A

NUMERICAL METHODS FOR IMPROVED DECOUPLED SAMPLING OF GAUSSIAN PROCESSES

A.1 INTRODUCTION

Sampling from Gaussian processes (GPs) is not only crucial in its own right but also plays a pivotal role in various downstream tasks, notably Thompson sampling [252], as detailed in Section 2.5.2.4.

Using the standard approach, the computational cost scales cubically with the number of test points. Moreover, samples obtained through this method cannot be straightforwardly evaluated at arbitrary inputs, let alone optimised. To address these challenges, a common strategy involves utilising the weight-space approximation of GPs based on their spectral decomposition. However, this introduces its own issues, particularly when the number of training observations increases, leading to erratic extrapolations [29, 178, 272].

Recent work has proposed a hybrid approach that leverages a simple, effective, yet underutilised method for sampling from Gaussian conditionals [282, 284]. This method enables the combined use of the canonical basis and the spectral basis (also known as Fourier features), to generate samples efficiently. Notably, these samples can be obtained with a linear cost in the number of test points, and are easy to evaluate and optimise.

In existing works, the frequencies are selected through a straightforward MC approximation scheme. In this chapter, we explore the use of various numerical integration techniques to improve upon the selection mechanism. We provide a concise overview of approaches considered for the Fourier feature decomposition of stationary kernels, comparing their effectiveness in approximating the kernel matrix. Subsequently, we introduce variations to existing schemes, extending the applicability of decompositions to kernel classes beyond the SE kernel. We highlight a critical limitation is in an existing class of schemes based on Gaussian quadrature when dealing with kernels with small lengthscales. Specifically, small lengthscales result in highly oscillatory integrals that pose challenges for estimation through numerical methods. To address this, we consider a previously untapped techniques based on an extension of Newton-Cotes quadrature. Finally, we evaluate how the Fourier feature decompositions derived from the various numerical integration schemes impact the fidelity of the GP posterior samples.

A.2 DECOUPLED SAMPLING OF GAUSSIAN PROCESSES

We give a brief overview of the method proposed by Wilson et al. [282]. Recall that for practical purposes, a GP posterior at T query locations is simply a T-dimensional conditional Gaussian distribution. In general, consider jointly Gaussian random variables $\mathbf{a} \in \mathbb{R}^T$ and $\mathbf{b} \in \mathbb{R}^M$,

$$\begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_{\mathbf{a}} \\ \boldsymbol{\mu}_{\mathbf{b}} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{\mathbf{aa}} & \boldsymbol{\Sigma}_{\mathbf{ab}} \\ \boldsymbol{\Sigma}_{\mathbf{ba}} & \boldsymbol{\Sigma}_{\mathbf{bb}} \end{bmatrix} \right).$$

The distribution of $\mathbf{a}$ conditioned on $\mathbf{b} = \boldsymbol{\beta}$ is given by

$$p(\mathbf{a} \mid \mathbf{b} = \boldsymbol{\beta}) = \mathcal{N}(\mathbf{a} \mid \boldsymbol{\mu}_{\mathbf{a}|\mathbf{b}}, \boldsymbol{\Sigma}_{\mathbf{a}|\mathbf{b}}),$$

where the mean and covariance are given by

$$\boldsymbol{\mu}_{\mathbf{a}|\mathbf{b}} \triangleq \boldsymbol{\mu}_{\mathbf{a}} + \boldsymbol{\Sigma}_{\mathbf{ab}}\boldsymbol{\Sigma}_{\mathbf{bb}}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathbf{b}}), \quad \text{and} \quad \boldsymbol{\Sigma}_{\mathbf{a}|\mathbf{b}} \triangleq \boldsymbol{\Sigma}_{\mathbf{aa}} - \boldsymbol{\Sigma}_{\mathbf{ab}}\boldsymbol{\Sigma}_{\mathbf{bb}}^{-1}\boldsymbol{\Sigma}_{\mathbf{ba}}.$$

The standard approach to generating samples from $p(\mathbf{a} \mid \mathbf{b} = \boldsymbol{\beta})$ is to use a location-scale transform of normal random variables, i. e.,

$$\mathbf{a} = \boldsymbol{\mu}_{\mathbf{a}|\mathbf{b}} + \boldsymbol{\Sigma}_{\mathbf{a}|\mathbf{b}}^{1/2}\boldsymbol{\eta}, \quad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad \Leftrightarrow \quad \mathbf{a} \sim \mathcal{N}(\boldsymbol{\mu}_{\mathbf{a}|\mathbf{b}}, \boldsymbol{\Sigma}_{\mathbf{a}|\mathbf{b}}),$$

where $\boldsymbol{\Sigma}_{\mathbf{a}|\mathbf{b}}^{1/2}$ denotes the Cholesky factor of $\boldsymbol{\Sigma}_{\mathbf{a}|\mathbf{b}}$, whose calculation has a cost of $\mathcal{O}(T^3)$, and is precisely what makes the standard approach so computationally expensive.

Matheron's rule

A powerful alternative for sampling conditional Gaussian variables is Matheron's rule [117],

$$(\mathbf{a} \mid \mathbf{b} = \boldsymbol{\beta}) \overset{D}{=} \mathbf{a} + \boldsymbol{\Sigma}_{\mathbf{ab}}\boldsymbol{\Sigma}_{\mathbf{bb}}^{-1}(\boldsymbol{\beta} - \mathbf{b}), \tag{A.1}$$

where $\overset{D}{=}$ denotes equality *in distribution*. This is straightforward to verify. By computing the mean and covariance of this expression, we get

$$\mathbb{E}[\mathbf{a} + \boldsymbol{\Sigma}_{\mathbf{ab}}\boldsymbol{\Sigma}_{\mathbf{bb}}^{-1}(\boldsymbol{\beta} - \mathbf{b})] = \boldsymbol{\mu}_{\mathbf{a}} + \boldsymbol{\Sigma}_{\mathbf{ab}}\boldsymbol{\Sigma}_{\mathbf{bb}}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\mathbf{b}}) = \boldsymbol{\mu}_{\mathbf{a}|\mathbf{b}}$$

and

$$\begin{aligned} &\mathrm{Cov}[\mathbf{a} + \boldsymbol{\Sigma}_{\mathbf{ab}}\boldsymbol{\Sigma}_{\mathbf{bb}}^{-1}(\boldsymbol{\beta} - \mathbf{b})] \\ &\quad = \boldsymbol{\Sigma}_{\mathbf{aa}} - 2\boldsymbol{\Sigma}_{\mathbf{ab}}\boldsymbol{\Sigma}_{\mathbf{bb}}^{-1}\boldsymbol{\Sigma}_{\mathbf{ba}} + \boldsymbol{\Sigma}_{\mathbf{ab}}\boldsymbol{\Sigma}_{\mathbf{bb}}^{-1}\boldsymbol{\Sigma}_{\mathbf{bb}}\boldsymbol{\Sigma}_{\mathbf{bb}}^{-1}\boldsymbol{\Sigma}_{\mathbf{ba}} \\ &\quad = \boldsymbol{\Sigma}_{\mathbf{aa}} - \boldsymbol{\Sigma}_{\mathbf{ab}}\boldsymbol{\Sigma}_{\mathbf{bb}}^{-1}\boldsymbol{\Sigma}_{\mathbf{ba}} = \boldsymbol{\Sigma}_{\mathbf{a}|\mathbf{b}} \end{aligned}$$

respectively. Recall from Section 2.4.1 that a GP is a random function such that, at a finite set of locations $\mathbf{X}_*$, the vector $\mathbf{f}_* = f(\mathbf{X}_*)$ follows a Gaussian distribution. Specifically, if $f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$, then $\mathbf{f}_* \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_{**})$ where $\mathbf{K}_{**} \triangleq k(\mathbf{X}_*, \mathbf{X}_*)$ for some covariance function k. Further recall from Equation (2.16) that the posterior of an exact GP at test locations $\mathbf{X}_*$ given N observations $\mathbf{y}$ is $p(\mathbf{f}_* \mid \mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_{*|\mathbf{y}}, \boldsymbol{\Sigma}_{**|\mathbf{y}})$, where

$$\begin{aligned} \boldsymbol{\mu}_{*|\mathbf{y}} &\triangleq \mathbf{K}_{*\mathbf{f}}(\mathbf{K}_{\mathbf{ff}} + \beta^{-1}\mathbf{I})^{-1}\mathbf{y}, \\ \boldsymbol{\Sigma}_{**|\mathbf{y}} &\triangleq \mathbf{K}_{**} - \mathbf{K}_{*\mathbf{f}}(\mathbf{K}_{\mathbf{ff}} + \beta^{-1}\mathbf{I})^{-1}\mathbf{K}_{\mathbf{f}*}, \end{aligned} \tag{A.2}$$

and from Equation (2.22) that the conditional distribution of SVGP models at test locations $\mathbf{X}_*$ given inducing variables $\mathbf{u} \sim p(\mathbf{u})$ is $p(\mathbf{f}_* \mid \mathbf{u}) = \mathcal{N}(\boldsymbol{\mu}_{*|\mathbf{u}}, \boldsymbol{\Sigma}_{**|\mathbf{u}})$, where

$$\begin{aligned} \boldsymbol{\mu}_{*|\mathbf{u}} &\triangleq \mathbf{K}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{u}, \\ \boldsymbol{\Sigma}_{**|\mathbf{u}} &\triangleq \mathbf{K}_{**} - \mathbf{K}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}*}. \end{aligned} \tag{A.3}$$

Applying Matheron's rule from Equation (A.1) to these conditionals, we have, for exact GPs,

$$(\mathbf{f}_* \mid \mathbf{y}) \overset{D}{=} \mathbf{f}_* + \mathbf{K}_{*\mathbf{f}}(\mathbf{K}_{\mathbf{ff}} + \beta^{-1}\mathbf{I})^{-1}(\mathbf{y} - \mathbf{f}_N - \boldsymbol{\epsilon}),$$

and, for sparse GPs,

$$(\mathbf{f}_* \mid \mathbf{u}) \overset{D}{=} \mathbf{f}_* + \mathbf{K}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}(\mathbf{u} - \mathbf{f}_M),$$

where $(\mathbf{f}_*, \mathbf{f}_N)$ and $(\mathbf{f}_*, \mathbf{f}_M)$ respectively are jointly sampled from the GP prior. The astute reader will recognise the absurdity of this approach, as it is in fact considerably more expensive than the conventional one. Specifically, jointly sampling from the prior incurs costs of $\mathcal{O}\left((T+N)^3\right)$ and $\mathcal{O}\left((T+M)^3\right)$, respectively. As we shall see, this is the paradox that Wilson et al. [282] managed to resolve.

