these contributions to be below the foregrounds of galactic binaries and extragalactic SOBBHs and BNSs.

We now focus on what we will assume to be the dominant contribution, the SOBBH and BNS foreground. The vast majority of these signals cannot be individually resolved by LISA [366–368] and, for the most part individual detections are possible for the few multi-band sources [369] (see [370], for a more accurate study of such sources). The best estimates for the populations of these objects are based on observations from ground-based detectors [371, 372]. Due to the relatively uniform distribution of the sources and the limited angular resolution of LISA, this component can be well modeled as an isotropic

[14]Keeping track of anisotropic nature of the signal, requires the analysis to be time-frequency. To consistently work this out, one would also need to keep track of the non-stationary nature of the noise (see e.g. [59, 357–360]) and the presence of gaps in the data.

SGWB signal with the power-law shape

$$h^2\Omega_{\rm GW}^{\rm Ext}(f) = h^2\Omega_{\rm Ext}\left(\frac{f}{1\rm mHz}\right)^{2/3}\,, \tag{5.16}$$

where $h^2\Omega_{\rm Ext}$ is the amplitude at 1 mHz. Recent observations by LIGO-Virgo-KAGRA collaboration estimate the magnitude of SGWB signal from SOBBHs and BNS as [371]

$$\Omega_{\rm Ext} = 7.2^{+3.3}_{-2.3}\times 10^{-10} \text{ at } f = 25\,\rm Hz\,. \tag{5.17}$$

In order to keep the notation compact, we define $\log_{10}(h^2\Omega_{\rm Ext}) \equiv \alpha_{\rm Ext}$. Extrapolating this amplitude to the LISA band [348], yields the fiducial value $\alpha_{\rm Ext} = -12.38$.

5.2 Analysis of the simulated data

In this section, we summarise the data analysis scheme implemented in the `SGWBinner` code (see Refs. [61, 62] for more details). Let us start with the generation of simulated data. Given the effective observation time $T_{\rm obs}$ and the number of segments N_d (which define the duration of each segment $\tau = T_{\rm obs}/N_d$), the code generates the data $\tilde{d}_i^s(f_k)$ $(s = 1, ..., N_d)$ segment-by-segment in the frequency-domain. For each frequency bin f_k (spanning $[3\times 10^{-5}, 0.5]$ Hz with spacing $\Delta f = 1/\tau$), N_d Gaussian realizations of the signal, noise, and foregrounds are generated with zero mean and variances defined by their respective PSDs. These data are then averaged over segments to define $\bar{D}_{ij}^k \equiv \sum_{s=1}^{N_d} d_i^s(f_k) d_j^{s*}(f_k)/N_d$, which gives an estimate of the total power at all frequencies. The next step consists of coarse-graining the data using inverse variance weighting. This results in a coarser set of frequency bins f_{ij}^k and a data set D_{ij}^k with weights n_{ij}^k, retaining similar statistical properties of the original dataset. Similarly to Refs. [62, 67, 328, 329, 332], we set $\tau = 11.4$ days ($\Delta f = 10^{-6}$ Hz), $N_d = 126$, and $T_{\rm obs} = 4$ years in our analysis.

The likelihood employed in the code reads [62]

$$\ln\mathcal{L}(D|\boldsymbol{\theta}) = \frac{1}{3}\ln\mathcal{L}_{\rm G}(D|\boldsymbol{\theta}) + \frac{2}{3}\ln\mathcal{L}_{\rm LN}(D|\boldsymbol{\theta})\,, \tag{5.18}$$

with

$$\ln\mathcal{L}_{\rm G}(D|\boldsymbol{\theta}) = -\frac{N_d}{2}\sum_{i\in\{\rm AET\}}\sum_k n_{ii}^k\left[\frac{\mathcal{D}_{ii}^{th}(f_{ii}^k,\boldsymbol{\theta}) - \mathcal{D}_{ii}^k}{\mathcal{D}_{ii}^{th}(f_{ii}^k,\boldsymbol{\theta})}\right]^2\,, \tag{5.19}$$

$$\ln\mathcal{L}_{\rm LN}(D|\boldsymbol{\theta}) = -\frac{N_d}{2}\sum_{i\in\{\rm AET\}}\sum_k n_{ii}^k\ln^2\left[\frac{\mathcal{D}_{ii}^{th}(f_{ii}^k,\boldsymbol{\theta})}{\mathcal{D}_{ii}^k}\right]\,, \tag{5.20}$$

where the index k runs over the coarse-grained data points and $\mathcal{D}_{ii}^{th}(f,\boldsymbol{\theta})$ denotes the theoretical predictions for the data, depending on some parameters $\boldsymbol{\theta}$. The model can be further expressed as $\mathcal{D}_{ii}^{th}(f,\boldsymbol{\theta}) \equiv \mathcal{R}_{ii}\,h^2\Omega_{\rm GW}(f,\boldsymbol{\theta}_{\rm cosmo},\boldsymbol{\theta}_{\rm fg}) + P_{N,ii}^{\Omega}(f,\boldsymbol{\theta}_{\rm n})$, with $\boldsymbol{\theta}_{\rm cosmo},\boldsymbol{\theta}_{\rm fg}$ and $\boldsymbol{\theta}_{\rm n}$ denoting the signal, foreground, and noise parameters, respectively. Notice that the diagonality of the AET basis has been exploited, and no cross terms appear in the likelihood. Given some priors $\pi(\boldsymbol{\theta})$ for the parameters, the posterior distribution reads

$$p(\boldsymbol{\theta}|D) \equiv \frac{\pi(\boldsymbol{\theta})\mathcal{L}(D|\boldsymbol{\theta})}{Z(D)}\,, \tag{5.21}$$

where $Z(D)$ is the model evidence defined as

$$Z(D) \equiv \int \mathrm{d}\boldsymbol{\theta}\, \pi(\boldsymbol{\theta}) \mathcal{L}(D|\boldsymbol{\theta}) \,. \tag{5.22}$$

To compare the validity of two different models $M_i(\boldsymbol{\theta}_i)$, each characterized by a set of parameters $\boldsymbol{\theta}_i$, we can use the evidence Z_i as a measure of the quality of the models given the data. The *Bayes factor* for two models i, j is defined as $B_{ij} \equiv Z_i/Z_j$. This Bayes factor can then be compared to the Jeffreys' scale [373] to determine which model is favored by the data.

As key functionalities, the `SGWBinner` code offers *i)* model-agnostic signal reconstruction and *ii)* template-based signal reconstruction. The former fits the signal in each frequency bin using a power-law template, i.e. the signal parameters are

$$\boldsymbol{\theta}_{\rm cosmo} = \{\alpha_1, n_{\mathrm{T},1}, \ldots, \alpha_n, n_{\mathrm{T},n}\} \ , \tag{5.23}$$

with n denoting the number of bins. The number and width of the bins are dynamically adjusted as described in Refs. [61, 62]. In practice, this method enables a preliminary identification of the spectral shape of the signal, which can guide the choice of the template for the template-based analysis. For the latter, the vector of parameters of the cosmological component $\boldsymbol{\theta}_{\rm cosmo}$ corresponds to the template parameters.

In this work, we assume the fiducial noise and foreground parameters to be

$$\boldsymbol{\theta}_{\rm n} = \{A_{\rm noise}, P_{\rm noise}\}, \quad \boldsymbol{\theta}_{\rm fg} = \{\alpha_{\rm Gal}, \alpha_{\rm Ext}\}, \tag{5.24}$$

while we assume that the other foreground parameters are known[15], and $\boldsymbol{\theta}_{\rm cosmo}$ is model dependent. Moreover, when fitting the simulated data, we use the same noise model applied to generate the data.[16] Both for the noise and foreground amplitudes, we assume Gaussian priors centered on their fiducial values. For the former, we set the standard deviation to be 20% of the fiducial mean value. For $\alpha_{\rm Gal}$ and $\alpha_{\rm Ext}$, we set the standard deviation to be 0.21 and 0.17, respectively. To sample the parameter space the code relies on the `Cobaya` [374] inference framework. To facilitate template-based analysis specifically for SIGW signals, we develop the dedicated `SIGWAY` code. As detailed in App. A, the `SIGWAY` code implements the parameterization of curvature perturbations discussed in Sec. 3 and performs the numerical computation of SIGW signals.

Finally, the code also supports Fisher analysis. In practice, the Fisher Information Matrix (FIM) can be computed by the continuous integral over the frequency range, expressed as

$$F_{ab} \equiv T_{\rm obs} \sum_{i \in \{\mathrm{AET}\}} \int_{f_{\rm min}}^{f_{\rm max}} \mathrm{d}f \ \left. \frac{\partial \ln \mathcal{D}_{ii}^{th}}{\partial \theta^a} \frac{\partial \ln \mathcal{D}_{ii}^{th}}{\partial \theta^b} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{\rm fid}} , \tag{5.25}$$

where $f_{\rm min}$ and $f_{\rm max}$ represent the detector's minimal and maximum measured frequencies, assumed to be $f_{\rm min} = 3 \times 10^{-5}\,\mathrm{Hz}$ and $f_{\rm max} = 0.5\,\mathrm{Hz}$ [54]. If non-trivial (log-)priors

[15] The effect of loosening this assumption on the signal reconstruction has been discussed in Ref. [332].

[16] We note that any differences between the instrumental noise and the model could introduce bias. This issue will have to be closely monitored in future upgrades of the code.

are included in the analysis, the code consistently adds their derivatives to Eq. (5.25) to obtain the full FIM. The relative uncertainty on the reconstruction parameters can then be estimated from the covariance $\mathrm{cov}_{ab} = F_{ab}^{-1}$. Given its computational efficiency, we also employed the FIM approach to assess the prospect of signal reconstruction with some level of accuracy. Note that in the case of SIGW signals, the FIM can be efficiently computed using the automatic differentiation feature of the JAX library [375] by applying the $\boldsymbol{\theta}$ derivative in Eq. (5.25) directly to $\mathcal{P}_\zeta$, before it is integrated to yield $\mathcal{P}_h$. We stress that the FIM formalism only works under the assumption that the likelihood is well approximated by a Gaussian distribution in the model parameters around the best fit (and that, when dealing with real data, the true values of the parameters lie within the region where the FIM is evaluated).

Finally, to complement the visualization of relative uncertainties in the parameter space, we will plot the signal-to-noise ratio (SNR) defined as

$$\mathrm{SNR} \equiv \sqrt{T_{\rm obs} \sum_{i\in\{\mathrm{AET}\}} \int_{f_{\rm min}}^{f_{\rm max}} \left(\frac{P^{\sigma}_{S,ii}}{P_{N,ii}}\right)^2 \mathrm{d}f} \,, \tag{5.26}$$

which scales linearly with the signal amplitude.

6 Results

In this section, we summarise our main results presenting different analyses based on the `SIGWAY` code outlined in Sec. 5.2 (see App. A for more details). We adopt the three approaches discussed in Sec. 3, namely *i)* binned spectrum agnostic approach; *ii)* template-based approach; *iii)* first principle USR model of inflation – all limited to the leading order SIGW and assuming an RD universe. We then consider specific examples including non-standard early universe evolution and non-Gaussianities.

We report results for both $\Omega_{\rm GW}$ and $\mathcal{P}_\zeta$. In the former case, we include the noise curves as well as the foregrounds as discussed in Sec. 5. Since the SIGW backgrounds we consider are emitted at very high redshift, when scales currently associated with mHz re-enter the Hubble sphere, they contribute to the energy budget in the early Universe and can affect cosmological observables as any other relativistic free-streaming component beyond the standard model. In particular, the SIGW contributes to the effective number of neutrino species as $N_{\rm eff} \equiv 3.044 + \Delta N_{\rm eff}^{\rm GW}$, with $\Delta N_{\rm eff}^{\rm GW} = \rho_{\rm GW}/\rho_{\nu,1}$ and $\rho_{\nu,1}$ is the energy density of a single neutrino species. Specifically, the total (integrated) GW abundance is $\Omega_{\rm GW}h^2 \simeq 1.6\cdot 10^{-6}\left(\Delta N_{\rm eff}^{\rm GW}/0.28\right)$ [10]. Measurements of the CMB [293] and Baryon Acoustic Oscillations (BAO) constrain $\Delta N_{\rm eff} \leq 0.28$ at 95% C.L. We report this bound for reference as shaded gray regions in the $\Omega_{\rm GW}h^2$ plots.

Strong primordial density perturbations can lead to the copious formation of PBHs with masses of the order of the horizon mass $M_H = 1.3\times 10^{-15} M_\odot \left[(k/\kappa_{\rm rm})/\mathrm{s}^{-1}\right]^{-2}$, where we kept track of the additional prefactor $\kappa_{\rm rm} \equiv k r_m \sim \mathcal{O}(3)$ that relates the perturbation scale to the characteristic perturbation size r_m at Hubble crossing [376–379]. Thus, the overproduction of dark matter in the form of PBHs in the asteroid mass range implies

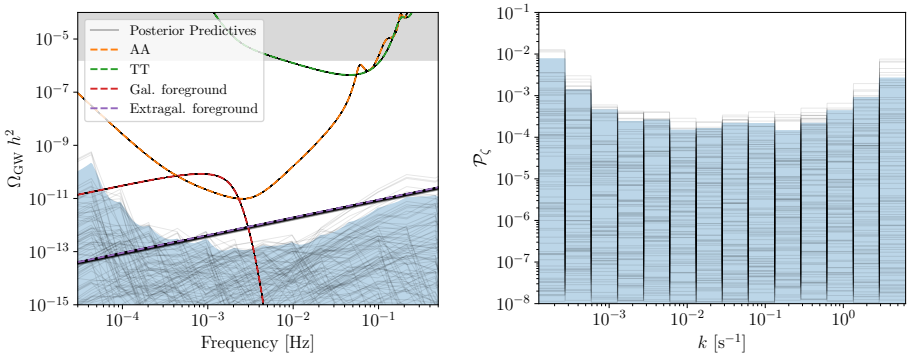


Figure 3. Posterior predictive distribution for both $\Omega_{\rm GW}h^2$ (left panel) and $\mathcal{P}_\zeta$ (right panel). We represent the binned reconstruction of Sec. 6.1 in the case without an injected signal. The posterior saturates the lower bound of the prior for the amplitudes A_i, due to the absence of a resolvable signal. We therefore can only set upper bounds on both $\Omega_{\rm GW}h^2$ and $\mathcal{P}_\zeta$. The light blue line shows the 95% credible intervals, while the pale black lines individual realisations of a signal sampled from the parameters' posterior distribution. The upper bound from $\Delta N_{\rm eff}$ is shown with a gray shading.

a bound $\mathcal{P}_\zeta \leq \mathcal{O}(10^{-2})$ [257, 380, 381] on the scalar curvature perturbations and thus also on the strength of the SIGW in the mHz frequency band. The abundance and the mass distribution of PBHs depend on the shape of the curvature power spectrum, non-Gaussianities, and on the equation of state of the universe during their formation [28], so does the implied upper bound on SIGWs.

6.1 Binned spectrum method

In Fig. 3 we report the constraints obtained with the binned method (see Sec. 3.1 and 4.4) when injecting no SIGW signal. This analysis forecasts the model-independent upper bounds on both the SIGW energy density spectrum (left panel) and the primordial curvature power spectrum (right panel) in case of no SGWB detection at LISA: this only relies on observational data, without assuming a specific signal model. For this analysis, we assume that the spectrum is divided into $N = 15$ bins. The free parameters in this model are

$$\boldsymbol{\theta}_{\rm cosmo} = \{A_1, \cdots, A_{15}\}. \tag{6.1}$$

In the left panel of Fig. 3 we indicate the posterior predictive distribution for $\Omega_{\rm GW}$ with the shaded light blue region, denoting the 95% credible interval (CI). The upper bound effectively reflects the LISA sensitivity, which falls around two orders of magnitude below the noise components in the AA channel because of the long observation time $T_{\rm obs} = 4$yr (see Sec. 5.2). In the right panel, we then show how the LISA sensitivity translates into the $\mathcal{P}_\zeta$ parameter space. The figure displays the upper bounds on $\mathcal{P}_\zeta$ across the range of momenta k considered, which is $k \in [1.26 \times 10^{-4}, 6.28]$/s. The posterior saturates at the lower edge of the prior on the amplitude parameters $A_i > 10^{-8}$ reflecting the absence of detectable power beyond the noise level. The posterior predictive bands illustrate that the

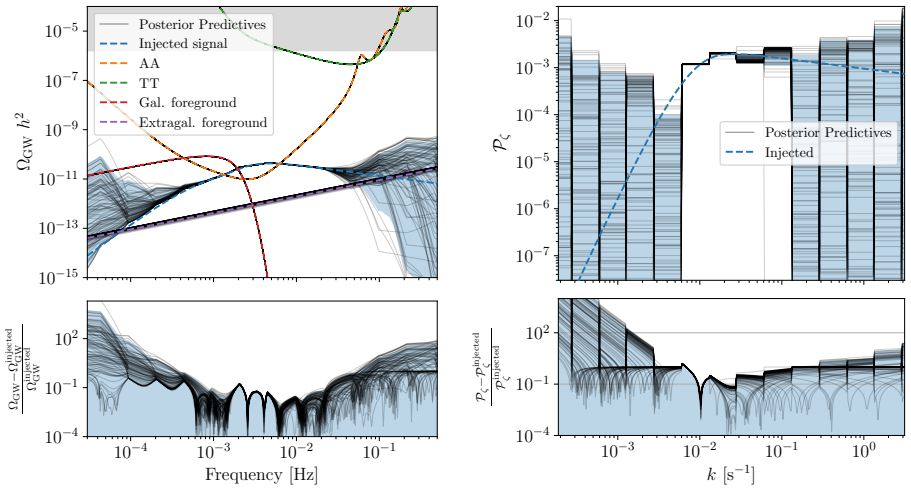


Figure 4. Same as Fig. 3, but simulating the observation of a signal obtained in the benchmark USR model scenario. The quantity $\mathcal{P}_\zeta$ is reconstructed with the model-independent binning method with 15 bins. The blue band in the upper panels shows the 90% (symmetric) credible interval, while the blue band in the bottom shows the 95% upper bound on the residuals.

method can constrain $\mathcal{P}_\zeta$ across several orders of magnitude in k. In the most sensitive range, this bound reaches $\mathcal{P}_\zeta \lesssim 2 \times 10^{-4}$.

This sensitivity is sufficient for probing a wide range of viable scenarios for asteroid mass PBHs. In particular, a non-detection of a SIGW by LISA would close the asteroid mass window for PBH dark matter formed from the collapse of moderately non-Gaussian curvature fluctuations, including models of PBHs from first-order phase transitions [382–384]. However, as with μ-distortion bounds on heavy PBHs [385–388], extremely strong non-Gaussianities could enhance PBH production and potentially allow evading these constraints. Such extreme scenarios and their theoretical consistency should be studied case by case.

In Fig. 4 we report the constraints obtained with the binned method when injecting the benchmark SIGW signal derived from the single field USR model of Sec 2.4. The curvature power spectrum has a BPL shape (see Eq. (3.7)), with a peak at around $\mathcal{P}_\zeta \sim 2 \cdot 10^{-3}$. Again, we use a template with $N = 15$ bins. In the left panel of Fig. 4, we show the injected SIGW signal (blue dashed line), along with the posterior predictive distribution (light blue band). Although the low number of bins reduces the frequency resolution of our model compared to the one achieved by LISA, the SIGW spectrum is well reconstructed, reaching a precision of the order of a few percent around the peak. At the edges of the observable range of frequencies, the blue bands widen up indicating a poor constraining power on the tail regions. The right panel of Fig. 4 indicates the SIGW bounds translate into four bins being well constrained in the range $k \sim [10^{-2}, 10^{-1}]$/s with around $\mathcal{O}(10)\%$

precision, while the other ones being subject to an upper bound of similar amplitude as in Fig. 3.

One could in principle enhance the frequency resolution by using a template with a larger number of bins, at the cost of drastically increasing the computational cost of the Bayesian MCMC inference. In App. B we discuss these issues in more detail.

6.2 Template based method

In this section, we present a forecast on reconstructing the SIGW signal using a template-based method, addressing different scenarios discussed in Sec. 3.2.

6.2.1 Smooth spectra

Lognormal scalar spectrum. The first signal injection we consider is a log-normal shape of $\mathcal{P}_\zeta$ as defined in Eq. (3.5) with the benchmark values defined in Eq. 3.6, which we report here $\log_{10} A_s = -2.5$, $\log_{10} \Delta = \log_{10}(0.5)$, $\log_{10}\left(k_*/\mathrm{s}^{-1}\right) = -2$. This produces a loud signal with the typical double peak being fully within the LISA band. This choice gives us a concrete measure of the precision that is achievable in measuring the SIGW background in this scenario for a reasonably loud signal. The free parameters for this approach are

$$\boldsymbol{\theta}_{\rm cosmo} = \{\log_{10} A_s, \log_{10} \Delta, \log_{10}\left(k_*/\mathrm{s}^{-1}\right)\}. \tag{6.2}$$

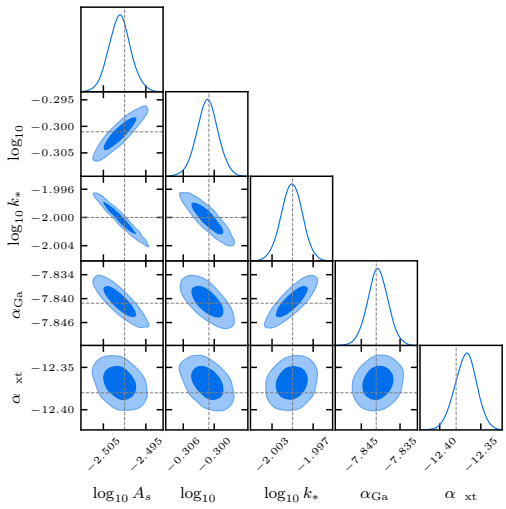


Figure 5. Corner plot with the posterior distribution for an injected LN spectrum $\mathcal{P}_\zeta$, as defined in Eqs. (3.5) and (3.6). k_* is expressed in 1/s units.

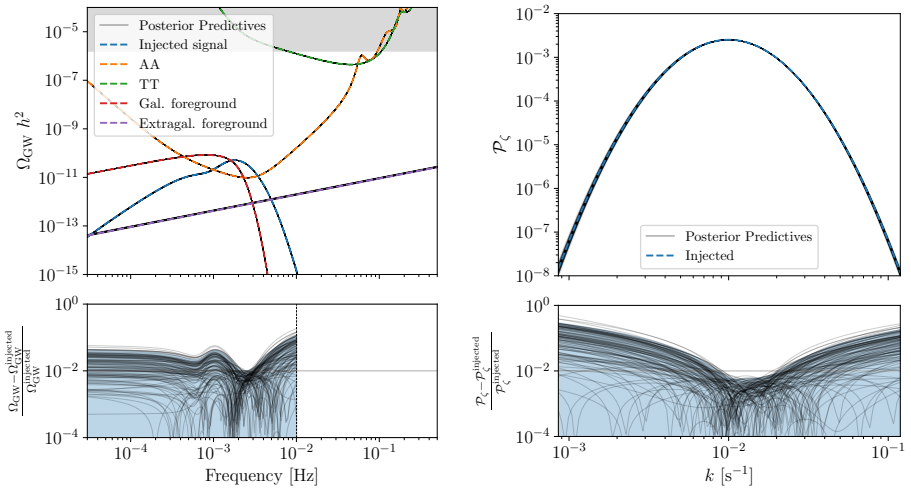


Figure 6. Same as Fig. 3, for an injected log-normal spectrum, as defined in Eqs. (3.5) and (3.6). In the left panel, we arbitrarily cut the posterior predictive where the signal falls below $\Omega_{\rm GW} \lesssim 10^{-15}$.

Fig. 5 shows the posterior distribution for each parameter of the LN template, alongside the ones describing the galactic/extragalactic foregrounds. We omit in these plots the posterior distributions for the noise parameters, as they are weakly correlated with the others in all cases. The injected values are indicated with a dashed gray line. As we can see, due to the relatively high SNR of the injected signal, the parameters of this template are very accurately reconstructed, with $\mathcal{P}_\zeta$ being reconstructed with a relative error of a few percent at its peak. The correlation between A_s and Δ originates from the definition of $\mathcal{P}_\zeta$. As customary in the literature [255], the amplitude at the peak is A_s/Δ, while A_s is the integrated power spectrum $\int_{-\infty}^{\infty} \mathcal{P}_\zeta \, \mathrm{d}\log k = A_s$. Therefore, A_s enters the power spectrum only through the ratio A_s/Δ, thus the positive correlation between the two parameters. Defining A_s to be the peak amplitude would avoid this degeneracy. The correlation of A_s, k_*, and $\alpha_{\rm Gal}$ is instead specific of our choice of fiducial parameters. As can be seen from the posterior predictive distribution in the left panel in Fig. 6, the injected signal is close enough to the galactic foreground that a slight increase in A_s with a decrease in k_* can be compensated by a small decrease in the background amplitude $\alpha_{\rm Gal}$. We expect that these correlations fade away with a larger injected k_*, when the signal and the galactic background are more distinct.

Fig. 6 shows the posterior predictive distribution for the SIGW (left panel) and curvature power spectrum (right panel). We see that the signal reconstruction achieves better than percent uncertainties on $\Omega_{\rm GW}$ and $\mathcal{P}_\zeta$ around the peak. The uncertainty on the low-frequency tail of $\Omega_{\rm GW}$ saturates at around a few percent, due to the universal behavior of the causality tail [389] sufficiently deep in the IR. See e.g. Appendix B of [67] for a dis-

cussion of logarithmic corrections to the IR tail of the SIGW spectrum. The uncertainty on the tails of $\mathcal{P}_\zeta$ remains low even at very small values, because of the rigid assumption about the LN template.

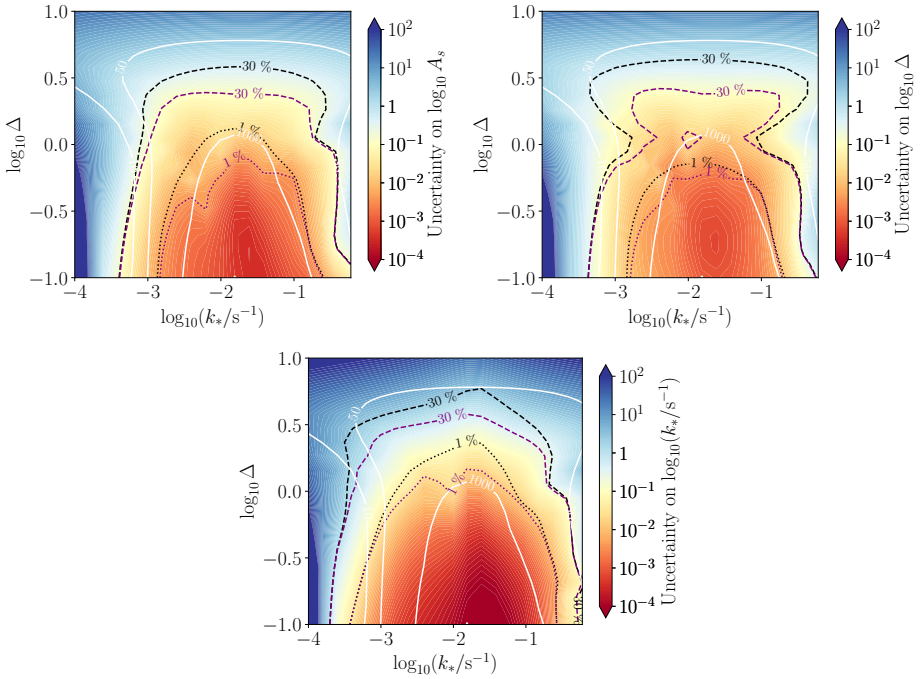


Figure 7. From left to right: Relative uncertainties of each of the parameters of the LN template, computed with an FIM forecast injecting an LN $\mathcal{P}_\zeta$ with fixed amplitude $\log_{10} A_s = -2.5$ and varying k_* and Δ. Black (purple) contours show uncertainties without (with) astrophysical foregrounds. The white line indicates SNR values.

