
Marginalising over Stationary Kernels with Bayesian Quadrature

Saad Hamid
University of Oxford

Sebastian Schulze
University of Oxford

Michael A. Osborne
University of Oxford

Stephen J. Roberts
University of Oxford

Abstract

Marginalising over families of Gaussian Process kernels produces flexible model classes with well-calibrated uncertainty estimates. Existing approaches require likelihood evaluations of many kernels, rendering them prohibitively expensive for larger datasets. We propose a Bayesian Quadrature scheme to make this marginalisation more efficient and thereby more practical. Through use of the maximum mean discrepancies between distributions, we define a kernel over kernels that captures invariances between Spectral Mixture (SM) Kernels. Kernel samples are selected by generalising an information-theoretic acquisition function for warped Bayesian Quadrature. We show that our framework achieves more accurate predictions with better calibrated uncertainty than state-of-the-art baselines, especially when given limited (wall-clock) time budgets.

1 INTRODUCTION

Gaussian Processes (GPs) (Rasmussen and Williams, 2006) are a rich class of models, which place probability distributions directly on classes of functions. Crucially, the success of these models is tied to the choice of their kernels. Analogously to architecture design in deep learning, kernels control the expressiveness and complexity of a GP model. If the correct kernel is chosen, GPs have shown the ability to make accurate predictions based on comparatively small training data sets. Beyond that, they natively provide predictive uncertainties with little additional computation. These confidence measurements can be vital for various downstream tasks such as decision-making in the real world.

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151. Copyright 2022 by the author(s).

If, however, the chosen kernel is misspecified, little probability mass will be assigned to the neighbourhood of the true function, resulting in a poor fit. Many commonly used kernels, such as the Radial Basis Function and Matérn kernels, have been shown to be universal kernels (Lugosi, 2006), meaning they can approximate arbitrary functions given sufficient data. Despite this they, in practice, encode strong and frequently task-inappropriate inductive biases, significantly delaying learning progress.

To increase model flexibility and avoid such pathologies practitioners typically define parameterised *kernel families*. The selection of a particular kernel is delayed until training and made data-dependent, e.g. via a maximum likelihood estimate (MLE). A fully Bayesian approach gains further expressivity and robustness by marginalising across the chosen family instead of settling on a single kernel instance for test-time predictions. This is particularly valuable when working with large, complex data sets that exhibit complicated interactions between data points.

Our approach aims to create a flexible regression model capable of representing arbitrary stationary kernels and efficiently learning complex functions from large data sets. We draw on several previously separate strands of research to achieve this.

The analysis of kernels in the spectral domain has been shown to be an effective tool in learning expressive kernels (Wilson and Adams, 2013). Due to Bochner (1959) a link between stationary kernel functions and frequency measures can be established. The former is a general class of kernels encompassing many popular families (e.g. RBF, Matérn, periodic). The link to the latter allows the construction of posterior distributions in a principled and computationally efficient way with little user intervention and opens up clear avenues for their interpretation.

Rather than settling on a specific kernel via MLE, we conduct inference in the spectral-kernel framework by marginalising the entire kernel family against parameter likelihoods. The resulting integrals are computationally intractable and solutions can only be approximated.

Previous approaches (Oliva et al., 2016; Benton et al., 2019; Simpson et al., 2021) based on Monte Carlo variants effectively average across likelihood evaluations of kernels randomly sampled from the posterior. Each likelihood evaluation requires an inversion of the kernel Gram matrix over all data points, making such approaches increasingly expensive as we scale to larger data sets.

To stay efficient even in large data, expensive likelihood settings, we adopt a model-based approach to integral approximation. Bayesian Quadrature (BQ) (O’Hagan, 1991; Rasmussen and Ghahramani, 2003) methods model the function to be integrated directly, to incorporate and exploit prior knowledge of regularities (e.g. smoothness and non-negativity of likelihood surfaces (Osborne et al., 2012a)). This enables more careful sample acquisition, yielding performance superior to MCMC methods in wall-clock time for moderate-dimensional problems (Gunter et al., 2014).

Throughout this paper, we make the following contributions. Firstly, the construction of a custom *hyper*-kernel based on the maximum mean discrepancy between distributions. This kernel is cheap to evaluate, whilst producing meaningful covariances between spectral densities, regardless of their parameterisation. Secondly, a BQ framework based on this kernel to conduct computationally efficient GP inference on large data sets. Thirdly, the derivation of an information-theoretic acquisition function for the selection of new parameter evaluations. Finally, we empirically demonstrate the viability of the resulting approach on several synthetic examples and real world data sets. Our algorithm chooses informative likelihood evaluations and achieves high performance (log-likelihoods) on limited computational budgets measured in wall-clock time.

2 BACKGROUND

2.1 Gaussian Processes

A Gaussian Process (GP) defines a probability distribution over the space of functions $f : \mathcal{X} \rightarrow \mathbb{R}$, such that for any finite subset $\mathcal{X}' \subset \mathcal{X}$, the vector $\{f(x)\}_{x \in \mathcal{X}'}$ is normally distributed. We denote such a distribution $f \sim \mathcal{GP}(m, k)$, where the mean $m(x) = E[f(x)]$ and kernel functions $k(x_1, x_2) = E[(f(x_1) - m(x_1))(f(x_2) - m(x_2))]$ encode our prior beliefs about the function. Analytic expressions for posterior mean and kernel of a process conditioned on a set of observations D under a Normal likelihood are readily available.¹

¹A detailed discussion of the GP framework can be found in Rasmussen and Williams (2006).

2.1.1 Spectral Covariance Functions

In this work the prior kernel functions $k(x_1, x_2)$ are assumed to be stationary. With slight abuse of notation, they take the form:

$$k(x_1, x_2) = k(|x_1 - x_2|) = k(\rho), \quad (1)$$

such that two function values covary only depending on their separation ρ and not their location.

Bochner’s Theorem (Bochner, 1959) states that $k(\rho)$ is a positive-definite function on \mathbb{R} (and valid kernel) if and only if its Fourier transform, $S(\omega)$, is a positive spectral density, i.e.:

$$k(\rho) = \int e^{2\pi i \rho \omega} dS(\omega). \quad (2)$$

As a consequence, any parameterisation of a set of spectral measures induces a corresponding parameterisation of stationary kernels.²

2.1.2 Kernels based on distance metrics

Distance-based kernels which often take the form $k(x_1, x_2) = \lambda^2 \exp(-\frac{d(x_1, x_2)^q}{l^2})$ (λ , and l being hyper-parameters), are valid, i.e. positive definite, for $q = 1$ if and only if the underlying metric $d(x_1, x_2)$ is conditionally negative definite (Jayasumana et al., 2015; Feragen et al., 2015), meaning that it must give rise to matrices, D , such that $c^T D c < 0, \forall c : \sum_i c_i = 0$. For $q = 2$ the underlying metric must be Hilbertian, meaning that there must be an isometry between the metric space defining d and a Hilbert space (Feragen et al., 2015).

2.2 Bayesian Quadrature

Bayesian inference in machine learning frequently involves computation of intractable integrals of the form:

$$Z = \int f(x)p(x)dx \quad , \quad (3)$$

where $p(x)$ is a known prior density and $f(x)$ a (likelihood-)function. Bayesian quadrature is a model-based approach to approximately evaluating such integrals by modelling f as a GP. Since GPs are closed under affine transforms the posterior over Z is then Gaussian. For suitable kernel-prior combinations the quadrature weights can be evaluated analytically.

While the maintenance of the GP model requires some computational effort, it enables the generalisation of sample information across the integration domain. Consequently, samples can be chosen in a targeted fashion

²For $k(\rho)$ to be real, we only consider symmetric $S(\omega)$.

and BQ has been shown to be a competitive, more evaluation-efficient alternative to Monte Carlo methods, particularly in moderate-dimensional domains.

Recent work (Osborne et al., 2012a; Gunter et al., 2014; Chai and Garnett, 2019) further increases sample efficiency by encoding model-constraints such as the positivity of likelihood functions using warped GPs. This work builds upon WSABI (Gunter et al., 2014) in particular, which places a GP prior over the square-root of f and is described in more detail in Appendix A.