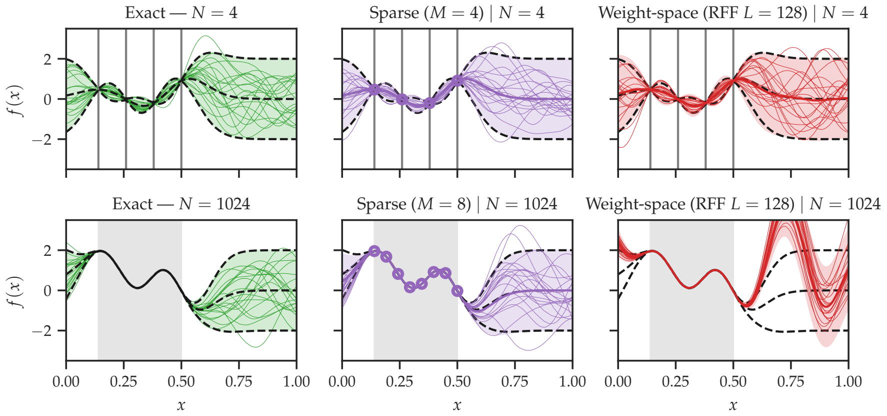

Figure A.1: An illustration of the *variance starvation* phenomenon. Across the columns, we have a comparison of various GP posteriors and their samples, given $n = 4$ (top) and $n = 1,024$ (bottom) observations at locations indicated by the shaded regions. A reproduction of the figures originating from Wilson et al. [282].

Let's consider the weight-space approximation described in Section 2.4.3. Recall from Equation (2.37) that the posterior weight density is $p(\mathbf{w} \mid \mathbf{y}) = \mathcal{N}(\boldsymbol{\mu}_{\mathbf{w}|\mathbf{y}}, \boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}})$, where

$$\begin{aligned} \boldsymbol{\mu}_{\mathbf{w}|\mathbf{y}} &\triangleq (\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \beta^{-1}\mathbf{I})^{-1} \boldsymbol{\Phi}^\top \mathbf{y}, \\ \boldsymbol{\Sigma}_{\mathbf{w}|\mathbf{y}} &\triangleq \beta^{-1}(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \beta^{-1}\mathbf{I})^{-1}. \end{aligned}$$

Applying Matheron's rule of Equation (A.1) to the weight-space posterior, we have

$$(\mathbf{w} \mid \mathbf{y}) \overset{D}{=} \mathbf{w} + \boldsymbol{\Phi}^\top (\boldsymbol{\Phi}\boldsymbol{\Phi}^\top + \beta^{-1}\mathbf{I})^{-1}(\mathbf{y} - \boldsymbol{\Phi}\mathbf{w} - \boldsymbol{\epsilon}).$$

variance starvation

While possible to sample from efficiently, as alluded to previously, this approach is beset by the general limited expressiveness of finite-dimensional feature maps, which can hamper ability to extrapolate predictions at test time. In particular, for Fourier feature decompositions, this is a phenomenon known as *variance starvation*, whereby extrapolations become erratic as the number of observations N increases. The intuition behind this is that although the Fourier basis is suited for representing stationary GPs, the posterior is generally non-stationary. See Figure A.1 for an illustration of the variance starvation phenomenon.

Wilson et al. [282] seek to combine the best of both worlds, by leveraging the strength of the Fourier basis $\boldsymbol{\phi}(\,\cdot\,)$ at representing stationary priors [200], and the strength of the canonical basis $k(\,\cdot\,, \mathbf{z})$ at representing the data [24].

The decoupled sampling approach for sparse GPs is

$$(\mathbf{f}_* \mid \mathbf{u}) \overset{D}{\approx} \boldsymbol{\Phi}_*\mathbf{w} + \mathbf{K}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\left(\mathbf{u} - \boldsymbol{\Phi}\mathbf{w}\right), \tag{A.4}$$

and, for exact GPs, is

$$(\mathbf{f}_* \mid \mathbf{y}) \overset{D}{\approx} \boldsymbol{\Phi}_*\mathbf{w} + \mathbf{K}_{*\mathbf{f}}(\mathbf{K}_{\mathbf{ff}} + \beta^{-1}\mathbf{I})^{-1}\left(\mathbf{y} - \boldsymbol{\Phi}\mathbf{w} - \boldsymbol{\epsilon}\right). \tag{A.5}$$

It's important to emphasise that these are in fact only *approximately* equal in distribution. To understand precisely how they differ, let us compute their moments. We focus on the case of sparse GPs in Equation (A.4), the mean and covariance of which are

$$\mathbb{E}[\boldsymbol{\Phi}_*\mathbf{w} + \mathbf{K}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\left(\mathbf{u} - \boldsymbol{\Phi}\mathbf{w}\right)] = \mathbf{K}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{u} = \boldsymbol{\mu}_{*|\mathbf{u}}$$

and

$$\begin{aligned} &\operatorname{Cov}[\boldsymbol{\Phi}_*\mathbf{w} + \mathbf{K}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\left(\mathbf{u} - \boldsymbol{\Phi}\mathbf{w}\right)] \\ &\quad = \boldsymbol{\Phi}_*\boldsymbol{\Phi}_*^\top - 2\mathbf{K}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}_*^\top + \mathbf{K}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}*} \\ &\quad \approx \mathbf{K}_{**} - \mathbf{K}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}*} = \boldsymbol{\Sigma}_{**|\mathbf{u}} \end{aligned} \tag{A.6}$$

We see that mean is exactly equal to the $\boldsymbol{\mu}_{*|\mathbf{u}}$ of Equation (A.3), but the covariance is only equal to $\boldsymbol{\Sigma}_{**|\mathbf{u}}$ if

$$\boldsymbol{\Phi}_*\boldsymbol{\Phi}_*^\top = \mathbf{K}_{**}, \quad \boldsymbol{\Phi}\boldsymbol{\Phi}_*^\top = \mathbf{K}_{\mathbf{u}*}, \quad \text{and} \quad \boldsymbol{\Phi}\boldsymbol{\Phi}^\top = \mathbf{K}_{\mathbf{uu}},$$

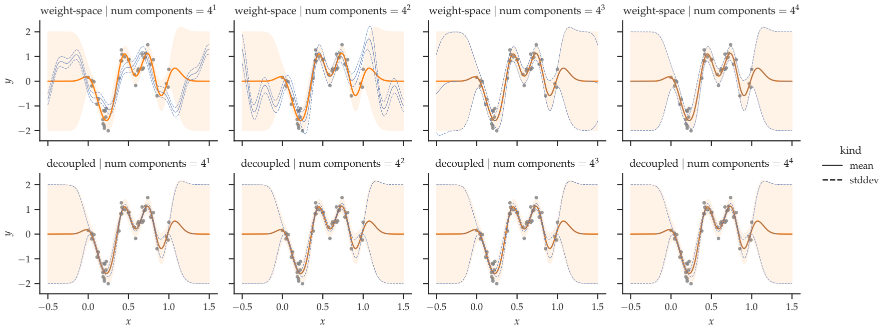

Figure A.2: Posterior predictive distributions.

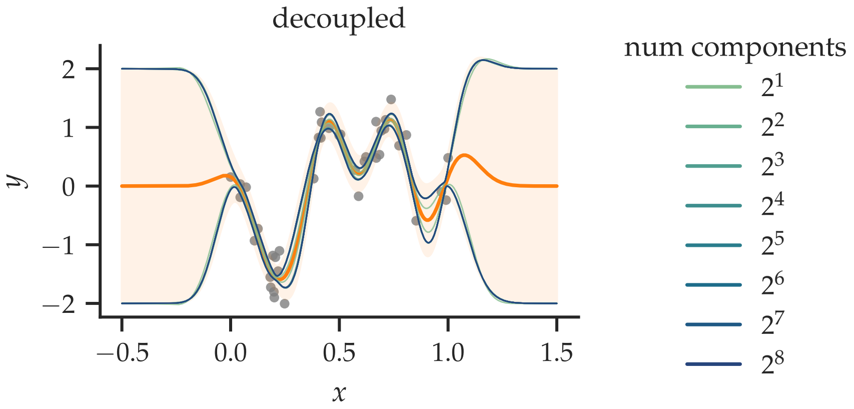

Figure A.3: Posterior predictive distributions from the decoupled approach, overlayed on top of one another.

which are satisfied when $k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\phi}(\mathbf{x}')$ for all $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. In other words, we have equality in distribution in Equations (A.4) and (A.5) when the kernel approximation is exact. Thus seen, the quality of decoupled pathwise samples relies crucially on the quality of the kernel approximation in Equation (2.38) itself. In this chapter, we explore various methods from the classical literature on numerical integration [50] to tighten this approximation.

A.3 NUMERICAL INTEGRATION FOR GP PRIOR APPROXIMATIONS

A multitude of *numerical integration* methods [50] can readily be deployed to compute the expectation in Equation (2.44),

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}_{p(\omega)}[\varphi_\omega(\mathbf{x})^\top \varphi_\omega(\mathbf{x}')] \\ &\approx \sum_{i=1}^{L} \alpha_i \left(\varphi_{\xi_i}(\mathbf{x})^\top \varphi_{\xi_i}(\mathbf{x}') \right), \end{aligned}$$

where $\boldsymbol{\xi}_i$ are referred to as the *abscissas*, or, *nodes*, and α_i the *weights*, or, *coefficients*. Let us define the mapping $\boldsymbol{\varphi} : \mathbb{R}^D \to \mathbb{R}^{L'}$,

$$\boldsymbol{\varphi}(\mathbf{x}) \triangleq \begin{bmatrix} \sqrt{\alpha_1} \cos \boldsymbol{\xi}_1^\top \mathbf{x} \\ \vdots \\ \sqrt{\alpha_L} \cos \boldsymbol{\xi}_L^\top \mathbf{x} \\ \sqrt{\alpha_1} \sin \boldsymbol{\xi}_1^\top \mathbf{x} \\ \vdots \\ \sqrt{\alpha_L} \sin \boldsymbol{\xi}_L^\top \mathbf{x} \end{bmatrix}, \tag{A.7}$$

where $L' = 2L$. We therefore have

$$\boldsymbol{\varphi}(\mathbf{x})^\top \boldsymbol{\varphi}(\mathbf{x}') = \sum_{i=1}^{L} a_i \left(\varphi_{\boldsymbol{\xi}_i}(\mathbf{x})^\top \varphi_{\boldsymbol{\xi}_i}(\mathbf{x}') \right) \approx k(\mathbf{x}, \mathbf{x}').$$

We can view $\boldsymbol{\varphi}(\mathbf{x}), \boldsymbol{\varphi}(\mathbf{x}')$ as a factorisation, or, decomposition, of the kernel $k(\mathbf{x}, \mathbf{x}')$. Thus, we refer to $\boldsymbol{\varphi}$ as a *Fourier feature decomposition* of k.