In Fig. 7, we scan the parameter space in k_* and Δ estimating the relative uncertainties on all LN parameters using the FIM method. We fix the amplitude of $\mathcal{P}_\zeta$ to $\log_{10} A_s = -2.5$. The SIGW amplitude scales like $\Omega_{\rm GW} \sim A_s^2$, and in the high SNR limit we expect the uncertainties on the parameter to scale inversely $\sim 1/A_s^2$. The results highlight the great sensitivity that is achievable on a SIGW background if the peak lies around the peak sensitivity of LISA, $k_* \sim 10^{-3} - 10^{-1}\,\mathrm{s}^{-1}$. In that range, the width Δ for an LN scalar spectrum can be measured with an accuracy of order 10% or better if $\Delta \lesssim \mathcal{O}(1)$. Notice from Fig. 7 that the purple contours, marking the sensitivity on the primordial SIGW background accounting for astrophysical foregrounds, degrade when k_* coincides with the expected peak of the white-dwarfs (WD) galactic foreground, and the two GW backgrounds are less distinguishable [390].

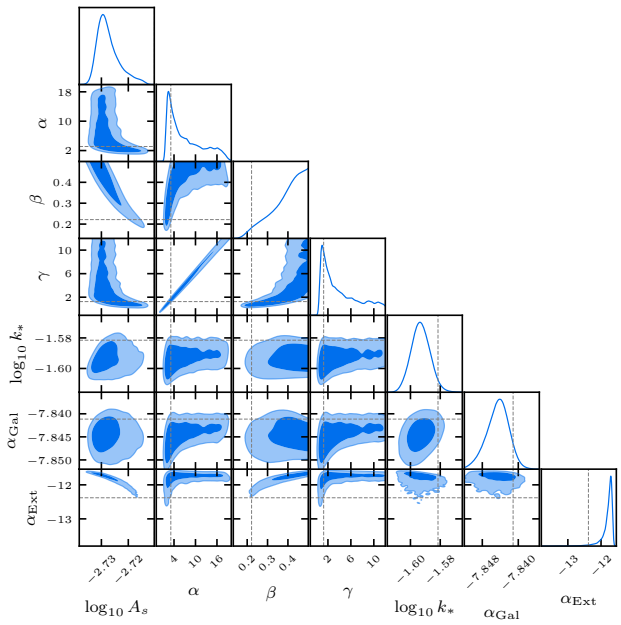


Figure 8. Same as Fig. 5, but for a recovered BPL curvature power spectrum assuming an injected signal motivated by the USR benchmark model. k_* is expressed in 1/s units.

Broken power law. The second injected signal is a BPL that is derived from the USR model discussed in Sec. 2.4 and 3.3, with input parameters as in Eqs. (3.7) and (3.8): $\log_{10} A_s = -2.71$, $\log_{10}(k_*/\mathrm{s}^{-1}) = -1.58$, $\alpha = 3.11$, $\beta = 0.221$, $\gamma = 1.25$. We reconstruct the signal using a BPL template. The results of the reconstruction of the signal using the USR model will be discussed below. The free parameters for this approach are

$$\boldsymbol{\theta}_{\rm cosmo} = \{\log_{10} A_s, \log_{10}\left(k_*/\mathrm{s}^{-1}\right), \alpha, \beta, \gamma\}. \tag{6.3}$$

In Fig. 8 we show a corner plot of the reconstructed parameters of the broken power-law template, while Fig. 9 displays the posterior predictive distribution for the SIGW (left panel) and curvature power spectrum (right panel). While the amplitude of the $\mathcal{P}_\zeta$ peak A_s and the peak position k_* are well reconstructed, the infrared (IR) spectral index α and the smoothing coefficient γ are poorly constrained, and α and γ appear to be very degenerate as the IR tail of the signal is hidden by the galactic foreground. This can be seen in the corner plot of Fig. 8 as well as from the right panel of Fig. 9, where the slope for $k < k_*$ has a large uncertainty. In order to reconstruct α and β with some precision, one would need to be sensitive to the tails of the signal outside the peak region. This would be only possible for a much larger signal, which nevertheless would have to compete with stringent bounds from PBH overproduction. With the relatively low SNR injected here, changes in the tilt can be traded for a smoother turnover around the peak, and vice versa.

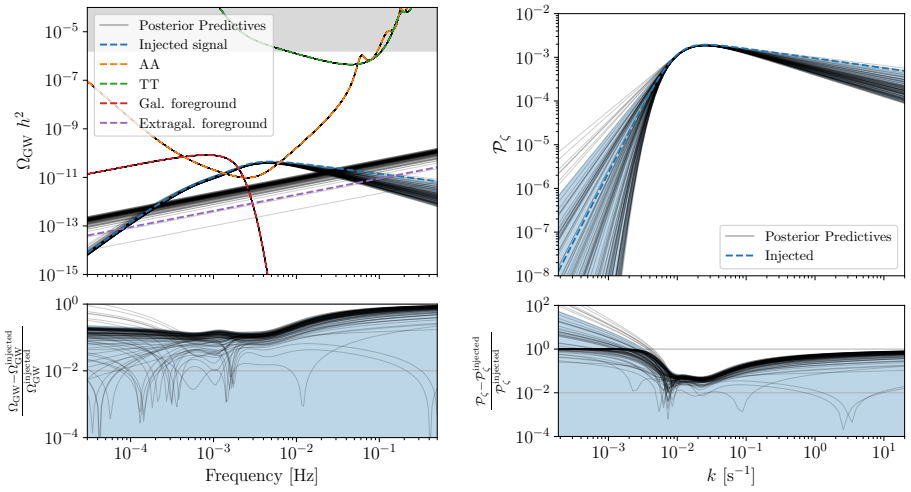


Figure 9. Same as Fig. 3, but for an injected signal from a BPL $\mathcal{P}_\zeta$, recovered using the BPL template. The injected parameters are motivated by the USR benchmark model.

Furthermore, there is a residual correlation between β and γ, although it is less pronounced. Biases in A_s, β, and $\alpha_{\rm Ext}$ are also evident due to their degeneracy in the high-frequency tail. Specifically, there is a tendency to reconstruct higher foreground values compared to the UV part of the SIGW spectrum. Importantly, we have verified that this bias is not an artifact introduced by the additional degeneracies induced by γ, which is correlated with both tilts and amplitude. This conclusion is supported by tests we ran with γ fixed to its injected value.

Notably, this bias does not appear when the same signal is reconstructed using the USR model. The USR model is inherently less flexible, and its UV tilt is better constrained, thereby mitigating the impact of degeneracies.

In Fig. 10 we show FIM estimates of uncertainties on the tilt parameters depending on the injected BPL shape and k_*. Tilts and γ can be independently resolved only if one observes with sufficient SNR the tail of the signal. Otherwise, a shallower (steeper) tilt can be traded off for a smoother (faster) transition. For this reason, in part of the parameter space explored in Fig. 10, we would obtain ill-conditioned FIM. In order to avoid this, only in this case we remove γ from the parameters of the FIM and fix it to the injected value.

The left panel of Fig. 10 shows the absolute uncertainty achievable on α in the (k_*, α) plane, and the right panel shows the same for β. We checked that these uncertainties do not depend on the injected value for the other tilt parameter. The slope of the IR tail of the GW spectrum is only mildly dependent on α, as discussed before, so most of the sensitivity comes from the signal in the frequency range around the peak. For this reason, the uncertainty on α reduces to 0.1 or better only if the SIGW is well within the LISA range, and on the right of the galactic WD foreground ($10^{-2} < k_*/s^{-1} < 10^{-1}$). Still,

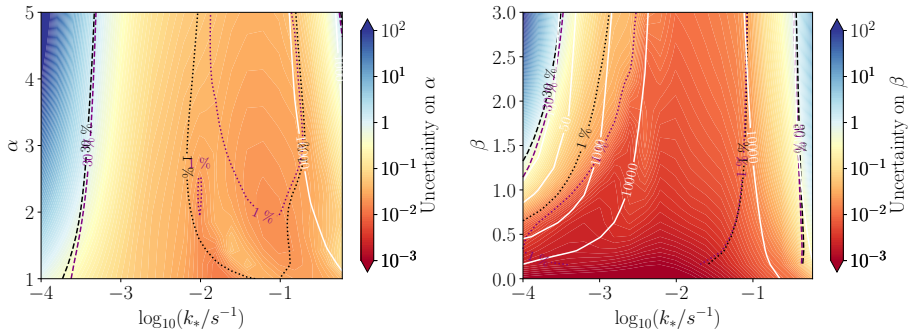


Figure 10. Uncertainty on α (left panel) and β (right panel) computed with the FIM approach for the broken power law $\mathcal{P}_\zeta$. The remaining parameters injected are those of the benchmark scenario in Eq. (3.8). We fix γ to the injection to remove the degeneracies with the two tilts in the small SNR regions of the parameter space.

it is very interesting to notice that steep values of the tilt, higher than $\alpha \simeq 1.5$, which are fully covered by the causality tail $\Omega_{\rm GW} \sim f^3$ [389, 391, 392] in the SIGW spectrum (up to log-corrections), can be well constrained, as information on the tilt is still retained in the shape of the double peak feature of the signal close to the dominant peak. The sensitivity to β (right panel of Fig. 10) is instead much better, as it determines the UV slope $\Omega_{\rm GW}(f) \sim f^{-2\beta}$. Therefore, β cannot be measured with an uncertainty smaller than 0.1 only if the SIGW lies outside LISA's peak range ($k_* > 10^{-1}\,{\rm s}^{-1}$) or if $\beta \gtrsim -2$, where the SIGW background falls too quickly in the UV.

6.2.2 Spectra with oscillations

Turns in multi-field inflation. As a benchmark example of a primordial feature in the power spectrum, we analyze a signal arising from turns in multi-field space as introduced in Sec. 3.2.2. The free parameters for this analysis are

$$\boldsymbol{\theta}_{\rm cosmo} = \{\log_{10} A_s, \log_{10}\left(k_*/{\rm s}^{-1}\right), \delta, \eta_\perp, F\}. \tag{6.4}$$

In Fig. 11 we show the posterior distributions for key parameters governing the sharp-turn scenario in multi-field inflation, along with the foreground parameters. The injection assumes the benchmark scenario where the parameters controlling the template (3.9) are fixed as in Eq. (3.12). We see that in this case, the parameter F is constrained to be close to unity with better than percent precision, showing the high sensitivity to the template oscillations. In this case, the signal amplitude and central scale k_* are weakly correlated, while the former is still positively correlated to both δ and $\eta_\perp$ which control the enhancement factor. Due to the ideal location of the SIGW peak, we also observe weak correlations between the foreground parameters and the signal parameters.

Figure 12 shows the posterior predictive distribution for $\Omega_{\rm GW}$ and $\mathcal{P}_\zeta$. In this example, the main peak of the SGWB lies within the LISA sensitivity band and above both astro-

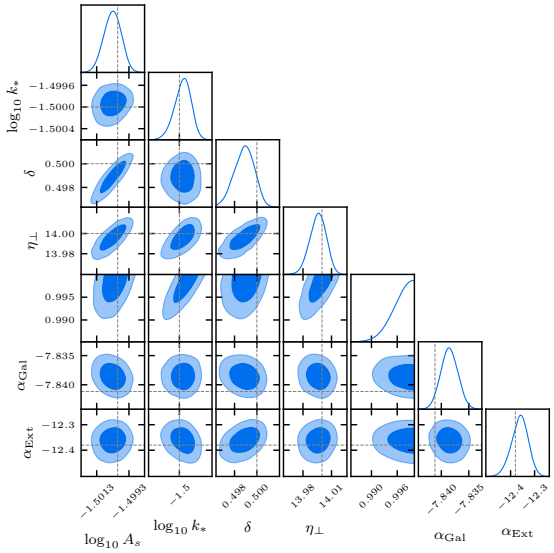


Figure 11. Same as Fig. 5, but for a simulated signal motivated by the multi-field scenario with sharp turns from (3.9). k_* is expressed in 1/s units.

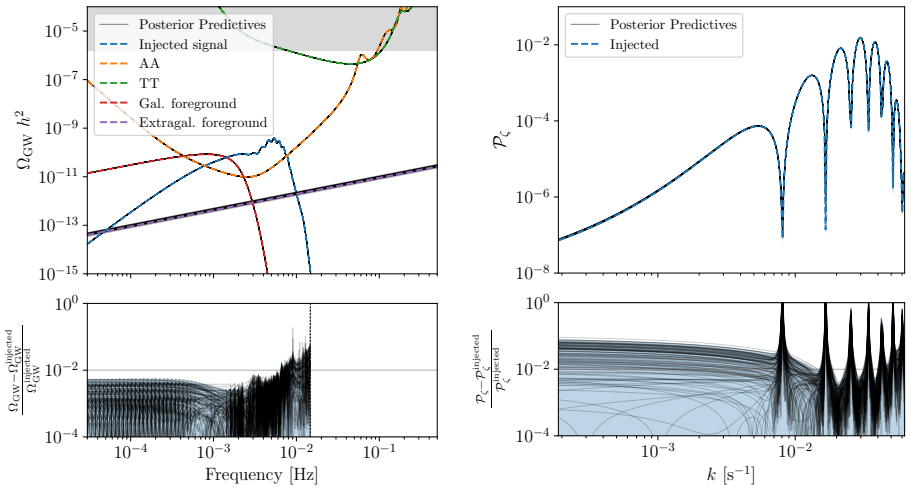


Figure 12. Same as Fig. 3, for the signal generated through a multi-field scenario with sharp turns from Eq. (3.9).

physical foregrounds. As a result, both the shape and amplitude of the peak in $\Omega_{\rm GW}$, along with the $O(20\%)$ modulations, are reconstructed at the percent level. Since the frequency

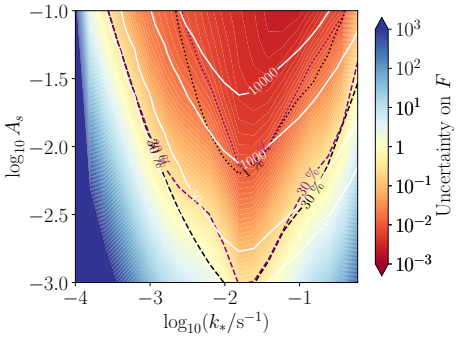


Figure 13. Fisher analysis for the oscillation template from multi-field inflation with turns. We vary k_* and A_s while keeping the remaining parameters fixed to the benchmark values (3.12).

of these modulations is linked to the oscillations in $\mathcal{P}_\zeta$ through the assumed thermal history at horizon re-entry, the oscillations are also reconstructed with high accuracy–see the right bottom plot. In this fortunate case, it would be possible to pinpoint the duration and strength of the field-space turn, as well as the inflationary time scale of the phenomenon. The latter is related to the oscillation frequency, as it is customary from the sharp feature phenomenon, while the former two can be disentangled by combining the peak amplitude, its location, and the frequency of the modulations.

Finally, the results of a Fisher analysis, highlighting the uncertainty in the parameter F associated with the oscillatory behavior, are presented in Fig. 13. There, we vary the power spectrum amplitude A_s and the position of the main peak, while keeping the other parameters fixed to the benchmark values discussed just above. This simplification is useful for illustrative purposes, as the parameters in the current model are not independent. Notably, when the signal is centered near the LISA sweet spot at $\log_{10}(k_*/s^{-1}) \simeq -2$, the oscillations can be accurately detected even with a moderate enhancement such as $\log_{10} A_s \simeq -3$.

Rapid transitions between SR and USR phases. The other injected spectrum with oscillatory features is characteristic of single field models with fast transitions from an SR to a USR phase described by Eqs. (3.13), (3.14). The free parameters for this analysis are

$$\boldsymbol{\theta}_{\rm cosmo} = \{\log_{10} A_s, \log_{10}\left(k_*/\mathrm{s}^{-1}\right), \nu_{\rm I}, \nu_{\rm II}, F\}. \tag{6.5}$$

We consider the benchmark scenario with the input parameters listed Eq. (3.18): $\log_{10} A_s = -2.58$, $\log_{10}(k_*/\mathrm{s}^{-1}) = -2.02$, $\nu_I = 1.95$, $\nu_{II} = 1.61$, $\gamma = 1.67$, $F = 1$. We perform the MCMC Bayesian inference modelling of the signal using the template (3.14), which allows us to turn on the oscillations smoothly by varying the parameter F from 0 to 1. The value $F = 0$ corresponds to a featureless BPL similar to the one considered above. Furthermore, we fix γ as the strong degeneracy between α and γ (see Fig. 8) makes sampling challenging and since our main concern lies in constraining F.

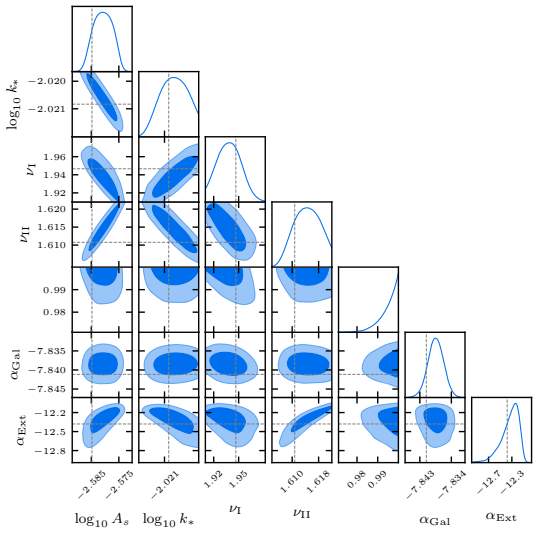


Figure 14. Same as Fig. 5, but for a simulated signal from Eq. (3.14). k_* is expressed in $1/s$ units. Note the injected value $F = 1$ is not visible sitting at the edge of the plot.

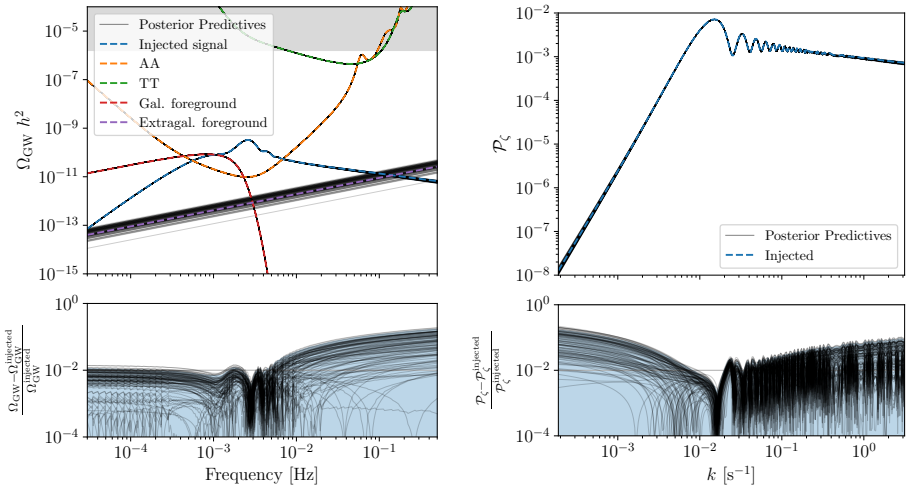


Figure 15. Same as Fig. 3, for the signal generated through a fast transition from SR to USR from (3.14). Both $\mathcal{P}_\zeta$ and $\Omega_{\rm GW}$ are reconstructed very well.

In Fig. 14 we show the posterior distribution for each parameter of the signal and foregrounds. First of all, we see the parameter F controlling the relevance of the oscillations

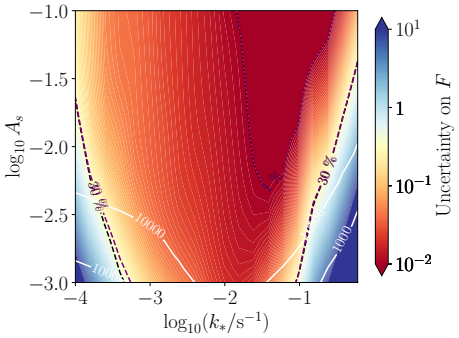


Figure 16. Absolute uncertainty on F estimated using the FIM for the oscillation template from a sharp transition between SR and USR. We vary $\log_{10} k_*$ and $\log_{10} A_s$ while keeping the remaining parameters fixed to the benchmark values (3.18).

over the smooth BPL is very tightly constrained around unity. This tells us the presence of oscillations can be resolved with high accuracy for such a high-SNR signal. The sensitivity to oscillations is mainly driven by the dominant peak, as we will discuss in the following. We also find tight correlations between the parameters, which are non-trivially connected in the signal template (3.13). In particular, we observe a strong correlation between the BPL tilts and the amplitude, due to the large impact of the former on the overall amplitude of the dominant peak. The negative correlation in the (A_s, k_*) plane is probably induced by the way the dominant peak, contributing to most of the SNR, can be adjusted, as one could lower the characteristic scale by enhancing the amplitude.

In Fig. 15 we show the posterior predictive distribution for both $\Omega_{\rm GW}$ and $\mathcal{P}_\zeta$. The presence of a dominant peak at scales around k_* in the right plot leads to a distinctive large enhancement of the SIGW signal around peak frequencies seen in the left plot. Additional oscillations in the SIGW spectrum can be observed at larger frequencies, although the second-order emission soon washes out further oscillations in the UV tail. The residuals of $\Omega_{\rm GW}$ show that the IR tail of the signal is reconstructed at around the percent level, with a flat behavior due to the causality tail dominating the IR. On the other hand, the relative deviation grows larger than $\mathcal{O}(10)\%$ percent in the UV part of the plot, due to the finite precision at reconstructing $\nu_{\rm II}$. The best accuracy is obtained around the peak, as expected. Correspondingly, in the right plot, $\mathcal{P}_\zeta$ is reconstructed with better than percent accuracy around the peak, while the reconstruction degrades in both tails. The oscillations are reconstructed for a few cycles in k, while they are lost in the UV as seen in the bottom panel showing the relative deviation from the injected signal. The envelope of the out-of-phase oscillations behaves following the underlying power-law tail. Note that the tails are reconstructed better than in the pure BPL scenario, as in this template (3.14), the shape of the dominant peak also brings information on the parameter $\nu_{\rm I,II}$ controlling the tails.

We perform a Fisher analysis focusing on the uncertainty on F, by varying $\log_{10} A_s$ and $\log_{10} k_*$, see Fig. 16. The oscillations are very well recovered, to $\mathcal{O}(10^{-2})$ uncertainty,

if the peak of the signal falls within the LISA band. This is because the oscillations in $\mathcal{P}_\zeta$ translate into oscillations mainly around the peak in the SGWB, as visible in Fig. 15.

6.3 Single field USR inference

In Fig. 17 we show the reconstruction capability of the USR model parameters of Sec. 2.4, obtained by running a MCMC Bayesian inference using the USR inflationary model with free parameters

$$\boldsymbol{\theta}_{\rm cosmo} = \{\lambda, v, b_l, b_f\}. \tag{6.6}$$

controlling the inflaton potential. We assume that the CMB scale crosses the Hubble sphere $N = 58$ e-folds before the end of inflation. We therefore avoid modelling the reheating era, and postopone its inclusion for future work (see e.g. [393]). The input values determining the injected signal were introduced in Eqs. (2.4) and (2.6), but we report them here for convenience: $\lambda = 1.47312 \times 10^{-6}$, $v = 0.19688$, $b_l = 0.71223$, $b_f = 1.87 \times 10^{-5}$.

We understand the results as follows. The height of the peak in $\mathcal{P}_\zeta$ is proportional to λ and it is also sensitive to the tuning of b_f.[17] This results in the negative correlation between λ and b_f. The self-coupling λ can be constrained, even though with only $\mathcal{O}(1)$ precision, because it controls the slope of the potential and therefore the SR parameter η_H *before* the inflection point, which determines the growth of $\mathcal{P}_\zeta$ before the peak as discussed below Eq. (3.7). The parameters b_l and v appear to be strongly correlated, meaning that the linear term in b_l/v in Eq. (2.4) gives the dominant dependence on b_l in the potential. The galactic background is well reconstructed due to its large magnitude, while the extragalactic one is completely hidden by the USR signal.

Figure 18 shows the posterior predictives in $\Omega_{\rm GW}h^2$ and in $\mathcal{P}_\zeta$. It is interesting to compare the right panel of this figure with that of Fig. 9, which is obtained with the same injected signal but a different template for the reconstruction. In the present case, the spectrum of scalar perturbations $\mathcal{P}_\zeta$ is reconstructed with excellent precision, even if LISA is sensitive only to the peak. This comes from the fact that the spectrum for the USR model has a universal slope $\sim k^4$ in the IR, whereas the IR slope is a free parameter for the BPL model.

The relatively large uncertainty on the overall potential amplitude $V(\phi)$ in Fig. 19 is due to the degeneracies between the overall scale $V_0 \sim \lambda v^4$ and the parameter b_f controlling the enhancement. As we are only constraining the enhanced part of the spectrum, there is a tight correlation between λ and b_f. Adding information from CMB data in the inference would reduce this uncertainty by adding an independent constraint on V_0.

Comparison between different methods. We can compare the performance of different methods when fitting the same injected signal, which is taken to be the USR benchmark scenario. In Fig. 20 we show the upper bound at 95% C.L. on the relative difference between the posterior predictive distribution and the injected signal for the binned, template-based, and ab initio USR approaches.

[17] With other potential parameters fixed, we did find the approximate behavior $\mathcal{P}_\zeta \propto (1 - b_f/b_{f,*})^{-n}$ in the parameter region supporting peaked $\mathcal{P}_\zeta$. Here, $b_{f,*}$ and $n > 2$ are parameters that depend on the remaining parameters of the potential. Such scaling is observed in other models [123].

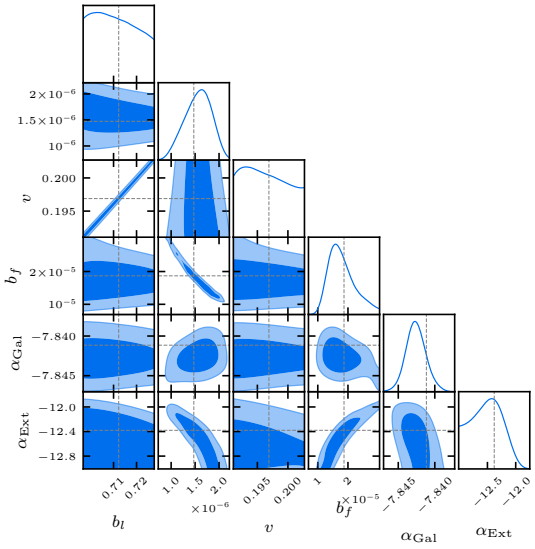


Figure 17. Same as Fig. 5, but for the USR reconstruction of the benchmark scenario.