3 OUR METHOD: MASKERADE

In the following we present a flexible, data efficient Gaussian Process framework and show how to conduct Bayesian inference within it.

3.1 The generative model

Without loss of generality, we assume input dimensions of the training data to be rescaled to the interval $[0, 1]$ and outputs to be normalised to have zero mean and unit variance.

By this construction, a zero mean prior for our GP model is the most logical choice. As mentioned above, kernels are only assumed to be stationary. Following Wilson and Adams (2013) and Oliva et al. (2016), we parameterise spectral densities through Gaussian Mixture Models (GMMs) to induce a corresponding parameterisation of stationary kernels. In the limit of infinitely many components GMMs can approximate any spectral density to arbitrary precision.

The kernel, matching a spectral density parameterised by θ , can be recovered via (2). To create only real-valued kernels, we reflect all mixtures at the origin. We also note, that an additional output-scale $k(0)$ is required to cover the space of (unnormalised) measures and recover arbitrary stationary kernels. Since data has unit variance, however, we omit this in the following.

To perform Bayesian inference and marginalise across stationary kernels, we define a hyperprior $p(\theta)$ (see next section). A graphical representation of the resulting generative model is shown in Figure 1.

3.1.1 The Hyperprior $p(\theta)$

The definition of our Bayesian model necessitates the choice of a suitable hyperprior. We see this is an advantage to the practitioner, as they are able to include existing knowledge to reduce the space of kernels needing to be explored. In the absence of such information, we motivate a set of heuristics based on general properties of the training data.

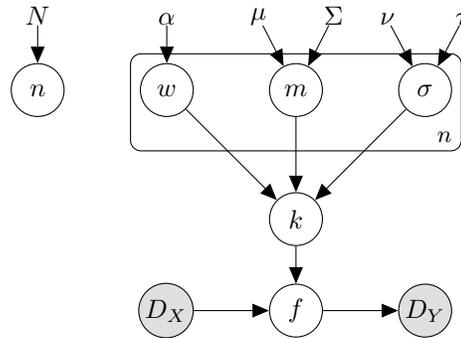


Figure 1: Bayesian Network of our model. The number of mixture components n is drawn from a uniform prior; α parameterises a Dirichlet distribution over mixture weights w ; μ and Σ parameterise Gaussians over mixture means; ν and τ parameterise Log-Normal distributions over mixture scales. Not shown are hyperparameters of the hyper-kernel, λ and l .

The prior over the number of mixture components n is uniform up to some maximum N , limiting the size of the overall parameter space. The n components are weighted against one another through a Dirichlet prior.

Mixture means are drawn from a normal distribution with zero mean and standard deviation $F_s/5$ so that negligible mass lies above the Nyquist frequency F_s of the data. Recall that the Nyquist frequency is the highest frequency directly identifiable from a data set.³

Mixture scales can be seen as approximate inverse lengthscales of the kernel. Accordingly their log-normal prior is set such that most of its mass lies between the inverse of the maximum distance between data points and the Nyquist frequency with a mean of $\log(1/|D_x|)$ and a scale of $\log(\sqrt{2})$ (where $|D_x|$ is the size of the data window).

We stress that the prior outlined above is, in our view, the widest reasonable prior, and that performance is likely to be improved considerably by specifying narrower priors, if possible, especially for high dimensional datasets.

3.2 Posterior Inference

Since we do not know the full parameterisation of our GP model, we marginalise out the unknown kernel parameters. Accordingly, the predictive distribution at some location x_* under the above model, after observ-

³For unevenly-spaced data we choose F_s as the Nyquist frequency of a fictitious dataset with sampling frequency equal to the inverted average distance between datapoints.

ing data $D = \{x_D, y_D\}$, is given by:

$$\begin{aligned} p(y_* | x_*, D) &= \int p(y_* | x_*, \theta, D) p(\theta | D) d\theta \\ &= \int p(y_* | x_*, \theta, D) \frac{p(D | \theta) p(\theta)}{p(D)} d\theta \\ &= \frac{\int p(y_* | x_*, \theta, D) p(D | \theta) p(\theta) d\theta}{\int p(D | \theta) p(\theta) d\theta}, \end{aligned} \quad (4)$$

where $\theta = (n, \{w_i, m_i, \sigma_i\}_{i=1}^n)$ is the parameterisation of a spectral GMM with n components, each described by a weight w_i , mean m_i and scale σ_i . The predictive distributions $p(y_* | x_*, \theta, D)$ and likelihoods $p(D | \theta)$ for the spectral kernel corresponding to a particular θ are given by the GP framework.

Unfortunately, both integrals in (4) are intractable and require approximation. We note that a major cost here is the evaluation of Gaussian Process likelihoods. These involve an inversion of the kernel Gram matrix of the entire data set whose computational cost increases quickly with data set size (typically cubically). To reduce the number of likelihood evaluations required, we propose to adopt a Bayesian Quadrature approach.

3.3 Bayesian Quadrature

We begin by constructing GP models of the functions $p(y_* | x_*, \theta, D) p(D | \theta)$ and $p(D | \theta)$ that are to be integrated against the prior. Since the locations x^* of test points are unknown at training time, inference separates into two steps.

During training, a ‘‘pseudo-dataset’’ \mathfrak{D} of decompositions and matching likelihood evaluations (and model evidences respectively) is compiled through active sampling (see Section 3.3.2) to shape our surrogate models and improve approximation quality. Hereby, we follow Gunter et al. (2014) and enforce positivity of the likelihood surrogate by placing a GP prior over the square-root of the likelihood function $L(\theta) = P(D | \theta)$:

$$z = \sqrt{2(L(\theta) - \epsilon)} \sim \mathcal{GP}(0, \kappa), \quad (5)$$

where $\kappa(\theta_1, \theta_2)$ is a kernel between spectral decompositions (see section 3.3.1) and ϵ is a small non-negative constant.

Under this model $L(\theta)$ is not a GP, but can be approximated with one using a linear transformation of the surrogate. The posterior mean and covariance functions $\mathbf{m}_{\mathfrak{D}}(\theta)$ and $\mathfrak{K}_{\mathfrak{D}}(\theta_1, \theta_2)$ then remain functions of the kernel κ and the likelihood observations. To ensure a good fit to the actual likelihood function, hyperparameters are re-optimised after each evaluation. Since GPs are closed under continuous bounded linear transformations (Bogachev, 1961), we obtain a Gaussian

posterior over the model evidence – the integral in the denominator of (4). The mean of this Gaussian takes the form $z^T Q z$, where Q are the BQ weights – a full derivation is given in Appendix A.

To make a prediction, we compute the integral in the numerator of (4) in the same fashion. $p(y_* | x_*, \theta, D)$ is a GP posterior. Since we employ the same kernel κ to measure similarity between spectral densities, and we reuse the same likelihood evaluations from the training stage, the quadrature weights are identical and only z differs, allowing for efficient computation of the desired quantities.

3.3.1 The Hyper-Kernel

Noting that the densities under consideration are GMMs with a limited number of components, it becomes apparent that a suitable *hyper*-kernel, κ , has to overcome two obstacles. Firstly, it needs to cope with varying numbers of parameters, or otherwise assign a predefined covariance between GMMs with differing numbers of components. Secondly, it should capture a number of invariances in the parameterisation of GMMs (e.g. reordering of components, subdivision or recombination of components etc.). A naïve approach for the construction of such a kernel would base distances between densities on the Euclidean distance between their parameters and fails to achieve either desiderata.

Instead we construct a kernel based on the distance between the spectral densities represented by those parameters. Accordingly, we seek a conditionally-negative definite or Hilbertian metric between distributions. The maximum mean discrepancy meets our criteria, and is cheap to compute.