A.3.1 *Monte Carlo Estimation*

Let us consider the simple case of MC integration, where $\alpha_i \triangleq 1/L$ and $\boldsymbol{\xi}_i \triangleq \boldsymbol{\omega}^{(i)}$, with $\boldsymbol{\omega}^{(i)} \sim p(\boldsymbol{\omega})$. More explicitly,

$$k(\mathbf{x}, \mathbf{x}') \approx \frac{1}{L} \sum_{i=1}^{L} \varphi_{\boldsymbol{\omega}^{(i)}}(\mathbf{x})^\top \varphi_{\boldsymbol{\omega}^{(i)}}(\mathbf{x}'), \quad \text{where} \quad \boldsymbol{\omega}^{(i)} \sim p(\boldsymbol{\omega}).$$

The corresponding Fourier feature decomposition is then

$$\boldsymbol{\varphi}(\mathbf{x}) \triangleq \sqrt{\frac{2}{L'}} \begin{bmatrix} \cos {\boldsymbol{\omega}^{(1)}}^\top \mathbf{x} \\ \vdots \\ \cos {\boldsymbol{\omega}^{(L'/2)}}^\top \mathbf{x} \\ \sin {\boldsymbol{\omega}^{(1)}}^\top \mathbf{x} \\ \vdots \\ \sin {\boldsymbol{\omega}^{(L'/2)}}^\top \mathbf{x} \end{bmatrix}, \tag{A.8}$$

where $\boldsymbol{\omega}^{(i)} \sim p(\boldsymbol{\omega})$, which we refer to as the MC Fourier features or, more commonly, RFF [200, 201].

phase-shifted cosine features

Let us define $\phi_{(\boldsymbol{\omega},b)} : \mathbb{R}^D \to \mathbb{R}$ to be, as before, the projection in some random direction $\boldsymbol{\omega} \sim p(\boldsymbol{\omega})$, but shifted by some $b \sim \mathcal{U}[0, 2\pi]$,

$$\phi_{(\boldsymbol{\omega},b)}(\mathbf{x}) \triangleq \sqrt{2} \cos\left(\boldsymbol{\omega}^\top \mathbf{x} + b\right). \tag{A.9}$$

We take the product of $\phi_{(\boldsymbol{\omega},b)}$ evaluated at inputs $\mathbf{x}$ and $\mathbf{x}'$ to get

$$\begin{align} \phi_{(\boldsymbol{\omega},b)}(\mathbf{x}) \phi_{(\boldsymbol{\omega},b)}(\mathbf{x}') &= 2 \cos\left(\boldsymbol{\omega}^\top \mathbf{x} + b\right) \cos\left(\boldsymbol{\omega}^\top \mathbf{x}' + b\right) \tag{A.10} \\ &= \cos\left(\boldsymbol{\omega}^\top (\mathbf{x} + \mathbf{x}') + 2b\right) + \cos\left(\boldsymbol{\omega}^\top (\mathbf{x} - \mathbf{x}')\right), \tag{A.11} \end{align}$$

where, in the last line, we've used the *product-to-sum* trigonometric identity (see Appendix A.A for details). By virtue of the periodicity of sinusoids, taking the expectation of Equation (A.10) erases the first term of Equation (A.11), giving

$$\begin{aligned}
&\mathbb{E}_{p(\boldsymbol{\omega},b)}[\phi_{(\boldsymbol{\omega},b)}(\mathbf{x})\phi_{(\boldsymbol{\omega},b)}(\mathbf{x}')] \\
&\quad= \mathbb{E}_{p(\boldsymbol{\omega})}[\cos(\boldsymbol{\omega}^\top(\mathbf{x}-\mathbf{x}'))] + \cancel{\mathbb{E}_{p(\boldsymbol{\omega},b)}[\cos(\boldsymbol{\omega}^\top(\mathbf{x}+\mathbf{x}')+2b)]} \\
&\quad= k(\mathbf{x},\mathbf{x}').
\end{aligned}$$

See Appendix A.B for details. Hence, the product in Equation (A.10) is also an unbiased estimator of the kernel. For brevity, we shall write $\phi_i(\mathbf{x})$ to signify $\phi_{(\boldsymbol{\omega}^{(i)},b^{(i)})}(\mathbf{x})$ for $\boldsymbol{\omega}^{(i)} \sim p(\boldsymbol{\omega})$ and $b^{(i)} \sim \mathcal{U}[0,2\pi]$. The analogous Fourier feature decomposition $\boldsymbol{\phi} : \mathbb{R}^D \to \mathbb{R}^L$ is given by

$$\boldsymbol{\phi}(\mathbf{x}) \triangleq \sqrt{\frac{2}{L}} \begin{bmatrix} \cos\left({\boldsymbol{\omega}^{(1)}}^\top \mathbf{x} + b^{(1)}\right) \\ \vdots \\ \cos\left({\boldsymbol{\omega}^{(L)}}^\top \mathbf{x} + b^{(L)}\right) \end{bmatrix} = \frac{1}{\sqrt{L}} \begin{bmatrix} \phi_1(\mathbf{x}) \\ \vdots \\ \phi_L(\mathbf{x}) \end{bmatrix}. \qquad \text{(A.12)}$$

We refer to this Fourier feature decomposition, originally proposed by Rahimi and Recht [200], as the *phase-shifted cosine* variant of RFF. Both MC estimators outlined in this section introduces error that decays at the rate of $\mathcal{O}(L^{-1/2})$, which, notably, is independent of the input dimensionality. A theoretical comparison of the Fourier feature decompositions of Equations (A.8) and (A.12) is given by Sutherland and Schneider [248], who report that for the SE kernel, the latter produces strictly higher variance and results in worse bounds.

A.3.2 *Quasi-Monte Carlo*

We can readily improve upon the convergence of MC by employing quasi Monte Carlo (QMC), which uses deterministic low-discrepancy sequences to construct samples. We refer to this family of Fourier feature decompositions as quasi-random Fourier features (QRFF) [5, 286].

In particular, QMC approximates following integral over the unit hypercube,

$$\int_{[0,1]^D} f(\mathbf{u})\,\mathrm{d}\mathbf{u} \approx \frac{1}{L}\sum_{i=1}^{L} f(\mathbf{u}^{(i)}), \qquad \text{(A.13)}$$

by sequentially constructing the samples $\mathbf{u}^{(i)}$ deterministically using low-discrepancy sequences, thereby ameliorating the undesirable effects of samples forming clusters that commonly occurs when sampling independently at random. The interested reader may wish to refer to the manuscripts by Caflisch [27] and Dick, Kuo, and Sloan [57] for a more complete treatment of the topic.

To approximate multi-dimensional integrals with a Gaussian mea- *multi-dimensional Gaussian integrals* sure over $\mathbb{R}^D$, we can apply a change-of-variables based on the Gaussian inverse cumulative distribution function (CDF), or *quantile function*, to reduce it to an integral in the form of Equation (A.13). Suppose we have a multivariate Gaussian density $q(\boldsymbol{\omega})$. Then we can write

$$\int_{\mathbb{R}^D} q(\boldsymbol{\omega}) f(\boldsymbol{\omega}) \,\mathrm{d}\boldsymbol{\omega} = \int_{[0,1]^D} f(\Phi^{-1}(\mathbf{u})) \,\mathrm{d}\mathbf{u}, \tag{A.14}$$

where $\Phi^{-1} : [0,1]^D \to \mathbb{R}$ is the quantile function of q.

importance sampling

For non-Gaussian densities $p(\boldsymbol{\omega})$ in general, we can utilize *importance sampling* to cast our problem into the Gaussian integral of Equation (A.14), by using a Gaussian $q(\boldsymbol{\omega})$ as the *proposal* distribution,

$$\begin{aligned}\int_{\mathbb{R}^D} p(\boldsymbol{\omega}) f(\boldsymbol{\omega}) \,\mathrm{d}\boldsymbol{\omega} &= \int_{\mathbb{R}^D} q(\boldsymbol{\omega}) \left(\frac{p(\boldsymbol{\omega})}{q(\boldsymbol{\omega})} f(\boldsymbol{\omega}) \right) \mathrm{d}\boldsymbol{\omega} \\ &= \int_{[0,1]^D} r(\Phi^{-1}(\mathbf{u})) f(\Phi^{-1}(\mathbf{u})) \,\mathrm{d}\mathbf{u},\end{aligned}$$

where $r(\boldsymbol{\omega}) \triangleq {}^{p(\boldsymbol{\omega})}/_{q(\boldsymbol{\omega})}$ is the *importance weight*, or *likelihood ratio*.

From this, we arrive at the following Fourier feature decomposition,

$$\begin{aligned}k(\mathbf{t}, \mathbf{0}) &\approx \frac{1}{L} \sum_{i=1}^{L} r(\Phi^{-1}(\mathbf{u}^{(i)})) \cos\left(\Phi^{-1}(\mathbf{u}^{(i)}) \cdot \mathbf{t}\right) \\ &= \boldsymbol{\varphi}(\mathbf{x})^\top \boldsymbol{\varphi}(\mathbf{x}'),\end{aligned}$$

where

$$\boldsymbol{\varphi}(\mathbf{x}) \triangleq \sqrt{\frac{2}{L'}} \begin{bmatrix} \sqrt{r(\Phi^{-1}(\mathbf{u}^{(1)}))} \cos\left(\Phi^{-1}(\mathbf{u}^{(1)}) \cdot \mathbf{x}\right) \\ \vdots \\ \sqrt{r(\Phi^{-1}(\mathbf{u}^{(L'/2)}))} \cos\left(\Phi^{-1}(\mathbf{u}^{(L'/2)}) \cdot \mathbf{x}\right) \\ \sqrt{r(\Phi^{-1}(\mathbf{u}^{(1)}))} \sin\left(\Phi^{-1}(\mathbf{u}^{(1)}) \cdot \mathbf{x}\right) \\ \vdots \\ \sqrt{r(\Phi^{-1}(\mathbf{u}^{(L'/2)}))} \sin\left(\Phi^{-1}(\mathbf{u}^{(L'/2)}) \cdot \mathbf{x}\right) \end{bmatrix}.$$

A.3.3 *Quadrature*

We now introduce quadrature Fourier features (QFF) [6, 49, 175, 178]. We first restrict our attention to the one-dimensional case and defer our discussion of the multi-dimensional case when we introduce the multi-dimensional generalisations of numerical quadrature, sometimes referred to as *cubature*.

Gaussian-Christoffel quadrature; Gaussian quadrature

A quadrature formula that approximates the following integral by a finite sum

$$\int_a^b w(u) f(u) \,\mathrm{d}u \approx \sum_{i=1}^{L} \alpha_i f(\xi_i) \tag{A.15}$$

is called a *Gauss-Christoffel quadrature formula* (or simply a *Gaussian quadrature formula*) if it has maximum degree of exactness, i. e., if Equation (A.15) is an exact equality whenever f is a polynomial of degree $2L-1$ [78]. We refer to ξ_i as a the *Christoffel abscissas* and α_i the *Christoffel weights* associated with the weight function $w(u)$. The case of $w(u) \triangleq 1$ on the interval $[-1,1]$ was first studied by Gauss [77], and is now referred to as Gauss-Legendre quadrature. Other classical cases are associated with the names of Jacobi, Laguerre, and Hermite. The formulation based on orthogonal polynomials was advanced by Jacobi [109]. A comprehensive though possibly now outdated review of the topic of Gauss-Christoffel quadrature can be found in the landmark survey of Gautschi [79].