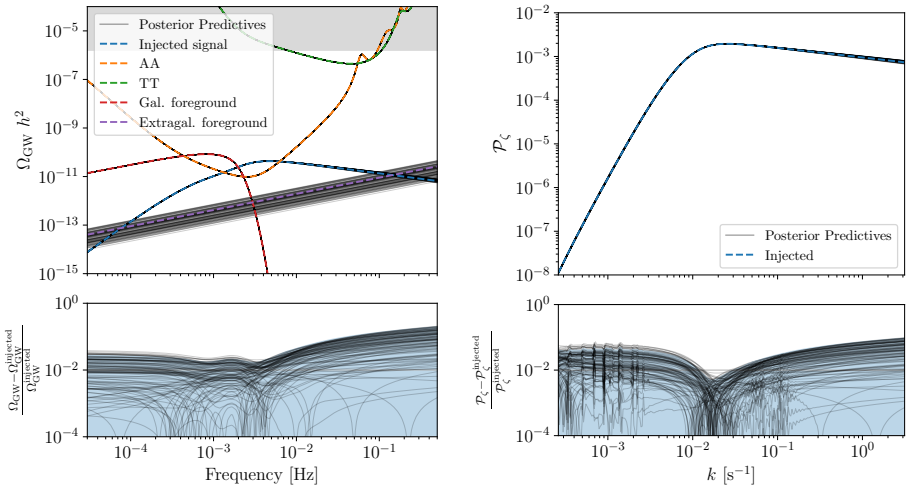


Figure 18. Same as Fig. 3, but for the USR reconstruction of the benchmark scenario.

We observe that the binned method provides a competitive constraint on $\Omega_{\rm GW}h^2$ in the central frequencies close to the peak (barring oscillations induced by the poor resolution associated with choosing 15 bins). However, the constraint quickly degrades at both ends,

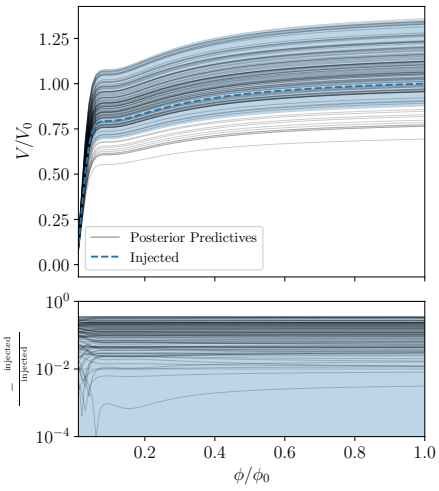


Figure 19. Posterior predictive distribution of the potential $V(\phi)$ for the USR reconstruction of the benchmark scenario.

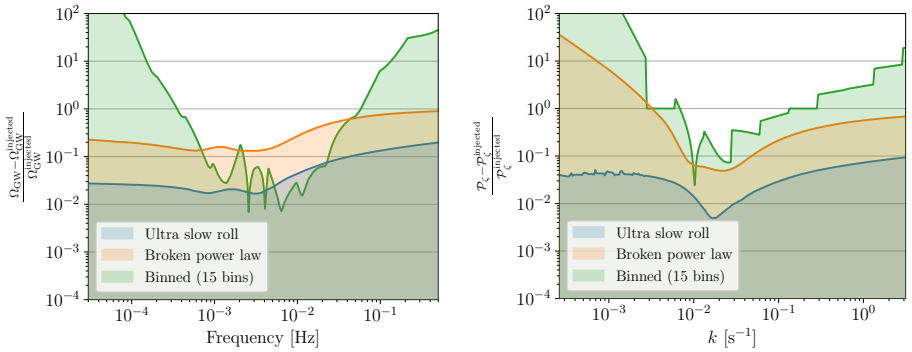


Figure 20. Comparison of the residuals between the three different methods for recovering the signal injected assuming the benchmark USR model.

due to the unconstrained curvature spectral amplitude there. The template-based method (assuming a priori a SIGW from a BPL scalar power spectrum) improves the reconstruction of the tails, but results in an overall loss of precision of a factor $\mathcal{O}(6)$ with respect to the posterior predictive derived assuming the USR scenario. This is most probably due to the larger number of parameters in the BPL template compared to the USR model, and the known degeneracy between γ and the two tilts around the peak.

Also, $\mathcal{P}_\zeta$, shown in the right panel, is most tightly constrained when using the USR model, showing the effectiveness and robustness of our analysis pipeline in reconstructing the injected high SNR signal. The BPL template gives an intermediate result on the IR tail,

which, however, degrades much faster than USR, due to the limited information on the tail of the SIGW, which is mostly controlled by the causality tail. Finally, the binned method gives a worse reconstruction of both the IR and UV tail, due to the small information within the LISA band about these regimes, and the independence assumed in this method between the central bins (best constrained) and the ones on the sides.

Overall, the binned method proves to be a powerful approach to explore the interpretation of a primordial background at LISA within a more agnostic approach. The comparison with specific SIGW templates does not significantly outperform the binned method for the range of frequencies around the peak, which are the best constrained by LISA. However, consistently with expectations, adopting the correct USR model provides greater accuracy in capturing the features of the signal, leading to a more precise reconstruction. These findings demonstrate the power of also adopting inference analyses based on explicit *ab initio* models (of which USR is just an example) that could outperform traditional template-based approaches. This, of course, assumes one can identify the best early universe model through model comparison. We will come back to discussing how to compare different scenarios in Sec. 7.

6.4 Non standard thermal histories

Using information on the SIGW spectrum, LISA would be able to challenge the vanilla assumption that the SIGW was emitted during a RD era. As discussed in Sec. 4, the kernels entering the computation of the SIGW spectrum bring information about the equation of state around the epoch of SIGW emission.

A sudden transition from eMD era to the RD era. We exemplify this case by showing how LISA can constrain the SIGWs emitted within an alternative thermal history by considering an early period of matter domination (eMD). We further assume sudden reheating, as introduced in Sec. 4.3. As we discussed, during this eMD epoch Φ does not decay, leading to an enhancement of the SIGW spectra around the scale $k \gtrsim 1/\eta_{\rm R}$. We take as a benchmark a nearly scale-invariant spectrum $\mathcal{P}_\zeta$, with a cutoff at placed at $k_{\rm max}$. We simplistically describe the spectrum as

$$\mathcal{P}_\zeta(k) = A_s \Theta\left(k_{\rm max} - k\right), \tag{6.7}$$

with benchmark parameters

$$A_s = 2.1 \times 10^{-9}, \qquad \eta_{\rm R} = 2000\,{\rm s}, \qquad k_{\rm max} = 0.06\,{\rm s}^{-1}. \tag{6.8}$$

Therefore the free parameters to examine in this case are

$$\boldsymbol{\theta}_{\rm cosmo} = \{A_s, k_{\rm max}, \eta_{\rm R}\}. \tag{6.9}$$

Our phenomenological parametrization should be regarded as a toy model, with the UV cut-off scale $k_{\rm max}$ introduced to ensure perturbativity, as assumed when computing the SIGW. For this reason, given the fact that the energy density contrast grows linearly with the scale factor during a MD era, i.e. $\delta\rho/\rho \propto a$, one can associate $k_{\rm max}$ as the scale at which

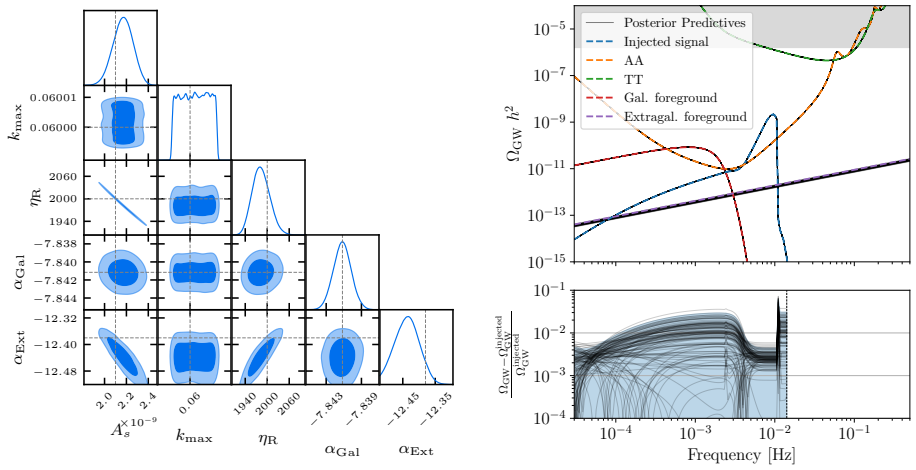


Figure 21. Left panel: Same as Fig. 5, but for the case of a nearly scale-invariant power spectrum and a sudden eMD to RD transition. $k_{\rm max}$ and $\eta_{\rm R}$ are expressed in units of s^{-1} and s, respectively. Right panel: Corresponding posterior predictive distribution for $\Omega_{\rm GW}h^2$. The notation used matches the one in Fig. 3.

the power spectrum of density contrast becomes unity, i.e. $\mathcal{P}_\delta(k_{\rm max}) = 1$ [302, 303, 394], although our actual choice is slightly more restrictive. One then can easily understand why $k_{\rm max}$ depends on the scale we are probing, as the source is largest for modes that spent the most time within the horizon during the eMD era. While we do not model the non-linear part of the spectrum, it may lead to further observational signatures [211, 395–397]. Finally, the template (6.7) can be made more realistic by introducing a smooth cut-off.

The parameters of this template are accurately reconstructed as shown in Fig. 21 (left panel). We notice that $\eta_{\rm R}$ and A_s are strongly correlated, as the duration of the MD signal directly controls the growth of perturbations emitting SIGWs, which is therefore degenerate with the primordial amplitude. The cut-off scale is also strongly constrained, with a marginalised posterior distribution which is flat within a narrow range of scales corresponding to the resolution adopted in this forecast. It should be kept in mind, however, that more realistic spectra would feature a smoother drop-off, alongside a contribution from non-linear scales not included here, thus jeopardizing the relevance of the constraining power on $k_{\rm kmax}$. In the right panel of Fig. 21, we show the posterior predictive distribution for the SIGW. We find the reconstruction to be accurate up to the cutoff scale (better than a few %). The SIGW spectrum is reconstructed well in the large-scale approximation, with the resonant amplification improving accuracy by an order of magnitude. The resonant peak is reconstructed with a larger accuracy due to its milder model dependence and due to its tilt being controlled by the resonant conditions (see discussion around Eq. (4.23)). Moreover, that part of the signal appears with a larger SNR in the LISA detector. The associated $\mathcal{P}_\zeta$ is accurately reconstructed as a flat spectrum with a maximum relative error

of order 10%.

Our numerical pipeline can also be applied to other scenarios for the thermal history of the early Universe. Of particular interest would be the study of time-dependent EoS parameters on the SIGWs, like in the case of smooth-crossovers [398, 399], analogous to the QCD phase transition.

6.5 Non-Gaussian effects on SIGWs

As discussed in Sec. 4.5, the tensor power spectrum of SIGWs receives contributions from the four-point correlation function of curvature perturbations. This contribution can be split into disconnected and connected terms. While the disconnected one depends only on the scalar power spectrum, the connected part arises from the primordial trispectrum, i.e. it is sensitive to primordial NG. And, as stressed in Sec. 4.5, for local NG, $\tau_{\rm NL}$ would be the key observable to extract a constraint on NG from SIGWs. However, when the curvature perturbation originates from a single fluctuating degree of freedom beyond the inflaton, the parameters $\tau_{\rm NL}$ and $f_{\rm NL}$ are connected as measures of higher-order correlations in the curvature perturbation, satisfying the relation $\tau_{\rm NL} = \left(\frac{6}{5} f_{\rm NL}\right)^2$, which saturates the Suyama-Yamaguchi inequality [400]. This relation has relevant implications for SIGWs, generated by models characterized by local-type NG, since observational constraints on one parameter can indirectly provide bounds on the other, assuming a given model. In the following analysis, we adopt the strategy of performing the analysis considering $f_{\rm NL}$ as a parameter of the model and assume the shape (4.38) of the full power spectrum. We then discuss the implications for $\tau_{\rm NL}$ that derive from the constraints on $f_{\rm NL}$. In this case, we assume the curvature power spectrum to have a LN profile, (see Eq. 3.5)[18] and as free parameters we use

$$\boldsymbol{\theta}_{\rm cosmo} = \{\log_{10} A_s, \log_{10} \Delta, \log_{10}\left(k_*/{\rm s}^{-1}\right), f_{\rm NL} \equiv 5/6\sqrt{\tau_{\rm NL}}\}. \tag{6.10}$$

The left panel of Fig. 22 shows the absolute uncertainty associated with $f_{\rm NL}$ varying the parameter $f_{\rm NL}$ against $\log_{10}(k_*)$ computed using the FIM method. We fix the amplitude of $\mathcal{P}_\zeta$ to $\log_{10} A_s = -2$ and the width to $\log_{10}\Delta = -0.75$. The dashed black (purple) contour lines represent the relative percentage error associated with $f_{\rm NL}$ when astrophysical foregrounds are not included (or are included). The white vertical line indicates the SNR. Notice that for small and large values of k_*, $f_{\rm NL}$ exhibits higher uncertainties, whereas the intermediate range of $k_* \sim 10^{-2} - 10^{-1}\, s^{-1}$ shows the minimal uncertainties for $f_{\rm NL}$. This suggests that tight constraints on $f_{\rm NL}$ can only be achieved within this specific range of k_*. Outside of this range, the errors increase notably, indicating less reliable measurements for $f_{\rm NL}$. Even in the optimal case, the reconstruction of $f_{\rm NL}$ only reaches the percent level for large $f_{\rm NL} \gtrsim 12.5$. Notice that, in the presence of foregrounds, the accuracy on $f_{\rm NL}$

[18] Note that our choice of the LN $\mathcal{P}_\zeta$ assumes that the dimensionless primordial curvature fluctuations – including the higher order term coming from the trispectrum (the left-hand side of Eq. (4.38) multiplied by $k^3/2\pi^2$) – describe a lognormal. This practically isolates the effect of computing $\mathcal{P}_\zeta$ including non-Gaussian contributions *ab-initio* from the effect caused by a NG contribution on the computation of $\Omega_{\rm GW}$ given $\mathcal{P}_\zeta$. We will also compare our results to the LN case where only the Gaussian contribution is considered.

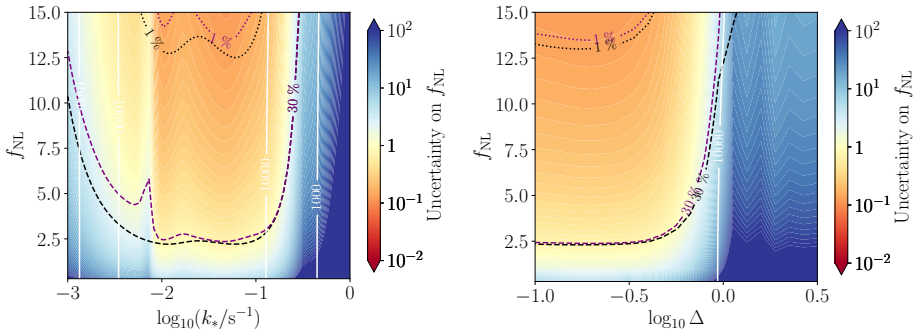


Figure 22. *Left panel:* Absolute uncertainties of each of the parameters in the case of an injected signal which includes NG and assuming the LN template with parameters fixed to $\log_{10} A_s = -2$ and $\log_{10}\Delta = -0.75$. *Right panel:* Same as the left panel, but varying Δ instead of k_*, for an injected signal which includes NG and assuming the LN template with parameters fixed to $\log_{10} A_s = -2$ and $\log_{10}(k_*/s^{-1}) = -1.4$.

slightly degrades, in particular when k_* coincides with the expected peak of the galactic foreground, i.e. when $k_* \sim 10^{-2.2}\,{\rm s}^{-1}$.

Similarly, the right panel of Fig. 22 shows the FIM absolute uncertainty associated with $f_{\rm NL}$ varying the parameter $f_{\rm NL}$ against $\log_{10}\Delta$. We fix the amplitude of $\mathcal{P}_\zeta$ to $\log_{10} A_s = -2$ and the peak scale to the optimal location $\log_{10}(k_*/{\rm s}^{-1}) = -1.4$. In this case, the uncertainty when estimating $f_{\rm NL}$ is lower ($\lesssim 30\%$) in the region of $\log_{10}\Delta$ below -0.3, while it significantly degrades for larger widths. This suggests that the most stringent measurements of $f_{\rm NL}$ will be obtained for relatively narrow curvature power spectra.

Given the relation between $f_{\rm NL}$ and $\tau_{\rm NL}$, improvements in the precision of the former directly translate into tighter constraints on the latter. The FIM analysis shows that percent-level accuracy on $f_{\rm NL}$ is achievable only for large $f_{\rm NL}$, which in turn would correspond to a percent-level constraint on $\tau_{\rm NL}$. In favorable scenarios –where the peak scale k_* lies within $10^{-2} - 10^{-1}\,{\rm s}^{-1}$ and the spectral width $\log_{10}\Delta$ is relatively narrow – uncertainties in $f_{\rm NL}$ are minimal, restricting the allowed range of $\tau_{\rm NL}$. Conversely, when $f_{\rm NL}$ is less precisely determined, $\tau_{\rm NL}$ remains poorly constrained.

The corner plot in Fig. 23 illustrates the posterior distributions of the SIGW signal parameters $\{\log_{10} A_s,\ \log_{10}(k_*/s^{-1}),\ \log_{10}\Delta\}$, including the primordial NG parameter $f_{\rm NL}$, alongside the extragalactic and galactic background amplitude energy densities $\log_{10}(h^2\Omega_{\rm Ext}) \equiv \alpha_{\rm Ext}$, $\log_{10}(h^2\Omega_{\rm Gal}) \equiv \alpha_{\rm Gal}$. As in the other cases, we omit the LISA noise parameters $A_{\rm noise}, P_{\rm noise}$ from the corner plot as they are tightly constrained and weakly correlated with the rest. As first benchmark, we injected a template with $\log_{10} A_s = -2$, $\log_{10}\Delta = -0.75$, $\log_{10}(k_*/{\rm s}^{-1}) = -2.3$ and $f_{\rm NL} = 1$. On the right panel of Fig. 23, we report the corresponding reconstructed $\Omega_{\rm GW}h^2$. For comparison, we also plot the GW energy density in the Gaussian case. Due to the low value of $f_{\rm NL}$ chosen, the reconstructed and Gaussian curves are almost superimposed, showing that the effects of NG are quite

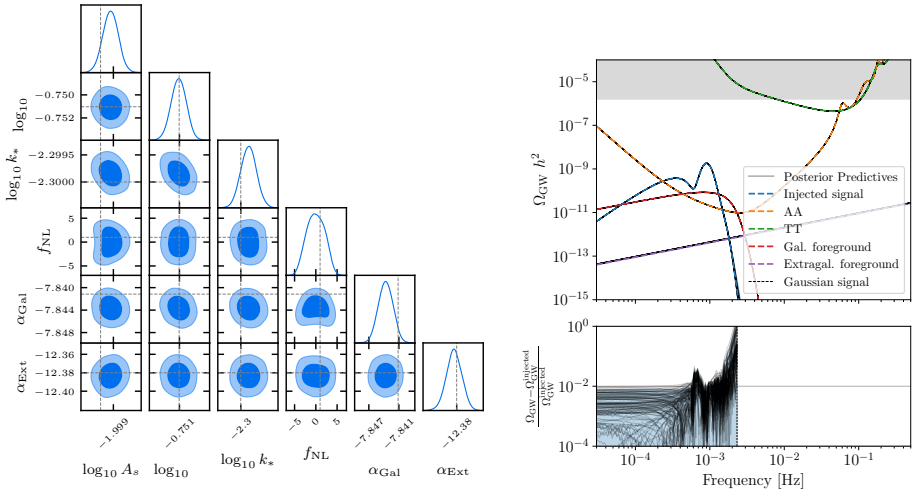


Figure 23. Posterior distribution for an injected signal which includes NG, $f_{\rm NL} = 1$ and assuming the LN with $\log_{10} A_s = -2$, $\log_{10} \Delta = -0.75$, $\log_{10}(k_*/\mathrm{s}^{-1}) = -2.3$.

mild in this case.

The marginalized posterior for $f_{\rm NL}$ exhibits a broad posterior distribution, ranging from -5 to 5. This distribution indicates a large uncertainty in $f_{\rm NL}$ spanning roughly ± 3 at the 68% C.L. and is compatible with $f_{\rm NL} = 0$. This suggests a limited constraining power of LISA on $f_{\rm NL}$. For such a signal, the posterior shows a bimodal structure, that arises because $f_{\rm NL}$ enters quadratically in the GW spectral energy density through the trispectrum. As the observed value suggests compatibility with a Gaussian primordial distribution, the broad posterior distribution also implies that ruling out moderate NG will be challenging, emphasizing the need for improved precision or additional data to refine these estimates. The other parameters, such as the log-amplitude of the seed power spectrum ($\log_{10} A_s$), show a remarkable reconstruction, in line with the results of Sec. 6.2.1. Note however that we are injecting different values of $\{\log_{10} A_s,\ \log_{10}(k_*/s^{-1}),\ \log_{10} \Delta\}$. For a comparison to the fully Gaussian case with the same injection in $\mathcal{P}_\zeta$ see Fig. 30. The effect of adding the small NG correction $f_{\rm NL} = 1$ on the recoverability of power spectral parameters $\{\log_{10} A_s,\ \log_{10}(k_*/s^{-1}),\ \log_{10} \Delta\}$ and foreground parameters $\{\alpha_{\rm Gal}, \alpha_{\rm Ext}\}$ is small in this case, indicating that a small NG contribution does not spoil the reconstruction of curvature power spectra parameters. This also indirectly supports our choice of not including NG corrections in the benchmark USR scenario discussed in Sec. 3.3, which is characterized by $f_{\rm NL} \simeq 0.09$. However, notice the visible non-zero correlation between $f_{\rm NL}$ and both $\{\log_{10} A_s, \alpha_{\rm Gal}\}$. The joint posterior $f_{\rm NL}$ - $\log_{10} A_s$ reflects the multimodality induced by the double peak structure of $f_{\rm NL}$. Nevertheless, LISA can still strongly constrain the amplitude of the non-linear power spectrum. Hence the detection of the GWB is not strongly influenced by the primordial NG, which is beneficial for simplifying the analysis

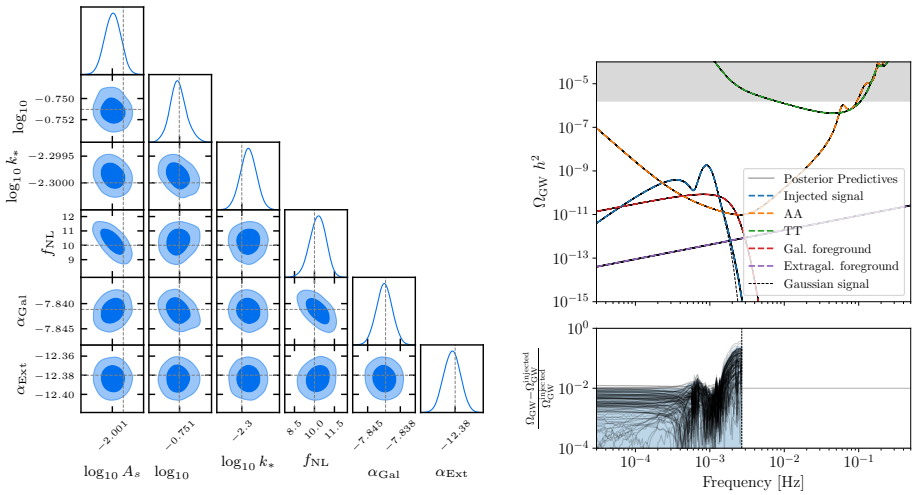


Figure 24. Same as Fig. 23 with $f_{\rm NL} = 10$. Note that there is a second, identical mode at $f_{\rm NL} = -10$ as only the square of $f_{\rm NL}$ enters $\Omega_{\rm GW}$. We omitted this mode.

when considering models predicting small NGs.

The posterior distributions for astrophysical foreground amplitudes are broader and show limited correlation with signal parameters. In contrast, instrumental noise parameters are tightly constrained and largely independent of signal estimation. This decoupling ensures robust estimation of the signal amplitude, enhancing the reliability of SGWB detection despite noise uncertainties.

Overall, the joint distributions show only weak correlations, indicating that each of these parameters can be inferred with a good degree of independence from the others.

The situation changes when considering a larger value of $f_{\rm NL}$, as shown in Fig. 24. Specifically, we run the MCMC with the following injected template with $\log_{10} A_s = -2$, $\log_{10}\Delta = -0.75$, $\log_{10}(k_*/{\rm s}^{-1}) = -2.3$ and $f_{\rm NL} = 10$, differing from the previous benchmark only in the choice of $f_{\rm NL}$. On the right panel, we report the reconstructed $\Omega_{\rm GW}h^2$ as well as the Gaussian counterpart for comparison. Given the higher value of $f_{\rm NL}$, the differences are now more evident, resulting in an enhancement of the UV tail, but also in a slightly higher peak and more smoothed minimum (not visible in Fig. 24). Now, the marginalized posterior distribution for $f_{\rm NL}$ shows only a narrow spread indicating that the estimate of $f_{\rm NL}$ is precise. The absolute value of $f_{\rm NL}$ has a mean reconstructed value that varies about ± 0.8 at the 68% C.L. which makes it incompatible with $f_{\rm NL} = 0$.

As with the previous case, the non-linear power spectral parameters are tightly constrained and no significant bias is observed with respect to the injected values. Both k_* and Δ are very weakly correlated with $f_{\rm NL}$, suggesting that their reconstruction is not heavily influenced when $f_{\rm NL}$ is large. However, $\log_{10} A_s$ shows a larger anti-correlation with the $f_{\rm NL}$. An anti-correlation also appears between $f_{\rm NL}$ and $\alpha_{\rm Gal}$, probably induced by

the similar IR shape behavior. The tighter constraints associated with such a signal imply that higher levels of NG are easier to constrain, yielding clearer and more reliable effects.

The independence between $f_{\rm NL}$ and other parameters (except for $\log_{10} A_s$), indicated by the weak correlations, suggests robustness in parameter estimation. This means that uncertainties in $f_{\rm NL}$ do not drastically affect the inference of other parameters, resulting in more precise parameter constraints compared to the $f_{\rm NL} = 1$ case. When $f_{\rm NL}$ is larger, the signal is more distinct, allowing for setting more stringent constraints on primordial NG. Note that the prior in Fig. 24 is restricted to positive values of $f_{\rm NL}$. Similarly to Fig. 23 the posterior distribution has a second mode at $f_{\rm NL} = -10$ since it only enters quadratically in the signal.