We obtain the following kernel:

$$\kappa(\theta_1, \theta_2) = \lambda^2 \exp\left(-\frac{d(\theta_1, \theta_2)^q}{l^2}\right), \quad (6)$$

where $\lambda > 0$ is an output-scale, $l > 0$ is a length-scale parameter, $q = \{1, 2\}$ and d is a maximum mean discrepancy between the spectral densities (GMMs) parameterised by θ_1 and θ_2 . The exact form of the MMD depends on the underlying kernel. Our experiments use the Energy Distance MMD (Feydy, 2020), given by

$$d(\theta_1, \theta_2) = w_1^T M_{12} w_2 - \frac{1}{2} w_1^T M_{11} w_1 - \frac{1}{2} w_2^T M_{22} w_2 \quad (7)$$

where w_1 and w_2 are vectors of component weights and M_{12} is a matrix of euclidean distances between all pairs of component parameters $\{(m_i, \sigma_i)\}_{i=1}^{n_1}$ and $\{(m_i, \sigma_i)\}_{i=1}^{n_2}$, but our framework allows for alternatives (e.g. the Gaussian MMD, which can also be computed analytically for GMMs). Note that, since the MMD

is analytically computed (rather than approximated through sampling as is usual in the MMD literature), it is a valid distance between unnormalised densities. The hyper-kernel therefore remains valid when kernels with arbitrary, non-unit output-scales are considered.

The only drawback is that integration of this kernel against the prior distribution is no longer possible analytically, so Monte Carlo integration is necessary. Arguably, however, the integration of the kernel against the prior lends itself better to such an approximation than that of the data likelihood. Intuitively, the likelihood function will have many sharp peaks for large datasets requiring samples to be drawn at very specific (unknown) locations for an accurate estimate. The kernel function by comparison is smoother, cheaper to evaluate and the resulting integral easier to approximate via random sampling from the prior. Our empirical results justify this approach, showing that we outperform competing methods despite the requirement for sampling.

3.3.2 Information-theoretic point acquisition

The GP surrogate is trained based on a set of sampled spectral distributions, $\mathfrak{D} = \{(\theta_i, L_i)\}$, and the likelihood values of their corresponding kernels. Since adding a new point θ^{new} to this set requires a computationally expensive likelihood evaluation, it should be chosen to maximise its informativeness w.r.t. the quantity we care about – the predictive distribution at test locations x_* .

As we do not expect the model to have access to the test locations at training time, we choose a different acquisition criterion for the new point. Instead, we aim to improve the approximation of the denominator of (4) – and thereby the fit of the posterior distribution over spectral mixtures – at training time. In Gunter et al. (2014) an acquisition function based on the uncertainty of the integrand is proposed. We extend the work of Gessner et al. (2020) and Ru et al. (2018) to propose an information-theoretic criterion instead.

We observe that the expected reduction in the entropy of our integral estimate after making an observation at sample location θ^s is given by

$$\alpha(\theta_*) = H(L_* | \mathfrak{D}, \theta_*) - E_{p(Z|\mathfrak{D}, \theta_*)}[H(L_* | \mathfrak{D}, \theta_*, Z)]. \quad (8)$$

Here L_* is the predicted observation at location θ_* and Z is the integral in the denominator of (4). Within the GP framework, the entropy of the predictive distribution $H(L_* | \mathfrak{D}, \theta_*)$ is the entropy of a Normal distribution and can be calculated in closed form. The second term turns out to be the expectation of a constant and the predictive distribution conditioned on Z can once more be computed in closed form. Gessner

et al. (2020) discuss this acquisition function in the context of ordinary and multi-source Bayesian Quadrature. We note that it is available for warped Bayesian Quadrature. In particular, the acquisition function can be computed analytically when the linearisation approximation is used, provided that the kernel integrals are analytic. This is because the composition of the approximation and integration remains an affine transformation. Therefore, the integral of the approximate warped surrogate is jointly Gaussian distributed with the unwarped surrogate.

The acquisition function (8) is also valid for selecting batches of a fixed size. Importantly, our combination of hyper-kernel and acquisition function allows for the selection of (batches of) GMMs that are maximally informative about the overall evidence, $\int p(D | \theta)p(\theta)d\theta$, rather than selecting GMMs for each parameter domain (corresponding to a fixed number of components) separately. This is because the kernel is able to define covariances between GMMs with different numbers of mixture components. Incorporating such information represents an improvement over Chai et al. (2019) whose framework, in this instance, would be naïve as it assumes that the GPs over each domain are independent.

3.4 Computational Complexity

During learning, each iteration requires optimising the hyperparameters of the surrogate GP and the acquisition function. The time complexity of hyperparameter optimisation is dominated by the cost of making likelihood evaluations, $\mathcal{O}(h^3)$ where h is the number of Spectral Mixture kernels evaluated thus far. Acquisition function optimisation incurs an initial cost of $\mathcal{O}(h^3 + mh^2 + m^2)$, where m is the number of Monte Carlo samples used to approximate the kernel integrals. Subsequent evaluations require $\mathcal{O}(bh^2)$ operations, where b is the batch size. The memory complexity during hyperparameter optimisation is dominated by the cost of storing the hyper-kernel evaluated between observed SM kernels $\mathcal{O}(h^2)$. During the initialisation of the acquisition function, $\mathcal{O}(m^2)$ memory is additionally required to store the hyper-kernel evaluated between the MC samples.

Once the evaluation budget is exhausted the quadrature weights can be computed in $\mathcal{O}(h^3 + mh^2 + m^2)$ time and $\mathcal{O}(h^2 + m^2)$ memory. Inference takes $\mathcal{O}(h|D|^3)$ operations, though this can be reduced significantly by disregarding terms that have low quadrature weight. The memory complexity of inference is $\mathcal{O}(h^2|D|^2)$ for a naïve implementation, and can be reduced to $\mathcal{O}(h|D|^2)$ by looping appropriately (note that this leaves the asymptotic time complexity unchanged).

4 RELATED WORK

4.1 Kernel learning

Early work in kernel learning focussed on searching over compositions (sums and products) of a set of basic covariance functions (Duvenaud et al., 2013). More recent work has proposed learning kernels using their spectral representations (Wilson and Adams, 2013; Kom Samo, 2017; Lázaro-Gredilla et al., 2010; Gal and Turner, 2015; Kom Samo and Roberts, 2015; Oliva et al., 2016; Remes et al., 2017; Jang et al., 2017; Ambrogioni and Maris, 2018; Tobar, 2018; Teng et al., 2019; Benton et al., 2019). Beyond the ability to capture more complex structures, this approach also lends itself to approximate inference, reducing learning time complexity (Gal and Turner, 2015; Rahimi and Recht, 2008; Hensman et al., 2018). A model similar to ours uses a Dirichlet Process prior over the number of components in a Gaussian Mixture Model (GMM) residing in the spectral domain (Oliva et al., 2016). Simpson et al. (2021) use Nested Sampling to marginalise over the parameters of a Spectral Mixture kernel with a fixed number of mixture components.

Inference in spectral models reduces to the marginalisation of the likelihood function against the posterior over kernels (or, equivalently, their spectral decompositions). The resulting integrals are intractable and have to be approximated – usually through the use of an appropriate MCMC scheme, such as Gibbs sampling, RJ-MCMC (Green, 1995) or elliptical slice sampling (Murray et al., 2010). Scaling MCMC marginalisation over the spectral kernel parameters up to large datasets is problematic. The computational cost of likelihood evaluations scales poorly in the size of the training data set and the collection of sufficiently many samples for an accurate Monte-Carlo estimate becomes prohibitively expensive. As above, BQ may be a more appropriate choice in these settings.

The recently proposed baselines we will compare against in our experiments include:

Variational Sparse Spectrum Gaussian Process

Gal and Turner (2015) which performs Variational Inference over a sparse spectrum approximation.

Bayesian Nonparametric Kernel Learning

Oliva et al. (2016) which learns the posterior distribution over Spectral Mixture kernels via a parameterisation of spectral densities based on GMMs and a Gibbs sampling scheme.

Functional Kernel Learning Benton et al. (2019) which places a GP prior over the kernel’s spectral density and infers a posterior over kernels using an Elliptical Slice Sampling scheme.

Various approximations to reduce the cost of likelihood evaluations have been proposed in the literature, such as PCG (Cutajar et al., 2016) and BBMM (Gardner et al., 2018) – modified conjugate gradient methods – and Random Fourier Features (Rahimi and Recht, 2008). This is orthogonal to our investigation and our framework is optionally capable of leveraging both of these for scalability.