A.3.3.1 *Gauss-Hermite Quadrature*

The weight function serves to help factor out unruly behaviour in the integrand. Particularly relevant is the case of Gauss-Hermite quadrature, in which the weight function of interest is $w(u) \triangleq e^{-u^2}$ and the interval of integration is $(-\infty,\infty)$. That is, we're interested in approximating integrals of the form

Gauss-Hermite quadrature

$$\int_{-\infty}^{\infty} e^{-u^2} f(u)\,\mathrm{d}u \tag{A.16}$$

The nodes ξ_i are roots of $H_L(u)$, the Hermite polynomial of degree L, and the associated weights α_i are given by

$$\alpha_i \triangleq \frac{2^{L-1}L!\sqrt{\pi}}{L^2[H_{L-1}(\xi_i)]^2}.$$

It is not hard to appreciate the power of this quadrature formula, for it is trivial to apply it to the calculation of expectations under Gaussian distributions, a quantity upon which many problems in statistical ML rely. In particular, we are often interested in computing the expected value of $f(\omega)$ under $p(\omega) = \mathcal{N}(\omega \mid \mu, \sigma^2)$,

Gaussian expectations

$$\begin{aligned}\mathbb{E}_{p(\omega)}[f(\omega)] &= \int_{-\infty}^{\infty} \mathcal{N}(\omega \mid \mu, \sigma^2) f(\omega)\,\mathrm{d}\omega \qquad (A.17)\\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{\infty} e^{-\left(\frac{\omega-\mu}{\sqrt{2}\sigma}\right)^2} f(\omega)\,\mathrm{d}\omega.\end{aligned}$$

Simply by making a change-of-variable $u = \frac{\omega-\mu}{\sqrt{2}\sigma} \Leftrightarrow \omega = \sqrt{2}\sigma u + \mu$, we can rewrite Equation (A.17) in the form of Equation (A.16),

$$\mathbb{E}_{p(\omega)}[f(\omega)] = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-u^2} g(u)\,\mathrm{d}u, \tag{A.18}$$

where we've defined $g(u) \triangleq f(\sqrt{2}\sigma u + \mu)$. We thus have the following quadrature formula:

$$\mathbb{E}_{p(\omega)}[f(\omega)] \approx \frac{1}{\sqrt{\pi}} \sum_{i=1}^{L} \alpha_i g(\xi_i).$$

Kernel decomposition of the SE kernel based on Gauss-Hermite quadrature

Recall that by Equation (2.41), assuming its spectral density $p(\omega)$ is even symmetric, we can express a stationary kernel $k(t,0)$ as the expected value of function $f(\omega) = \cos(\omega t)$ under $p(\omega)$. Let us first focus on the case of the SE kernel, the spectral density of which is given in Equation (2.40) as a Gaussian, $p(\omega) = \mathcal{N}\left(\omega \mid 0, \ell^{-2}\right)$. Therefore, by Equation (A.18), with $g(u) = f\left(\sqrt{2}u/\ell\right) = \cos\left(\sqrt{2}ut/\ell\right)$, we can write

$$
\begin{aligned}
k(t,0) &= \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} e^{-u^2} \cos\left(\frac{\sqrt{2}ut}{\ell}\right) \mathrm{d}u && \text{(A.19)} \\
&\approx \frac{1}{\sqrt{\pi}} \sum_{i=1}^{L} \alpha_i \cos\left(\frac{\sqrt{2}\xi_i t}{\ell}\right). && \text{(A.20)}
\end{aligned}
$$

By Equation (2.43), we have

$$
\begin{aligned}
\cos\left(\frac{\sqrt{2}\xi_i t}{\ell}\right) &= \cos\left(\frac{\sqrt{2}\xi_i (x - x')}{\ell}\right) \\
&= \varphi_{\left(\sqrt{2}\xi_i/\ell\right)}(x)^\top \varphi_{\left(\sqrt{2}\xi_i/\ell\right)}(x'),
\end{aligned}
$$

where $\varphi_{(\cdot)}(x)$ is defined in Equation (2.42). Accordingly, as in Equation (A.7), we have the Fourier feature decomposition $\boldsymbol{\varphi} : \mathbb{R} \to \mathbb{R}^{L'}$,

$$
\boldsymbol{\varphi}(x) \triangleq \frac{1}{\sqrt[4]{\pi}} \begin{bmatrix} \sqrt{\alpha_1} \cos\left(\frac{\sqrt{2}\xi_1 x}{\ell}\right) \\ \vdots \\ \sqrt{\alpha_L} \cos\left(\frac{\sqrt{2}\xi_L x}{\ell}\right) \\ \sqrt{\alpha_1} \sin\left(\frac{\sqrt{2}\xi_1 x}{\ell}\right) \\ \vdots \\ \sqrt{\alpha_L} \sin\left(\frac{\sqrt{2}\xi_L x}{\ell}\right) \end{bmatrix}.
$$

Extending this to kernels with non-Gaussian spectral densities can be done using the importance sampling technique described in the preceding section.

A.3.3.2 *Gauss-Legendre Quadrature*

Gauss-Legendre quadrature

Let us consider the classical case of Gauss-Legendre quadrature, in which the weight function of interest is $w(u) \triangleq 1$ and the interval of integration is $[-1, 1]$. That is, we're interested in approximating integrals of the form

$$
\int_{-1}^{1} f(u) \,\mathrm{d}u
$$

The nodes ξ_i are roots of $P_L(u)$, the Legendre polynomial of degree L normalized to give $P_L(1) = 1$, and the associated weights a_i are given by

$$
\alpha_i \triangleq \frac{2}{\left(1 - \xi_i^2\right) \left[P_L'(\xi_i)\right]^2}.
$$

An integral over $[a,b]$ must be changed into an integral over $[-1,1]$ before applying Gauss-Legendre quadrature,

Change of interval

$$\int_a^b f(\omega)\,\mathrm{d}\omega = \int_{T(-1)}^{T(1)} f(\omega)\,\mathrm{d}\omega = \int_{-1}^{1} f(T(u))T'(u)\,\mathrm{d}u,$$

where $T : [a,b] \to [-1,1]$ is a differentiable function with a continuous derivative. In particular, for integration over the infinite interval, we can use the substitution $T(u) = \tan\left(\frac{\pi}{2}u\right)$ to give

$$\int_{-\infty}^{\infty} f(\omega)\,\mathrm{d}\omega = \frac{\pi}{2}\int_{-1}^{1} \frac{f\left(\tan\left(\frac{\pi}{2}u\right)\right)}{\cos^2\left(\frac{\pi}{2}u\right)}\,\mathrm{d}u,$$

where we've used $T'(u) = \frac{\pi}{2}\frac{1}{\cos^2\left(\frac{\pi}{2}u\right)}$. A myriad other choices are also available. For instance, Mutný and Krause [179] use $T(u) \triangleq \cot\left(\frac{\pi}{2}(u+1)\right)$ to give

$$T'(u) = -\frac{\pi}{2}\frac{1}{\sin^2\left(\frac{\pi}{2}(u+1)\right)},$$

and

$$\int_{-\infty}^{\infty} f(\omega)\,\mathrm{d}\omega = \frac{\pi}{2}\int_{-1}^{1} \frac{f\left[\cot\left(\frac{\pi}{2}(u+1)\right)\right]}{\sin^2\left(\frac{\pi}{2}(u+1)\right)}\,\mathrm{d}u.$$

Recall that in Gauss-Hermite quadrature, our integrand of interest is $f(\omega) = \cos(\omega t)$. That is, the contribution of the spectral density $p(\omega)$ is absorbed into the weight function. In contrast, when using Gauss-Legendre quadrature, our integrand explicitly includes the contribution from the spectral density, $f(\omega) = p(\omega)\cos(\omega t)$. Therefore, we can directly incorporate non-Gaussian spectral densities $p(\omega)$ without needing to resort to importance sampling. However, unlike in Gauss-Hermite quadrature, we will not be able to isolate potentially deleterious effects of the spectral density from our approximation.

All in all, we have the Fourier feature decomposition $\boldsymbol{\varphi} : \mathbb{R} \to \mathbb{R}^{L'}$,

$$\boldsymbol{\varphi}(x) \triangleq \sqrt{\frac{\pi}{2}}\begin{bmatrix} \sqrt{\frac{\alpha_1 p\left(\tan\left(\frac{\pi}{2}\xi_1\right)\right)}{\cos^2\left(\frac{\pi}{2}\xi_1\right)}}\cos\left(\tan\left(\frac{\pi}{2}\xi_1\right)\cdot x\right) \\ \vdots \\ \sqrt{\frac{\alpha_L p\left(\tan\left(\frac{\pi}{2}\xi_L\right)\right)}{\cos^2\left(\frac{\pi}{2}\xi_L\right)}}\cos\left(\tan\left(\frac{\pi}{2}\xi_L\right)\cdot x\right) \\ \sqrt{\frac{\alpha_1 p\left(\tan\left(\frac{\pi}{2}\xi_1\right)\right)}{\cos^2\left(\frac{\pi}{2}\xi_1\right)}}\sin\left(\tan\left(\frac{\pi}{2}\xi_1\right)\cdot x\right) \\ \vdots \\ \sqrt{\frac{\alpha_L p\left(\tan\left(\frac{\pi}{2}\xi_L\right)\right)}{\cos^2\left(\frac{\pi}{2}\xi_L\right)}}\sin\left(\tan\left(\frac{\pi}{2}\xi_L\right)\cdot x\right) \end{bmatrix}.$$

