Integrated Fourier Features for Fast Sparse Variational Gaussian Process Regression

Talay M Cheema
Department of Engineering
University of Cambridge, UK
tmc49@cam.ac.uk

Abstract

Sparse variational approximations are popular methods for scaling up inference in Gaussian processes to larger datasets. For N training points, exact inference has $O(N^3)$ cost; with $M << N$ features, sparse variational methods have $O(NM^2)$ cost. Recently, methods have been proposed using harmonic features; when the domain is spherical, the resultant method has $O(M^3)$ cost, but in the common case of a Euclidean domain, previous methods do not avoid the $O(N)$ scaling and are generally limited to a fairly small class of kernels. In this work, we propose integrated Fourier features, with which we can obtain $O(M^3)$ cost, and the method can easily be applied to any covariance function for which we can easily evaluate the spectral density. We provide convergence results, and synthetic experiments showing practical performance gains.

1 Introduction

Gaussian processes (GPs) are probabilistic models widely used in machine learning applications where predictive uncertainties are important – for example, in active learning, Bayesian optimisation, or for risk-aware forecasts. The hyperparameters of these models are often trained by maximising the marginal likelihood, so it is important that this quantity can be evaluated fairly cheaply. Yet, for N datapoints, the time cost is $O(N^3)$, which is prohibitively large for many datasets of interest.

One popular approach for improving scalability is to use a sparse variational approximation, wherein $M < N$ inducing features are used as a compact representation, and a lower bound of the log marginal likelihood is maximised, which reduces the cost to $O(NM^2)$. In the conjugate setting, where the noise model is additive white Gaussian, the variational distribution is available in closed form (Titsias, 2009), and for various different choices of inducing features, under reasonable conditions and for commonly used covariance functions, it can be shown that the lower bound converges to the log marginal likelihood with $M << N$ as $M, N \to \infty$ (Burt et al., 2019, 2020).

The scaling with N is problematic – one popular way to avoid this is to use batches of data (Hensman et al., 2015). But in the conjugate setting, this greatly increases the number of iterations needed, as we must directly learn the variational distribution which is otherwise available in closed form. In the case of zonal kernels on spherical domains, the $O(N)$ part of the computation can be taken outside of the optimisation loop (Dutordoir et al., 2020). In this work, we propose integrated Fourier features (IFF), a method which achieves the same in the Euclidean case.

In Section 2 we review variational GP regression in the conjugate setting and related work. In Section 3 we present our IFF method, the complexity analysis, and the main convergence results. Finally in Section 4 we show experimentally our method does indeed train faster that standard sparse GP regression on synthetic data.

2022 NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems.

2 Background

In the conjugate setting, the probabilistic model for Gaussian process regression is

$$f \sim \mathcal{GP}(0, k) \quad y_n = f(x_n) + \rho_n \quad \rho_n \sim \mathcal{N}(0, \sigma^2) \tag{1}$$

for $n \in \{1 : N\}$, with $x_n \in \mathbb{R}^D$ and $\rho_n, y \in \mathbb{R}$, with the covariance function $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}_{\geq 0}$ symmetric and positive definite. We use K_{xx} for the $N \times N$ matrix with $[K_{xx}]_{nn'} = k(x_n, x_{n'})$. The posterior predictive at some collection of inputs x_* and marginal likelihood are as follows, where $x = x_{1:N}, y = y_{1:N}$ (Rasmussen & Williams, 2006, Chapter 2).

$$p(f(x_*)|x, y) = \mathcal{N}(f(x_*)|K_{*x}(K_{xx} + \sigma^2 I)^{-1} y, K_{**} - K_{*x}(K_{xx} + \sigma^2 I)^{-1} K_{x*}) \tag{2}$$

$$e^{\mathcal{L}} = p(y|x) = \mathcal{N}(y|0, K_{xx} + \sigma^2 I). \tag{3}$$

We optimise the latter with respect to the covariance function's parameters. For each evaluation of the (log) marginal likelihood, we need to compute the (log) determinant $|K_{xx} + \sigma^2 I|$ and the quadratic form $y^\top (K_{xx} + \sigma^2 I)^{-1} y$, both of which have $O(N^3)$ computational cost in general.