4.2 Bayesian Quadrature

Recent work in BQ (Osborne et al., 2012a; Gunter et al., 2014; Chai and Garnett, 2019) has proposed the use of warped GPs (Snelson et al., 2004) to incorporate a priori known model-constraints into the surrogate. This work builds upon WSABI (Gunter et al., 2014) in particular, which places a GP prior over the square-root of the integrand. The acquisition function developed in Section 3.3.2 also applies to this setting and is an extension of the information-theoretic scheme discussed in Gessner et al. (2020). Osborne et al. (2012b) also proposed a BQ framework to infer ratios of integrals, which is relevant to the computation of marginalised posteriors. However, the combination with WSABI and our framework introduces further intractable integral terms, which must be approximated with Monte Carlo sampling and would prohibitively raise the cost of inference. Additionally, the work of Xi et al. (2018) and Chai et al. (2019) is related as it concerns the use of BQ for computing correlated integrals, and for automatic model selection.

4.3 Gaussian Processes on Spaces of Measures

The BQ integrand model requires the specification of a positive-definite kernel across the integration domain – the space of spectral densities. While numerous metrics and divergences have been proposed to measure differences between probability distributions, many do not give rise to valid kernels.

Recent work has used Wasserstein distances (Bachoc et al., 2018) to define Gaussian Processes on a space of measures. However, Wasserstein distances are only Hilbertian in 1D (Peyré and Cuturi, 2019) so extensions to multidimensional distributions rely on Hilbert space embeddings of optimal transport maps to a reference distribution (Bachoc et al., 2019). Furthermore, evaluating Wasserstein distances can be expensive as it involves finding the solution to an optimal transport problem. The Independence kernel based on the Sinkhorn distance (Cuturi, 2013) has been shown to be a valid positive definite kernel, but the distance between identical distributions may not be zero. An elegant alternative that satisfies our desiderata are maximum mean discrepancies (MMDs), which are defined

in terms of kernel mean embeddings. MMDs are Hilbertian, so are a valid metric for kernels of the form (6) (Muandet et al., 2017).

5 RESULTS

In the following we empirically assess the performance of our model on a variety of tasks.

5.1 Experiment setup

The experiments were conducted on a single Nvidia Titan V GPU. In all cases data is preprocessed so that the training targets have unit variance, and the training inputs lie in the range $[0, 1]$ in all dimensions. We then report the Root Mean Squared Errors (RMSE) of the mean function of the predictive posterior as well as the Log-Likelihoods (LL) of the predictive posterior on held out test data after training each model for a fixed training time budget. Since the cost of likelihood evaluations depends on the dataset size, so does the training budget. The numbers given indicate either mean performance or mean performance and standard deviation.

We compare our approach of MARGINALISING SPECTRAL KERNELS AS DENSITIES (MASKERADE)⁴ to models previously proposed in the literature including VSSGP (Gal and Turner, 2015), BaNK (Oliva et al., 2016), and FKL (Benton et al., 2019). Results for the first two algorithms are directly taken from the papers. FKL hyperparameters were set to the defaults for the `spectralgp`⁵ Python package, with the ω_{max} (maximum frequency) argument modified to match the Nyquist frequency of the dataset.

We limit the parameter space MASKERADE considers to GMMs with up to 5 components ($N=5$). The concentration of the Dirichlet prior over weights is set to $\alpha = 1$. The remaining priors are chosen as described in section 3.1.1. We initialise the BQ surrogate GPs with likelihood evaluations of parameters randomly chosen from the prior and thereafter acquire additional evaluations in batches of size 20. LBFGS is used to optimise both the acquisition function as well as the parameters of the hyper-kernel (which happens every 5 batches).

5.2 Medium scale data sets

We begin by fitting MASKERADE to the Solar dataset (Lean, 2004) to compare against VSSGP. We replicate the setup from Gal and Turner (2015), withholding 5 sets of length 20 as the test set.

Table 1: Test set RMSE for the Solar Irradiance Data Set. Results for VSSGP are taken from Gal and Turner (2015).

	VSSGP	MASKERADE
RMSE	0.41	0.17

We further examine the algorithms’ performance on four datasets from the UCI Machine Learning Repository (Dua and Graff, 2017): Yacht Hydrodynamics (308 instances, 6 input dimensions) (Lopez, 2013), Auto MPG (398 instances, 8 input dimensions) (Quinlan, 1993), Concrete Compressive Strength (1030 instances, 8 input dimensions) (Yeh, 2007), and Airfoil Self-Noise (1503 instances, 5 input dimensions) (Lopez, 2014).

To compare against FKL we follow Benton et al. (2019) and average over 10 runs, randomly partitioning the data into 90/10 train/test sets every time. These results are presented in Table 2.

Table 2: Posterior log likelihood of and RMSE on the test set for UCI MLR datasets. Results for FKL taken from Benton et al. (2019).

DATASET	LL	
	FKL	MASKERADE
YACHT	15.703 \pm 8.233	45.324 \pm 12.731
AUTOMPG	-98.942 \pm 6.135	-52.168 \pm 1.172
CONCRETE	-384.242 \pm 140.779	-145.41 \pm 55.84
AIRFOIL	-270.073 \pm 28.424	-135.09 \pm 15.76
DATASET	RMSE	
	FKL	MASKERADE
YACHT	0.193 \pm 0.13	0.172 \pm 0.094
AUTOMPG	2.838 \pm 0.374	1.058 \pm 0.057
CONCRETE	3.781 \pm 0.501	0.719 \pm 0.145
AIRFOIL	1.378 \pm 0.176	0.565 \pm 0.104

Similarly, to compare against BaNK we follow Oliva et al. (2016) and perform 3 repeats of 5-fold cross-validation on the Concrete Compressive Strength and Airfoil Self-Noise datasets, for which we present results in Table 3.

Table 3: Test set MSE on two UCI data sets. Results for BaNK taken from Oliva et al. (2016).

DATASET	BANK	MASKERADE
CONCRETE	0.1195 \pm 0.0108	0.6889 \pm 0.0168
AIRFOIL	0.3359 \pm 0.0354	0.3244 \pm 0.0214

⁴<https://github.com/saadhamidml/maskerade>

⁵<https://github.com/wjmaddox/spectralgp>

5.3 Large scale data sets

Finally we compare MASKERADE against FKL on two large datasets, with a limited time budget of 20 minutes for learning. We average over 10 runs with a 90/10 train/test split and report results in Table 4. MASKERADE once more achieves superior performance to FKL on the same time budget. The datasets are the Sterling Broad Based Exchange Rate Bank of England (2020) (data from 1990-01-02–2020-10-08, totalling 7781 data points) and the UCI 3D Road Network Kaul (2013) (10,000 data point subset of 250,000 data points. We choose the first 10,000 instances with a unique OSM-ID) for which the task is to predict altitude from latitude and longitude.

Table 4: Posterior log likelihood of and RMSE on the test set for FKL and MASKERADE fit to two large datasets. Both methods were given a time budget of 20 minutes for training.

DATASET	LL	
	FKL	MASKERADE
STERLING	-476.98 ± 342.94	162.34 ± 20.952
ROAD ALT	-3195.9 ± 68.029	-1102.4 ± 45.0945

DATASET	RMSE	
	FKL	MASKERADE
STERLING	1.381 ± 1.011	1.132 ± 0.057
ROAD ALT	5.809 ± 0.6478	0.729 ± 0.065

5.4 Ablation Study

We conduct an ablation study to verify the effectiveness of two components of our model in Table 5. Firstly, we verify that marginalising across the kernel family does indeed lead to better predictive performance. As a baseline, we compare against a GP using Spectral Mixture kernel (SM) whose hyperparameters were optimised via gradient descent to maximise the likelihood of the training data. Secondly, we compare our proposed acquisition function against uncertainty sampling (Gunter et al., 2014) and random point acquisition. We set an evaluation budget of 1000 and evaluate performance on the Airline Passenger Makridakis et al. (1998) and Mauna Loa Atmospheric CO₂ Concentration NOAA, USA (2020) datasets. We average over 10 runs with a 90/10 train/test split. In both cases, our chosen approach compares favourably to alternative designs.