The error of a Gaussian quadrature formula is as follows [242],

Gaussian quadrature error analysis

$$\int_a^b w(u)f(u)\,\mathrm{d}u - \sum_{i=1}^{L} a_i f(\xi_i) = \frac{f^{(2L)}(\theta)}{(2L)!}\langle p_L, p_L\rangle, \tag{A.21}$$

for some $\theta \in (a, b)$ where p_L is a monic orthogonal polynomial of degree L, and $\langle \cdot, \cdot \rangle$ is the scalar product associated with the weight function $w(u)$,

$$\langle p, q \rangle = \int_a^b w(u) p(u) q(u) \, \mathrm{d}u.$$

Cubature: quadrature in multiple dimensions

Let us now consider quadrature for functions of several variables,

$$f(\mathbf{u}) = f(u_1, \ldots, u_D).$$

For weight functions w that factorize as

$$w\left(\|\mathbf{u}\|_2\right) = \prod_{d=1}^{D} w(u_d), \tag{A.22}$$

we have

$$\begin{aligned}
\int_{\mathbb{R}^D} w\left(\|\mathbf{u}\|_2\right) f(\mathbf{u}) \, \mathrm{d}\mathbf{u} &= \int \cdots \int \prod_{d=1}^{D} w(u_d) f(u_1, \ldots, u_D) \, \mathrm{d}u_1 \cdots \mathrm{d}u_D \\
&= \int w(u_D) \left(\int w(u_{D-1}) \cdots \left(\int w(u_1) f(u_1, \ldots, u_D) \, \mathrm{d}u_1 \right) \cdots \mathrm{d}u_{D-1} \right) \mathrm{d}u_D \\
&\approx \sum_{i=1}^{L_D} \alpha_i^{(D)} \left(\int w(u_{D-1}) \cdots \left(\int w(u_1) f(u_1, \ldots, \xi_i^{(D)}) \, \mathrm{d}u_1 \right) \cdots \mathrm{d}u_{D-1} \right) \\
&\vdots \\
&\approx \sum_{i_1=1}^{L_1} \cdots \sum_{i_{D-1}=1}^{L_{D-1}} \sum_{i_D=1}^{L_D} \alpha_{i_1}^{(1)} \cdots \alpha_{i_{D-1}}^{(D-1)} \alpha_{i_D}^{(D)} f(\xi_{i_1}^{(1)}, \ldots, \xi_{i_{D-1}}^{(D-1)}, \xi_{i_D}^{(D)}),
\end{aligned} \tag{A.23}$$

where $\xi_i^{(d)}$ and $\alpha_i^{(d)}$ are the $L_d > 0$ abscissa and weights for quadrature along the dth dimension. In other words, for weight functions that satisfy Equation (A.22), we can decompose its multi-dimensional quadrature through repeated application of one-dimensional quadrature along each dimension.

The nested sum of Equation (A.23) can be written as a single sum over the elements of the D-ary Cartesian product of the quadrature nodes along each dimension $(\xi_1^{(1)} \cdots \xi_{L_1}^{(1)}), \ldots, (\xi_1^{(D)} \cdots \xi_{L_D}^{(D)})$, of which there are in total $\prod_{d=1}^{D} L_d$. Assuming for simplicity that $L_d = L$ for all $d = 1, \ldots, D$ and some $L > 0$, then there are a total of of L^D quadrature nodes. That is, the number of nodes grows exponentially in the input dimensionality.

The weight function in Gauss-Legendre quadrature trivially satisfies Equation (A.22). It is easy to verify that it is also satisfied by the weight function in Gauss-Hermite quadrature. Namely, for $w(u) \triangleq e^{-u^2}$, we have

$$e^{-\|\mathbf{u}\|_2^2} = e^{-\sum_{d=1}^{D} u_d^2} = \prod_{d=1}^{D} e^{-u_d^2}.$$

A.3.3.3 *Newton-Cotes Quadrature*

Let us now consider an alternative to Gaussian quadrature, namely, the quadrature rules of Newton and Cotes, which is obtained by replacing the integrand with a suitable interpolating polynomial $P(u)$. Consult the text of Stoer and Bulirsch [242] for a more complete treatment of the subject. Consider a uniform partition of the closed interval $[a, b]$ with

Newton-Cotes quadrature

$$\xi_i \triangleq a + ih,$$

and step width $h \triangleq \frac{b-a}{m}$, for some integer $m > 0$, and let P_m be the interpolating polynomial of degree m or less with

$$P_m(\xi_i) = f_i \triangleq f(\xi_i)$$

for $i = 0, 1, \ldots, m$. Lagrange's interpolation formula gives

$$P_m(u) \triangleq \sum_{i=0}^{m} f_i \mathcal{L}_i(u), \qquad \mathcal{L}_i(u) \triangleq \prod_{\substack{j=0 \\ j \neq i}}^{m} \frac{u - \xi_j}{\xi_i - \xi_j}.$$

Integration gives

$$\int_a^b P_m(u)\,\mathrm{d}u = h \sum_{i=0}^{m} \alpha_i f(\xi_i)$$

where the weights α_i are some function strictly of m, and crucially not dependent on the integrand f, nor on the boundaries of the interval, a, b.

In the case of $m = 2$, we obtain the approximation

Simpson's rule

$$\int_a^b f(u)\,\mathrm{d}u \approx \int_a^b P_2(u)\,\mathrm{d}u = \frac{h}{3}(f(\xi_0) + 4f(\xi_1) + f(\xi_2)),$$

which is commonly known as *Simpson's rule.*

Consider a step width $h > 0$ such that

$$b = a + L'h$$

for some positive even integer $L' = 2L, L > 0$. We can apply Simpson's rule to each subinterval $[\xi_{2k-2}, \xi_{2k-1}, \xi_{2k}]$, where $k = 1, \ldots, L$.

Compound Simpson's rule

$$\begin{aligned}\int_a^b f(u)\,\mathrm{d}u &= \sum_{k=1}^{L} \int_{\xi_{2k-2}}^{\xi_{2k}} f(u)\,\mathrm{d}u \\ &\approx \frac{h}{3} \sum_{k=1}^{L} [f(\xi_{2k-2}) + 4f(\xi_{2k-1}) + f(\xi_{2k})] \qquad \text{(A.24)} \\ &\triangleq \mathcal{S}[f].\end{aligned}$$

We can rearrange by odd and even terms to get

$$\mathcal{S}[f] = \frac{h}{3}\left(4\sum_{k=1}^{L} f(\xi_{2k-1}) + 2\sum_{k=1}^{L} f(\xi_{2k}) + f(\xi_0) - f(\xi_{L'})\right).$$

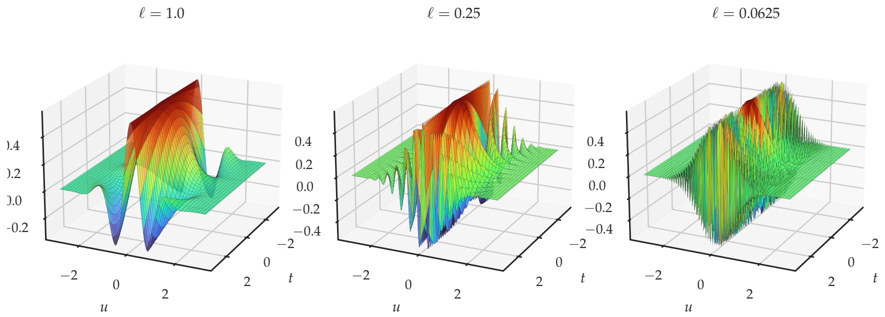

Figure A.4: The integrand $H(u) = e^{-u^2} \cos\left(\frac{\sqrt{2}ut}{\ell}\right)$ which becomes increasingly oscillatory as the lengthscale decreases $\ell = 4^{-k}$ where $k = 0, 1, 2$.

We can also write

$$\mathcal{S}[f] = \sum_{i=0}^{L'} \gamma_i f(\xi_i)$$

where

$$\gamma_i = \begin{cases} \frac{4}{3}h & i \text{ odd}, \\ \frac{2}{3}hc_i & i \text{ even}. \end{cases} \quad \text{and} \quad c_i \quad = \begin{cases} \frac{1}{2} & i = 0 \text{ or } i = L', \\ 1 & 0 < i < L'. \end{cases}$$

for $i = 0, \ldots, L'$. Denoting the odd and even terms as

$$S_{\text{odd}} \triangleq \sum_{k=1}^{L} f(\xi_{2k-1}), \tag{A.25}$$

$$\begin{aligned} S_{\text{even}} &\triangleq \sum_{k=1}^{L} f(\xi_{2k}) + \frac{f(\xi_0) - f(\xi_{L'})}{2} \\ &= \sum_{k=0}^{L} f(\xi_{2k}) - \frac{f(\xi_0) + f(\xi_{L'})}{2}, \end{aligned} \tag{A.26}$$

respectively, we can simplify Equation (A.24) as

$$\mathcal{S}[f] = \frac{h}{3}\left(4 \cdot S_{\text{odd}} + 2 \cdot S_{\text{even}}\right). \tag{A.27}$$

Newton-Cotes error analysis

A.3.3.4 *Filon's rule for highly-oscillatory integrals*

Highly-oscillatory integrals

Recall from Equation (A.19) that the integrand in which we're interested is

$$F(u) = e^{-u^2} \cos\left(\frac{\sqrt{2}ut}{\ell}\right) \tag{A.28}$$

See Figure A.4 for surface plots of this function at varying settings of ℓ. More broadly, consider integrals of the form

$$\int_a^b g(u)\cos(ut)\,\mathrm{d}u, \quad \text{and} \quad \int_a^b g(u)\sin(ut)\,\mathrm{d}u, \tag{A.29}$$

or, more generally,

$$\int_a^b g(u)e^{iut}\,\mathrm{d}u, \tag{A.30}$$

of which Equation (A.19) is clearly an instance.

Filon's rule

An extension of Simpson's rule that is aimed at dealing with highly-oscillatory integrals, known as Filon's rule [68]. Consider integrands of the form

$$f(u) = g(u)\cos(ut)$$

Let $\theta \triangleq ht$ so that ${}^1/_t = {}^h/_\theta$. We have

$$\begin{aligned}
&\int_a^b g(u)\cos(ut)\,\mathrm{d}u = \sum_{k=1}^{L} \int_{\xi_{2k-2}}^{\xi_{2k}} g(u)\cos(ut)\,\mathrm{d}u \\
&\approx \frac{1}{t}\left[\frac{4}{\theta}\left(\frac{\sin\theta}{\theta} - \cos\theta\right)\sum_{k=1}^{L} f(\xi_{2k-1})\right. \\
&+\frac{1}{\theta}\left(1+\cos^2\theta - \frac{2\cos\theta\sin\theta}{\theta}\right)\left(2\sum_{k=1}^{L} f(\xi_{2k}) + f(\xi_0) - f(\xi_{L'})\right) \\
&\left.+\left(1+\frac{\cos\theta\sin\theta}{\theta} - \frac{2\sin^2\theta}{\theta^2}\right)[g(\xi_{L'})\sin(\xi_{L'}t) - g(\xi_0)\sin(\xi_0 t)]\right] \\
&= \frac{h}{\theta^3}\left[4\left(\sin\theta - \theta\cos\theta\right)\sum_{k=1}^{L} f(\xi_{2k-1})\right. \\
&+\left(\theta(1+\cos^2\theta) - 2\cos\theta\sin\theta\right)\left(2\sum_{k=1}^{L} f(\xi_{2k}) + f(\xi_0) - f(\xi_{L'})\right) \\
&\left.+\left(\theta^2 + \theta\cos\theta\sin\theta - 2\sin^2\theta\right)[g(\xi_{L'})\sin(\xi_{L'}t) - g(\xi_0)\sin(\xi_0 t)]\right] \\
&\triangleq \mathcal{F}[g]
\end{aligned} \tag{A.31}$$

Using Equations (A.25) and (A.26), we can simplify Equation (A.31) to

$$\begin{aligned}
\mathcal{F}[g] =& h\left[\alpha\cdot S_{\text{odd}} + \beta\cdot S_{\text{even}}\right. \\
&\left.+\gamma\left(g(\xi_{L'})\sin(\xi_{L'}t) - g(\xi_0)\sin(\xi_0 t)\right)\right]
\end{aligned} \tag{A.32}$$

We can also write

$$\mathcal{F}[g] = \sum_{i=0}^{L'} \gamma_i g(\xi_i)$$
$$+\gamma\,(g(\xi_{L'})\sin(\xi_{L'}t) - g(\xi_0)\sin($$

where

$$\gamma_i = \begin{cases} \alpha h & i \text{ odd} \\ \beta h c_i & i \text{ even} \end{cases},$$

$$c_i = \begin{cases} \frac{1}{2} & i = 0 \text{ or } i = L' \\ 1 & 0 < i < L' \end{cases}.$$

for $i = 0, \ldots, L'$.