For the amplitude A_s considered in this case, the imprints due to possible inaccuracies in accounting for the full non-Gaussian behavior for some models of inflation are expected to be negligible when $f_{\rm NL} = 1$, but could be substantial when $f_{\rm NL} = 10$, as recently argued by [68]. For this analysis, we neglected those refinements. We further stress that for the enhanced amplitude of the power spectrum considered here, $f_{\rm NL} = 10$ represents much larger deviations from Gaussianity than on CMB scales, because the expansion parameter determining the relative size of the trispectrum versus power spectrum is $\tau_{\rm NL} \cdot \mathcal{P}_\zeta$. It is of order one in the current context, while less than 10^{-4} on CMB scales.

Concerning the implications for $\tau_{\rm NL}$, large uncertainties in $f_{\rm NL}$ directly translate into poor constraints on $\tau_{\rm NL}$. For example, for $f_{\rm NL} = 1$ with a quite broad uncertainty of ± 3, $\tau_{\rm NL}$ could span from values close to zero (if $f_{\rm NL} \approx 0$) up to $\simeq 23$ (if $f_{\rm NL} \approx 4$), making it challenging to clearly identify a primordial NG signal. In this range, even moderate NGs become difficult to distinguish from a Gaussian spectrum. Without improved precision on $f_{\rm NL}$, the corresponding $\tau_{\rm NL}$ will remain poorly determined, limiting our ability to discriminate between different levels of primordial NG. When $f_{\rm NL} = 10$, providing a more pronounced non-Gaussian signal, the corresponding $\tau_{\rm NL} = \left(\frac{6}{5} \cdot 10\right)^2 = 144$ is now much more tightly constrained. Since the uncertainty in $f_{\rm NL}$ is roughly ± 0.8, $\tau_{\rm NL}$ varies up to a $\pm 15\%$ range. This tighter range is obviously better than the scenario with small $f_{\rm NL}$. Hence, larger $f_{\rm NL}$ values significantly improve our ability to determine $\tau_{\rm NL}$, allowing LISA to better distinguish between different levels of primordial NG in the SGWB.

Finally, it is important to highlight that while Planck provides constraints that are very close to zero [254], indicating no evidence for primordial NG at large scales, the analysis we are performing for LISA focuses on NG at much smaller scales. LISA's ability to provide tight constraints on $f_{\rm NL}$ suggests that GW detection could play a crucial role in refining our understanding of primordial NG, particularly in scenarios where the signal is expected to be strong. In addition, the sensitivity of LISA to different scales compared to Planck provides an important cross-check, helping to verify any scale dependence for $f_{\rm NL}$ [401, 402]. Overall, while Planck remains a benchmark for CMB-based constraints on $f_{\rm NL}$ at large scales, LISA shows the potential of GW detectors to significantly advance the search for, and the characterization of primordial NG.

7 Testing the scalar-induced hypothesis

In this section, we outline a procedure to test the compatibility of the SIGW hypothesis with a possible SGWB detection. So far, our analysis assumed that the cosmological contribution of the SGWB originates from SIGWs. There are, however, many alternative sources of the SGWB that originate from different physical processes in the early universe. Our goal is to offer a practical approach for assessing the validity of the hypothesis explored in this work—namely, whether or not a hypothetically detected signal originates from enhanced scalar fluctuations of inflationary origin. To this end, we focus on two illustrative scenarios that are distinct in nature, leaving a detailed comparison of various early-universe signals—which is beyond the scope of this paper—for future work. We use the evidence (5.22) as an estimator for model selection. Specifically, given an injected signal, we consider different reconstruction techniques, for which we can compute the (log) evidence using the nested sampler `PolyChord` [403, 404].

It is important to note that the Bayes factor is a global estimator: it not only assesses the goodness of fit of a model to the data but also incorporates information about the prior volume and its compression as the prior transitions to the posterior. Additionally, given a similar fit to data, it naturally favors simpler models—those with fewer parameters—over more complex ones.

In the following analysis, the different approaches follow two distinct philosophies, depending on whether they assume, or not, that the signal originates from scalar-induced GWs:

1. Using the `SGWBinner` code [61, 62] we can use both a template-based [67], as well as a model-agnostic approach to reconstruct the signal in $\Omega_{\rm GW}$. This does not assume the underlying physics. The templates we use to fit the model in this case are informed by the injection, that is known to us. Of course, with real LISA data, these will be several shapes that are informed by physical processes that can potentially generate SGWBs. On the other hand, the binned approach of the `SGWBinner` divides the frequency space into bins by SNR and then fits a power law within each bin. This results in an agnostic reconstruction of $\Omega_{\rm GW}$.

2. By contrast, the various techniques presented in the previous sections assume that the signal we are considering is coming from SIGWs. Somewhat like with the `SGWBinner`, we can reconstruct the SGWB with the `SIGWAY` in two different ways. One option is to use a template-based approach (see Sec. 3.2), in which a template for $\mathcal{P}_\zeta$ is specified. A second possibility is to use the $\mathcal{P}_\zeta$-agnostic (still assuming SIGW to be the source of the SGWB) binned approach, as described in Sec. 3.1.[19] We choose the

[19] Let us note that there is a slight difference in the implementation of the binned $\mathcal{P}_\zeta(k)$ approach, compared to the binned $\Omega_{\rm GW}$. In the latter approach, the `SGWBinner` code dynamically selects the optimal number of bins before the nested sampling, based on the Akaike information criterium [405] (we refer the interested reader to the discussion around Eq. (3.6) of [61] for more details). On the other hand, such a feature has not been implemented in the `SIGWAY` code, as there are fundamental difficulties to attempting a similar approach in $\mathcal{P}_\zeta$-space (see App. B for a detailed discussion) so the number of bins for $\mathcal{P}_\zeta(k)$ has to be chosen by hand.

Package	Method	Case 1: not SIGW signal $\log\left[Z(\Omega_{\rm GW}^{\rm LN})\right]/10^4$	Case 2: SIGW signal $\log\left[Z(\mathcal{P}_\zeta^{\rm LN})\right]/10^4$
`SGWBinner`	Template	**0.5444**	−0.5278
	Binned	−0.5625	−0.5479
`SIGWAY`	Template	−25.6078	**−0.5203**
	Binned	−25.7934	−2.712

Table 1. log Bayes factors (normalized to a reference value 10^4) comparing the SGWB reconstruction (either using the injected model as a template, or a model-agnostic binned method) with the SIGW reconstruction for two signals: *i)* a log-normal power spectrum in $\Omega_{\rm GW}$ which cannot be generated by SIGW (within the assumptions we are working in), *ii)* a log-normal power spectrum in $\mathcal{P}_\zeta$ which generates detectable SIGW. In bold we show the Bayes factors for the recovery which assumes the injected template. As expected, they are the best reconstructions for each injection.

number of bins $N_{\rm bins} = 40$. For simplicity, in this Section, we assume the SIGW to be produced during the radiation-dominated era.

For illustrative purposes, we simulate the following two qualitatively different signals:

- **Case 1. Not SIGWs.** The first is a narrow lognormal in $\Omega_{\rm GW}(f)$. This injection serves as an example of a signal that cannot be produced by SIGWs, assuming the modes reenter during RD. In this case, regardless of how narrow the peak in $\mathcal{P}_\zeta$ is, the generated SIGW will always exhibit the so-called "causality tail" proportional to f^3. As a benchmark for this injection, we chose to reproduce the main peak of the double-peak background shown in the top-left panel of Fig. 11 in [67], with a slightly lowered amplitude. This amounts to choosing the following parameters: $\log_{10}(h^2\Omega_*) = -9.5$, $\log_{10}(f_*/{\rm Hz}) = -2.21$, $\log_{10}(\rho) = -1.10$ in Eq. (2.8) of [67]. We will henceforth call this signal $\Omega_{\rm GW}^{\rm LN}$.

- **Case 2. SIGWs.** The second injection is instead derived from a SIGW scenario. We inject a lognormal power spectrum of curvature perturbations, see Eq. (3.5) and compute the resulting SGWB numerically. The power spectrum parameters used were: $\log_{10} A_s = -2.3$, $\log_{10}\Delta = -0.70$, $\log_{10}(k_*/{\rm s}^{-1}) = -1.5$. The resulting shape in the SGWB can be described through the double-peak template in Eq. (2.10) of [67], with parameters $\{\log_{10}(h^2\Omega_*), \log_{10}(f_*/{\rm Hz}), \beta, \kappa_1, \kappa_2, \rho, \gamma\} = \{-9.5, -5, 0.242, 0.456, 1.234, 0.08, 6.91\}$. With this choice, the main peak of the SIGW coincides with the injection of **Case 1**.

We fit each of the two injections using the four models specified above. The results of our analysis are summarized in Fig. 25 (see also Fig. 31 in App. D showing the reconstruction including foregrounds), and Tab. 1, which we now comment in order.

Let us begin with **Case 1**. The models that perform the worst in terms of model selection are the two based on the SIGW hypothesis (marked by the row `SIGWAY`). As shown in the left panels of Fig. 25, both the lognormal and binned $\mathcal{P}_\zeta(k)$ models attempt

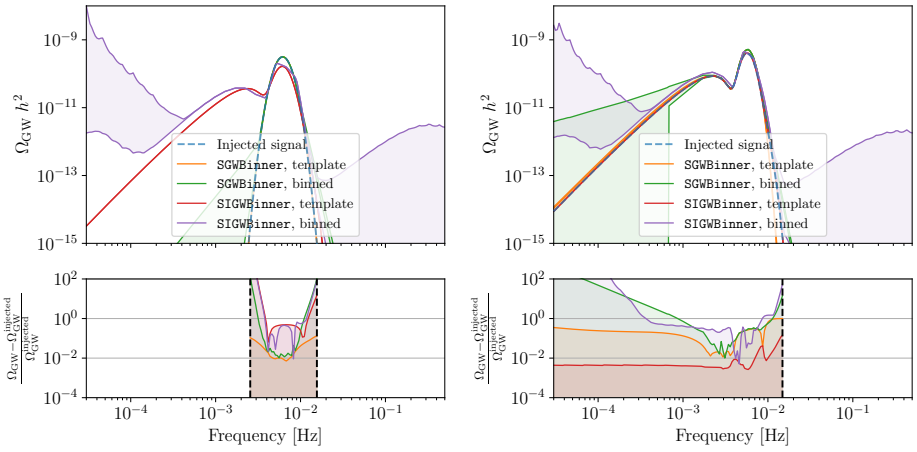


Figure 25. Reconstruction of $\Omega_{\rm GW}$ for case 1 (left) and case 2 (right). We only show the reconstruction of the injected cosmological contribution to the SGWB. It is clearly visible that the two models that assume SIGWs cannot reconstruct the narrow log-normal peak in case 1 and therefore lead to much lower evidence. We denote the wrapper to perform the analysis using `SIGWAY` as `SIGWBinner` in this plot.

to fit the lognormal SGWB using the primary peak of the SIGW. However, the secondary peak at lower frequencies severely undermines the fit to the data, resulting in a very poor likelihood value. According to the Jeffreys' scale (see Sec. 5.2), these models are decisively ruled out when compared to the two alternative hypotheses. As expected the the binned $\mathcal{P}_\zeta(k)$ performs worst, due to its significantly larger number of parameters.

We also fit the same injection using a lognormal template for $\Omega_{\rm GW}$ (which corresponds to the true injection) and a binned $\Omega_{\rm GW}$ reconstruction with `SGWBinner`. As illustrated in Fig. 25, both models reconstruct the signal very well, closely matching the injection. Furthermore, they yield very similar best-fit likelihood values with a preference for the lognormal template (matching the injected signal) due to its smaller number of parameters.

The main takeaway from Case 1 is that if a similar signal was detected, we could conclude with very high statistical significance that the signal does not have a scalar-induced origin.

We now discuss **Case 2**. In this case, all the models considered successfully capture the injected signal, which consistently falls within the reconstructed contours. Unlike the previous case, the Bayes factors are closer together, while still being orders of magnitude apart (we would like to stress that the differences quoted in Tab. 1 are $\log(Z)/10^4$). Even here, the worst performing models are the free reconstructions – whether in $\Omega_{\rm GW}$ or $\mathcal{P}_\zeta(k)$ – as they introduce a large number of parameters despite achieving a good fit. However, their flexibility and agnostic nature make them useful in real data analysis, as they do not require specifying a particular template. Once the main features of the signal are identified, Tab. 1 demonstrates the power of specifying the model.

This injection is detected so well that even using a template for $\Omega_{\rm GW}$ developed in [67, Eq. (2.10)] to accurately parameterize a SGWB induced by scalar perturbations, the model is decisively ruled out when compared to the true lognormal-in-$\mathcal{P}_\zeta(k)$ injection. This outcome arises both because the template in [67] is described by seven free parameters compared to three for the lognormal $\mathcal{P}_\zeta(k)$, and because the fit of the latter is slightly better. Despite the former template accurately approximating the signal, it remains a phenomenological model rather than the true description. It is also interesting to stress that, in case 2, the template-based `SIGWAY` method performs better than the template-based `SGWBinner` in reconstructing $\Omega_{\rm GW}$, as can be seen in the right column of Fig. 25. This improvement arises because the assumption of SIGWs enforces a more restricting shape for $\Omega_{\rm GW}$, characterized by a smaller number of parameters. In contrast, the `SGWBinner` imposes weaker restrictions in the reconstruction. On the other hand, when comparing the agnostic approaches, we see that the former gives slightly better $\Omega_{\rm GW}$ reconstructions around the peak of the signal, with an oscillatory behavior of the residuals due to the finite resolution of the binned approach (40 bins). This result stresses the importance of adopting optimal modeling of the eventual cosmological scenario when reconstructing the signal (see also App. D).

All in all, the results of this section, although based on two illustrative examples, demonstrate that LISA has the potential to confirm the scalar-induced nature of the SGWB with high statistical significance. These results confirm that the true models (lognormal $\Omega_{\rm GW}(f)$ and lognormal $\mathcal{P}_\zeta(k)$) achieve the best Bayes factors when appropriately matched. Alternative and binned models consistently show inferior fits, highlighting the distinctiveness of the injected signals.

However, is important to stress, that not all SGWB that may appear in LISA lead to such a clear difference between signals that can or cannot be generated by SIGW. The characteristic double-peak structure that we observe with the injected $\mathcal{P}_\zeta^{\rm LN}$ signal is only measurable by LISA if *(a)* the peak in $\mathcal{P}_\zeta$ is sufficiently narrow and *(b)* both peaks in $\Omega_{\rm GW}$ happen to fall within the sensitivity of LISA. On the other hand, there are many potential shapes for $\mathcal{P}_\zeta$ where these conditions are not met. In these cases the SIGW signal can easily mimic one expected from other cosmological sources, potentially making it much harder to rule out models. We will leave a more detailed discussion of this for future work.

8 Conclusions

In this work, we investigated the potential of the LISA detector for reconstructing the SGWB sourced by second-order scalar perturbations. Three approaches were explored: A binned spectrum reconstruction, template-based methods, and a direct modeling approach rooted in first-principles scenarios (taken to be the single-field USR inflationary model for presentation purposes).

Our results demonstrate that the direct modeling approach yields the tightest constraints on the primordial curvature power spectrum $\mathcal{P}_\zeta$, particularly capturing both the IR and UV tails of the signal with better precision than alternative methods, due to the stronger prior information inevitably included in the fit. This highlights the power of incor-

porating *ab initio* physics into signal reconstruction pipelines to leverage the constraining power of LISA observations at their best.

The binned spectrum reconstruction approach is complementary and proved effective in providing model-independent upper bounds on $\mathcal{P}_\zeta$ when the cosmological contribution to the SGWB is below the sensitivity. Capturing the overall shape of the SIGW spectrum with this approach proved to be difficult, due to a combination of missing SNR towards the edges of the LISA window and strong degeneracies between the bins when choosing a large number of them. Despite these shortcomings, it is possible – if a cosmological contribution to the SGWB is detected – to tell apart signals that can be SIGW from those that cannot, by Bayesian model selection.

In comparison, the template-based methods provided a more consistent reconstruction across frequencies, though they inherently rely on prior assumptions about the shape of the spectrum. The complementary strengths of these approaches suggest that an optimal reconstruction strategy would involve their combined use, with model-dependent templates guiding reconstructions and binned methods offering flexibility in capturing unanticipated spectral features or in setting bounds that are agnostic on the spectral shape.

We also examined the impact of going beyond the simplest vanilla cosmological scenarios on the SIGW reconstruction investigating the sensitivity of SIGW signals to early-universe physics. In particular, we included in our analysis the study of the effect of the transition from early matter-dominated to radiation-dominated eras, as well as the role of non-Gaussianity in the SIGW spectrum. Future research will include the study of the effect of a time-dependent equation-of-state parameter on the SIGW spectrum, as would be generated in the case of a smooth crossover [398, 399], such as in the QCD phase transition. Overall our analysis demonstrated how SIGW searches in LISA will provide constraints that vastly outperform those deduced from the effective number of relativistic species $\Delta N_{\rm eff}$ and PBH overproduction bounds. In this regard, SIGW searches will also be an invaluable tool for probing the asteroid mass window of PBH dark matter.

Looking forward, several key avenues remain open for future work. On the phenomenological side, it remains an open question how well LISA will be able to constrain primordial non-Gaussianity or non-standard thermal histories while allowing for a fully non-parametric curvature power spectrum. In this work we have only quoted template-based constraints on these effects, e.g. in Figures 21 and 23, which do not consider possible degeneracies with the shape of the spectrum.

Concerning the binned approach to reconstructing $\mathcal{P}_\zeta$, a more mature method that incorporates assumptions about the smoothness of the spectrum of scalar perturbations and addresses the computational difficulties with the binning will be needed in the future. It is likely that – even allowing for non-Gaussianities and non-standard thermal histories – some shapes of the SGWB cannot be scalar-induced and can be confirmed as signatures of directly sourced tensor perturbations, even without identifying a specific early-universe source.

On the theory side, future work should consider expanding the scope of single-field scenarios by exploring more general actions in the Jordan frame, incorporating non-minimal coupling and non-canonical kinetic terms (e.g. Eq. 2.1 of [83]). Also, going beyond single-

field USR models, first-principle multi-field inflationary scenarios merit investigation, as discussed in Sec. 2.2. In some cases, multi-field models can effectively be reduced to single field descriptions, making some of the techniques developed here already applicable, and enabling simpler parameter space scans as done in [406, 407]. A reverse engineering approach could be particularly valuable, where inflationary dynamics are modeled based on a minimal set of parameters, and the corresponding inflationary potential is reconstructed within single- or multi-field frameworks, as demonstrated in Refs. [268, 408]. Expanding the framework for computing the non-Gaussian signatures predicted in most SIGW models beyond the lognormal template in Sec. 4.5 could serve as a diagnostic tool for breaking degeneracies in cases where the SGWB spectrum alone is insufficient. Also, implementing the binned approach in scenarios with NGs could also allow us to reduce the computational costs of these analyses. While in this work we only considered the monopole signal, as non-Gaussianities may impact the large scale SIGW anisotropies, it would be interesting to include information from higher order in the multiple expansion of power in the sky [316, 409–411]. Finally, integrating these advanced modeling and reconstruction techniques into the global fit pipeline of LISA, as well as incorporating measurements from other experiments, will be essential for unlocking the mission's full potential in probing the early universe's cosmological landscape.

Acknowledgments. We thank Nicola Bartolo, Gaetano Luciano, Marco Merchand, Sabino Matarrese, and Toni Riotto for discussions and interactions in an early stage of this project. We acknowledge the LISA Cosmology Working Group members for seminal discussions. We especially thank the authors of refs. [67, 328, 329, 332] for the collaborative developments of the `SGWBinner` code upon which we built to produce many of the forecasts we presented in this work. The research of JF is supported by the grant PID2022-136224NB-C22, funded by MCIN/AEI/10.13039/501100011033/FEDER, UE, and by the grant/ 2021-SGR00872. RR was supported by an appointment to the NASA Postdoctoral Program at the NASA Marshall Space Flight Center, administered by Oak Ridge Associated Universities under contract with NASA. The work of EM is supported by the Italian Ministry of University and Research grant Rita Levi-Montalcini "New directions in axion cosmology". EM acknowledges the support of Istituto Nazionale di Fisica Nucleare (INFN) through the *iniziativa specifica* TAsP. JE and GN acknowledge support by the ROMFORSK grant project no. 302640. TP acknowledges the contribution of the COST Action CA21136 "Addressing observational tensions in cosmology with systematics and fundamental physics (CosmoVerse)" and of the INFN Sezione di Napoli *iniziativa specifica* QGSKY. He acknowledges as well financial support from the Foundation for Education and European Culture in Greece. DR is supported by the UZH Postdoc Grant 2023 Nr. FK-23-130. During the course of this work, S.RP and DW were supported by the European Research Council under the European Union's Horizon 2020 research and innovation programme (grant agreement No 758792, Starting Grant project GEODESI). The work of AG is supported by the UKRI AIMLAC CDT, funded by grant EP/S023992/1. The work of AG, GT, and IZ is partially funded by the STFC grants ST/T000813/1 and ST/X000648/1. The work of HV was supported by the Estonian Research Council grants PSG869 and RVTT7 and the

Center of Excellence program TK202. M.Pe acknowledges support from Istituto Nazionale di Fisica Nucleare (INFN) through the Theoretical Astroparticle Physics (TAsP) project and from the MIUR Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) Bando 2022 - grant 20228RMX4A, funded by the European Union - Next generation EU, Mission 4, Component 1, CUP C53D23000940006. JK acknowledges support from the JSPS Overseas Research Fellowships and the INFN TAsP project. GP acknowledges partial financial support by ASI Grant No. 2016-24-H.0. AR acknowledges support from Istituto Nazionale di Fisica Nucleare (INFN) through the *iniziativa specifica* TEONGRAV and by the project BIGA - "Boosting Inference for Gravitational-wave Astrophysics" funded by the MUR Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) Bando 2022 - grant 20228TLHPE - CUP I53D23000630006.

Authors' contribution. JE: Main developer of `SIGWAY`. Implementation and optimization of techniques discussed in App. A, coding the interface to `SGWBinner`. Co-coding different thermal histories. Running analyses and producing figures for all results except Secs. 6.5 and 7. Writing Secs. 6 and Appendices. Reviewing all draft. AG: Co-coding different thermal histories. Writing Secs. 4.3 and 6.4. Reviewing Sec. 3. GF: Proposing and coordinating the project with RR, including defining goals and managing tasks. Co-coding parts of `SIGWAY`, mainly on SIGW computations and the USR module. Writing and Reviewing all sections of the draft. TP: Co-coding kernel functions and different thermal histories. Writing Sec. 6.4 and parts of Secs. 2.1, 2.2, 2.3, 3.2 and 8. Reviewing Secs. 6.1, 6.2 and 6.3. MPe: Devising methods for computations in Sec. 3.3 and developing the algorithm for the binned spectrum approach, contributing to the code implementation. Writing Secs. 3.1 and 4.4. Reviewing Secs. 1, 2, 5, and 8. GP: Co-coding SIGW computations with NGs in `SIGWAY`. Running and producing Figs. for Sec. 6.5. Writing App. A.4 and Sec. 4. Reviewing Sec. 4. MPi: Co-coding SIGW computation in the `SIGWAY` and interfacing with `SGWBinner`. Writing Sec. 5. AR: Co-coding SIGW computations with NGs in `SIGWAY`. Running and producing Figs. for Sec. 6.5. Writing Sec. 6.5. Reviewing Sec. 1. RR: Proposing and coordinating the project with GF, including defining goals and managing tasks. Co-coding the USR module. Writing and Reviewing all sections of the draft. GT: Developing the algorithm for the binned spectrum approach, contributing to the code implementation. Writing Secs. 3.1 and 4.4. Reviewing the draft. MB: Running the analysis and writing of Sec. 7. JF: Running the analysis and writing of Sec. 7. Reviewing Secs. 7, 3.2.2, 2.2 and Sec. 2. JK: Coding `SGWBinner`. Writing Sec. 5. Reviewing Sec. 3. EM: Contributing to an early version of the USR code. Writing Secs. 2.1, 2.2, 2.4, 3.2, 3.3, 6.2, and 6.3. Reviewing all draft. GN: Writing Sec. 1. Reviewing all draft. DR: Writing Secs. 3 (intro), 4.1, 4.2, and 6.2.1. Reviewing Secs. 4 and 7. SRP: Devising and interpreting Sec. 4.5 and 6.5. Writing Sec. 4.5 and related elements in Secs. 4 and 6.5. Reviewing all draft. HV: Devising Secs. 2 and 3. Writing Secs. 2, 3, and 6.2. Reviewing Secs. 1, 6.4, 6.5, and 8. DW: Devising and interpreting Sec. 4.5. Reviewing Secs. 4.5, 1, and 8. IZ: Writing parts of Sec. 4 and reviewing Secs. 2, 3, and 4.

A SIGWAY code: technicalities

In this appendix, we describe some technical aspects of the `SIGWAY` code developed for the analysis performed in this work.

A.1 Perturbations in USR scenarios: code structure

Given a potential $V(\phi)$ inducing a USR phase, the curvature power spectrum is computed in three steps that are described below. $\mathcal{P}_\zeta(k)$ is then interpolated and $\Omega_{\rm GW}h^2$ is computed as described in A.2.

Using the notation described in Sec. 3.3, the inputs in the code are:

- the inflaton potential $V(\phi)$;
- the number of e-folds from when CMB modes exit the Hubble horizon to the end of inflation $N_{\rm CMB\to end}$.
- The initial conditions ϕ_0 and $\pi_0 = \phi_0'$.

The code then automatically defines the dimensionless variables in Eq. (3.19). For definiteness, in this paper, we have fixed $N_{\rm CMB\to end} = 58$. Fixing the number of e-folds from the CMB scale to the end of inflation effectively allows us to set the correspondence between N and wavenumbers k. Fixing $N_{\rm CMB}$ also implicitly fixes the thermal history of reheating and subsequent phases. We do not model these eras for simplicity, but they would be fixed in a complete USR+reheating model. Also, we checked that the initial SR attractor would quickly pull the field space trajectory to the background evolution, and thus one could also assume negligible initial velocity for simplicity.

The computation then proceeds as follows:

1. Solve for the the background evolution using Eqs. (3.21). We evolve the dynamics until the first SR condition is violated ($\epsilon_H = 1$), collecting $x(N), y(N)$ and $h(N)$. We denote the number of e-folds at the end of inflation $N_{\rm end}$.

2. Compute the relation between the wave number k and the number of e-folds N at Hubble crossing using

$$k = k_{\rm CMB}\frac{h(N)}{h_{\rm CMB}}\exp(N - N_{\rm CMB}), \tag{A.1}$$

   where $k_{\rm CMB} = 0.05/\,\text{Mpc}$, $N_{\rm CMB} = N_{\rm end} - N_{\rm CMB\to end}$ and $h_{\rm CMB} = h(N_{\rm CMB})$. With this, we can compute $\mathcal{P}_\zeta(k)$ in the SR approximation according to Eq. (3.25).

3. We are only interested in computing the spectrum of curvature perturbations beyond the SR approximation for modes of relevance for LISA. Once we have computed $k(N)$, a set of modes covering the LISA frequency band $\sim (10^{-5} - 10^{-1})\,\text{Hz}$ is selected and those modes are evolved according to Eq. (3.31) from $N_{\rm in}$ to $N_{\rm out}$, where $N_{\rm in}$ and $N_{\rm out}$ can be set by the user. We found $N_{\rm in} = N - 3$ and $N_{\rm out} = N + 7$, where N is the horizon crossing time of the mode, to be long enough for the modes to freeze out

as there is no super horizon evolution in these models once USR has ended. Notice that the USR can only last about $\mathcal{O}(3)$ e-folds without making perturbations grow beyond the validity of perturbation theory. Additionally, for simple shapes of the potential, we find that evolving ~ 100 modes in k and interpolating between them yields sufficient precision.