Table 5: Test set RMSE for the Airline Passenger and Mauna Loa datasets. MASKERADE-I indicates the use of our proposed information theoretic acquisition function, MASKERADE-U indicates uncertainty sampling Gunter et al. (2014), and MASKERADE-R indicates random sampling under the prior.

DATASET	SM	MASKERADE-I
AIRPASS	0.283 ± 0.000	0.236 ± 0.044
MAUNA LOA	0.802 ± 0.685	0.243 ± 0.000

	MASKERADE-U	MASKERADE-R
	0.274 ± 0.104	0.280 ± 0.140
	0.246 ± 0.126	0.360 ± 0.121

6 DISCUSSION

We have introduced a novel framework for kernel learning for Gaussian Processes that seeks to marginalise over Spectral Mixture Kernels using Bayesian Quadrature. Specifically we use a maximum mean discrepancy as a metric underlying an exponential kernel to define a Gaussian Process on a space of GMMs, and show how this can be efficiently computed. This elegantly incorporates invariances between Spectral Mixture Kernels into our model. Additionally we show that an information-theoretic acquisition function is applicable for warped Bayesian Quadrature, and how to use it within our framework. We empirically evaluate our method on several datasets and find that it is competitive with state-of-the-art baselines.

The key limitation of our method is due to the curse of dimensionality. The most scalable instantiation of our framework handles multi-dimensionality by using GMMs in which each Gaussian has a diagonal covariance structure. This means that the number of parameters for an SM kernel with M mixtures is $M(1 + 2D)$, which grows faster than the number of data dimensions, D . This limits the effectiveness of any Bayesian Quadrature regime. Note that Monte Carlo methods also suffer in high dimensional spaces, especially if likelihood evaluations are expensive. (Our method relies on MC integration of kernel integrals, but these are considerably easier to compute than the marginalisation integrals because our method only requires samples from the priors over hyperparameters rather than the posteriors, and the kernel is much cheaper to evaluate than the likelihood.) Our approach is most beneficial for inference based on large, low dimensional datasets.

The societal impacts of this work will depend on the problems that practitioners apply it to, as it is a general-purpose framework.

Acknowledgements

The authors would like to thank anonymous reviewers for their detailed and constructive feedback. S.H. acknowledges funding from EPSRC, and S.S. is supported by an I-CASE studentship funded by EPSRC and Dyson.

References

- Luca Ambrogioni and Eric Maris. Integral Transforms from Finite Data: An Application of Gaussian Process Regression to Fourier Analysis. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics*, pages 217–225, 2018.
- Francois Bachoc, Alexandra Suvorikova, David Ginsbourger, Jean-Michel Loubes, and Vladimir Spokoiny. Gaussian processes with multidimensional distribution inputs via optimal transport and Hilbertian embedding. In *arXiv:1805.00753 [stat]*, 2019.
- François Bachoc, Fabrice Gamboa, Jean-Michel Loubes, and Nil Venet. A Gaussian Process Regression Model for Distribution Inputs. *IEEE Transactions on Information Theory*, 64(10):6620–6637, 2018.
- Bank of England. Broad Effective Exchange Rate Index, Sterling, 2020. URL <https://www.quandl.com/data/BOE/XUDLBK82-Broad-Effective-Exchange-Rate-Index-Sterling-jan-2005-100>.
- Gregory W. Benton, Wesley J. Maddox, Jayson P. Salkey, Julio Albinati, and Andrew Gordon Wilson. Function-Space Distributions over Kernels. In *Advanced in Neural Information Processing Systems*, 2019.
- Salomon Bochner. *Lectures on Fourier Integrals*, volume 42. Princeton University Press, 1959.
- Vladimir Igorevich Bogachev. *Gaussian Measures*. 1961.
- Henry Chai and Roman Garnett. Improving Quadrature for Constrained Integrands. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 2019.
- Henry Chai, Jean-Francois Ton, Michael A Osborne, and Roman Garnett. Automated Model Selection with Bayesian Quadrature. In *36th International Conference on Machine Learning*, pages 1555–1564, 2019.
- Kurt Cutajar, Michael Osborne, John Cunningham, and Maurizio Filippone. Preconditioning Kernel Matrices. In *International Conference on Machine Learning*, pages 2529–2538. PMLR, 2016.
- Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300, 2013.
- Dheeru Dua and Casey Graff. UCI Machine Learning Repository, 2017. URL <http://archive.ics.uci.edu/ml>.
- David Duvenaud, James Robert Lloyd, Roger Grosse, Joshua B. Tenenbaum, and Zoubin Ghahramani. Structure Discovery in Nonparametric Regression through Compositional Kernel Search. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.
- Aasa Feragen, Francois Lauze, and Søren Hauberg. Geodesic Exponential Kernels: When Curvature and Linearity Conflict. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3032–3042, 2015.
- Jean Feydy. Geometric data analysis, beyond convolutions. *PhD thesis, Paris-Saclay University*, page 94, 2020.
- Yarin Gal and Richard Turner. Improving the Gaussian process sparse spectrum approximation by representing uncertainty in frequency inputs. In *32nd International Conference on Machine Learning*, volume 1, pages 655–664, 2015.
- Jacob R. Gardner, Geoff Pleiss, David Bindel, Kilian Q. Weinberger, and Andrew Gordon Wilson. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. In *Advances in Neural Information Processing Systems*, 2018.
- Alexandra Gessner, Javier Gonzalez, and Maren Mahsereci. Active multi-information source bayesian quadrature. In Ryan P. Adams and Vibhav Gogate, editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 712–721, Tel Aviv, Israel, 22–25 Jul 2020. PMLR.
- Peter J Green. Reversible jump markov chain monte carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.
- Tom Gunter, Michael A. Osborne, Roman Garnett, Philipp Hennig, and Stephen J. Roberts. Sampling for Inference in Probabilistic Models with Fast Bayesian Quadrature. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, volume 4, pages 2789–2797, 2014.
- James Hensman, Nicolas Durrande, and Arno Solin. Variational Fourier features for Gaussian processes. *Journal of Machine Learning Research*, 18:1–52, 2018.
- Phillip A Jang, Andrew Loeb, Matthew Davidow, and Andrew G Wilson. Scalable Levy Process Priors for Spectral Kernel Learning. In *Advanced in Neural Information Processing Systems*, pages 3941–3950, 2017.