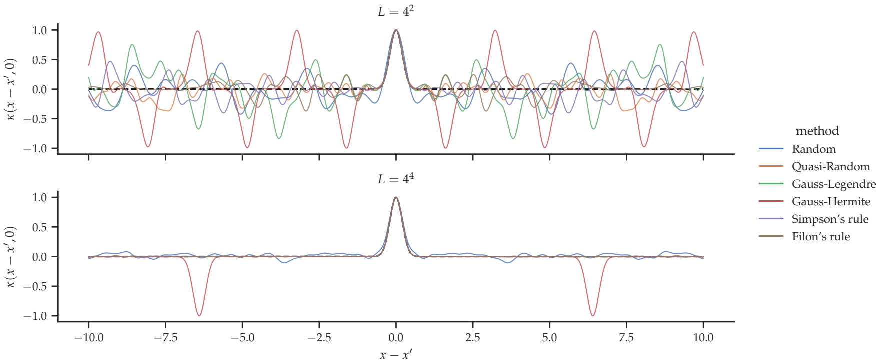

Figure A.5: SE kernel with variance 1 and lengthscale $\ell = 1/5$, and various approximations thereof, visualized on the domain $[-3, 3]$. In this domain, apart from Gauss-Hermite quadrature, the difference between quadrature methods is virtually indistinguishable for $L = 4^4 = 256$.

where

$$\alpha \triangleq \frac{4}{\theta^3}\left[\sin\theta - \theta \cdot \cos\theta\right],$$
$$\beta \triangleq \frac{2}{\theta^3}\left[\theta \cdot (1 + \cos^2\theta) - 2\cos\theta\sin\theta\right],$$
$$\gamma \triangleq \frac{1}{\theta^3}\left[\theta^2 + \theta \cdot \cos\theta\sin\theta - 2\sin^2\theta\right].$$

Now, by expanding α, β, and γ in powers of θ, we get

$$\alpha = \frac{4}{3} - \frac{2\theta^2}{15} + \frac{\theta^4}{210} - \frac{\theta^6}{11340} + \frac{\theta^8}{997920} - \frac{\theta^{10}}{129729600} + \cdots$$
$$\beta = \frac{2}{3} + \frac{2\theta^2}{15} - \frac{4\theta^4}{105} + \frac{2\theta^6}{567} - \frac{4\theta^8}{22275} + \frac{4\theta^{10}}{675675} - \cdots$$
$$\gamma = \frac{2\theta^3}{45} - \frac{2\theta^5}{315} + \frac{2\theta^7}{4725} - \frac{8\theta^9}{467775} + \frac{4\theta^{11}}{8513505} - \cdots$$

It is clear to see that as $\theta \to 0$ (or, equivalently, as $t \to 0$) we have

$$\alpha \to \frac{4}{3}, \qquad \beta \to \frac{2}{3}, \qquad \gamma \to 0.$$

In other words, Filon's rule of Equation (A.32) reduces exactly to Simpson's rule of Equation (A.27). This suggests that for sufficiently small values of t, Simpson's rule is just as good as Filon's rule when it comes to dealing with highly-oscillatory integrals. Of course, another way to look at it is that Filon's rule may actually be no better than Simpson's rule in this setting.

A.3.4 *Other Approaches*

orthogonal random features (ORF) [40, 41, 290].

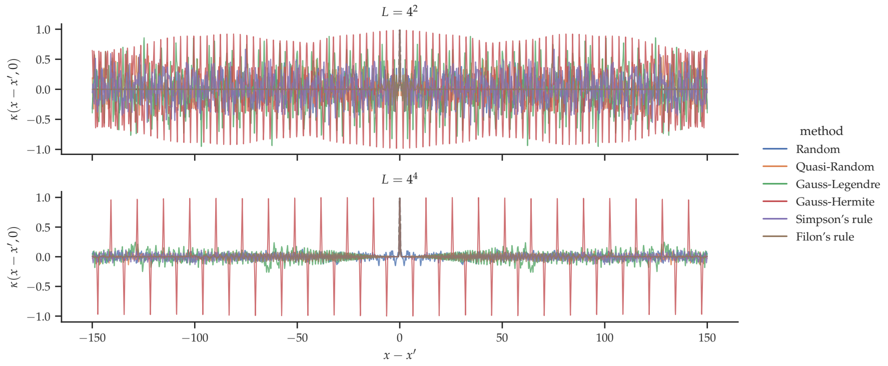

Figure A.6: SE kernel with variance 1 and lengthscale $\ell = 1/5$, and various approximations thereof, visualized on the domain $[-150, 150]$. The advantages of Filon's rule appear only to be realized when $|x - x'| > 100$ where the spurious oscillations begin to attenuate.

Let us define matrix $\mathbf{W}$ to be the collection of L frequencies sampled from the kernel's spectral density, $\mathbf{W} = \begin{bmatrix}\boldsymbol{\omega}_1 \cdots \boldsymbol{\omega}_L\end{bmatrix}^\top \in \mathbb{R}^{L \times D}$. We can write Equation (A.8) as

$$\boldsymbol{\varphi}(\mathbf{x}) \triangleq \sqrt{\frac{2}{L'}} \begin{bmatrix} \cos(\mathbf{W}\mathbf{x}) \\ \sin(\mathbf{W}\mathbf{x}) \end{bmatrix},$$

By Equation (A.8) and Table 2.1, the matrix $\mathbf{W}$ of RFF can be written as

$$\mathbf{W} = \mathbf{M}^{-\frac{1}{2}}\mathbf{G}$$

where $\mathbf{G}$ is a Gaussian random matrix. For ORF, we assume $L = D$ so that $\mathbf{G}$ is a square matrix, and set

$$\mathbf{W} \triangleq \mathbf{M}^{-\frac{1}{2}}\mathbf{S}\mathbf{Q},$$

where $\mathbf{Q}$ is the orthogonal matrix such that $\mathbf{QR} = \mathbf{G}$ for some upper triangular matrix $\mathbf{R}$, and $\mathbf{S} = \mathrm{diag}(s_1, \cdots, s_D)$ with $s_i \sim \chi_D$, where χ_D denotes the χ-distribution[1] with D degrees of freedom. The transformation by $\mathbf{S}$ has the effect of making the rows of $\mathbf{SQ}$ and $\mathbf{G}$ identically distributed.

A number of relevant approaches, such as Fastfood [137], À la Carte [289] and the Nyström approximation [280, 288], have been excluded from the scope this work.

A.4 EXPERIMENTS

A.4.1 *Prior Approximation*

We're interested in the *relative error*, defined as the Frobenius norm of the difference between the kernel's exact Gram matrix $\mathbf{K}$, and

[1] not to be confused with the χ^2-distribution

its approximation based on an Fourier feature decomposition $\mathbf{\Phi}\mathbf{\Phi}^\top$, normalized by the Frobenius norm of $\mathbf{K}$,

$$\text{relative error} \triangleq \frac{\|\mathbf{K} - \mathbf{\Phi}\mathbf{\Phi}^\top\|_F}{\|\mathbf{K}\|_F}.$$

We restrict our focus to the Fourier feature decompositions that have been outlined in this report: RFF and its phase-shifted cosine variant, QFF, specifically its variants based on Gaussian quadrature (Gauss-Legendre, Gauss-Hermite), and Newton-Cotes quadrature (Simpson's rule), QRFF with Sobol sequences, and finally, ORF. We consider a number of datasets, namely, MOTORCYCLE, IRIS, DIABETES, BOSTON, WINE, and BREAST CANCER, and look at two kernels: the SE and Matérn-5/2 kernels both with decreasing lengthscales $\ell = 4^{-k}$ for $k = 0, 1, 2$.

See Figures A.7 and A.8 for results on the SE and Matérn-5/2 kernels, respectively. For methods with an inherent source of randomness, we report the mean and 95% confidence interval across 25 repetitions.

For the SE kernel, the picture is clear: for problems of moderate dimensionality (say, $D < 5$), Gaussian quadrature methods are far more efficient than any competing method. Furthermore, in the case of $D = 1$, the deleterious effects of small lengthscales are barely noticeable. In all settings of the lengthscale, kernel decompositions based on quadrature rapidly converges to the exact kernel, and require orders of magnitude fewer features.

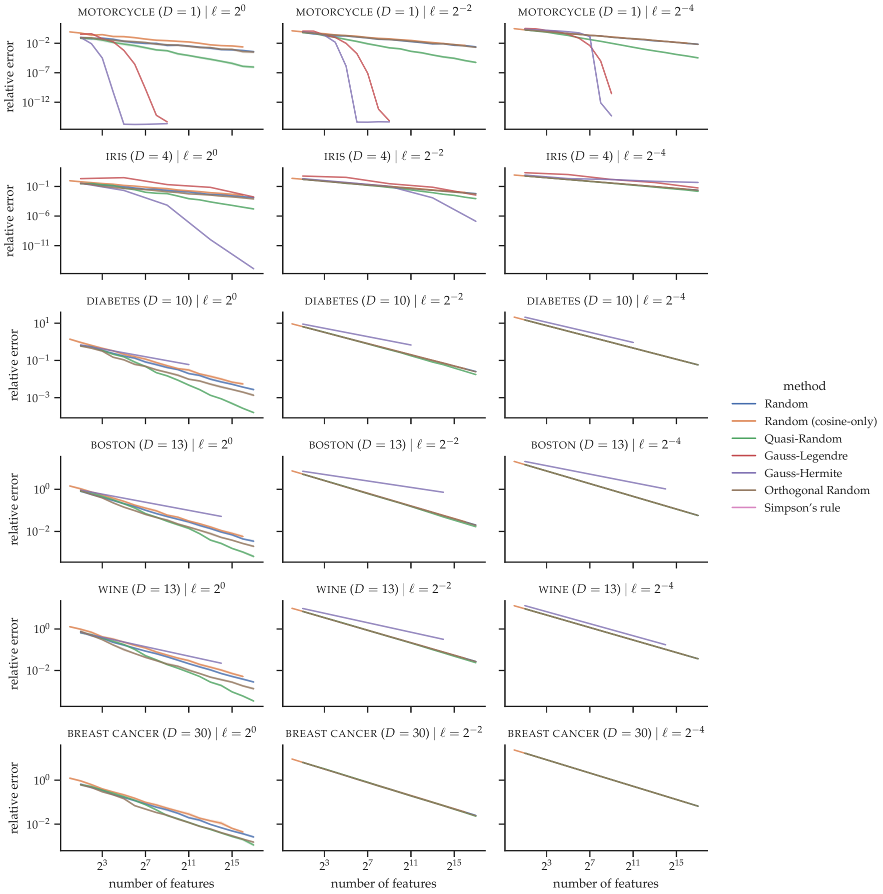

Figure A.7: Comparing the efficiency of various Fourier feature decompositions for the SE kernel.