For the variational approximation, we construct an approximate posterior $q(f) = \int p(f|u)q(u)du \approx p(f|y)$ where $u = u_{1:M}$ is a collection of inducing features with prior distribution $p(u)$. We maximise a lower bound on the log marginal likelihood (D_{KL} is the KL divergence).

$$\mathcal{F} = \int q(f) \log \frac{p(y, f, u)}{q(f)} df = \mathcal{L} - D_{KL}(q(f)||p(f)) \leq \mathcal{L} \tag{4}$$

Now, u_m is chosen to be a linear functional ϕ_m of f (Lázaro-Gredilla & Figueiras-Vidal, 2009), denoted $\langle \phi_m, f \rangle \in \mathbb{C}$ with associated parameter z_m, in order that u is Gaussian a priori. For example, in standard sparse GP regression, $u_m = f(z_m)$. Let ϕ_m^* be such that $\langle \phi_m^*, f \rangle = \langle \phi_m, f \rangle^*$ and K such that $[K\phi_m](x_*) = \langle \phi_m^*, k(x_*, \cdot) \rangle$. Then (Bogachev, 1998, Chapter 2; Lifshits, 2012)

$$\langle \phi_m, f \rangle \sim \mathcal{N}(0, \langle \phi_m, K\phi_m \rangle), \quad \mathbb{E}[\langle \phi_m, f \rangle \langle \phi_{m'}, f \rangle] = \langle \phi_m, K\phi_{m'} \rangle, \quad \mathbb{E}[f(x_*)\langle \phi_m, f \rangle] = \langle \phi_m^*, k(x_*, \cdot) \rangle$$

and for convenience define

$$c(z_m, x_*) = \langle \phi_m^*, k(x_*, \cdot) \rangle = c^*(x_*, z_m) = \langle \phi_m, k(\cdot, x_*) \rangle \tag{5}$$

$$\bar{k}(z_m, z_{m'}) = \langle \phi_m, K\phi_{m'} \rangle = \langle \phi_m, c(z_m, x_*) \rangle = \langle \phi_m^*, c(x_*, z_m) \rangle \tag{6}$$

and the corresponding matrices $C, \bar{K}$ according to similar convention as used for K. Then $p(u) = \mathcal{N}(0, \bar{K}_{zz})$, and the optimal $q(u)$ is available in closed form as

$$q(u) \sim \mathcal{N}(\mu_u, \Sigma_u) \text{ with } \quad \Sigma_u^{-1} = \bar{K}_{zz}^{-1}(\bar{K}_{zz} + \sigma^{-2} C_{zx} C_{zx}^*)\bar{K}_{zz}^{-1}, \; \mu_u = \sigma^{-2}\Sigma_u C_{zx} y \tag{7}$$

with corresponding training objective (Titsias, 2009)

$$\mathcal{F}(\mu_u, \Sigma_u) = \log \mathcal{N}(y|0, \; C_{zx}^* \bar{K}_{zz}^{-1} C_{xz} + \sigma^2 I) - \frac{1}{2}\sigma^{-2} \mathrm{tr}(\bar{K}_{zz} - C_{zx}^* \bar{K}_{zz}^{-1} C_{xz}) \tag{8}$$

wherein the key matrix to invert and compute the log determinant of is of the form $(\bar{K}_{zz} + \sigma^2 C_{zx} C_{zx}^*)$ (by rearranging the log-normal terms), which is only $M \times M$. The dominant cost is $O(NM^2)$ to form $C_{zx} C_{zx}^*$. In the classic case of sparse regression, $c = \bar{k} = k$. By choosing the linear functionals carefully, we aspire to simplify the computations.