- Sadeep Jayasumana, Richard Hartley, Mathieu Salzmann, Hongdong Li, and Mehrtaash Harandi. Kernel Methods on Riemannian Manifolds with Gaussian RBF Kernels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(12):2464–2477, 2015.
- Manohar Kaul. 3D Road Network (North Jutland, Denmark) Data Set, 2013. URL [https://archive.ics.uci.edu/ml/datasets/3D+Road+Network+\(North+Jutland,+Denmark\)#](https://archive.ics.uci.edu/ml/datasets/3D+Road+Network+(North+Jutland,+Denmark)#).
- Yves-Laurent Kom Samo. *Advances in Kernel Methods*. PhD thesis, University of Oxford, 2017.
- Yves-Laurent Kom Samo and Stephen Roberts. Generalized Spectral Kernels. In *arXiv:1506.02236 [stat]*, 2015.
- J. Lean. Solar Irradiance Reconstruction. IGBP PAGES/World Data Center for Paleoclimatology, Data Contribution Series # 2004-035. NOAA/NGDC Paleoclimatology Program, Boulder CO, USA. 2004. URL ftp://ftp.ncdc.noaa.gov/pub/data/paleo/climate_forcing/solar_variability/lean2000_irradiance.txt.
- Roberto Lopez. Yacht Hydrodynamics Data Set, 2013. URL <https://archive.ics.uci.edu/ml/datasets/yacht+hydrodynamics>.
- Roberto Lopez. Airfoil Self-Noise Data Set, 2014. URL <https://archive.ics.uci.edu/ml/datasets/airfoil+self-noise>.
- Gabor Lugosi. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.
- Miguel Lázaro-Gredilla, Joaquin Quiñonero-Candela, Carl Edward Rasmussen, and Aníbal R Figueiras-Vidal. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11: 1–17, 2010.
- S. Makridakis, S. Wheelwright, and R. Hyndman. Airline Passenger Data Set, 1998. URL <https://github.com/robjhyndman/fma/blob/master/data/airpass.rda>.
- Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel Mean Embedding of Distributions: A Review and Beyond. *Foundations and Trends in Machine Learning*, 10(1): 1–141, 2017.
- Iain Murray, Ryan Prescott Adams, and David J C MacKay. Elliptical slice sampling. *Journal of Machine Learning Research*, 9:541–548, 2010.
- NOAA, USA. Mauna Loa Atmospheric CO₂ Observations, 2020. URL ftp://afftp.cmdl.noaa.gov/ccg/co2/trends/co2_mm_gl.csv.
- A. O’Hagan. Bayes–Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245–260, 1991.
- Junier B Oliva, Avinava Dubey, Andrew G Wilson, Barnabas Poczos, Jeff Schneider, and Eric P Xing. Bayesian nonparametric kernel-learning. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pages 1078–1086, 2016.
- Michael A Osborne, David Duvenaud, Roman Garnett, Carl E Rasmussen, Stephen J Roberts, and Zoubin Ghahramani. Active learning of model evidence using Bayesian quadrature. In *Advances in Neural Information Processing Systems 26*, pages 46–54, 2012a.
- Michael A Osborne, Roman Garnett, Stephen J Roberts, Christopher Hart, Suzanne Aigrain, and Neale P Gibson. Bayesian quadrature for ratios. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics*, pages 832–840, 2012b.
- Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. *Foundations and Trends in Machine Learning*, 11(5):1–257, 2019.
- Ross Quinlan. Auto MPG Data Set, 1993. URL <https://archive.ics.uci.edu/ml/datasets/auto+mpg>.
- Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pages 1177–1184. Curran Associates, Inc., 2008.
- Carl Rasmussen and Zoubin Ghahramani. Bayesian monte carlo. In *Advanced in Neural Information Processing Systems 16*, 2003.
- Carl Rasmussen and Christopher Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.
- Sami Remes, Markus Heinonen, and Samuel Kaski. Non-Stationary Spectral Kernels. In *Advances in Neural Information Processing Systems*, pages 4642–4651, 2017.
- Binxin Ru, Mark McLeod, Diego Granzoli, and Michael A. Osborne. Fast Information-theoretic Bayesian Optimisation. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.
- Fergus Simpson, Vidhi Lalchand, and Carl Rasmussen. Marginalised spectral mixture kernels with nested sampling. In *Advanced in Neural Information Processing Systems*, volume 34, 2021.
- Edward Snelson, Zoubin Ghahramani, and Carl E Rasmussen. Warped Gaussian processes. In *Advances in Neural Information Processing Systems 17*, pages 1–8, 2004.

- Tong Teng, Jie Chen, Yehong Zhang, and Kian Hsiang Low. Scalable Variational Bayesian Kernel Selection for Sparse Gaussian Process Regression. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, 2019.
- Felipe Tobar. Bayesian Nonparametric Spectral Estimation. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, 2018.
- Andrew Gordon Wilson and Ryan Prescott Adams. Gaussian Process Kernels for Pattern Discovery and Extrapolation. In *30th International Conference on Machine Learning*, pages 2104–2112, 2013.
- Xiaoyue Xi, François-Xavier Briol, and Mark Girolami. Bayesian quadrature for multiple related integrals. In *35th International Conference on Machine Learning*, volume 12, pages 8533–8564, 2018. URL <http://arxiv.org/abs/1801.04153>.
- I-Cheng Yeh. Concrete Compressive Strength Data Set, 2007. URL <https://archive.ics.uci.edu/ml/datasets/concrete+compressive+strength>.

Supplementary Material: Marginalising over Stationary Kernels with Bayesian Quadrature

A SUMMARY OF WSABI

For completeness we include a summary of the Warped Sequential Active Bayesian Integration (WSABI) algorithm proposed by Gunter et al. (2014), upon which our work builds heavily.

Recall that we are interested in computing integrals of the form $\int f(\theta)d\pi(\theta)$, where $f(\theta)$ is non-negative. We enforce this constraint by modelling $g(\theta) = \sqrt{2(f(\theta) - \epsilon)} \sim \mathcal{GP}(0, \kappa)$. This induces a non-central Chi-squared distribution over $f(\theta)$.

Now, denote by Θ the set of observation locations, the matching transformed observations by $z = \sqrt{2(f(\Theta) - \epsilon)}$, the covariance between all pairs of observations as $\kappa_{\Theta\Theta}$, the covariance between an arbitrary point θ and the set of observations as $\kappa_{\theta\Theta}$ and the kernel between two arbitrary points as $\kappa_{\theta_1\theta_2}$. Additionally, let $\mathfrak{D} = (\Theta, z)$.

Further, recall that the posterior over $g(\theta)$ conditioned on z is given by $\mathcal{GP}(\mu_{\mathfrak{D}}, \Sigma_{\mathfrak{D}})$ with:

$$\begin{aligned}\mu_{\mathfrak{D}} &= \kappa_{\theta\Theta}\kappa_{\Theta\Theta}^{-1}z \\ \Sigma_{\mathfrak{D}} &= \kappa_{\theta_1\theta_2} - \kappa_{\theta_1\Theta}\kappa_{\Theta\Theta}^{-1}\kappa_{\Theta\theta_2}^T\end{aligned}$$

By taking a Taylor expansion around the posterior mean of $g(\theta)$, $\mu_{\mathfrak{D}}$, we can approximate

$$\begin{aligned}f(g) &\approx f(\mu_{\mathfrak{D}}) + (g - \mu_{\mathfrak{D}}) \left. \frac{df}{dg} \right|_{g=\mu_{\mathfrak{D}}} \\ &= \epsilon - \frac{1}{2}\mu_{\mathfrak{D}}^2 + \mu_{\mathfrak{D}}g.\end{aligned}$$

As this is a linear transform of g , we can approximate the distribution over f as $\mathcal{GP}(\mathfrak{m}_{\mathfrak{D}}, \mathfrak{K}_{\mathfrak{D}})$ with moments

$$\begin{aligned}\mathfrak{m}_{\mathfrak{D}}(\theta) &= \epsilon + \frac{1}{2}\mu_{\mathfrak{D}}(\theta)^2, \\ \mathfrak{K}_{\mathfrak{D}}(\theta_1, \theta_2) &= \mu_{\mathfrak{D}}(\theta_1)\Sigma_{\mathfrak{D}}(\theta_1, \theta_2)\mu_{\mathfrak{D}}(\theta_2).\end{aligned}$$

Then the first two moments of the integral of interest are given by

$$\begin{aligned}\mathbb{E}\left[\int f(\theta)d\pi(\theta)\right] &= \int \mathbb{E}[f(\theta)]d\pi(\theta) = \int \left(\epsilon + \frac{\mu_{\mathfrak{D}}(\theta)^2}{2}\right)d\pi(\theta) \\ &= \epsilon + \frac{1}{2}\left(z^T\kappa_{\Theta\Theta}^{-1}\int\kappa_{\theta\Theta}^T\kappa_{\theta\Theta}d\pi(\theta)\kappa_{\Theta\Theta}^{-1}z\right) \\ &= \epsilon + \frac{1}{2}z^TQz.\end{aligned}$$

and

$$\begin{aligned}\text{Cov}\left[\int f(\theta)d\pi(\theta)\right] &= \iint\mu_{\mathfrak{D}}(\theta)\Sigma_{\mathfrak{D}}(\theta, \theta')\mu_{\mathfrak{D}}(\theta')d\pi(\theta)d\pi(\theta') \\ &= z^T\kappa_{\Theta\Theta}^{-1}\left(\iint\kappa_{\theta\theta'}\kappa_{\theta\Theta}^T\kappa_{\Theta\theta'}d\pi(\theta)d\pi(\theta') - \int\kappa_{\theta\Theta}^T\kappa_{\theta\Theta}d\pi(\theta)\kappa_{\Theta\Theta}^{-1}\int\kappa_{\theta'\Theta}^T\kappa_{\theta'\Theta}d\pi(\theta')\right)\kappa_{\Theta\Theta}^{-1}z\end{aligned}$$

Where the quadrature weights, $Q = \kappa_{\Theta}^{-1} \int \kappa_{\theta}^T \kappa_{\theta} d\pi(\theta) \kappa_{\Theta}^{-1}$, only depend on the particular kernel choice and locations of already observed function values.