In the case of $D = 4$, We see that Gauss-Hermite still outperforms all other methods for settings of the lengthscale above $\ell = 2^{-4}$, though perhaps less dramatically,

On the other hand, for the same dimensionality, Gauss-Legendre already begins to perform worse than all other methods. In fact, in all remaining problems (where dimensionality $D > 5$), Gauss-Legendre performs orders of magnitude worse (by orders of 10^2 or more). For the sake of readability, we've omitted the results of Gauss-Legendre and Simpson as these distort the scale of the plot dramatically.

As we move on to higher dimensions, particular in $D = 10$, the performance of Gauss-Hermite quadrature already degrades so significantly that it has become the worst of all competing methods. Furthermore, it becomes practically inapplicable beyond $D = 13$. In dimensionality $D \geq 10$ where Gauss-Hermite quadrature is still feasible, we see that its curves are truncated. This is because the errors are reported for just two settings of number of features. Recall that the number of features in multidimensional quadrature is L^D for some $L > 0$. When $D \geq 10$, for any setting of $L > 2$, this clearly becomes prohibitively large. Therefore, in such high-dimensional problems, we are restricted to setting $L \leq 2$. However, this is amounts to computing up to just two abscissa along the real line and then taking their D-ary Cartesian power to form D-dimensional quadrature nodes. Thus seen, it is no surprise that it fails to yield good results.

Outside of quadrature methods, we see that Quasi-random performs consistently well, in both low- and high-dimensional regimes. In low dimensions, it is second only to quadrature methods; in high dimensions, it consistently outperforms all competing methods. Therefore, it's safe to conclude that for the SE kernel in low-dimensional settings, one should prefer Gauss-Hermite QFF and in high-dimensional settings one should resort to QRFF.

For the Matérn-5/2 kernel, the story is quite similar, with one major exception: Quasi-random performs considerably worse and with far higher variance, particularly in high dimensions. Recall that to extend QRFF to kernels non-Gaussian spectral densities we are required to resort to importance sampling. Although this still results in an unbiased estimator, the variance is now a function of the likelihood ratio $r(\cdot)$, which is prone to taking on large values in high-dimensional settings.

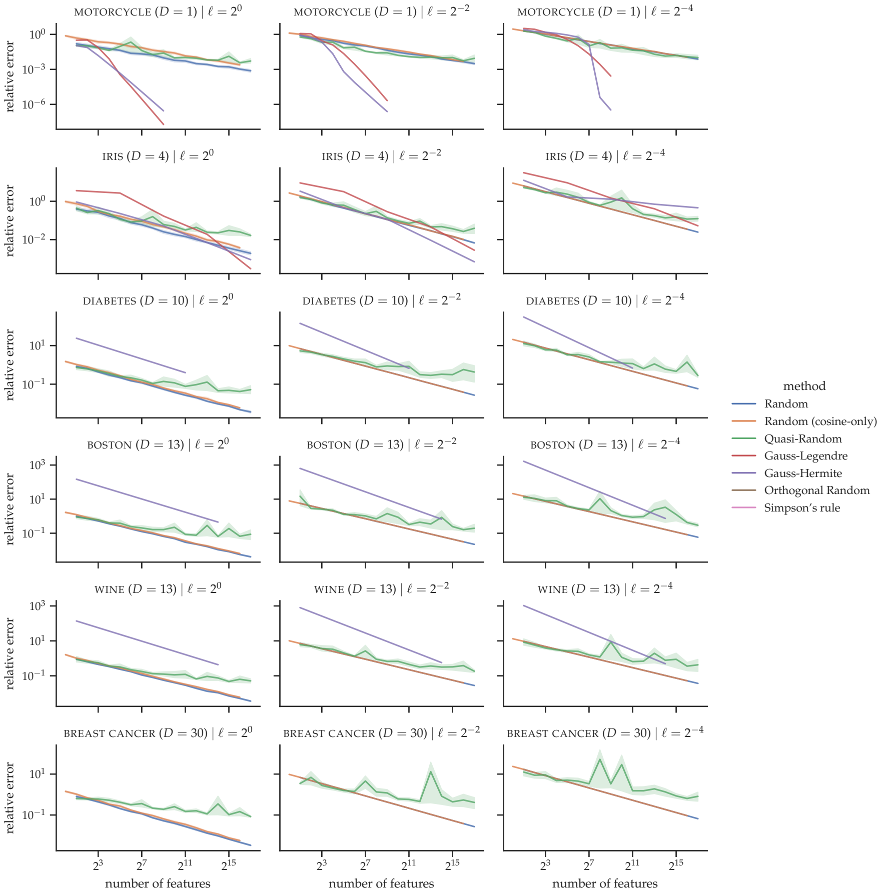

Figure A.8: Comparing the efficiency of various Fourier feature decompositions for the Matérn-5/2 kernel.

A.4.2 *Posterior Sample Approximation*

To assess the quality of posterior samples, we follow the approach of Wilson et al. [282], namely, by measuring the 2-Wasserstein distance [157] between the exact GP posterior and an empirical distribution constructed from samples.

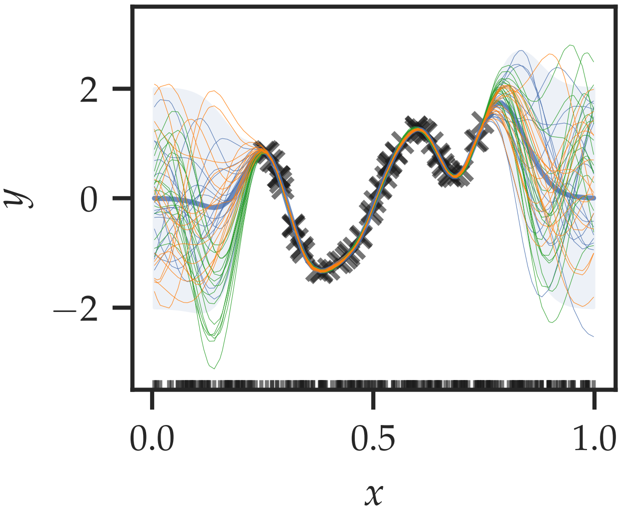

Figure A.9: An example a synthetic problem in 1D. In this illustration, there are $N = 2^6$ crosses ('$\times$') which represent the observations, and $T = 2^8$ vertical notches along the horizontal axis which represent the test, or query, points. The observations are generated using a GP with a SE kernel with unit variance and lengthscale $\ell = 2^{-4}$, while the test points are sampled uniformly at random. The *blue* curves are samples drawn from the exact GP posterior at the test points. Similarly, the *green* curves are 2^4 samples drawn from the weight-space approximate posterior, and the *orange* curves are samples generated with decoupled sampling. The kernel approximations are based on an RFF decomposition using $L = 256$ samples.

Toy datasets are synthesized as follows. The N training locations $\mathbf{X}$ are sampled uniformly at random and their corresponding observations are generated from the prior $\mathbf{y} \sim \mathcal{GP}(\mathbf{0}, \mathbf{K}_{\mathbf{ff}} + \beta^{-1}\mathbf{I})$ with observation noise variance $\beta^{-1} = 10^{-3}$, using the SE kernel with unit amplitude and lengthscales of decreasing order $\ell = 4^{-k}$ for $k = 0, 1, 2$. Likewise, the T test locations $\mathbf{X}_*$ are sampled uniformly at random from $\mathcal{U}[0,1]^{T \times D}$, where we set $T = 2^6 = 64$. The above is repeated to generate D-dimensional datasets for $D = 2^0, \ldots, 2^4$. See Figure A.9 for an example problem in one dimension.

Consistent with the findings of Wilson et al. [282], we observe the decoupled sampling scheme to be robust against variance starvation. In particular, the distance remains largely the same irrespective of the training size N. Consequently, we only report results for the setting $N = 2^7 = 128$.

To eliminate confounding factors, we restrict our attention to exact GPs using the SE kernel with known and fixed hyperparameters, i. e. the hyperparameters that were used to synthesize the observed data. In total, $2^{12} = 4,096$ samples of $\mathbf{f}_* \mid \mathbf{y}$ are used as unbiased estimates $(\hat{\boldsymbol{\mu}}_{*|\mathbf{f}}, \hat{\boldsymbol{\Sigma}}_{**|\mathbf{f}})$ of the exact posterior moments $(\boldsymbol{\mu}_{*|\mathbf{f}}, \boldsymbol{\Sigma}_{**|\mathbf{f}})$ given in Equation (A.2). The 2-Wasserstein distances are then computed based on these moments,

$$\mathcal{W}_2 \left(\mathcal{N}(\boldsymbol{\mu}_{*|\mathbf{f}}, \boldsymbol{\Sigma}_{**|\mathbf{f}}), \mathcal{N}(\hat{\boldsymbol{\mu}}_{*|\mathbf{f}}, \hat{\boldsymbol{\Sigma}}_{**|\mathbf{f}}) \right)^2 .$$

This is computed both for samples from the weight-space approximate posterior, and for samples generated with decoupled sampling, shown in Figures A.10 and A.11, respectively.

We restrict our focus to the Fourier feature decompositions that have been outlined in this chapter: RFF, QFF, specifically its variants based on Gaussian quadrature (Gauss-Legendre, Gauss-Hermite), and Newton-Cotes quadrature (Simpson's rule), QRFF with Sobol sequences, and finally, ORF.

Lastly, we report, for each combination of dimensionality D and kernel lengthscale ℓ, the mean and 95% confidence interval across 5 repetitions.

As expected, for the weight-space approximation as pictured in Figure A.10, having a tighter approximation of the Fourier feature decomposition to the kernel seems to have a large positive effect. Particularly, we see that in low dimensionalities ($D < 5$) with sufficiently large lengthscales, the distances are considerably lower when using QFF with Gauss-Hermite quadrature.