Fourier features When k is stationary, the Fourier features $\langle \phi_m, f \rangle = \int f(x) e^{-i2\pi x z_m} dx$ seem appealing. Indeed, this yields independent features, but with infinite variance (Lifshits, 2012, Chapter 3), so unsuitable for conditioning. Modifications to Fourier features include applying a Gaussian window (Lázaro-Gredilla & Figueiras-Vidal, 2009) which gives finite variance but highly co-dependent features, and Variational Fourier Features (VFF) (Hensman et al., 2017) which sets $\langle \phi_m, f \rangle$ to an RKHS inner product between the harmonics and f on a compact subset of $\mathbb{R}^D$. This gives diagonal + low-rank structure in $\bar{K}_{zz}$ for low order Matérn kernels, but this is not easily extended to other covariance functions. Generally, these and comporable methods suffer a significant bottleneck due to the $O(NM^2)$ scaling (Burt et al., 2020).

Spherical harmonics Recently, Dutordoir et al (2020) used spherical harmonic features for zonal kernels on the sphere. In this case the inducing features are well defined and independent; moreover, the $O(NM^2)$ part of the work can be taken outside of the optimisation loop, reducing the cost to $O(M^3)$. We seek to replicate this property in the Euclidean case.

3 Integrated Fourier features

We propose to average Fourier features over a finite width, with disjoint intervals, to produce tractable features. We now show this eliminates the infinite variance issue. We use $s(\xi) = \int k(r)e^{-i2\pi r}dr$ for the spectral density of k, where $k(r) = k(x, x+r)$. Furthermore, we scale the integrand down by $\sqrt{s}$ and assume that the spectral density is approximately constant over the integration width. We focus on 1D inputs for now.

$$\langle \phi_m, f \rangle = \varepsilon^{-1} \int_{z_m-\varepsilon/2}^{z_m+\varepsilon/2} s^{-\frac{1}{2}}(\xi) \int f(x)e^{-i2\pi\xi x}dxd\xi \tag{9}$$

$$c(x_*, z_m) = \langle \phi_m, k(\cdot, x_*) \rangle = \varepsilon^{-1} \int_{z_m-\varepsilon/2}^{z_m+\varepsilon/2} s^{-\frac{1}{2}}(\xi) \int k(x, x_*)e^{-i2\pi\xi x}\, dxd\xi = \varepsilon^{-1} \int_{z_m-\varepsilon/2}^{z_m+\varepsilon/2} s^{-\frac{1}{2}}(\xi)s(\xi)e^{-i2\pi\xi x_*}d\xi$$
$$\approx \sqrt{s(z_m)}e^{-i2\pi z_m x_*} \tag{10}$$

$$\bar{k}(z_m, z_{m'}) = \langle \phi^*_m, c(\cdot, z_{m'}) \rangle = \varepsilon^{-2} \int_{z_m-\varepsilon/2}^{z_m+\varepsilon/2} \int_{z_{m'}-\varepsilon/2}^{z_{m'}+\varepsilon/2} s^{-\frac{1}{2}}(\xi)s^{\frac{1}{2}}(\xi') \int e^{i2\pi x_*(\xi-\xi')}\, dx_* d\xi' d\xi$$
$$= \varepsilon^{-2} \int_{z_m-\varepsilon/2}^{z_m+\varepsilon/2} \int_{z_{m'}-\varepsilon/2}^{z_{m'}+\varepsilon/2} s^{-\frac{1}{2}}(\xi)s^{\frac{1}{2}}(\xi')\delta(\xi-\xi')d\xi' d\xi = \varepsilon^{-1}\delta_{m-m'} \tag{11}$$

In the last case, we use the sifting property of the delta function to elimminate the inner integral. This yields all the terms we need for the Equation (8).Now, let S_z be a diagonal matrix with $[S_z]_{mm} = s(z_m)$, and let $\tilde{C}_{zx} = S_z^{\frac{1}{2}} C_{zx}$. Then the matrix we need to invert or take the log determinant of is

$$\bar{K}_{zz} + \sigma^2 C_{zx}C^*_{zx} = \varepsilon^{-1}I + \sigma^2 C_{zx}C^*_{zx} = S_z^{\frac{1}{2}}(\varepsilon^{-1}S_z^{-1} + \tilde{C}_{zx}\tilde{C}^*_{zx})S_z^{\frac{1}{2}}. \tag{12}$$