For the numerator of Equation (4) in the main text we simply substitute $z_* = \sqrt{2(p(y_* | x_*, D, \theta)p(D | \theta) - \epsilon)}$ for $z = \sqrt{2(p(D | \theta) - \epsilon)}$. The predictive posterior is then a weighted sum of products of Gaussians.

B A REMARK ON POSTERIOR INFERENCE (4)

The predictive posterior given in (4) is typically viewed through the heirarchical Bayesian framework:

$$\begin{aligned} p(y_* | x_*, D) &= \dots = \frac{\int p(y_* | x_*, \Theta, D) p(D | \Theta) p(\Theta) d\Theta}{\int p(D | \Theta) p(\Theta) d\Theta} \\ &= \frac{\iint p(y_* | x_*, \theta^{(n)}, D) p(D | \theta^{(n)}) p(\theta^{(n)}, n) d\theta^{(n)} dn}{\iint p(D | \theta^{(n)}) p(\theta^{(n)}, n) d\theta^{(n)} dn} \\ &= \frac{\sum_{n=1}^N p(n) \int p(y_* | x_*, \theta^{(n)}, D) p(D | \theta^{(n)}) p(\theta^{(n)}) d\theta^{(n)}}{\sum_{n=1}^N p(n) \int p(D | \theta^{(n)}) p(\theta^{(n)}) d\theta^{(n)}}, \end{aligned} \tag{9}$$

where $\theta^{(n)}$ is the description of a GMM with exactly n components. For our experiments we choose uniform priors over n , which cancel (note that this is not an inherent limitation of our method).

A naïve approach performs the integrals over $\theta^{(n)}$ separately before computing the sums over n . The ability of MASKERADE to share information across models enables the selection of acquisitions which are maximally informative about the *sum of integrals rather than each individual integral separately*, improving the convergence rate. Additionally, accounting for the covariance between integrals over $\theta^{(n)}$ improves the quality of the posterior over the result of the summations.

B.1 DERIVATION OF INFORMATION THEORETIC SAMPLE ACQUISITION

The acquisition function is the expected information gained about the integral by making a set of likelihood observations

$$\alpha(\theta_*^{(n)}) = \int (H[Z | z, \Theta] - H[Z | z_*, \theta_*^{(n)}, z, \Theta]) p(z_* | \theta_*^{(n)}, z, \Theta) dz_*$$

Due to the symmetry of the mutual information, we can swap the role of Z and z_* [52]

$$\begin{aligned} \alpha(\theta_*^{(n)}) &= H[z_* | z, \Theta] - \int H[z_* | \theta_*^{(n)}, Z, z, \Theta] p(Z | z, \Theta) dZ \\ &= H[z_* | z, \Theta] - H[z_* | \theta_*^{(n)}, Z, z, \Theta] \end{aligned}$$

where H is differential entropy (always of Gaussians). The second line follows because the entropy of the Gaussian $p(z_* | \theta_*^{(n)}, Z, z, \Theta)$ does not depend on the value of Z .

B.2 Monte Carlo approximation to the kernel integrals

The kernel integrals are approximated using Quasi Monte Carlo sampling under the prior. Table 6 shows the effect of varying the number of QMC samples.

Table 6: The effect of number of Monte Carlo samples used to estimate the kernel integrals (i.e. the those described in Section A). For a MASKERADE model that marginalises over up to 5 mixture components we randomly sample sets of 100, 500 and 1000 hyperparameters from the prior, and compute their corresponding likelihoods on the UCI Airfoil Self-Noise dataset. We then infer the model evidence with WSABI using the MASKERADE hyper-kernel. For a given number of MC samples, we repeat this five times and report the mean and SEM for the posterior mean of the model evidence. We observe that the estimate for the model evidence is not highly sensitive to the number of monte carlo samples used to compute the kernel integrals.

No. LIKELIHOOD OBSERVATIONS	NO. OF MC SAMPLES		
	100	1000	10000
100	0.02546 ± 0.00149	0.02542 ± 0.00020	0.02526 ± 0.00012
500	0.00369 ± 0.00035	0.00379 ± 0.00013	0.00358 ± 0.00003
1000	0.00333 ± 0.00095	0.00292 ± 0.00031	0.00292 ± 0.00006

C FULL ALGORITHM DESCRIPTION

For completeness we once more give the full generative model:

$$\begin{aligned}
 n &\sim \text{Uniform}(\{1, 2, \dots, N\}), \\
 w &\sim \text{Dirichlet}(\alpha), \\
 m_{1..n} &\sim \mathcal{N}(\mu, \Sigma), \\
 \sigma_{1..n} &\sim \text{Log-Normal}(\nu, \tau), \\
 S(\omega) &= \sum_{j=1}^n \frac{w_j}{2} (\mathcal{N}(\omega; m_j, \sigma_j) + \mathcal{N}(\omega; -m_j, \sigma_j)), \\
 k(x, x') &= \int e^{2\pi i |x-x'| \omega} S(\omega) d\omega, \\
 f &\sim \mathcal{GP}(0, k).
 \end{aligned} \tag{10}$$

Below, we provide a summary of the MASKERADE framework in Algorithm 1. A schematic outlining our method is made available in Figure 2.

Algorithm 1 Pseudocode for both the learning and prediction phases of our algorithm.

```

obtain initial samples  $\Theta, z$  ▷  $z$  are likelihood evaluations at  $\Theta$ .
 $\lambda, l \leftarrow \text{argmax}_{\lambda, l} p(z | \Theta, \lambda, l)$  ▷ Optimise BQ Surrogate
while  $i \geq 0$  do
   $\{\theta_s^{(n)}\}_{1:b} \leftarrow \text{argmax}_{\{\theta_s^{(n)}\}_{1:b}} \alpha(\{\theta_s^{(n)}\}_{1:b})$  ▷ Optimise acquisition function
  append  $\{\theta_s^{(n)}\}_{1:b}$  to  $\Theta$ .
  append likelihood( $\{\theta_s^{(n)}\}_{1:b}$ ) to  $z$ .
   $\lambda, l \leftarrow \text{argmax}_{\lambda, l} p(z | \Theta, \lambda, l)$  ▷ Optimise BQ surrogate
   $i \leftarrow i - 1$ 
end while
 $Q \leftarrow \text{compute\_bq\_weights}(\Theta, z, \lambda, l, \phi)$  ▷  $Q$  defined as in Appendix on WSABI above.  $\phi$  are all prior parameters.
for  $\theta_s^{(n)}$  in  $\Theta$  do
   $\mu_*, \Sigma_* \leftarrow \text{predict}(x_*, D, \theta_s^{(n)})$ 
end for
return predictive_posterior( $Q, \{\mu_*\}, \{\Sigma_*\}, z$ ) ▷ See Appendix on WSABI for details.