4. Lastly, $\mathcal{P}_\zeta(k)$ is computed without resorting to the SR approximation using Eq. (3.32). This quantity is then passed to the algorithm computing $\Omega_{\rm GW}h^2$ which can then be passed to the LISA likelihood.

Fig. 26 shows the evolution of a number of modes in $\mathcal{P}_\zeta$ as a function of N across the USR phase. As this algorithm is called many times when sampling the LISA likelihood, we are using the package `diffrax` [412] to solve the inflationary perturbation equation of motion.

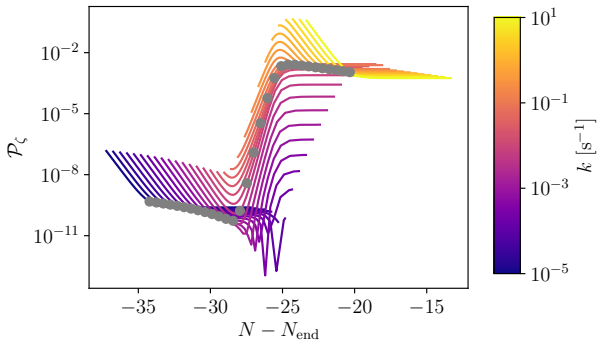


Figure 26. Power spectrum as a function of N for different modes k indicated by the color palette, close to the onset of the USR phase $N - N_{\rm end} \simeq -28$. This plot shows the evolution and freezing out of modes for different values of k for the potential from Eq. (2.4). The evolution of each mode is traced for around $\Delta N \simeq 10$ e-folds from within the Hubble sphere to sufficiently after Hubble crossing and freezing. The time of Hubble crossing is marked with gray dots.

A.2 Computation of SIGWs from the spectrum of curvature perturbations

For a given shape of $\mathcal{P}_\zeta(k)$, computing $\Omega_{\rm GW}h^2$ is relatively straightforward by evaluating the double integral in Eq. (4.20). The main concern here is making this computation as fast as possible, as this dominates the time it takes for each likelihood evaluation. To take full advantage of vectorization, we compute s, t, k on a grid, where s is linearly spaced, and t, k are logarithmically spaced. The integration in t is convergent for any realistic shape of $\mathcal{P}_\zeta$ as both the value of $\overline{I^2(k,s,t)}$ tends towards 0 for large t and there needs to be some cutoff controlling the amplitude of $\mathcal{P}_\zeta$ at large momenta not to violate BBN bounds. If $\mathcal{P}_\zeta(k)$ features a scale after which it drops rapidly, it can be advantageous to define a custom $t(k)$ and compute the grid as $s, t(k), k$.

After evaluating the integrand in Eq. (4.20) on the grid, the integral is computed with Simpson-integration. Depending on the shape of $\mathcal{P}_\zeta$, we found a number of points in the grid around $N_s \sim 10 - 100$, $N_t \sim 300 - 1000$ and $N_k \sim 100$ would give sufficient precision.

To take full advantage of threading, we use `JAX` and just-in-time compilation [375] for computing the integrand and performing the integration itself. Altogether, this results in a significant speedup over other publicly available codes (e.g. [413]). We record wall-clock times of $\sim 10^{-2}\,\mathrm{s}$ for calculating 100 values of $\Omega_{\rm GW}$ for $\mathcal{P}_\zeta$ containing only `JAX`-native functions. In the case discussed in Sec. 6.2.2 this time rises to about 1 s. Crucially, this speedup allows us to sample the posterior distribution efficiently and obtain good MCMC convergence with a laptop on timescales of $\mathcal{O}$(hours).

A.3 Computation of SIGWs using binned coefficients

In the case where we bin $\mathcal{P}_\zeta(p)$ in momentum space, $\Omega_{\rm GW}(k)$ can be computed in a straightforward manner through Eq. (4.27). The most computationally expensive part of this is the computation of the coefficients $\Omega_{\rm GW}^{(i,j)}(k)$. Luckily we can precompute these coefficients on a grid and save them as a $N_k \times N_p \times N_p$ tensor $K_{ij}^k \equiv \Omega_{\rm GW}^{(i,j)}(k)$. At runtime, we then compute a $N_p \times N_p$ tensor of $B = A \otimes A$, with which we can conveniently re-write Eq. (4.27) as

$$\Omega_{\rm GW}^k = \Omega_{ij}^k B^{ij}\,, \tag{A.2}$$

where we used Einstein sum convention. The reason for doing this arguably very simple conversion is that the matrix equation is fully vectorizable with `JAX`.

Using this trick we are able to compute $\Omega_{\rm GW}$ from $\mathcal{P}_\zeta$ in $\lesssim 10^{-3}\,\mathrm{s}$ for $N_k = N_p = 50$, thus making inference possible despite a large number of parameters to sample.

A.4 Computation of SIGWs including primordial NGs

In this appendix, we provide additional technical details regarding the MCMC analyses that resulted in Figures 24 and 23, as well as the parameter scan leading to Fig. 22. As reported in Sec. 4.5, the trispectrum arising from the local expansion Eq. (4.37) leads to additional contributions to the SIGW spectrum. In particular, from the connected part, one obtains the following two terms

$$\begin{aligned}\Omega_{\rm GW}(k,\eta)|_{\rm t} = \frac{1}{12\pi}\left(\frac{k}{aH}\right)^2\left(\frac{3}{5}f_{\rm NL}\right)^2 \int_0^\infty dt_1 \int_{-1}^1 ds_1 \int_0^\infty dt_2 \int_{-1}^1 ds_2 \\ \times \int_0^{2\pi} d\varphi_{12}\cos 2\varphi_{12}\frac{u_1 v_1}{(u_2 v_2)^2}\frac{1}{w_{a,12}^3}\overline{\tilde{J}(u_1,v_1,x)\tilde{J}(u_2,v_2,x)} \\ \times \mathcal{P}_{\zeta_g}(v_2 k)\mathcal{P}_{\zeta_g}(u_2 k)\mathcal{P}_{\zeta_g}(w_{a,12}k)\,,\end{aligned} \tag{A.3}$$

and

$$\Omega_{\rm GW}(k,\eta)|_{\rm u} = \frac{1}{12\pi}\left(\frac{k}{aH}\right)^2\left(\frac{3}{5}f_{\rm NL}\right)^2\int_0^\infty dt_1\int_{-1}^1 ds_1\int_0^\infty dt_2\int_{-1}^1 ds_2 \times\int_0^{2\pi} d\varphi_{12}\cos 2\varphi_{12}\frac{u_1u_2}{(v_1v_2)^2}\frac{1}{w_{b,12}^3}\overline{\tilde{J}(u_1,v_1,x)\tilde{J}(u_2,v_2,x)} \times\mathcal{P}_{\zeta_g}(v_1k)\mathcal{P}_{\zeta_g}(v_2k)\mathcal{P}_{\zeta_g}(w_{b,12}k)\,, \tag{A.4}$$

were the integration variables t_i and s_i are defined as in Eq. (4.19). To keep the equation concise, some terms have been left expressed as functions of u_i and v_i, but they have to be intended as depending on the integration variables t_i and s_i. Moreover, we introduced $\tilde{J}(u_i,v_i,x) = v_i^2k^2\sin^2\theta I(u_i,v_i,x)$, with $I(u,v,x)$ the integration kernel defined in the main text and $w_{a,12}$ and $w_{b,12}$, defined as

$$w_{a,12} = [v_1^2+v_2^2-2v_1v_2(\cos\theta_1\cos\theta_2+\sin\theta_1\sin\theta_2\cos\varphi_{12})]^{1/2}\,, \tag{A.5}$$

and

$$w_{b,12} = \big[1+v_1^2+v_2^2+2v_1v_2(\cos\theta_1\cos\theta_2+\sin\theta_1\sin\theta_2\cos\varphi_{12}) - 2v_1\cos\theta_1-2v_2\cos\theta_2\big]^{1/2}\,. \tag{A.6}$$

The sine and cosine functions are related to the integration variables by

$$\cos\theta_i = \frac{1-s_i(1+t_i)}{t_i-s_i+1}\,,\qquad \sin^2\theta_i = \frac{(1-s_i^2)t_i(2+t_i)}{(t_i-s_i+1)^2}\,. \tag{A.7}$$

As shown in Eqs. (A.3) and (A.4), the evaluation of the NG corrections requires a 5 dimensional integration for each of the frequencies at which the final GW spectrum is evaluated. However, $f_{\rm NL}$ and A_s are multiplicative parameters and the effect of k_* just results in a shift of the spectrum along the k-axis. Hence, once the spectrum is evaluated for a fixed width Δ, it can be used for different values of the parameters reported above, without requiring further evaluation. When Δ is varied, instead, a new evaluation of the spectrum is required each time. Hence, just a single evaluation of the spectrum would require relatively little time, but the evaluation of the whole spectrum for each point of the MCMC would notably slow down the run, making it difficult to get the final posterior in a reasonable time, also considering the presence of other parameters in the MCMC evaluation.

For this reason, to speed up the evaluation we proceed as follows: we numerically pre-compute a grid of NG contributions to the GW spectra as a function of frequency for different widths, in order to explore the range $\log_{10}\Delta\in[-1,1]$. This grid is then used to obtain the NG corrections to the spectrum corresponding to any value of Δ in the range considered, by interpolating them from the pre-computed ones. In detail, to get the spectrum corresponding to those values of $\bar{\Delta}$ not present in the grid, we first search for $\Delta_{\rm max}$ and $\Delta_{\rm min}$, respectively immediately above and below $\bar{\Delta}$. Then we compute the interpolated spectrum by Taylor expanding around these values, obtaining

$$\Omega_{\rm GW}(\bar{\Delta}) = \Omega_{\rm GW}(\Delta_{\rm max})w_{\rm min}+\Omega_{\rm GW}(\Delta_{\rm min})(1-w_{\rm min})\,, \tag{A.8}$$

with

$$w_{\min} = \frac{(\bar{\Delta} - \Delta_{\min})}{(\Delta_{\max} - \Delta_{\min})} \,. \tag{A.9}$$

For the evaluation of the scans that require a Fisher forecast and hence the derivatives with respect to the parameters, we proceed in a similar way. We pre-compute a grid of derivatives[20] in the range $\log_{10}\Delta \in [-1, 1]$ and then we interpolate as explained above.

A.5 Inference

Once $\Omega_{\rm GW}$ has been computed by the `SIGWAY`, the resulting spectrum is interpolated in log-space and passed to the `SGWBinner` which computes the posterior distribution according to Eq. (5.21). `Cobaya` [374, 414] is used as an inference-framework. We use different samplers for Monte Carlo sampling depending on the dimensionality, requirements, and structure of the posterior surface:

- The inferences in Figs. 3, 4, 8, 9, 11, 12, and 21 have been run using the nested sampler `nessai` [415–417].
- Figs. 5, 6, 19, 23, and 24 have been obtained using `Cobaya`'s `CosmoMC` [418, 419] MCMC, where we started the chains at the injected values and injected FIM estimates of the covariance matrix to speed up convergence.
- The evidences in Sec. 7 have been computed with the nested sampler `PolyChord` [403, 404]
- The inference for Figs. 14 and 15 was performed using the active learning algorithm `GPry` [420, 421] due to the prohibitively slow speed of computing $\mathcal{P}_\zeta$ stemming from the Bessel functions in its equation.

All corner plots have been created with `GetDist` [422]. In the corner plots, we omitted showing the marginalised constraints on $A_{\rm noise}$ and $P_{\rm noise}$, as in all cases they were well constrained and showed weak degeneracies with the signal parameters. To give an idea of how tightly these parameters tend to be constrained, Fig. 27 shows a corner plot including the constraints on $A_{\rm noise}$ and $P_{\rm noise}$ for the injected lognormal $\mathcal{P}_\zeta$ (see Eq. (3.5)).

B Challenges with binned analyses and a large number of bins

The binned approach to performing the double integration going from $\mathcal{P}_\zeta$ to $\Omega_{\rm GW}$ introduced in Sec. 4.4 – while in principle extremely powerful at reconstructing any SIGW spectrum without model-dependence – unfortunately suffers from some crucial shortcomings as will be explained in this section.

For the sake of illustration, we will only consider the case where all modes reenter during radiation domination (see Sec. 4.2) where the kernel is k-independent. The situation

[20] Note that when taking the derivative with respect to A_s and $f_{\rm NL}$ the integrals remain unchanged, hence we consider the same pre-computed NG contributions used in the MCMC runs. When taking the derivative with respect to k_* or Δ, instead, since the integrand is varied, we pre-compute a grid for each of these two derivatives.

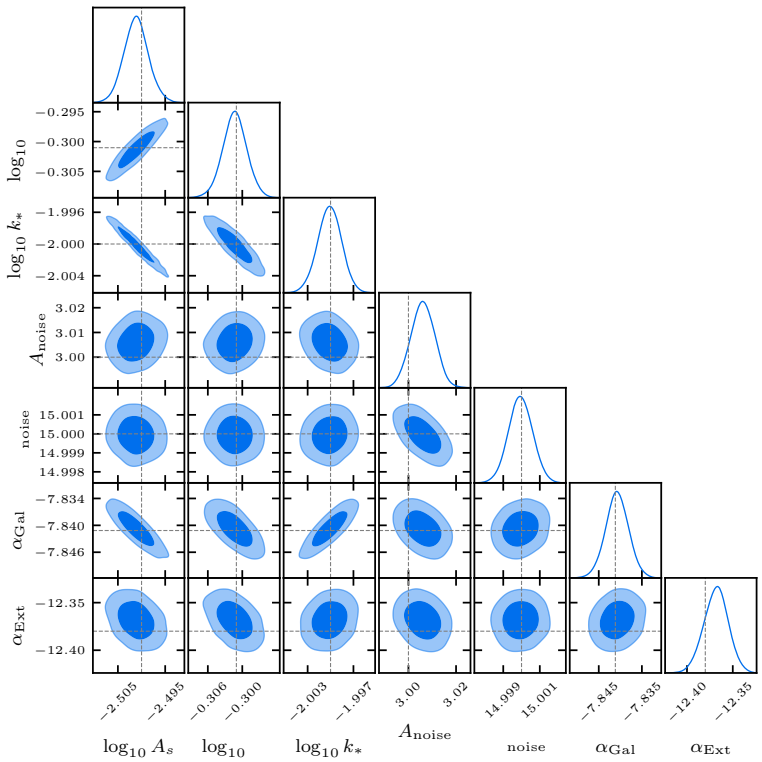


Figure 27. Same as Fig. 5 but showing the marginalised contours for all sampled parameters including the noise parameters $A_{\rm noise}, P_{\rm noise}$. There is a relatively mild degeneracy between $A_{\rm noise}$ and $\alpha_{\rm Gal}$ but no correlations between the signal parameters A_s, Δ, k_* and the noise parameters. We found similar correlations (or the lack thereof) for all other injections.

changes a bit if the kernel has a k-dependence such as is the case during an early matter domination era (see Sec. 4.3), however, our main arguments remain unchanged.

It is clear from the structure of the integral in Eq. (4.20), that a single wavenumber k in $\mathcal{P}_\zeta$ affects multiple frequencies in $\Omega_{\rm GW}$. An easy way to understand this is to consider a $\mathcal{P}_\zeta$ that is sufficiently close to a monochromatic source $\mathcal{P}_\zeta(k) = A_s\delta(k - k_*)$. In our binned approach this translates to one single bin A_* being non-zero. Fig. 28 shows three such spectra with 100 bins, where each one contains a single non-zero bin $A_\star$ (bin nr. 85, 86, 87 in this case). It is evident from this figure that if the peak towards k_* is not resolved, and only the causality tail in the IR regime enters the LISA sensitivity, these spectra become entirely degenerate. This means that there is not necessarily a unique mapping $\Omega_{\rm GW} \mapsto \mathcal{P}_\zeta$. In other words, the power from the bins is "leaking" into adjacent bins.

In reality, this means that, for many bins and towards low SNR, the binned recon-

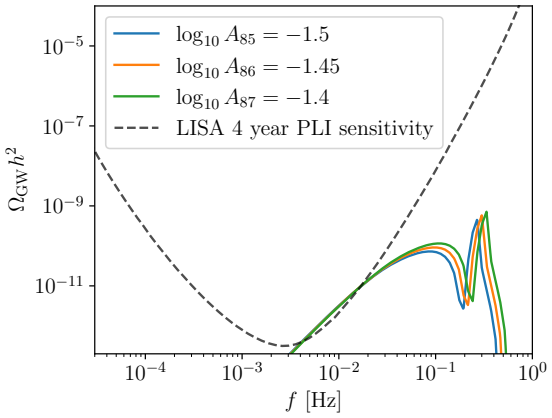


Figure 28. Three different spectra in $\Omega_{\rm GW}$ generated with the binned approach with 100 bins where for each spectrum only one of the bins is non-zero. The black dashed line shows the approximate power law integrated sensitivity of LISA assuming a 4-year mission. It is clear from this picture that the three adjacent bins shown are entirely degenerate when trying to resolve them with LISA as the peaks are well outside the sensitivity and the causality tails generated are exactly the same.

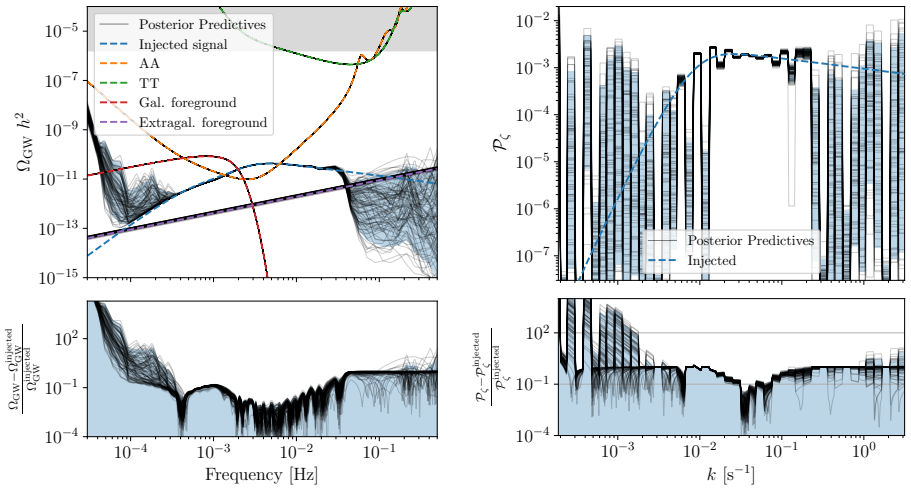


Figure 29. Same as Fig. 3, but for an injected BPL spectrum following the benchmark USR model and using 50 bins.

structed $\mathcal{P}_\zeta$ is highly degenerate, and valid reconstructions include "oscillations" between the bins as lower power in one bin can be compensated by higher power in an adjacent one.

The aforementioned degeneracies are not unexpected and are really just a feature of the physical properties of the process of scalar-induced gravitational waves. However, they

do induce some practical complications. In an ideal template-agnostic pipeline, we would want to perform inference on the $N_{\rm bins}$ bins in $\mathcal{P}_\zeta$ to reconstruct the signal with the highest evidence parameterization. Due to the large (non-linear) degeneracies between the bins, the likelihood is far from Gaussian and the FIM approximation is invalid, making MC-sampling necessary. Sampling over this space is very computationally challenging due to *(a)* the very narrow degeneracies *(b)* the resulting large number of posterior modes and *(c)* the high dimensionality of the parameter space. In practice, this leads to overconfidence in the reconstruction, as some posterior modes are inevitably missed or underexplored by the MC sampler. Fig. 29 shows the binned reconstruction of the USR model injection from Sec. 2.4 with 50 bins. The oscillation effect is clearly visible in the low SNR region, where $\mathcal{P}_\zeta$ oscillates between high power and low power, thus overconstraining certain bins. These degeneracies are partially broken by fewer bins, as visible in Fig. 3.

This leaves us in a dilemma: we would like to bin $\mathcal{P}_\zeta$ as finely as possible to increase the frequency resolution of the template, but as one increases $N_{\rm bins}$ the posterior becomes much more difficult to sample. In our tests, we found $N_{\rm bins} = 15$ to be reliable in terms of convergence for the BPL signal (Fig. 3) and $N_{\rm bins} = 40$ for the injected lognormal signals in $\Omega_{\rm GW}$ and $\mathcal{P}_\zeta$ (Fig. 25) that occupy less of the frequency range. However, it is clear that this low number of bins cannot reconstruct the shape of $\mathcal{P}_\zeta$ with high fidelity.

Luckily, this problem does not appear when no signal is present, as the posterior distribution in A_i becomes a simple upper bound, an unimodal structure that is easy to map by a nested sampler, even in high dimensions. We can therefore conclude that the upper bounds obtained by this method are reliable even with many bins.

Future work on this approach could include studying improved bases for the reconstructed bins (a basis of Gaussians or other wide kernels in P_ζ may be less multi-modal), or improving sampling by manually adding jump proposals to the degenerate modes of the posterior in a given basis (e.g. as is done for LISA black hole binary sampling in `BBHx` [56]).

C Testing the resolvability of Non-Gaussian corrections: Additional plots

Fig. 30 shows a corner plot that was obtained by injecting a purely Gaussian SIGW signal with a lognormal shape in $\mathcal{P}_\zeta$ (see Eq. (3.5)) that is equivalent to the cases discussed in Figs. 23 and 24 with $f_{\rm NL} = \tau_{\rm NL} = 0$. The remaining injected parameters are the same as in Section 6.5: $\log_{10} A_s = -2$, $\log_{10} \Delta = -0.75$, $\log_{10}(k_*/\mathrm{s}^{-1}) = -2.3$. By comparing these results to the one obtained for $f_{\rm NL} = 1$ shown in Fig. 23 and $f_{\rm NL} = 10$ shown in Fig. 24, we see that the NG contribution neither significantly improves, nor worsens the constraints on the signal and foreground parameters.

D Testing the scalar-induced hypothesis: Additional plots

Figure 31 provides further insight into the quality of reconstruction when comparing the `SGWBinner` and the `SIGWAY` methods. The noise is accurately reconstructed in all cases, with only a slight underestimation in case 1 (left panel) using the `SIGWAY` template recon-

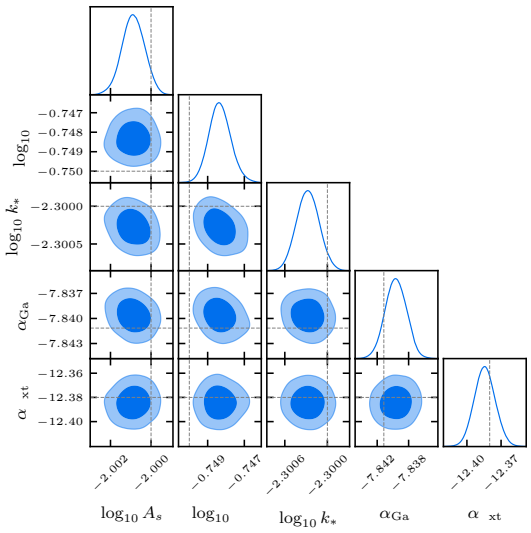


Figure 30. Same as Fig. 5 for the same injected parameters as in Sec. 6.5.

struction. In contrast, the extragalactic background is reconstructed less accurately with the `SGWBinner` compared to the other methods.

A particularly notable observation is that in case 1 (left column), the `SIGWAY` template method significantly underestimates both the extragalactic and galactic foregrounds. This underestimation can be attributed to the model compensating for excess power in the causality tail by reducing the power allocated to the foregrounds, due to the limited flexibility in the shape of $\Omega_{\rm GW}$ provided by the template. Interestingly, in case 2 (right column), the `SIGWAY` method – using the template or not – performs better than the `SGWBinner` in reconstructing the foregrounds. This improvement arises because the assumption of SIGWs enforces a specific shape for $\Omega_{\rm GW}$, which cannot be easily mimicked by the foregrounds. In contrast, the `SGWBinner` does not impose such restrictions during reconstruction. This result stresses the importance of adopting optimal modeling of the eventual cosmological signal even when reconstructing the astrophysical properties of the foreground sources.

References

[1] LISA collaboration, *Laser Interferometer Space Antenna*, 1702.00786.

[2] LISA Cosmology Working Group collaboration, *Cosmology with the Laser Interferometer Space Antenna*, *Living Rev. Rel.* **26** (2023) 5 [2204.05434].

[3] LISA collaboration, *New horizons for fundamental physics with LISA*, *Living Rev. Rel.* **25** (2022) 4 [2205.01597].

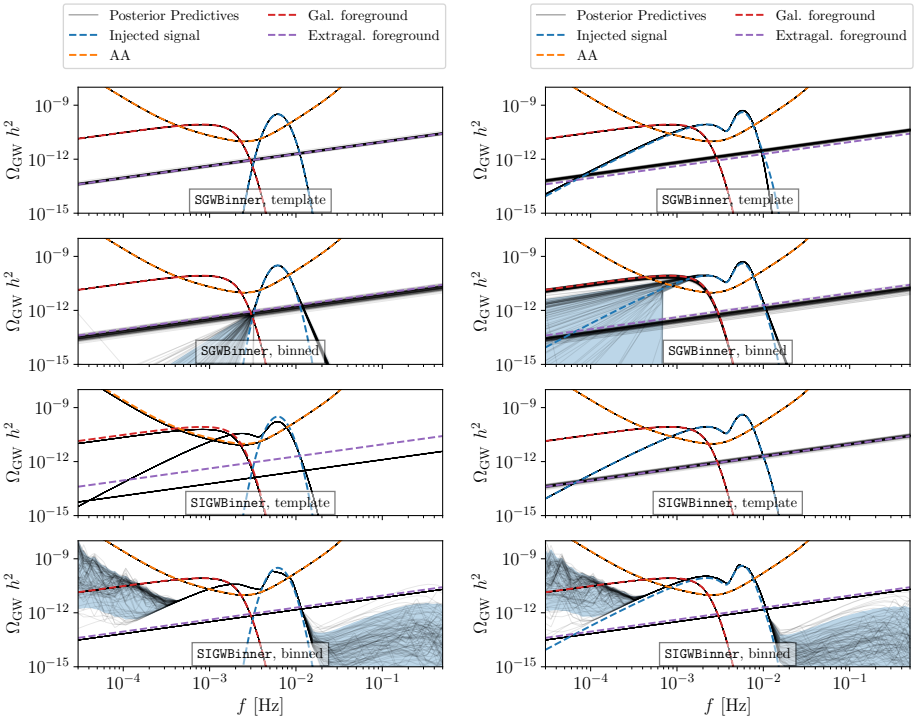


Figure 31. Reconstruction of $\Omega_{\rm GW}$ for case 1 (left) and case 2 (right) including the noise and foreground. See Sec. 7 for more details. The TT-channel component of the noise is above 10^{-7} and we therefore omit it. We denote the wrapper to perform the analysis using `SIGWAY` as `SIGWBinner` in this plot.

[4] LISA collaboration, *Astrophysics with the Laser Interferometer Space Antenna*, *Living Rev. Rel.* **26** (2023) 2 [2203.06016].