Crucially, subject to the approximation in Equation (10), $\tilde{C}_{zx}$ does not depend on the hyperparameters, so if we keep the inducing frequencies z and the parameter ε fixed, then the $O(NM^2)$ work of computing $\tilde{C}_{zx}\tilde{C}^*_{zx}$ can be done once outside of the loop. Indeed, all $O(N)$ work can be done outside of the loop (Appendix A), giving a cost per iteration of $O(M^3)$.

Despite the approximations we have introduced, the IFF lower bound converges to the log marginal likelihood as $M \to \infty$. The scaling of M with N depends on two parts: how quickly the spectral density decays, and how well our assumption that the integrand is constant over the interval holds up. The latter will dominate asymptotically unless the density has very heavy tails, leading approximately to $M \in O(\sqrt{N})$. See Appendix B for full results and proofs.

Theorem 3.1 (Convergence for large N with sub-Gaussian density)**.** *Assume that s has bounded first and second derivatives everywhere, and that we have a tail bound $\int_\xi^\infty \tilde{s}(\xi')d\xi' \in O(e^{-\xi})$. Select the inducing features ε apart centred on the origin, that is $z_m = (-(M+1)/2 + m)\varepsilon$, with M even. Let $\varepsilon \in O(M^{-1+a})$ for some $a \in (0, 1)$. Then if y is sampled from Equation (1) for any $\Delta, \delta > 0$, there exists $M_0\alpha > 0$ such that for $M \geq M_0$*

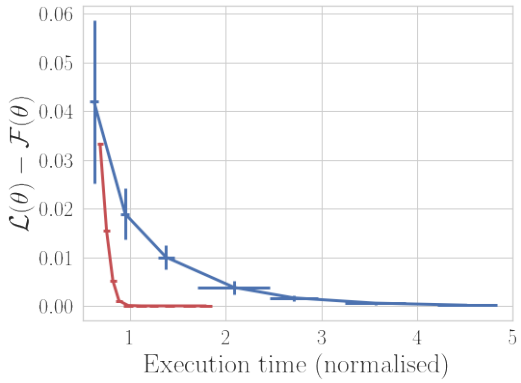
$$Pr[D_{KL}(q(f)\|p(f|y))/N > \Delta/N] \leq \delta \iff M \leq \left(\frac{\alpha}{\Delta\delta}N\right)^{\frac{1}{2-3a}}$$

Note that taking $a \to 0$ leads to $M \in O(\sqrt{N})$. In practice, we use the real valued features which are cosine and sine components for positive frequencies only, which is essentially equivalent to the above.

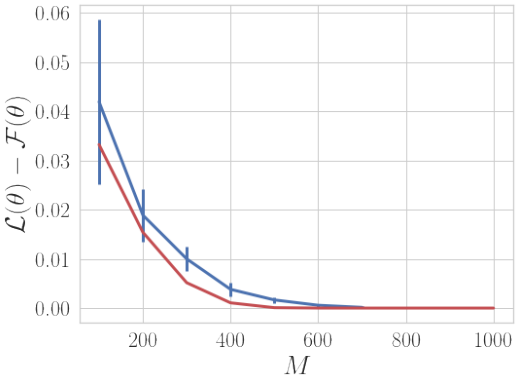
In higher dimensions, we can use a regular grid of frequencies. We modify the definition to $\langle \phi_m, f \rangle = \varepsilon^{-D} \int_{z_{m1}-\varepsilon/2}^{z_{m1}+\varepsilon/2} \ldots \int_{z_{mD}-\varepsilon/2}^{z_{mD}+\varepsilon/2} s^{-\frac{1}{2}}(\xi) \int f(x)e^{-i2\pi\xi x}dxd\xi_D \ldots d\xi_1$. Then, if we require P features in one dimension, we will need $M \in O(P^D)$ in D dimensions, which will work for low values of D but become disadvantageous for large D.