On the other hand, for the samples generated with decoupled sampling as pictured in Figure A.11, the distances are far less discernible from one another.

In the weight-space view, neither the mean nor the variance match that of the exact posterior. However, as we improve the kernel approximation (specifically, as we double the number of quadrature nodes), we observe a dramatic improvement in the approximation. In the last two panes (reading from left to right) with 2^7 and 2^8 nodes, the difference is virtually indistinguishable to the naked eye. In contrast, in decoupled pathwise sampling, we have equality *in expectation* – that is, the mean match up regardless of the quality of the kernel approximation. On the other hand, we do not have equality *in distribution*, so the variance is still dependent on the quality of the kernel approximation. We see that in the very beginning, with a few nodes, it already does a fairly good job of approximating the variance outside the regions in which the the observations are located. On the other hand, inside such regions it appears to severely underestimate the variance. What worse is that it doesn't seen to get better as we improve the kernel approximation. Indeed, doubling the number of nodes does not seem to change anything.

Increasing the number of nodes does seem to help, but only up to a point. Beyond 2^8 nodes, it is doubtful whether the approximation will improve. It is also unknowable, as this is around the limits of numerical precision. Yet, even with this amount of nodes, the understimation of the variance still persists.

We note, however, that it is difficult to draw conclusions using the 2-Wasserstein distance with empirical distributions. Beyond the numerical stability issues, note that even with samples from the exact GP drawn using the conventional location-scale transform approach, the distance based on empirical estimates are on the order of 10^{-2} (in theory, it should be 0).

Alternatively, it may be worthwhile to instead consider the NLPD of the samples under the exact GP posterior, or using the KL divergence. In particular, the KL divergence between Gaussian distributions with the same mean $\mathbf{m}$ but different covariances $\mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma}_0)$ and $\mathcal{N}(\mathbf{m}, \boldsymbol{\Sigma}_1)$ is

$$\text{KL}\left[\mathcal{N}_0 \parallel \mathcal{N}_1\right] = \frac{1}{2}\left[\ln|\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{-1}| + \operatorname{tr}\left(\boldsymbol{\Sigma}_1^{-1}(\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_1)\right)\right]. \tag{A.33}$$

Recall from Equations (A.3) and (A.6) that the covariance of a decoupled pathwise sample from a sparse GP posterior is

$$\boldsymbol{\Sigma}_0 \triangleq \boldsymbol{\Phi}_*\boldsymbol{\Phi}_*^\top - 2\mathbf{K}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}_*^\top + \mathbf{K}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}*},$$

while that of a sparse GP posterior is

$$\boldsymbol{\Sigma}_1 \triangleq \boldsymbol{\Sigma}_{**|\mathbf{u}} = \mathbf{K}_{**} - \mathbf{K}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{K}_{\mathbf{u}*}.$$

Further, the mean of both is

$$\mathbf{m} \triangleq \mu_{*|\mathbf{u}} = \mathbf{K}_{*\mathbf{u}}\mathbf{K}_{\mathbf{uu}}^{-1}\mathbf{u}.$$

Taking the KL divergence between Gaussians with these means and covariances is an attractive alternative to the 2-Wasserstein distance described above, since it is more numerically stable and can be computed analytically without resorting to empirical estimates. In fact, this author perceives no good reason to use empirical estimates, let alone the 2-Wasserstein distance, when we have the exact moments $\mathbf{m}$, $\boldsymbol{\Sigma}_0$, and $\boldsymbol{\Sigma}_1$ readily available to us.

A.5 SUMMARY

We motivated the work summarised in this chapter by showing that the quality of decoupled pathwise samples still depends crucially on the quality of the kernel approximation.

We conducted a survey of existing Fourier feature decompositions for approximating stationary kernels to provide a better understanding

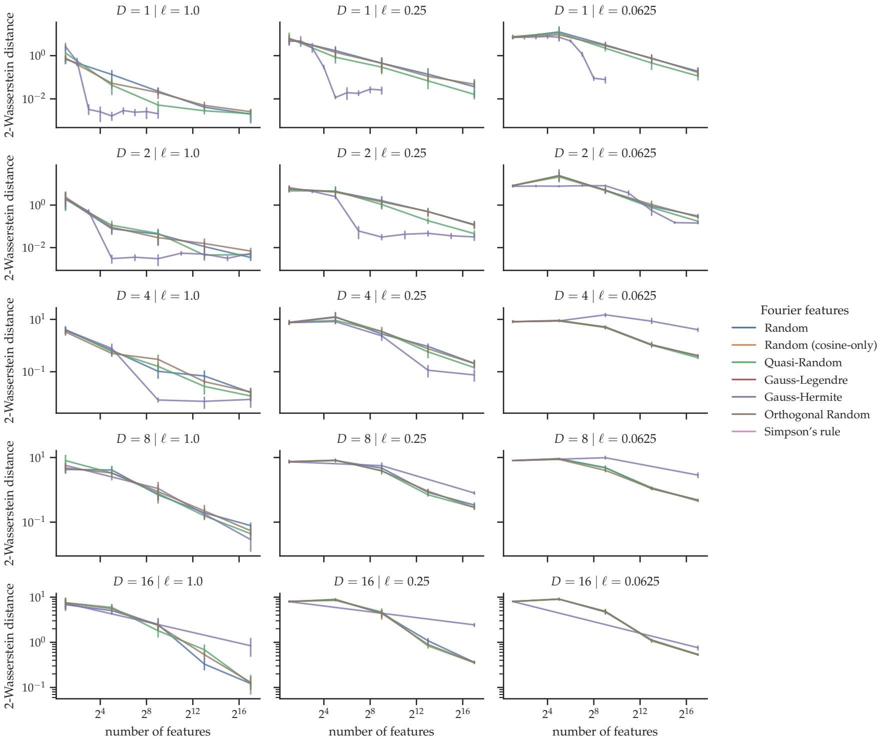

Figure A.10: Weight-space approximate posterior samples (SE kernel).

of the tightness of these various approximations. In doing so, we also made variations on existing schemes to construct new decompositions, or expanded the applicability of existing decompositions to classes of kernels beyond the SE kernel.

We also highlighted a significant shortcoming with an existing class of schemes, namely Gaussian quadrature, in dealing with small lengthscales. Small lengthscales in effect lead to highly-oscillatory integrals that are difficult to approximate. Unfortunately, efforts to ameliorate these shortcomings came up short, as it is not analytically possible to factorize the approximation into an inner product of feature maps. Furthermore, there is evidence to suggest that the benefits of more sophisticated schemes to deal with high-oscillations are only realised at scales well outside the input domains in which we're typically interested for practical purposes.

Lastly, contrary to our motivating hypothesis, we did not find a strong positive correlation between tightness of kernel approximation and quality of decoupled pathwise samples. However, we also underscored the potential flaws of the existing methods to assess sample quality, and emphasised that our empirical findings be taken with a pinch of salt. We suggest in future work that more attention be devoted to devising more principled methods for assessing sample quality.

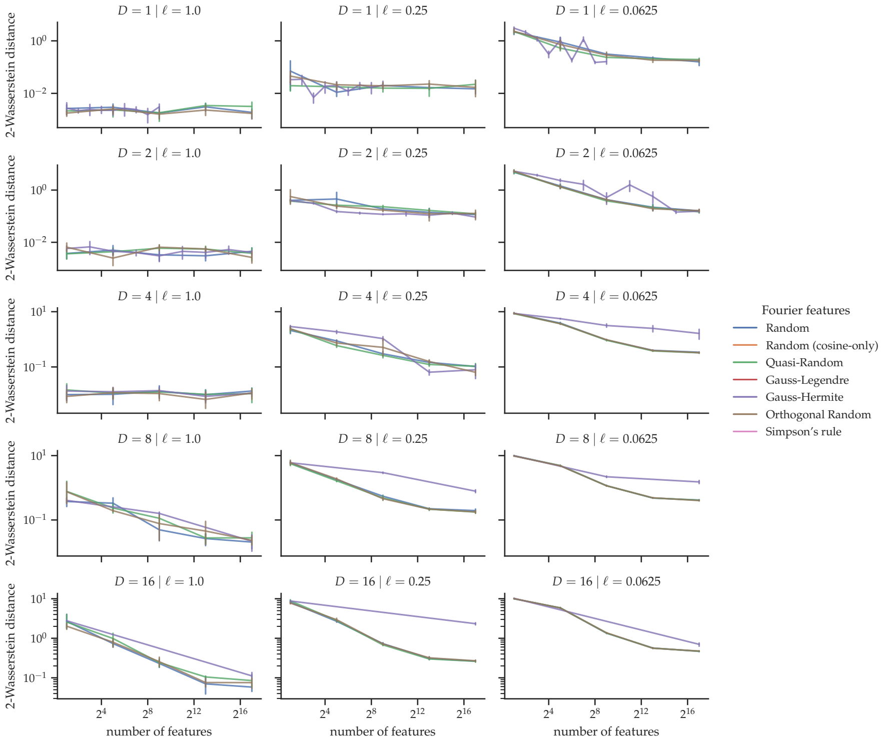

Figure A.11: Samples generated using decoupled sampling (se kernel).

ADDENDUM

A.A PRODUCT-TO-SUM IDENTITY

The product-to-sum identity, which follows as an immediate consequence of Equation (2.52), is given by

$$2\cos\alpha\cos\beta = \cos(\alpha+\beta) + \cos(\alpha-\beta). \tag{A.34}$$

A.B ZERO IN EXPECTATION

By the *law of total expectation*, we can rewrite the expectation as

$$\begin{aligned}&\mathbb{E}_{p(\boldsymbol{\omega},b)}[\cos(\boldsymbol{\omega}^\top(\mathbf{x}+\mathbf{x}')+2b)]\\ &\quad= \mathbb{E}_{p(\boldsymbol{\omega})}\left[\mathbb{E}[\cos(\boldsymbol{\omega}^\top(\mathbf{x}+\mathbf{x}')+2b) \mid \boldsymbol{\omega}]\right].\end{aligned}$$

For notational convenience, we set $\theta \triangleq \boldsymbol{\omega}^\top(\mathbf{x}+\mathbf{x}')$. The inner expectation evaluates to

$$\begin{aligned}\mathbb{E}[\cos(\theta+2b) \mid \boldsymbol{\omega}] &= \int_0^{2\pi} \cos(\theta+2b)p(b)\,\mathrm{d}b\\ &= \frac{1}{2\pi}\int_0^{2\pi} \cos(\theta+2b)\,\mathrm{d}b\\ &= \frac{1}{2\pi}\sin(\theta+2b)\Big|_0^{2\pi}\\ &= \frac{1}{2\pi}[\sin(\theta+4\pi)-\sin(\theta)] = 0\end{aligned}$$

since the sine function is 2π-periodic, i. e., $\sin(\theta+2\pi\cdot k) = \sin(\theta)$ for any integer k.