[5] N. Bartolo et al., *Science with the space-based interferometer LISA. IV: Probing inflation with gravitational waves*, *JCAP* **12** (2016) 026 [1610.06481].

[6] LISA Cosmology Working Group collaboration, *Gravitational waves from inflation in LISA: reconstruction pipeline and physics interpretation*, *JCAP* **11** (2024) 032 [2407.04356].

[7] M. Maggiore, *Gravitational Waves. Vol. 1: Theory and Experiments*. Oxford University Press, 2007, 10.1093/acprof:oso/9780198570745.001.0001.

[8] M. C. Guzzetti, N. Bartolo, M. Liguori and S. Matarrese, *Gravitational waves from inflation*, *Riv. Nuovo Cim.* **39** (2016) 399 [1605.01615].

[9] R.-G. Cai, Z. Cao, Z.-K. Guo, S.-J. Wang and T. Yang, *The Gravitational-Wave Physics*, *Natl. Sci. Rev.* **4** (2017) 687 [1703.00187].

[10] C. Caprini and D. G. Figueroa, *Cosmological Backgrounds of Gravitational Waves*, *Class.*

Quant. Grav. **35** (2018) 163001 [1801.04268].

[11] K. Tomita, *Evolution of Irregularities in a Chaotic Early Universe*, *Prog. Theor. Phys.* **54** (1975) 730.

[12] S. Matarrese, O. Pantano and D. Saez, *A General relativistic approach to the nonlinear evolution of collisionless matter*, *Phys. Rev. D* **47** (1993) 1311.

[13] S. Matarrese, O. Pantano and D. Saez, *General relativistic dynamics of irrotational dust: Cosmological implications*, *Phys. Rev. Lett.* **72** (1994) 320 [astro-ph/9310036].

[14] S. Matarrese, S. Mollerach and M. Bruni, *Second order perturbations of the Einstein-de Sitter universe*, *Phys. Rev. D* **58** (1998) 043504 [astro-ph/9707278].

[15] V. Acquaviva, N. Bartolo, S. Matarrese and A. Riotto, *Second order cosmological perturbations from inflation*, *Nucl. Phys. B* **667** (2003) 119 [astro-ph/0209156].

[16] S. Mollerach, D. Harari and S. Matarrese, *CMB polarization from secondary vector and tensor modes*, *Phys. Rev. D* **69** (2004) 063002 [astro-ph/0310711].

[17] C. Carbone and S. Matarrese, *A Unified treatment of cosmological perturbations from super-horizon to small scales*, *Phys. Rev. D* **71** (2005) 043508 [astro-ph/0407611].

[18] K. N. Ananda, C. Clarkson and D. Wands, *The Cosmological gravitational wave background from primordial density perturbations*, *Phys. Rev. D* **75** (2007) 123518 [gr-qc/0612013].

[19] D. Baumann, P. J. Steinhardt, K. Takahashi and K. Ichiki, *Gravitational Wave Spectrum Induced by Primordial Scalar Perturbations*, *Phys. Rev. D* **76** (2007) 084019 [hep-th/0703290].

[20] G. Domènech, *Scalar Induced Gravitational Waves Review*, *Universe* **7** (2021) 398 [2109.01398].

[21] Y. B. Zel'dovich and I. D. Novikov, *The Hypothesis of Cores Retarded during Expansion and the Hot Cosmological Model*, *Sov. Astron.* **10** (1967) 602.

[22] S. Hawking, *Gravitationally collapsed objects of very low mass*, *Mon. Not. Roy. Astron. Soc.* **152** (1971) 75.

[23] B. J. Carr and S. W. Hawking, *Black holes in the early Universe*, *Mon. Not. Roy. Astron. Soc.* **168** (1974) 399.

[24] B. J. Carr, *The Primordial black hole mass spectrum*, *Astrophys. J.* **201** (1975) 1.

[25] G. F. Chapline, *Cosmological effects of primordial black holes*, *Nature* **253** (1975) 251.

[26] N. Bartolo, V. De Luca, G. Franciolini, A. Lewis, M. Peloso and A. Riotto, *Primordial Black Hole Dark Matter: LISA Serendipity*, *Phys. Rev. Lett.* **122** (2019) 211301 [1810.12218].

[27] N. Bartolo, V. De Luca, G. Franciolini, M. Peloso, D. Racco and A. Riotto, *Testing primordial black holes as dark matter with LISA*, *Phys. Rev. D* **99** (2019) 103521 [1810.12224].

[28] LISA Cosmology Working Group collaboration, *Primordial black holes and their gravitational-wave signatures*, 2310.19857.

[29] A. Gangui, F. Lucchin, S. Matarrese and S. Mollerach, *The Three point correlation function of the cosmic microwave background in inflationary models*, *Astrophys. J.* **430** (1994) 447 [astro-ph/9312033].

[30] S. Matarrese, L. Verde and R. Jimenez, *The Abundance of high-redshift objects as a probe of non-Gaussian initial conditions*, *Astrophys. J.* **541** (2000) 10 [astro-ph/0001366].

[31] N. Bartolo, S. Matarrese and A. Riotto, *Nongaussianity from inflation*, *Phys. Rev. D* **65** (2002) 103505 [hep-ph/0112261].

[32] J. M. Maldacena, *Non-Gaussian features of primordial fluctuations in single field inflationary models*, *JHEP* **05** (2003) 013 [astro-ph/0210603].

[33] N. Bartolo, E. Komatsu, S. Matarrese and A. Riotto, *Non-Gaussianity from inflation: Theory and observations*, *Phys. Rept.* **402** (2004) 103 [astro-ph/0406398].

[34] X. Chen, *Primordial Non-Gaussianities from Inflation Models*, *Adv. Astron.* **2010** (2010) 638979 [1002.1416].

[35] C. T. Byrnes and K.-Y. Choi, *Review of local non-Gaussianity from multi-field inflation*, *Adv. Astron.* **2010** (2010) 724525 [1002.3110].

[36] D. Wands, *Local non-Gaussianity from inflation*, *Class. Quant. Grav.* **27** (2010) 124002 [1004.0818].

[37] S. Renaux-Petel, *Primordial non-Gaussianities after Planck 2015: an introductory review*, *Comptes Rendus Physique* **16** (2015) 969 [1508.06740].

[38] A. Achúcarro et al., *Inflation: Theory and Observations*, 2203.08128.

[39] T. Nakama, J. Silk and M. Kamionkowski, *Stochastic gravitational waves associated with the formation of primordial black holes*, *Phys. Rev. D* **95** (2017) 043511 [1612.06264].

[40] J. Garcia-Bellido, M. Peloso and C. Unal, *Gravitational Wave signatures of inflationary models from Primordial Black Hole Dark Matter*, *JCAP* **09** (2017) 013 [1707.02441].

[41] C. Unal, *Imprints of Primordial Non-Gaussianity on Gravitational Wave Spectrum*, *Phys. Rev. D* **99** (2019) 041301 [1811.09151].

[42] R.-g. Cai, S. Pi and M. Sasaki, *Gravitational Waves Induced by non-Gaussian Scalar Perturbations*, *Phys. Rev. Lett.* **122** (2019) 201101 [1810.11000].

[43] R.-G. Cai, S. Pi, S.-J. Wang and X.-Y. Yang, *Resonant multiple peaks in the induced gravitational waves*, *JCAP* **05** (2019) 013 [1901.10152].

[44] H. V. Ragavendra, P. Saha, L. Sriramkumar and J. Silk, *Primordial black holes and secondary gravitational waves from ultraslow roll and punctuated inflation*, *Phys. Rev. D* **103** (2021) 083510 [2008.12202].

[45] C. Yuan and Q.-G. Huang, *Gravitational waves induced by the local-type non-Gaussian curvature perturbations*, *Phys. Lett. B* **821** (2021) 136606 [2007.10686].

[46] P. Adshead, K. D. Lozanov and Z. J. Weiner, *Non-Gaussianity and the induced gravitational wave background*, *JCAP* **10** (2021) 080 [2105.01659].

[47] M. W. Davies, P. Carrilho and D. J. Mulryne, *Non-Gaussianity in inflationary scenarios for primordial black holes*, *JCAP* **06** (2022) 019 [2110.08189].

[48] K. T. Abe, R. Inui, Y. Tada and S. Yokoyama, *Primordial black holes and gravitational waves induced by exponential-tailed perturbations*, *JCAP* **05** (2023) 044 [2209.13891].

[49] S. Garcia-Saenz, L. Pinol, S. Renaux-Petel and D. Werth, *No-go theorem for scalar-trispectrum-induced gravitational waves*, *JCAP* **03** (2023) 057 [2207.14267].

[50] S. Garcia-Saenz, Y. Lu and Z. Shuai, *Scalar-induced gravitational waves from ghost inflation and parity violation*, *Phys. Rev. D* **108** (2023) 123507 [2306.09052].

[51] J.-P. Li, S. Wang, Z.-C. Zhao and K. Kohri, *Complete analysis of the background and anisotropies of scalar-induced gravitational waves: primordial non-Gaussianity f_{NL} and g_{NL} considered*, *JCAP* **06** (2024) 039 [2309.07792].

[52] C. Yuan, D.-S. Meng and Q.-G. Huang, *Full analysis of the scalar-induced gravitational waves for the curvature perturbation with local-type non-Gaussianities*, *JCAP* **12** (2023) 036 [2308.07155].

[53] G. Perna, C. Testini, A. Ricciardone and S. Matarrese, *Fully non-Gaussian Scalar-Induced Gravitational Waves*, *JCAP* **05** (2024) 086 [2403.06962].

[54] M. Colpi et al., *LISA Definition Study Report*, 2402.07571.

[55] T. B. Littenberg and N. J. Cornish, *Prototype global analysis of LISA data with multiple source types*, *Phys. Rev. D* **107** (2023) 063004 [2301.03673].

[56] M. L. Katz, N. Karnesis, N. Korsakova, J. R. Gair and N. Stergioulas, *An efficient GPU-accelerated multi-source global fit pipeline for LISA data analysis*, 2405.04690.

[57] S. H. Strub, L. Ferraioli, C. Schmelzbach, S. C. Stähler and D. Giardini, *Global analysis of LISA data with Galactic binaries and massive black hole binaries*, *Phys. Rev. D* **110** (2024) 024005 [2403.15318].

[58] M. Le Jeune and S. Babak, *Lisa data challenge sangria (ldc2a)*, Oct., 2022. 10.5281/zenodo.7132178.

[59] R. Rosati and T. B. Littenberg, *Prototype Stochastic Gravitational Wave Background Recovery in the LISA Global Fit Residual*, 2410.17180.

[60] N. Karnesis, M. Lilley and A. Petiteau, *Assessing the detectability of a Stochastic Gravitational Wave Background with LISA, using an excess of power approach*, *Class. Quant. Grav.* **37** (2020) 215017 [1906.09027].

[61] C. Caprini, D. G. Figueroa, R. Flauger, G. Nardini, M. Peloso, M. Pieroni et al., *Reconstructing the spectral shape of a stochastic gravitational wave background with LISA*, *JCAP* **11** (2019) 017 [1906.09244].

[62] R. Flauger, N. Karnesis, G. Nardini, M. Pieroni, A. Ricciardone and J. Torrado, *Improved reconstruction of a stochastic gravitational wave background with LISA*, *JCAP* **01** (2021) 059 [2009.11845].

[63] M. Pieroni and E. Barausse, *Foreground cleaning and template-free stochastic background extraction for LISA*, *JCAP* **07** (2020) 021 [2004.01135].

[64] Q. Baghi, N. Karnesis, J.-B. Bayle, M. Besançon and H. Inchauspé, *Uncovering gravitational-wave backgrounds from noises of unknown shape with LISA*, *JCAP* **04** (2023) 066 [2302.12573].

[65] F. Pozzoli, R. Buscicchio, C. J. Moore, F. Haardt and A. Sesana, *Weakly parametric approach to stochastic background inference in LISA*, *Phys. Rev. D* **109** (2024) 083029 [2311.12111].

[66] A. Dimitriou, D. G. Figueroa and B. Zaldivar, *Fast likelihood-free reconstruction of gravitational wave backgrounds*, *JCAP* **09** (2024) 032 [2309.08430].

[67] LISA Cosmology Working Group collaboration, *Gravitational waves from inflation in LISA: reconstruction pipeline and physics interpretation*, *JCAP* **11** (2024) 032 [2407.04356].

[68] A. J. Iovino, S. Matarrese, G. Perna, A. Ricciardone and A. Riotto, *How Well Do We Know the Scalar-Induced Gravitational Waves?*, 2412.06764.

[69] A. A. Starobinsky, *A New Type of Isotropic Cosmological Models Without Singularity*, *Phys. Lett. B* **91** (1980) 99.

[70] A. H. Guth, *The Inflationary Universe: A Possible Solution to the Horizon and Flatness Problems*, *Phys. Rev. D* **23** (1981) 347.

[71] A. D. Linde, *A New Inflationary Universe Scenario: A Possible Solution of the Horizon, Flatness, Homogeneity, Isotropy and Primordial Monopole Problems*, *Phys. Lett. B* **108** (1982) 389.

[72] A. Albrecht and P. J. Steinhardt, *Cosmology for Grand Unified Theories with Radiatively Induced Symmetry Breaking*, *Phys. Rev. Lett.* **48** (1982) 1220.

[73] Planck collaboration, *Planck 2018 results. X. Constraints on inflation*, *Astron. Astrophys.* **641** (2020) A10 [1807.06211].

[74] P. Ivanov, P. Naselsky and I. Novikov, *Inflation and primordial black holes as dark matter*, *Phys. Rev. D* **50** (1994) 7173.

[75] W. H. Kinney, *A Hamilton-Jacobi approach to nonslow roll inflation*, *Phys. Rev. D* **56** (1997) 2002 [hep-ph/9702427].

[76] S. Inoue and J. Yokoyama, *Curvature perturbation at the local extremum of the inflaton's potential*, *Phys. Lett. B* **524** (2002) 15 [hep-ph/0104083].

[77] W. H. Kinney, *Horizon crossing and inflation with large eta*, *Phys. Rev. D* **72** (2005) 023515 [gr-qc/0503017].

[78] J. Martin, H. Motohashi and T. Suyama, *Ultra Slow-Roll Inflation and the non-Gaussianity Consistency Relation*, *Phys. Rev. D* **87** (2013) 023514 [1211.0083].

[79] H. Motohashi and W. Hu, *Primordial Black Holes and Slow-Roll Violation*, *Phys. Rev. D* **96** (2017) 063503 [1706.06784].

[80] O. Özsoy and G. Tasinato, *On the slope of the curvature power spectrum in non-attractor inflation*, *JCAP* **04** (2020) 048 [1912.01061].

[81] A. Karam, N. Koivunen, E. Tomberg, V. Vaskonen and H. Veermäe, *Anatomy of single-field inflationary models for primordial black holes*, *JCAP* **03** (2023) 013 [2205.13540].

[82] J. Garcia-Bellido and E. Ruiz Morales, *Primordial black holes from single field models of inflation*, *Phys. Dark Univ.* **18** (2017) 47 [1702.03901].

[83] K. Kannike, L. Marzola, M. Raidal and H. Veermäe, *Single Field Double Inflation and Primordial Black Holes*, *JCAP* **09** (2017) 020 [1705.06225].

[84] G. Ballesteros and M. Taoso, *Primordial black hole dark matter from single field inflation*, *Phys. Rev. D* **97** (2018) 023501 [1709.05565].

[85] C. Germani and T. Prokopec, *On primordial black holes from an inflection point*, *Phys. Dark Univ.* **18** (2017) 6 [1706.04226].

[86] J. M. Ezquiaga, J. Garcia-Bellido and E. Ruiz Morales, *Primordial Black Hole production in Critical Higgs Inflation*, *Phys. Lett. B* **776** (2018) 345 [1705.04861].

[87] H. Di and Y. Gong, *Primordial black holes and second order gravitational waves from ultra-slow-roll inflation*, *JCAP* **07** (2018) 007 [1707.09578].

[88] M. P. Hertzberg and M. Yamada, *Primordial Black Holes from Polynomial Potentials in Single Field Inflation*, *Phys. Rev. D* **97** (2018) 083509 [1712.09750].

[89] S. Rasanen and E. Tomberg, *Planck scale black hole dark matter from Higgs inflation*, *JCAP* **01** (2019) 038 [1810.12608].

[90] M. Cicoli, V. A. Diaz and F. G. Pedro, *Primordial Black Holes from String Inflation*, *JCAP* **06** (2018) 034 [1803.02837].

[91] O. Özsoy, S. Parameswaran, G. Tasinato and I. Zavala, *Mechanisms for Primordial Black Hole Production in String Theory*, *JCAP* **07** (2018) 005 [1803.07626].

[92] T.-J. Gao and Z.-K. Guo, *Primordial Black Hole Production in Inflationary Models of Supergravity with a Single Chiral Superfield*, *Phys. Rev. D* **98** (2018) 063526 [1806.09320].

[93] V. Atal, J. Garriga and A. Marcos-Caballero, *Primordial black hole formation with non-Gaussian curvature perturbations*, *JCAP* **09** (2019) 073 [1905.13202].

[94] V. Atal, J. Cid, A. Escrivà and J. Garriga, *PBH in single field inflation: the effect of shape dispersion and non-Gaussianities*, *JCAP* **05** (2020) 022 [1908.11357].

[95] S. S. Mishra and V. Sahni, *Primordial Black Holes from a tiny bump/dip in the Inflaton potential*, *JCAP* **04** (2020) 007 [1911.00057].

[96] G. Ballesteros, J. Rey and F. Rompineve, *Detuning primordial black hole dark matter with early matter domination and axion monodromy*, *JCAP* **06** (2020) 014 [1912.01638].

[97] I. Dalianis, A. Kehagias and G. Tringas, *Primordial black holes from α-attractors*, *JCAP* **01** (2019) 037 [1805.09483].

[98] N. Bhaumik and R. K. Jain, *Primordial black holes dark matter from inflection point models of inflation and the effects of reheating*, *JCAP* **01** (2020) 037 [1907.04125].

[99] M. Drees and Y. Xu, *Overshooting, Critical Higgs Inflation and Second Order Gravitational Wave Signatures*, *Eur. Phys. J. C* **81** (2021) 182 [1905.13581].

[100] I. Dalianis and G. Tringas, *Primordial black hole remnants as dark matter produced in thermal, matter, and runaway-quintessence postinflationary scenarios*, *Phys. Rev. D* **100** (2019) 083512 [1905.01741].

[101] G. Ballesteros, J. Rey, M. Taoso and A. Urbano, *Primordial black holes as dark matter and gravitational waves from single-field polynomial inflation*, *JCAP* **07** (2020) 025 [2001.08220].

[102] D. V. Nanopoulos, V. C. Spanos and I. D. Stamou, *Primordial Black Holes from No-Scale Supergravity*, *Phys. Rev. D* **102** (2020) 083536 [2008.01457].

[103] L. Iacconi, H. Assadullahi, M. Fasiello and D. Wands, *Revisiting small-scale fluctuations in α-attractor models of inflation*, *JCAP* **06** (2022) 007 [2112.05092].

[104] I. D. Stamou, *Mechanisms of producing primordial black holes by breaking the $SU(2,1)/SU(2) \times U(1)$ symmetry*, *Phys. Rev. D* **103** (2021) 083512 [2104.08654].

[105] L. Wu, Y. Gong and T. Li, *Primordial black holes and secondary gravitational waves from string inspired general no-scale supergravity*, *Phys. Rev. D* **104** (2021) 123544 [2105.07694].

[106] K.-W. Ng and Y.-P. Wu, *Constant-rate inflation: primordial black holes from conformal weight transitions*, *JHEP* **11** (2021) 076 [2102.05620].

[107] K. Rezazadeh, Z. Teimoori, S. Karimi and K. Karami, *Non-Gaussianity and secondary gravitational waves from primordial black holes production in α-attractor inflation*, *Eur. Phys. J. C* **82** (2022) 758 [2110.01482].

[108] Q. Wang, Y.-C. Liu, B.-Y. Su and N. Li, *Primordial black holes from the perturbations in the inflaton potential in peak theory*, *Phys. Rev. D* **104** (2021) 083546 [2111.10028].

[109] B.-M. Gu, F.-W. Shu, K. Yang and Y.-P. Zhang, *Primordial black holes from an inflationary potential valley*, *Phys. Rev. D* **107** (2023) 023519 [2207.09968].

[110] D. Frolovsky, S. V. Ketov and S. Saburov, *E-models of inflation and primordial black holes*, *Front. in Phys.* **10** (2022) 1005333 [2207.11878].

[111] M. Cicoli, F. G. Pedro and N. Pedron, *Secondary GWs and PBHs in string inflation: formation and detectability*, *JCAP* **08** (2022) 030 [2203.00021].

[112] A. Ghoshal, A. Moursy and Q. Shafi, *Cosmological probes of grand unification: Primordial black holes and scalar-induced gravitational waves*, *Phys. Rev. D* **108** (2023) 055039 [2306.04002].

[113] Y.-F. Cai, X.-H. Ma, M. Sasaki, D.-G. Wang and Z. Zhou, *One small step for an inflaton, one giant leap for inflation: A novel non-Gaussian tail and primordial black holes*, *Phys. Lett. B* **834** (2022) 137461 [2112.13836].

[114] K. Inomata, E. McDonough and W. Hu, *Amplification of primordial perturbations from the rise or fall of the inflaton*, *JCAP* **02** (2022) 031 [2110.14641].

[115] K. Inomata, E. McDonough and W. Hu, *Primordial black holes arise when the inflaton falls*, *Phys. Rev. D* **104** (2021) 123553 [2104.03972].

[116] K. Kefala, G. P. Kodaxis, I. D. Stamou and N. Tetradis, *Features of the inflaton potential and the power spectrum of cosmological perturbations*, *Phys. Rev. D* **104** (2021) 023506 [2010.12483].

[117] I. Dalianis, G. P. Kodaxis, I. D. Stamou, N. Tetradis and A. Tsigkas-Kouvelis, *Spectrum oscillations from features in the potential of single-field inflation*, *Phys. Rev. D* **104** (2021) 103510 [2106.02467].

[118] J. Yokoyama, *Chaotic new inflation and formation of primordial black holes*, *Phys. Rev. D* **58** (1998) 083510 [astro-ph/9802357].

[119] R. Saito, J. Yokoyama and R. Nagata, *Single-field inflation, anomalous enhancement of superhorizon fluctuations, and non-Gaussianity in primordial black hole formation*, *JCAP* **06** (2008) 024 [0804.3470].

[120] E. Bugaev and P. Klimai, *Large curvature perturbations near horizon crossing in single-field inflation models*, *Phys. Rev. D* **78** (2008) 063515 [0806.4541].

[121] C. Fu, P. Wu and H. Yu, *Primordial black holes and oscillating gravitational waves in slow-roll and slow-climb inflation with an intermediate noninflationary phase*, *Phys. Rev. D* **102** (2020) 043527 [2006.03768].

[122] V. Briaud and V. Vennin, *Uphill inflation*, *JCAP* **06** (2023) 029 [2301.09336].

[123] A. Karam, N. Koivunen, E. Tomberg, A. Racioppi and H. Veermäe, *Primordial black holes and inflation from double-well potentials*, *JCAP* **09** (2023) 002 [2305.09630].

[124] R.-G. Cai, Z.-K. Guo, J. Liu, L. Liu and X.-Y. Yang, *Primordial black holes and gravitational waves from parametric amplification of curvature perturbations*, *JCAP* **06** (2020) 013 [1912.10437].

[125] G. Tasinato, *An analytic approach to non-slow-roll inflation*, *Phys. Rev. D* **103** (2021) 023535 [2012.02518].

[126] Z. Zhou, J. Jiang, Y.-F. Cai, M. Sasaki and S. Pi, *Primordial black holes and gravitational waves from resonant amplification during inflation*, *Phys. Rev. D* **102** (2020) 103527 [2010.03537].

[127] Z.-Z. Peng, C. Fu, J. Liu, Z.-K. Guo and R.-G. Cai, *Gravitational waves from resonant amplification of curvature perturbations during inflation*, *JCAP* **10** (2021) 050 [2106.11816].

[128] K. Inomata, M. Braglia, X. Chen and S. Renaux-Petel, *Questions on calculation of primordial power spectrum with large spikes: the resonance model case*, *JCAP* **04** (2023) 011 [2211.02586].

[129] J. Fumagalli, S. Bhattacharya, M. Peloso, S. Renaux-Petel and L. T. Witkowski, *One-loop infrared rescattering by enhanced scalar fluctuations during inflation*, *JCAP* **04** (2024) 029 [2307.08358].

[130] A. Caravano, K. Inomata and S. Renaux-Petel, *Inflationary Butterfly Effect: Nonperturbative Dynamics from Small-Scale Features*, *Phys. Rev. Lett.* **133** (2024) 151001 [2403.12811].

[131] D. Frolovsky, S. V. Ketov and S. Saburov, *Formation of primordial black holes after Starobinsky inflation*, *Mod. Phys. Lett. A* **37** (2022) 2250135 [2205.00603].

[132] G. Ballesteros, J. Beltran Jimenez and M. Pieroni, *Black hole formation from a general quadratic action for inflationary primordial fluctuations*, *JCAP* **06** (2019) 016 [1811.03065].

[133] C. Fu, P. Wu and H. Yu, *Primordial Black Holes from Inflation with Nonminimal Derivative Coupling*, *Phys. Rev. D* **100** (2019) 063532 [1907.05042].

[134] C. Fu, P. Wu and H. Yu, *Scalar induced gravitational waves in inflation with gravitationally enhanced friction*, *Phys. Rev. D* **101** (2020) 023529 [1912.05927].

[135] S. Heydari and K. Karami, *Primordial black holes in nonminimal derivative coupling inflation with quartic potential and reheating consideration*, *Eur. Phys. J. C* **82** (2022) 83 [2107.10550].

[136] S. Heydari and K. Karami, *Primordial black holes ensued from exponential potential and coupling parameter in nonminimal derivative inflation model*, *JCAP* **03** (2022) 033 [2111.00494].

[137] S. Kawai and J. Kim, *Primordial black holes from Gauss-Bonnet-corrected single field inflation*, *Phys. Rev. D* **104** (2021) 083545 [2108.01340].

[138] R. Arya, *Formation of Primordial Black Holes from Warm Inflation*, *JCAP* **09** (2020) 042 [1910.05238].

[139] A. Ashoorioon, A. Rostami and J. T. Firouzjaee, *EFT compatible PBHs: effective spawning of the seeds for primordial black holes during inflation*, *JHEP* **07** (2021) 087 [1912.13326].

[140] M. Bastero-Gil and M. S. Díaz-Blanco, *Gravity waves and primordial black holes in scalar warm little inflation*, *JCAP* **12** (2021) 052 [2105.08045].

[141] O. Özsoy and Z. Lalak, *Primordial black holes as dark matter and gravitational waves from bumpy axion inflation*, *JCAP* **01** (2021) 040 [2008.07549].

[142] M. Solbi and K. Karami, *Primordial black holes and induced gravitational waves in k-inflation*, *JCAP* **08** (2021) 056 [2102.05651].

[143] M. Solbi and K. Karami, *Primordial black holes formation in the inflationary model with field-dependent kinetic term for quartic and natural potentials*, *Eur. Phys. J. C* **81** (2021) 884 [2106.02863].