```

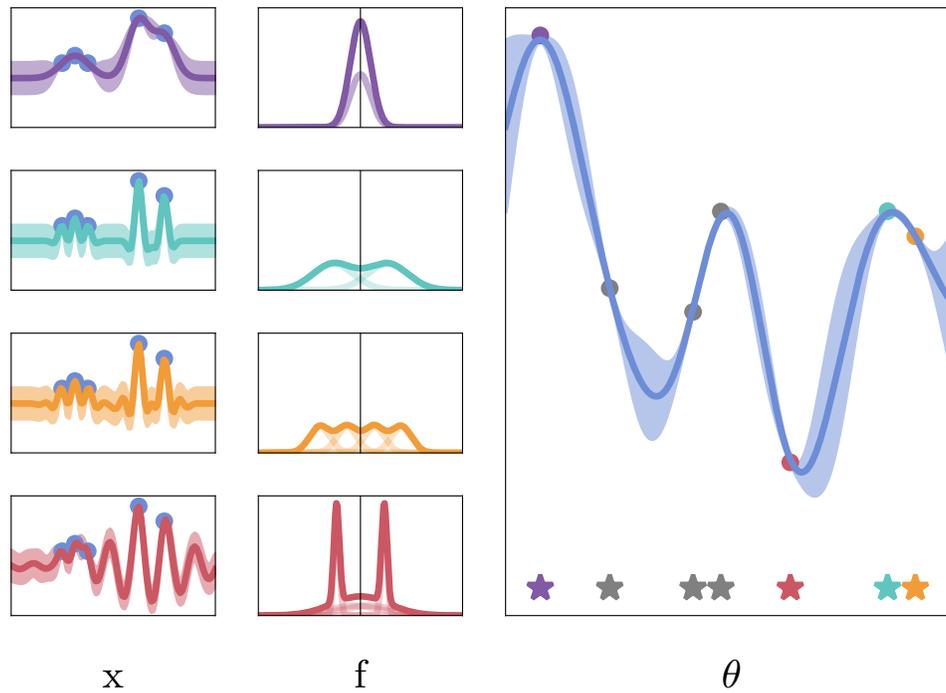


Figure 2: A schematic representation of the inference procedure under our model. The left column shows the GP posterior for different Spectral Mixture Kernels on the same dataset. The centre column shows the SM kernels in their spectral domain. The right column illustrates a GP posterior on a space indexed by the spectral densities of the SM kernels. BQ can then be used to marginalise over SM kernels which may have different numbers of mixture components.

D QUALITATIVE ANALYSIS

We qualitatively inspect the behaviour of MASKERADE in Figure 3 by examining the spectra of kernels assigned the highest weights in the posterior, and the effect of varying the number of mixture components that are marginalised over.

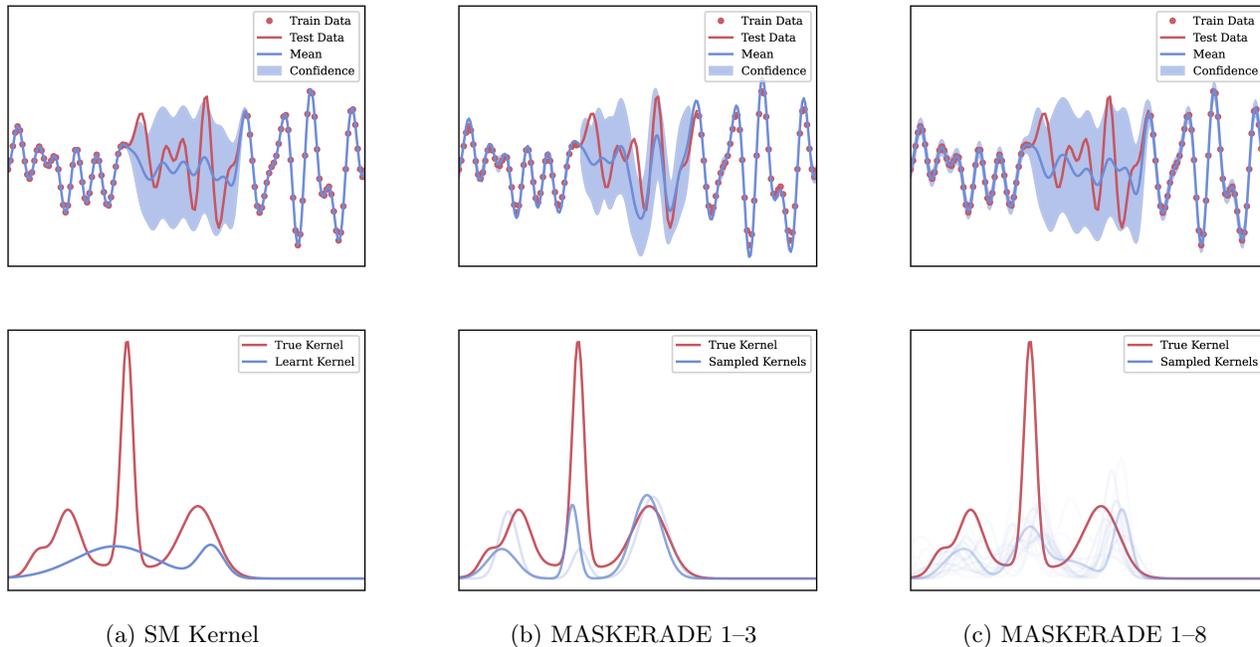


Figure 3: A comparison of a 5 component SM kernel with optimised hyperparameters, and two variants of MASKERADE – one that places a prior over up to 3 mixture components, and the other up to 8 mixture components – on a toy dataset drawn from a 5 component SM kernel. Each column plots attributes of the labelled model. The first row shows the posterior conditioned on the training data (for the MASKERADE models we show the moment-matched posterior). The second row shows the spectra (for positive frequencies) of the data generating kernel, and (for the 5 component SM kernel model) the optimised or (for the MASKERADE models) the sampled kernels. For the MASKERADE model, the opacity of a sampled kernel is proportional to the quadrature weight for all GP products of which that kernel is a part. (Recall that the posterior is a weighted sum of products of GP posteriors.) MASKERADE 1-3 is able to select samples near the data generating kernel, and therefore produce a posterior that generalises better than the other two models. Despite having the same number of mixture components as the data generating kernel, the optimised SM kernel sets the weights of 3 components to be very small. MASKERADE 1-8 struggles to explore its larger hyperparameter space with the same budget (500 likelihood evaluations) as MASKERADE 1-3. This can be seen by the fact that it spreads its posterior mass more evenly over a larger number of samples.

Next, we visually verify that our method improves upon baselines by plotting fits on the Mauna Loa dataset. These are shown in Figure 4. MASKERADE is parameterised to marginalise over 5 mixture components, with priors set as in Section 3.1.1. FKL uses the defaults in the `spectralgp`⁶ Python package, with the maximum frequency set to the Nyquist frequency of the Mauna Loa dataset. The Spectral Mixture kernel uses 5 mixtures, initialised using the empirical spectrum of the dataset.

⁶<https://github.com/wjmaddox/spectralgp>

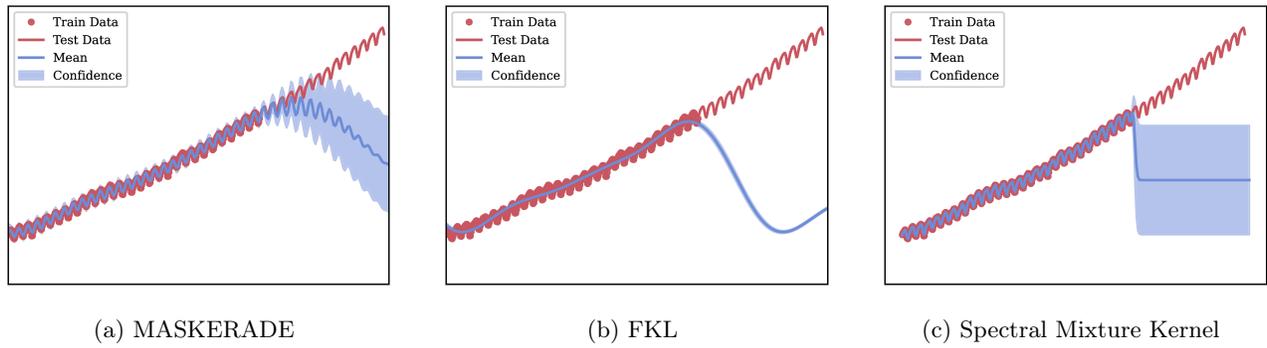


Figure 4: Posteriors for several methods on the Mauna Loa dataset (For MASKERADE and FKL we show moment matched posteriors). All methods struggle to model the linear trend, but MASKERADE is best able to extrapolate the periodic structure.