[144] Z. Teimoori, K. Rezazadeh, M. A. Rasheed and K. Karami, *Mechanism of primordial black holes production and secondary gravitational waves in α-attractor Galileon inflationary scenario*, *JCAP* **10** (2021) [2107.07620].

[145] M. Correa, M. R. Gangopadhyay, N. Jaman and G. J. Mathews, *Primordial black-hole dark matter via warm natural inflation*, *Phys. Lett. B* **835** (2022) 137510 [2207.10394].

[146] R. Kawaguchi and S. Tsujikawa, *Primordial black holes from Higgs inflation with a Gauss-Bonnet coupling*, *Phys. Rev. D* **107** (2023) 063508 [2211.13364].

[147] A. Poisson, I. Timiryasov and S. Zell, *Critical points in Palatini Higgs inflation with small non-minimal coupling*, *JHEP* **03** (2024) 130 [2306.03893].

[148] O. Özsoy and G. Tasinato, *Inflation and Primordial Black Holes*, *Universe* **9** (2023) 203 [2301.03600].

[149] J. Kristiano and J. Yokoyama, *Constraining Primordial Black Hole Formation from Single-Field Inflation*, *Phys. Rev. Lett.* **132** (2024) 221003 [2211.03395].

[150] A. Riotto, *The Primordial Black Hole Formation from Single-Field Inflation is Not Ruled Out*, 2301.00599.

[151] J. Kristiano and J. Yokoyama, *Note on the bispectrum and one-loop corrections in single-field inflation with primordial black hole formation*, *Phys. Rev. D* **109** (2024) 103541 [2303.00341].

[152] A. Riotto, *The Primordial Black Hole Formation from Single-Field Inflation is Still Not Ruled Out*, 2303.01727.

[153] H. Firouzjahi, *One-loop corrections in power spectrum in single field inflation*, *JCAP* **10** (2023) 006 [2303.12025].

[154] H. Firouzjahi and A. Riotto, *Primordial Black Holes and loops in single-field inflation*, *JCAP* **02** (2024) 021 [2304.07801].

[155] G. Franciolini, A. Iovino, Junior., M. Taoso and A. Urbano, *Perturbativity in the presence of ultraslow-roll dynamics*, *Phys. Rev. D* **109** (2024) 123550 [2305.03491].

[156] S.-L. Cheng, D.-S. Lee and K.-W. Ng, *Primordial perturbations from ultra-slow-roll single-field inflation with quantum loop effects*, *JCAP* **03** (2024) 008 [2305.16810].

[157] S. Maity, H. V. Ragavendra, S. K. Sethi and L. Sriramkumar, *Loop contributions to the scalar power spectrum due to quartic order action in ultra slow roll inflation*, *JCAP* **05** (2024) 046 [2307.13636].

[158] M. W. Davies, L. Iacconi and D. J. Mulryne, *Numerical 1-loop correction from a potential yielding ultra-slow-roll dynamics*, *JCAP* **04** (2024) 050 [2312.05694].

[159] G. Ballesteros and J. G. Egea, *One-loop power spectrum in ultra slow-roll inflation and implications for primordial black hole dark matter*, *JCAP* **07** (2024) 052 [2404.07196].

[160] J. Fumagalli, *Absence of one-loop effects on large scales from small scales in non-slow-roll dynamics*, 2305.19263.

[161] Y. Tada, T. Terada and J. Tokuda, *Cancellation of quantum corrections on the soft curvature perturbations*, *JHEP* **01** (2024) 105 [2308.04732].

[162] K. Inomata, *Superhorizon Curvature Perturbations Are Protected against One-Loop Corrections*, *Phys. Rev. Lett.* **133** (2024) 141001 [2403.04682].

[163] R. Kawaguchi, S. Tsujikawa and Y. Yamada, *Proving the absence of large one-loop corrections to the power spectrum of curvature perturbations in transient ultra-slow-roll inflation within the path-integral approach*, *JHEP* **12** (2024) 095 [2407.19742].

[164] J. Fumagalli, *Absence of one-loop effects on large scales from small scales in non-slow-roll dynamics II: Quartic interactions and consistency relations*, 2408.08296.

[165] J. Garcia-Bellido, A. D. Linde and D. Wands, *Density perturbations and black hole formation in hybrid inflation*, *Phys. Rev. D* **54** (1996) 6040 [astro-ph/9605094].

[166] M. Kawasaki and Y. Tada, *Can massive primordial black holes be produced in mild waterfall hybrid inflation?*, *JCAP* **08** (2016) 041 [1512.03515].

[167] R. Kallosh and A. Linde, *Dilaton-axion inflation with PBHs and GWs*, *JCAP* **08** (2022) 037 [2203.10437].

[168] M. Braglia, A. Linde, R. Kallosh and F. Finelli, *Hybrid α-attractors, primordial black holes and gravitational wave backgrounds*, *JCAP* **04** (2023) 033 [2211.14262].

[169] Y. Tada and M. Yamada, *Primordial black hole formation in hybrid inflation*, *Phys. Rev. D* **107** (2023) 123539 [2304.01249].

[170] Y. Tada and M. Yamada, *Stochastic dynamics of multi-waterfall hybrid inflation and formation of primordial black holes*, *JCAP* **11** (2023) 089 [2306.07324].

[171] G. A. Palma, S. Sypsas and C. Zenteno, *Seeding primordial black holes in multifield inflation*, *Phys. Rev. Lett.* **125** (2020) 121301 [2004.06106].

[172] J. Fumagalli, S. Renaux-Petel, J. W. Ronayne and L. T. Witkowski, *Turning in the landscape: A new mechanism for generating primordial black holes*, *Phys. Lett. B* **841** (2023) 137921 [2004.08369].

[173] J. Fumagalli, S. Renaux-Petel and L. T. Witkowski, *Oscillations in the stochastic gravitational wave background from sharp features and particle production during inflation*, *JCAP* **08** (2021) 030 [2012.02761].

[174] M. Braglia, D. K. Hazra, F. Finelli, G. F. Smoot, L. Sriramkumar and A. A. Starobinsky, *Generating PBHs and small-scale GWs in two-field models of inflation*, *JCAP* **08** (2020) 001 [2005.02895].

[175] M. Braglia, X. Chen and D. K. Hazra, *Probing Primordial Features with the Stochastic Gravitational Wave Background*, *JCAP* **03** (2021) 005 [2012.05821].

[176] S. Bhattacharya and I. Zavala, *Sharp turns in axion monodromy: primordial black holes and gravitational waves*, *JCAP* **04** (2023) 065 [2205.06065].

[177] V. Aragam, S. Paban and R. Rosati, *Primordial Stochastic Gravitational Wave Backgrounds from a Sharp Feature in Three-field Inflation*, 2304.00065.

[178] D. H. Lyth and D. Wands, *Generating the curvature perturbation without an inflaton*, *Phys. Lett. B* **524** (2002) 5 [hep-ph/0110002].

[179] C. Chen and Y.-F. Cai, *Primordial black holes from sound speed resonance in the inflaton-curvaton mixed scenario*, *JCAP* **10** (2019) 068 [1908.03942].

[180] A. D. Gow, T. Miranda and S. Nurmi, *Primordial black holes from a curvaton scenario with strongly non-Gaussian perturbations*, *JCAP* **11** (2023) 006 [2307.03078].

[181] L.-H. Liu and T. Prokopec, *Non-minimally coupled curvaton*, *JCAP* **06** (2021) 033 [2005.11069].

[182] S. Pi and M. Sasaki, *Primordial black hole formation in nonminimal curvaton scenarios*, *Phys. Rev. D* **108** (2023) L101301 [2112.12680].

[183] M. Kawasaki, N. Kitajima and T. T. Yanagida, *Primordial black hole formation from an axionlike curvaton model*, *Phys. Rev. D* **87** (2013) 063519 [1207.2550].

[184] K. Ando, K. Inomata, M. Kawasaki, K. Mukaida and T. T. Yanagida, *Primordial black holes for the LIGO events in the axionlike curvaton model*, *Phys. Rev. D* **97** (2018) 123512 [1711.08956].

[185] K. Ando, M. Kawasaki and H. Nakatsuka, *Formation of primordial black holes in an axionlike curvaton model*, *Phys. Rev. D* **98** (2018) 083508 [1805.07757].

[186] K. Inomata, M. Kawasaki, K. Mukaida and T. T. Yanagida, *NANOGrav Results and LIGO-Virgo Primordial Black Holes in Axionlike Curvaton Models*, *Phys. Rev. Lett.* **126** (2021) 131301 [2011.01270].

[187] N. Bartolo, S. Matarrese and A. Riotto, *On nonGaussianity in the curvaton scenario*, *Phys. Rev. D* **69** (2004) 043503 [hep-ph/0309033].

[188] N. Bartolo, S. Matarrese and A. Riotto, *Non-Gaussianity of Large-Scale Cosmic Microwave Background Anisotropies beyond Perturbation Theory*, *JCAP* **08** (2005) 010 [astro-ph/0506410].

[189] M. Sasaki, J. Valiviita and D. Wands, *Non-Gaussianity of the primordial perturbation in the curvaton model*, *Phys. Rev. D* **74** (2006) 103003 [astro-ph/0607627].

[190] K. Enqvist and T. Takahashi, *Signatures of Non-Gaussianity in the Curvaton Model*, *JCAP* **09** (2008) 012 [0807.3069].

[191] M. Kawasaki, N. Kitajima and S. Yokoyama, *Gravitational waves from a curvaton model with blue spectrum*, *JCAP* **08** (2013) 042 [1305.4464].

[192] G. Ferrante, G. Franciolini, A. Iovino, Junior. and A. Urbano, *Primordial black holes in the curvaton model: possible connections to pulsar timing arrays and dark matter*, *JCAP* **06** (2023) 057 [2305.13382].

[193] C. Chen, A. Ghoshal, G. Tasinato and E. Tomberg, *Stochastic Axion-like Curvaton: Non-Gaussianity and Primordial Black Holes Without Large Power Spectrum*, 2409.12950.

[194] N. Barnaby and M. Peloso, *Large Nongaussianity in Axion Inflation*, *Phys. Rev. Lett.* **106** (2011) 181301 [1011.1500].

[195] L. Sorbo, *Parity violation in the Cosmic Microwave Background from a pseudoscalar inflaton*, *JCAP* **06** (2011) 003 [1101.1525].

[196] R. Namba, M. Peloso, M. Shiraishi, L. Sorbo and C. Unal, *Scale-dependent gravitational waves from a rolling axion*, *JCAP* **01** (2016) 041 [1509.07521].

[197] M. Shiraishi, A. Ricciardone and S. Saga, *Parity violation in the CMB bispectrum by a rolling pseudoscalar*, *JCAP* **11** (2013) 051 [1308.6769].

[198] A. Linde, S. Mooij and E. Pajer, *Gauge field production in supergravity inflation: Local non-Gaussianity and primordial black holes*, *Phys. Rev. D* **87** (2013) 103506 [1212.1693].

[199] E. Bugaev and P. Klimai, *Axion inflation with gauge field production and primordial black holes*, *Phys. Rev. D* **90** (2014) 103501 [1312.7435].

[200] J. Garcia-Bellido, M. Peloso and C. Unal, *Gravitational waves at interferometer scales and primordial black holes in axion inflation*, *JCAP* **12** (2016) 031 [1610.03763].

[201] A. Caravano, E. Komatsu, K. D. Lozanov and J. Weller, *Lattice simulations of axion-U(1) inflation*, *Phys. Rev. D* **108** (2023) 043504 [2204.12874].

[202] D. G. Figueroa, J. Lizarraga, A. Urio and J. Urrestilla, *The strong backreaction regime in axion inflation*, 2303.17436.

[203] A. Caravano and M. Peloso, *Unveiling the nonlinear dynamics of a rolling axion during inflation*, 2407.13405.

[204] D. G. Figueroa, J. Lizarraga, N. Loayza, A. Urio and J. Urrestilla, *The non-linear dynamics of axion inflation: a detailed lattice study*, 2411.16368.

[205] L. Kofman, A. D. Linde and A. A. Starobinsky, *Reheating after inflation*, *Phys. Rev. Lett.* **73** (1994) 3195 [hep-th/9405187].

[206] L. Kofman, A. D. Linde and A. A. Starobinsky, *Towards the theory of reheating after inflation*, *Phys. Rev. D* **56** (1997) 3258 [hep-ph/9704452].

[207] F. Finelli and R. H. Brandenberger, *Parametric amplification of gravitational fluctuations during reheating*, *Phys. Rev. Lett.* **82** (1999) 1362 [hep-ph/9809490].

[208] B. A. Bassett, F. Tamburini, D. I. Kaiser and R. Maartens, *Metric preheating and limitations of linearized gravity. 2.*, *Nucl. Phys. B* **561** (1999) 188 [hep-ph/9901319].

[209] K. Jedamzik and G. Sigl, *On metric preheating*, *Phys. Rev. D* **61** (2000) 023519 [hep-ph/9906287].

[210] B. A. Bassett and F. Viniegra, *Massless metric preheating*, *Phys. Rev. D* **62** (2000) 043507 [hep-ph/9909353].

[211] K. Jedamzik, M. Lemoine and J. Martin, *Generation of gravitational waves during early structure formation between cosmic inflation and reheating*, *JCAP* **04** (2010) 021 [1002.3278].

[212] K. Jedamzik, M. Lemoine and J. Martin, *Collapse of Small-Scale Density Perturbations during Preheating in Single Field Inflation*, *JCAP* **09** (2010) 034 [1002.3039].

[213] J. Martin, T. Papanikolaou and V. Vennin, *Primordial black holes from the preheating instability in single-field inflation*, *JCAP* **01** (2020) 024 [1907.04236].

[214] J. Martin, T. Papanikolaou, L. Pinol and V. Vennin, *Metric preheating and radiative decay in single-field inflation*, *JCAP* **05** (2020) 003 [2002.01820].

[215] G. Ballesteros, J. Iguaz Juan, P. D. Serpico and M. Taoso, *Primordial black hole formation from self-resonant preheating?*, 2406.09122.

[216] B. A. Bassett, D. I. Kaiser and R. Maartens, *General relativistic preheating after inflation*, *Phys. Lett. B* **455** (1999) 84 [hep-ph/9808404].

[217] A. M. Green and K. A. Malik, *Primordial black hole production due to preheating*, *Phys. Rev. D* **64** (2001) 021301 [hep-ph/0008113].

[218] B. A. Bassett and S. Tsujikawa, *Inflationary preheating and primordial black holes*, *Phys. Rev. D* **63** (2001) 123503 [hep-ph/0008328].

[219] T. Suyama, T. Tanaka, B. Bassett and H. Kudoh, *Are black holes over-produced during preheating?*, *Phys. Rev. D* **71** (2005) 063507 [hep-ph/0410247].

[220] E. Torres-Lomas and L. A. Urena-LAlpez, *Primordial black hole production during preheating in a chaotic inflationary model*, *AIP Conf. Proc.* **1548** (2013) 238 [1308.1268].

[221] E. Torres-Lomas, J. C. Hidalgo, K. A. Malik and L. A. Ureña López, *Formation of subhorizon black holes from preheating*, *Phys. Rev. D* **89** (2014) 083008 [1401.6960].

[222] X.-X. Kou, C. Tian and S.-Y. Zhou, *Oscillon Preheating in Full General Relativity*, *Class. Quant. Grav.* **38** (2021) 045005 [1912.09658].

[223] C. Joana, *Gravitational dynamics in Higgs inflation: Preinflation and preheating with an auxiliary field*, *Phys. Rev. D* **106** (2022) 023504 [2202.07604].

[224] P. Adshead, J. T. Giblin, R. Grutkoski and Z. J. Weiner, *Gauge preheating with full general relativity*, *JCAP* **03** (2024) 017 [2311.01504].

[225] R. Easther, R. Flauger and J. B. Gilmore, *Delayed Reheating and the Breakdown of Coherent Oscillations*, *JCAP* **04** (2011) 027 [1003.3011].

[226] Y. Cui and E. I. Sfakianakis, *Detectable gravitational wave signals from inflationary preheating*, *Phys. Lett. B* **840** (2023) 137825 [2112.00762].

[227] R. H. Brandenberger, *The Matter Bounce Alternative to Inflationary Cosmology*, 1206.4196.

[228] Y.-F. Cai, D. A. Easson and R. Brandenberger, *Towards a Nonsingular Bouncing Cosmology*, *JCAP* **08** (2012) 020 [1206.2382].

[229] Y.-F. Cai, E. McDonough, F. Duplessis and R. H. Brandenberger, *Two Field Matter Bounce Cosmology*, *JCAP* **10** (2013) 024 [1305.5259].

[230] J.-W. Chen, J. Liu, H.-L. Xu and Y.-F. Cai, *Tracing Primordial Black Holes in Nonsingular Bouncing Cosmology*, *Phys. Lett. B* **769** (2017) 561 [1609.02571].

[231] S. Banerjee, T. Papanikolaou and E. N. Saridakis, *Constraining F(R) bouncing cosmologies through primordial black holes*, *Phys. Rev. D* **106** (2022) 124012 [2206.01150].

[232] T. Papanikolaou, S. Banerjee, Y.-F. Cai, S. Capozziello and E. N. Saridakis, *Primordial black holes and induced gravitational waves in non-singular matter bouncing cosmology*, *JCAP* **06** (2024) 066 [2404.03779].

[233] T. Papanikolaou, V. Vennin and D. Langlois, *Gravitational waves from a universe filled with primordial black holes*, *JCAP* **03** (2021) 053 [2010.11573].

[234] G. Domènech, C. Lin and M. Sasaki, *Gravitational wave constraints on the primordial black hole dominated early universe*, *JCAP* **04** (2021) 062 [2012.08151].

[235] G. Domènech, V. Takhistov and M. Sasaki, *Exploring evaporating primordial black holes with gravitational waves*, *Phys. Lett. B* **823** (2021) 136722 [2105.06816].

[236] G. Domènech, S. Passaglia and S. Renaux-Petel, *Gravitational waves from dark matter isocurvature*, *JCAP* **03** (2022) 023 [2112.10163].

[237] G. Domènech, *Cosmological gravitational waves from isocurvature fluctuations*, *AAPPS Bull.* **34** (2024) 4 [2311.02065].

[238] T. Papanikolaou, X.-C. He, X.-H. Ma, Y.-F. Cai, E. N. Saridakis and M. Sasaki, *New probe of non-Gaussianities with primordial black hole induced gravitational waves*, *Phys. Lett. B* **857** (2024) 138997 [2403.00660].

[239] X.-C. He, Y.-F. Cai, X.-H. Ma, T. Papanikolaou, E. N. Saridakis and M. Sasaki, *Gravitational waves from primordial black hole isocurvature: the effect of non-Gaussianities*, *JCAP* **12** (2024) 039 [2409.11333].

[240] T. Papanikolaou, C. Tzerefos, S. Basilakos and E. N. Saridakis, *Scalar induced gravitational waves from primordial black hole Poisson fluctuations in f(R) gravity*, *JCAP* **10** (2022) 013 [2112.15059].

[241] T. Papanikolaou, C. Tzerefos, S. Basilakos and E. N. Saridakis, *No constraints for f(T) gravity from gravitational waves induced from primordial black hole fluctuations*, *Eur. Phys. J. C* **83** (2023) 31 [2205.06094].

[242] C. Tzerefos, T. Papanikolaou, E. N. Saridakis and S. Basilakos, *Scalar induced gravitational waves in modified teleparallel gravity theories*, *Phys. Rev. D* **107** (2023) 124019 [2303.16695].

[243] G. Dvali, L. Eisemann, M. Michel and S. Zell, *Black hole metamorphosis and stabilization by memory burden*, *Phys. Rev. D* **102** (2020) 103523 [2006.00011].

[244] G. Domènech and M. Sasaki, *Gravitational wave hints black hole remnants as dark matter*, *Class. Quant. Grav.* **40** (2023) 177001 [2303.07661].

[245] G. Franciolini and P. Pani, *Stochastic gravitational-wave background at 3G detectors as a smoking gun for microscopic dark matter relics*, *Phys. Rev. D* **108** (2023) 083527 [2304.13576].

[246] S. Balaji, G. Domènech, G. Franciolini, A. Ganz and J. Tränkle, *Probing modified Hawking evaporation with gravitational waves from the primordial black hole dominated universe*, *JCAP* **11** (2024) 026 [2403.14309].

[247] G. Dvali, J. S. Valbuena-Bermúdez and M. Zantedeschi, *Memory burden effect in black holes and solitons: Implications for PBH*, *Phys. Rev. D* **110** (2024) 056029 [2405.13117].

[248] K. Kohri, T. Terada and T. T. Yanagida, *Induced Gravitational Waves probing Primordial Black Hole Dark Matter with Memory Burden*, 2409.06365.

[249] N. Bhaumik, M. R. Haque, R. K. Jain and M. Lewicki, *Memory burden effect mimics reheating signatures on SGWB from ultra-low mass PBH domination*, *JHEP* **10** (2024) 142 [2409.04436].

[250] G. Domènech and J. Tränkle, *From formation to evaporation: Induced gravitational wave probes of the primordial black hole reheating scenario*, 2409.12125.

[251] D. Wands, *Duality invariance of cosmological perturbation spectra*, *Phys. Rev. D* **60** (1999) 023507 [gr-qc/9809062].

[252] P. S. Cole, A. D. Gow, C. T. Byrnes and S. P. Patil, *Primordial black holes from single-field inflation: a fine-tuning audit*, *JCAP* **08** (2023) 031 [2304.01997].

[253] E. Madge, E. Morgante, C. Puchades-Ibáñez, N. Ramberg, W. Ratzinger, S. Schenk et al., *Primordial gravitational waves in the nano-Hertz regime and PTA data — towards solving the GW inverse problem*, *JHEP* **10** (2023) 171 [2306.14856].

[254] Planck collaboration, *Planck 2018 results. X. Constraints on inflation*, *Astron. Astrophys.* **641** (2020) A10 [1807.06211].

[255] S. Pi and M. Sasaki, *Gravitational Waves Induced by Scalar Perturbations with a Lognormal Peak*, *JCAP* **09** (2020) 037 [2005.12306].

[256] V. Dandoy, V. Domcke and F. Rompineve, *Search for scalar induced gravitational waves in the international pulsar timing array data release 2 and NANOgrav 12.5 years datasets*, *SciPost Phys. Core* **6** (2023) 060 [2302.07901].

[257] C. T. Byrnes, P. S. Cole and S. P. Patil, *Steepest growth of the power spectrum and primordial black holes*, *JCAP* **06** (2019) 028 [1811.11158].

[258] C.-Z. Li, C. Yuan and Q.-g. Huang, *Gravitational Waves Induced by Scalar Perturbations with a Broken Power-law Peak*, 2407.12914.

[259] J. Chluba, J. Hamann and S. P. Patil, *Features and New Physical Scales in Primordial Observables: Theory and Observation*, *Int. J. Mod. Phys. D* **24** (2015) 1530023 [1505.01834].

[260] A. Slosar et al., *Scratches from the Past: Inflationary Archaeology through Features in the Power Spectrum of Primordial Fluctuations*, *Bull. Am. Astron. Soc.* **51** (2019) 98 [1903.09883].

[261] A. A. Starobinsky, *Spectrum of adiabatic perturbations in the universe when there are singularities in the inflation potential*, *JETP Lett.* **55** (1992) 489.

[262] J. Fumagalli, S. Renaux-Petel and L. T. Witkowski, *Resonant features in the stochastic gravitational wave background*, *JCAP* **08** (2021) 059 [2105.06481].

[263] L. T. Witkowski, G. Domènech, J. Fumagalli and S. Renaux-Petel, *Expansion history-dependent oscillations in the scalar-induced gravitational wave background*, *JCAP* **05** (2022) 028 [2110.09480].

[264] J. Fumagalli, M. Pieroni, S. Renaux-Petel and L. T. Witkowski, *Detecting primordial features with LISA*, *JCAP* **07** (2022) 020 [2112.06903].

[265] S. Groot Nibbelink and B. J. W. van Tent, *Density perturbations arising from multiple field slow roll inflation*, hep-ph/0011325.

[266] S. Groot Nibbelink and B. J. W. van Tent, *Scalar perturbations during multiple field slow-roll inflation*, *Class. Quant. Grav.* **19** (2002) 613 [hep-ph/0107272].

[267] P. S. Cole, A. D. Gow, C. T. Byrnes and S. P. Patil, *Smooth vs instant inflationary transitions: steepest growth re-examined and primordial black holes*, *JCAP* **05** (2024) 022 [2204.07573].

[268] G. Franciolini and A. Urbano, *Primordial black hole dark matter from inflation: The reverse engineering approach*, *Phys. Rev. D* **106** (2022) 123519 [2207.10056].

[269] J. Ellis and D. Wands, *Inflation (2023)*, 2312.13238.

[270] M. Sasaki, *Large Scale Quantum Fluctuations in the Inflationary Universe*, *Prog. Theor. Phys.* **76** (1986) 1036.

[271] V. F. Mukhanov, *Quantum Theory of Gauge Invariant Cosmological Perturbations*, *Sov. Phys. JETP* **67** (1988) 1297.

[272] BICEP, Keck collaboration, *Improved Constraints on Primordial Gravitational Waves using Planck, WMAP, and BICEP/Keck Observations through the 2018 Observing Season*, *Phys. Rev. Lett.* **127** (2021) 151301 [2110.00483].

[273] S.-L. Cheng, D.-S. Lee and K.-W. Ng, *Power spectrum of primordial perturbations during ultra-slow-roll inflation with back reaction effects*, *Phys. Lett. B* **827** (2022) 136956 [2106.09275].

[274] J. Kristiano and J. Yokoyama, *Comparing sharp and smooth transitions of the second slow-roll parameter in single-field inflation*, *JCAP* **10** (2024) 036 [2405.12145].

[275] L. Iacconi, D. Mulryne and D. Seery, *Loop corrections in the separate universe picture*, *JCAP* **06** (2024) 062 [2312.12424].

[276] H. Motohashi and Y. Tada, *Squeezed bispectrum and one-loop corrections in transient constant-roll inflation*, *JCAP* **08** (2023) 069 [2303.16035].

[277] G. Tasinato, *Non-Gaussianities and the large —η— approach to inflation*, *Phys. Rev. D* **109** (2024) 063510 [2312.03498].

[278] G. Tasinato, *Large —η— approach to single field inflation*, *Phys. Rev. D* **108** (2023) 043526 [2305.11568].

[279] H. Firouzjahi, *Revisiting loop corrections in single field ultraslow-roll inflation*, *Phys. Rev. D* **109** (2024) 043514 [2311.04080].

[280] M. Braglia and L. Pinol, *No time to derive: unraveling total time derivatives in in-in perturbation theory*, *JHEP* **08** (2024) 068 [2403.14558].

[281] R. Kawaguchi, S. Tsujikawa and Y. Yamada, *Roles of boundary and equation-of-motion terms in cosmological correlation functions*, *Phys. Lett. B* **856** (2024) 138962 [2403.16022].

[282] D. Green and K. Gupta, *Soft Metric Fluctuations During Inflation*, 2410.11973.

[283] A. Caravano, G. Franciolini and S. Renaux-Petel, *Ultra-Slow-Roll Inflation on the Lattice: Backreaction and Nonlinear Effects*, 2410.23942.

[284] V. Atal and C. Germani, *The role of non-gaussianities in Primordial Black Hole formation*, *Phys. Dark Univ.* **24** (2019) 100275 [1811.07857].

[285] M. Biagetti, G. Franciolini, A. Kehagias and A. Riotto, *Primordial Black Holes from Inflation and Quantum Diffusion*, *JCAP* **07** (2018) 032 [1804.07124].

[286] E. Tomberg, *Stochastic constant-roll inflation and primordial black holes*, *Phys. Rev. D* **108** (2023) 043502 [2304.10903].

[287] G. Franciolini, A. Iovino, Junior., V. Vaskonen and H. Veermae, *The recent gravitational wave observation by pulsar timing arrays and primordial black holes: the importance of non-gaussianities*, 2306.17149.

[288] G. Ballesteros, T. Konstandin, A. Pérez Rodríguez, M. Pierre and J. Rey, *Non-Gaussian tails without stochastic inflation*, *JCAP* **11** (2024) 013 [2406.02417].

[289] P. Bari, A. Ricciardone, N. Bartolo, D. Bertacca and S. Matarrese, *Signatures of Primordial Gravitational Waves on the Large-Scale Structure of the Universe*, *Phys. Rev. Lett.* **129** (2022) 091301 [2111.06884].

[290] P. Bari, D. Bertacca, N. Bartolo, A. Ricciardone, S. Giardiello and S. Matarrese, *An analytical study of the primordial gravitational-wave-induced contribution to the large-scale structure of the Universe*, *JCAP* **07** (2023) 034 [2209.05329].

[291] P. Bari, N. Bartolo, G. Domènech and S. Matarrese, *Gravitational waves induced by scalar-tensor mixing*, *Phys. Rev. D* **109** (2024) 023509 [2307.05404].

[292] R. Picard and K. A. Malik, *Induced gravitational waves: the effect of first order tensor perturbations*, *JCAP* **10** (2024) 010 [2311.14513].

[293] Planck collaboration, *Planck 2018 results. VI. Cosmological parameters*, *Astron. Astrophys.* **641** (2020) A6 [1807.06209].

[294] S. Borsanyi et al., *Calculation of the axion mass based on high-temperature lattice quantum chromodynamics*, *Nature* **539** (2016) 69 [1606.07494].

[295] J. R. Espinosa, D. Racco and A. Riotto, *A Cosmological Signature of the SM Higgs Instability: Gravitational Waves*, *JCAP* **09** (2018) 012 [1804.07732].

[296] K. Kohri and T. Terada, *Semianalytic calculation of gravitational wave spectrum nonlinearly induced from primordial curvature perturbations*, *Phys. Rev. D* **97** (2018) 123532 [1804.08577].

[297] K. A. Malik and D. Wands, *Cosmological perturbations*, *Phys. Rept.* **475** (2009) 1 [0809.4944].

[298] V. De Luca, G. Franciolini, A. Kehagias and A. Riotto, *On the Gauge Invariance of Cosmological Gravitational Waves*, *JCAP* **03** (2020) 014 [1911.09689].

[299] K. Inomata and T. Terada, *Gauge Independence of Induced Gravitational Waves*, *Phys. Rev. D* **101** (2020) 023523 [1912.00785].

[300] C. Yuan, Z.-C. Chen and Q.-G. Huang, *Scalar induced gravitational waves in different gauges*, *Phys. Rev. D* **101** (2020) 063018 [1912.00885].

[301] G. Domènech and M. Sasaki, *Approximate gauge independence of the induced gravitational wave spectrum*, *Phys. Rev. D* **103** (2021) 063531 [2012.14016].

[302] K. Inomata, K. Kohri, T. Nakama and T. Terada, *Enhancement of Gravitational Waves Induced by Scalar Perturbations due to a Sudden Transition from an Early Matter Era to the Radiation Era*, *Phys. Rev. D* **100** (2019) 043532 [1904.12879].

[303] K. Inomata, K. Kohri, T. Nakama and T. Terada, *Gravitational Waves Induced by Scalar Perturbations during a Gradual Transition from an Early Matter Era to the Radiation Era*, *JCAP* **10** (2019) 071 [1904.12878].

[304] S. Kumar, H. Tai and L.-T. Wang, *Towards a Complete Treatment of Scalar-induced Gravitational Waves with Early Matter Domination*, 2410.17291.

[305] M. Pearce, L. Pearce, G. White and C. Balazs, *Gravitational wave signals from early matter domination: interpolating between fast and slow transitions*, *JCAP* **06** (2024) 021 [2311.12340].

[306] K. Inomata, M. Kawasaki, K. Mukaida, T. Terada and T. T. Yanagida, *Gravitational Wave Production right after a Primordial Black Hole Evaporation*, *Phys. Rev. D* **101** (2020) 123533 [2003.10455].

[307] Z. Yi, Z.-Q. You and Y. Wu, *Model-independent reconstruction of the primordial curvature power spectrum from PTA data*, *JCAP* **01** (2024) 066 [2308.05632].

[308] F. Kuhnel and I. Stamou, *Reconstructing Primordial Black Hole Power Spectra from Gravitational Waves*, 2404.06547.

[309] X.-X. Zeng, R.-G. Cai and S.-J. Wang, *Multiple peaks in gravitational waves induced from primordial curvature perturbations with non-Gaussianity*, 2406.05034.

[310] K. Inomata, *Bound on induced gravitational waves during inflation era*, *Phys. Rev. D* **104** (2021) 123525 [2109.06192].

[311] J. Fumagalli, G. A. Palma, S. Renaux-Petel, S. Sypsas, L. T. Witkowski and C. Zenteno, *Primordial gravitational waves from excited states*, *JHEP* **03** (2022) 196 [2111.14664].

[312] C. Unal, A. Papageorgiou and I. Obata, *Axion-gauge dynamics during inflation as the origin of pulsar timing array signals and primordial black holes*, *Phys. Lett. B* **856** (2024) 138873 [2307.02322].

[313] L. Iacconi and D. J. Mulryne, *Multi-field inflation with large scalar fluctuations: non-Gaussianity and perturbativity*, *JCAP* **09** (2023) 033 [2304.14260].

[314] V. Atal and G. Domènech, *Probing non-Gaussianities with the high frequency tail of induced gravitational waves*, *JCAP* **06** (2021) 001 [2103.01056].

[315] H. V. Ragavendra, *Accounting for scalar non-Gaussianity in secondary gravitational waves*, *Phys. Rev. D* **105** (2022) 063533 [2108.04193].

[316] J. A. Ruiz and J. Rey, *Gravitational waves in ultra-slow-roll and their anisotropy at two loops*, 2410.09014.

[317] R. Inui, C. Joana, H. Motohashi, S. Pi, Y. Tada and S. Yokoyama, *Primordial black holes and induced gravitational waves from logarithmic non-Gaussianity*, 2411.07647.

[318] J. W. Armstrong, F. B. Estabrook and M. Tinto, *Time-delay interferometry for space-based gravitational wave searches*, *The Astrophysical Journal* **527** (1999) 814.

[319] T. A. Prince, M. Tinto, S. L. Larson and J. W. Armstrong, *The LISA optimal sensitivity*, *Phys. Rev. D* **66** (2002) 122002 [gr-qc/0209039].

[320] D. A. Shaddock, *Operating LISA as a Sagnac interferometer*, *Phys. Rev. D* **69** (2004) 022001 [gr-qc/0306125].

[321] D. A. Shaddock, M. Tinto, F. B. Estabrook and J. W. Armstrong, *Data combinations accounting for LISA spacecraft motion*, *Phys. Rev. D* **68** (2003) 061303 [gr-qc/0307080].

[322] M. Tinto, F. B. Estabrook and J. W. Armstrong, *Time delay interferometry with moving spacecraft arrays*, *Phys. Rev. D* **69** (2004) 082001 [gr-qc/0310017].

[323] M. Vallisneri, *Geometric time delay interferometry*, *Phys. Rev. D* **72** (2005) 042003 [gr-qc/0504145].

[324] M. Muratore, D. Vetrugno and S. Vitale, *Revisitation of time delay interferometry combinations that suppress laser noise in LISA*, *Class. Quant. Grav.* **37** (2020) 185019 [2001.11221].

[325] M. Tinto and S. V. Dhurandhar, *Time-delay interferometry*, *Living Rev. Rel.* **24** (2021) 1.

[326] M. Muratore, D. Vetrugno, S. Vitale and O. Hartwig, *Time delay interferometry combinations as instrument noise monitors for LISA*, *Phys. Rev. D* **105** (2022) 023009 [2108.02738].

[327] O. Hartwig, M. Lilley, M. Muratore and M. Pieroni, *Stochastic gravitational wave background reconstruction for a nonequilateral and unequal-noise LISA constellation*, *Phys. Rev. D* **107** (2023) 123531 [2303.15929].

[328] LISA Cosmology Working Group collaboration, *Gravitational waves from first-order phase transitions in LISA: reconstruction pipeline and physics interpretation*, *JCAP* **10** (2024) 020 [2403.03723].

[329] LISA Cosmology Working Group collaboration, *Gravitational waves from cosmic strings in LISA: reconstruction pipeline and physics interpretation*, 2405.03740.

[330] W. Martens and E. Joffre, *Trajectory Design for the ESA LISA Mission*, 2101.03040.

[331] G. Mentasti, C. R. Contaldi and M. Peloso, *Probing the galactic and extragalactic gravitational wave backgrounds with space-based interferometers*, *JCAP* **06** (2024) 055 [2312.10792].

[332] J. Kume, M. Peloso, M. Pieroni and A. Ricciardone, *Assessing the Impact of Unequal Noises and Foreground Modeling on SGWB Reconstruction with LISA*, 2410.10342.

[333] O. Hartwig and M. Muratore, *Characterization of time delay interferometry combinations for the LISA instrument noise*, *Phys. Rev. D* **105** (2022) 062006 [2111.00975].

[334] C. J. Hogan and P. L. Bender, *Estimating stochastic gravitational wave backgrounds with Sagnac calibration*, *Phys. Rev. D* **64** (2001) 062002 [astro-ph/0104266].

[335] M. R. Adams and N. J. Cornish, *Discriminating between a Stochastic Gravitational Wave Background and Instrument Noise*, *Phys. Rev. D* **82** (2010) 022002 [1002.1291].

[336] M. Muratore, O. Hartwig, D. Vetrugno, S. Vitale and W. J. Weber, *Effectiveness of null time-delay interferometry channels as instrument noise monitors in LISA*, *Phys. Rev. D* **107** (2023) 082004 [2207.02138].

[337] N. J. Cornish and J. Crowder, *LISA data analysis using MCMC methods*, *Phys. Rev. D* **72** (2005) 043005 [gr-qc/0506059].

[338] M. Vallisneri, *A LISA Data-Analysis Primer*, *Class. Quant. Grav.* **26** (2009) 094024 [0812.0751].

[339] Mock LISA Data Challenge Task Force collaboration, *The Mock LISA Data Challenges: From Challenge 3 to Challenge 4*, *Class. Quant. Grav.* **27** (2010) 084009 [0912.0548].

[340] J. Alvey, U. Bhardwaj, V. Domcke, M. Pieroni and C. Weniger, *Simulation-based inference for stochastic gravitational wave background data analysis*, *Phys. Rev. D* **109** (2024) 083008 [2309.07954].

[341] T. Robson, N. J. Cornish and C. Liu, *The construction and use of LISA sensitivity curves*, *Class. Quant. Grav.* **36** (2019) 105011 [1803.01944].

[342] M. Armano, H. Audley, G. Auger, J. T. Baird, M. Bassan, P. Binetruy et al., *Sub-femto-g free fall for space-based gravitational wave observatories: Lisa pathfinder results*, *Phys. Rev. Lett.* **116** (2016) 231101.

[343] D. Quang Nam, Y. Lemière, A. Petiteau, J.-B. Bayle, O. Hartwig, J. Martino et al., *Time-delay interferometry noise transfer functions for LISA*, *Phys. Rev. D* **108** (2023) 082004 [2211.02539].

[344] R. Schneider, V. Ferrari, S. Matarrese and S. F. Portegies Zwart, *Low-frequency gravitational waves from cosmological compact binaries*, *Monthly Notices of the Royal Astronomical Society* **324** (2001) 797 [https://academic.oup.com/mnras/article-pdf/324/4/797/4142158/324-4-797.pdf].

[345] A. J. Farmer and E. S. Phinney, *The gravitational wave background from cosmological compact binaries*, *Mon. Not. Roy. Astron. Soc.* **346** (2003) 1197 [astro-ph/0304393].

[346] T. Regimbau, *The astrophysical gravitational wave stochastic background*, *Res. Astron. Astrophys.* **11** (2011) 369 [1101.2762].

[347] F. Pozzoli, S. Babak, A. Sesana, M. Bonetti and N. Karnesis, *Computation of stochastic background from extreme-mass-ratio inspiral populations for LISA*, *Phys. Rev. D* **108** (2023) 103039 [2302.07043].

[348] S. Babak, C. Caprini, D. G. Figueroa, N. Karnesis, P. Marcoccia, G. Nardini et al., *Stochastic gravitational wave background from stellar origin binary black holes in LISA*, *JCAP* **08** (2023) 034 [2304.06368].

[349] S. Staelens and G. Nelemans, *Likelihood of white dwarf binaries to dominate the astrophysical gravitational wave background in the mHz band*, *Astron. Astrophys.* **683** (2024) A139 [2310.19448].

[350] A. Toubiana, N. Karnesis, A. Lamberts and M. C. Miller, *The interacting double white dwarf population with LISA; stochastic foreground and resolved sources*, 2403.16867.

[351] C. R. Evans, I. Iben and L. Smarr, *Degenerate dwarf binaries as promising, detectable sources of gravitational radiation*, *Astrophys. J.* **323** (1987) 129.

[352] P. Bender and D. Hils, *Confusion noise level due to galactic and extragalactic binaries*, *Class. Quant. Grav.* **14** (1997) 1439.

[353] LIGO Scientific, Virgo collaboration, *Search for the isotropic stochastic background using data from Advanced LIGO's second observing run*, *Phys. Rev.* **D100** (2019) 061101 [1903.02886].

[354] S. Nissanke, M. Vallisneri, G. Nelemans and T. A. Prince, *Gravitational-wave emission from compact Galactic binaries*, *Astrophys. J.* **758** (2012) 131 [1201.4613].

[355] M. R. Adams and N. J. Cornish, *Detecting a Stochastic Gravitational Wave Background in the presence of a Galactic Foreground and Instrument Noise*, *Phys. Rev. D* **89** (2014) 022001 [1307.4116].

[356] M. Hindmarsh, D. C. Hooper, T. Minkkinen and D. J. Weir, *Recovering a phase transition signal in simulated LISA data with a modulated galactic foreground*, 2406.04894.

[357] Q. Baghi, N. Korsakova, J. Slutsky, E. Castelli, N. Karnesis and J.-B. Bayle, *Detection and characterization of instrumental transients in LISA Pathfinder and their projection to LISA*, *Phys. Rev. D* **105** (2022) 042002 [2112.07490].

[358] T. Robson and N. J. Cornish, *Detecting Gravitational Wave Bursts with LISA in the presence of Instrumental Glitches*, *Phys. Rev. D* **99** (2019) 024019 [1811.04490].

[359] P. A. Seoane et al., *The effect of mission duration on LISA science objectives*, *Gen. Rel. Grav.* **54** (2022) 3 [2107.09665].

[360] J. Alvey, U. Bhardwaj, V. Domcke, M. Pieroni and C. Weniger, *Leveraging Time-Dependent Instrumental Noise for LISA SGWB Analysis*, 2408.00832.

[361] R. Buscicchio, A. Klein, V. Korol, F. Di Renzo, C. J. Moore, D. Gerosa et al., *A test for LISA foreground Gaussianity and stationarity. I. Galactic white-dwarf binaries*, 2410.08263.

[362] N. Karnesis, A. Sasli, R. Buscicchio and N. Stergioulas, *Characterization of non-Gaussian stochastic signals with heavier-tailed likelihoods*, 2410.14354.

[363] N. Karnesis, S. Babak, M. Pieroni, N. Cornish and T. Littenberg, *Characterization of the stochastic signal originating from compact binary populations as measured by LISA*, *Phys. Rev. D* **104** (2021) 043019 [2103.14598].

[364] S. Hofman and G. Nelemans, *On the uncertainty of the White Dwarf Astrophysical Gravitational Wave Background*, 2407.10642.

[365] G. Boileau, T. Bruel, A. Toubiana, A. Lamberts and N. Christensen, *Gravitational Wave Background from Extragalactic Double White Dwarfs for LISA*, to appear.

[366] N. Seto and K. Kyutoku, *How many extragalactic stellar mass binary black holes will be detected by space gravitational-wave interferometers?*, *Mon. Not. Roy. Astron. Soc.* **514** (2022) 4669 [2201.02766].

[367] L. Lehoucq, I. Dvorkin, R. Srinivasan, C. Pellouin and A. Lamberts, *Astrophysical Uncertainties in the Gravitational-Wave Background from Stellar-Mass Compact Binary Mergers*, 2306.09861.

[368] K. Ruiz-Rocha, K. Holley-Bockelmann, K. Jani, M. Mapelli, S. Dunham and W. Gabella, *A Sea of Black Holes: Characterizing the LISA Signature for Stellar-Origin Black Hole Binaries*, 2407.21161.

[369] A. Sesana, *Prospects for Multiband Gravitational-Wave Astronomy after GW150914*, *Phys. Rev. Lett.* **116** (2016) 231102 [1602.06951].

[370] R. Buscicchio, J. Torrado, C. Caprini, G. Nardini, N. Karnesis, M. Pieroni et al., *Stellar-mass black-hole binaries in LISA: characteristics and complementarity with current-generation interferometers*, 2410.18171.

[371] KAGRA, VIRGO, LIGO Scientific collaboration, *Population of Merging Compact Binaries Inferred Using Gravitational Waves through GWTC-3*, *Phys. Rev. X* **13** (2023) 011048 [2111.03634].

[372] KAGRA, Virgo, LIGO Scientific collaboration, *Upper limits on the isotropic gravitational-wave background from Advanced LIGO and Advanced Virgo's third observing run*, *Phys. Rev. D* **104** (2021) 022004 [2101.12130].

[373] R. E. Kass and A. E. Raftery, *Bayes Factors*, *J. Am. Statist. Assoc.* **90** (1995) 773.

[374] J. Torrado and A. Lewis, *Cobaya: Code for Bayesian Analysis of hierarchical physical models*, *JCAP* **05** (2021) 057 [2005.05290].

[375] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin et al., *JAX: composable transformations of Python+NumPy programs*, 2018.

[376] C. Germani and I. Musco, *Abundance of Primordial Black Holes Depends on the Shape of the Inflationary Power Spectrum*, *Phys. Rev. Lett.* **122** (2019) 141302 [1805.04087].

[377] I. Musco, *Threshold for primordial black holes: Dependence on the shape of the cosmological perturbations*, *Phys. Rev. D* **100** (2019) 123524 [1809.02127].

[378] A. Escrivà, C. Germani and R. K. Sheth, *Universal threshold for primordial black hole formation*, *Phys. Rev. D* **101** (2020) 044022 [1907.13311].

[379] S. Young, *The primordial black hole formation criterion re-examined: Parametrisation, timing and the choice of window function*, *Int. J. Mod. Phys. D* **29** (2019) 2030002 [1905.01230].

[380] R. Saito and J. Yokoyama, *Gravitational wave background as a probe of the primordial black hole abundance*, *Phys. Rev. Lett.* **102** (2009) 161101 [0812.4339].

[381] A. D. Gow, C. T. Byrnes, P. S. Cole and S. Young, *The power spectrum on small scales: Robust constraints and comparing PBH methodologies*, *JCAP* **02** (2021) 002 [2008.03289].

[382] M. Lewicki, P. Toczek and V. Vaskonen, *Primordial black holes from strong first-order phase transitions*, *JHEP* **09** (2023) 092 [2305.04924].

[383] R.-G. Cai, Y.-S. Hao and S.-J. Wang, *Primordial black holes and curvature perturbations from false vacuum islands*, 2404.06506.

[384] M. Lewicki, P. Toczek and V. Vaskonen, *Black holes and gravitational waves from slow phase transitions*, 2402.04158.

[385] T. Nakama, B. Carr and J. Silk, *Limits on primordial black holes from μ distortions in cosmic microwave background*, *Phys. Rev. D* **97** (2018) 043525 [1710.06945].

[386] C. Ünal, E. D. Kovetz and S. P. Patil, *Multimessenger probes of inflationary fluctuations and primordial black holes*, *Phys. Rev. D* **103** (2021) 063519 [2008.11184].

[387] C. T. Byrnes, J. Lesgourgues and D. Sharma, *Robust μ-distortion constraints on primordial supermassive black holes from non-Gaussian perturbations*, *JCAP* **09** (2024) 012 [2404.18475].

[388] A. J. Iovino, G. Perna, A. Riotto and H. Veermäe, *Curbing PBHs with PTAs*, *JCAP* **10** (2024) 050 [2406.20089].

[389] A. Hook, G. Marques-Tavares and D. Racco, *Causal gravitational waves as a probe of free streaming particles and the expansion of the Universe*, *JHEP* **02** (2021) 117 [2010.03568].

[390] D. Racco and D. Poletti, *Precision cosmology with primordial GW backgrounds in presence of astrophysical foregrounds*, *JCAP* **04** (2023) 054 [2212.06602].

[391] C. Caprini, R. Durrer, T. Konstandin and G. Servant, *General Properties of the Gravitational Wave Spectrum from Phase Transitions*, *Phys. Rev. D* **79** (2009) 083519 [0901.1661].

[392] R. Allahverdi et al., *The First Three Seconds: a Review of Possible Expansion Histories of the Early Universe*, 2006.16182.

[393] S. Allegrini, L. Del Grosso, A. J. Iovino and A. Urbano, *Is the formation of primordial black holes from single-field inflation compatible with standard cosmology?*, 2412.14049.

[394] H. Assadullahi and D. Wands, *Gravitational waves from an early matter era*, *Phys. Rev. D* **79** (2009) 083511 [0901.0989].

[395] B. Eggemeier, J. C. Niemeyer, K. Jedamzik and R. Easther, *Stochastic gravitational waves from postinflationary structure formation*, *Phys. Rev. D* **107** (2023) 043503 [2212.00425].

[396] N. Fernandez, J. W. Foster, B. Lillard and J. Shelton, *Stochastic Gravitational Waves from Early Structure Formation*, *Phys. Rev. Lett.* **133** (2024) 111002 [2312.12499].

[397] L. E. Padilla, J. C. Hidalgo, K. A. Malik and D. Mulryne, *Detecting the Stochastic Gravitational Wave Background from Primordial Black Holes in Slow-reheating Scenarios*, 2405.19271.

[398] A. Escrivà, Y. Tada and C.-M. Yoo, *Primordial Black Holes and Induced Gravitational Waves from a Smooth Crossover beyond Standard Model*, 2311.17760.

[399] A. Escrivà, R. Inui, Y. Tada and C.-M. Yoo, *The LISA forecast on a smooth crossover beyond the Standard Model through the scalar-induced gravitational waves*, 2404.12591.

[400] T. Suyama and M. Yamaguchi, *Non-Gaussianity in the modulated reheating scenario*, *Phys. Rev. D* **77** (2008) 023505 [0709.2545].

[401] C. T. Byrnes, S. Nurmi, G. Tasinato and D. Wands, *Scale dependence of local f_NL*, *JCAP* **02** (2010) 034 [0911.2780].

[402] C. T. Byrnes, M. Gerstenlauer, S. Nurmi, G. Tasinato and D. Wands, *Scale-dependent non-Gaussianity probes inflationary physics*, *JCAP* **10** (2010) 004 [1007.4277].

[403] W. J. Handley, M. P. Hobson and A. N. Lasenby, *PolyChord: nested sampling for cosmology*, *Mon. Not. Roy. Astron. Soc.* **450** (2015) L61 [1502.01856].

[404] W. J. Handley, M. P. Hobson and A. N. Lasenby, *POLYCHORD: next-generation nested sampling*, *Mon. Not. Roy. Astron. Soc.* **453** (2015) 4384 [1506.00171].

[405] H. Akaike, *A new look at the statistical model identification*, *IEEE Trans. Automatic Control* **19** (1974) 716.

[406] S. R. Geller, W. Qin, E. McDonough and D. I. Kaiser, *Primordial black holes from multifield inflation with nonminimal couplings*, *Phys. Rev. D* **106** (2022) 063535 [2205.04471].

[407] W. Qin, S. R. Geller, S. Balaji, E. McDonough and D. I. Kaiser, *Planck constraints and gravitational wave forecasts for primordial black hole dark matter seeded by multifield inflation*, *Phys. Rev. D* **108** (2023) 043508 [2303.02168].

[408] G. Autieri and M. Redi, *Reconstructing the Inflaton Potential: Primordial Black Holes and Gravitational Waves in Slow Roll and Ultra Slow Roll Single Field Inflation*, 2408.12587.

[409] N. Bartolo, D. Bertacca, V. De Luca, G. Franciolini, S. Matarrese, M. Peloso et al., *Gravitational wave anisotropies from primordial black holes*, *JCAP* **02** (2020) 028 [1909.12619].

[410] LISA Cosmology Working Group collaboration, *Probing anisotropies of the Stochastic Gravitational Wave Background with LISA*, *JCAP* **11** (2022) 009 [2201.08782].

[411] J.-P. Li, S. Wang, Z.-C. Zhao and K. Kohri, *Primordial non-Gaussianity f_{NL} and anisotropies in scalar-induced gravitational waves*, *JCAP* **10** (2023) 056 [2305.19950].

[412] P. Kidger, *On Neural Differential Equations*, Ph.D. thesis, University of Oxford, 2021.

[413] L. T. Witkowski, *SIGWfast: a python package for the computation of scalar-induced gravitational wave spectra*, 2209.05296.

[414] J. Torrado and A. Lewis, "Cobaya: Bayesian analysis in cosmology." Astrophysics Source Code Library, record ascl:1910.019, Oct., 2019.

[415] M. J. Williams, *nessai: Nested sampling with artificial intelligence*, Feb., 2021. 10.5281/zenodo.4550693.

[416] M. J. Williams, J. Veitch and C. Messenger, *Nested sampling with normalizing flows for gravitational-wave inference*, *Phys. Rev. D* **103** (2021) 103006 [2102.11056].

[417] M. J. Williams, J. Veitch and C. Messenger, *Importance nested sampling with normalising flows*, 2302.08526.

[418] A. Lewis and S. Bridle, *Cosmological parameters from CMB and other data: A Monte Carlo approach*, *Phys. Rev.* **D66** (2002) 103511 [astro-ph/0205436].

[419] A. Lewis, *Efficient sampling of fast and slow cosmological parameters*, *Phys. Rev.* **D87** (2013) 103529 [1304.4473].

[420] J. El Gammal, N. Schöneberg, J. Torrado and C. Fidler, *Fast and robust bayesian inference using gaussian processes with gpry*, *Journal of Cosmology and Astroparticle Physics* **2023** (2023) 021.

[421] J. Torrado, N. Schöneberg and J. El Gammal, *Parallelized acquisition for active learning using monte carlo sampling*, 2023.

[422] A. Lewis, *GetDist: a Python package for analysing Monte Carlo samples*, 1910.13970.