4 Experiments

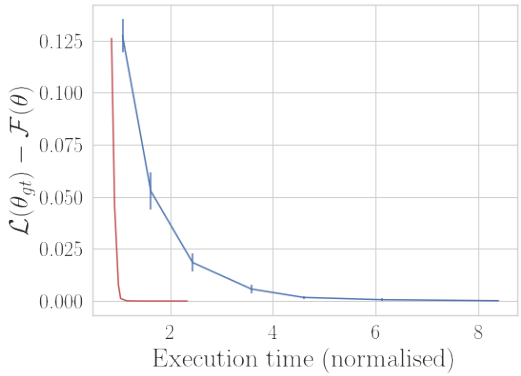
When adding more features, we can cover a higher proportion of the prior spectral density or reduce ε. In Theorem 3.1, note we have ε almost at $O(M^{-1})$, which suggests this part dominates, and



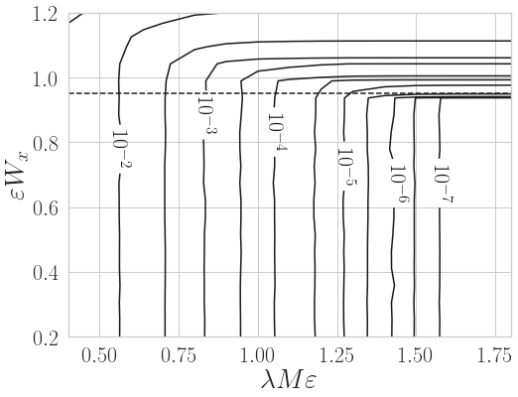
(a) KL



(b) KL



(c) Gap to ground-truth log marginal likelihood



(d) KL for different settings of M, ε.

Figure 1: (a-c) Comparing standard sparse Gaussian process regression (blue) to IFF (orange). For the standard approach, we initialise the inducing points with K-means and plot means and two standard deviations over 20 runs. The KL is generally lower with IFF for a given M (c), but the runtime is much lower (a, b). (d) Contour plot of the KL using the groundtruth hyperparameters for different settings of M, ε. W_x is the range of the data, and λ is the lengthscale. The dashed line marks $\varepsilon W_x = 0.95$.

the approximation in the cross-covariance (Equation (10)) should be valid for $\varepsilon << 1/(\pi|x_{\max}|)$ (Appendix B). We verify this numerically (Figure 1d) by plotting the gap between the IFF bound and the log marginal likelihood, varying the bandwidth covered and the size of ε relative to the inverse data width ($\approx 2x_{\max}$). The model and data generating process use a squared exponential kernel with lengthscale λ, so $s(\xi) \propto \mathcal{N}(0, (2\pi\lambda)^{-2})$. We see that in practice, as long as ε^{-1} is below around 95% of the inverse of the data width, the KL is not very sensitive to its value.

Now, we train with $N = 10^4$ using standard inducing points, kept fixed throughout and initialised by running K-means on the data. We compare this to the IFF method, with ε set to 95% of the data width. We initialise the hyperparameters to the same value in each case and optimise using LBFGS. IFF generally has slightly lower KL at the learnt optimum for any M (Figure 1b), but because the $O(NM^2)$ work is done only once, it is much faster to run (Figures 1a and 1c).

5 Conclusions

Integrated Fourier features offer a promising method for fast Gaussian process regression for large datasets. A key feature is that the $O(NM^2)$ part of the computation can be done outside of the loop, which leads to significant cost savings. These methods are immediately applicable to challenging spatial regression tasks, but a significant limitation is the need to increase M exponentially in D; alleviating this, and applying to real-world data, are important directions for future work.

References

Vladimir I Bogachev. Gaussian Measures. American Mathematical Society, 1998. ISBN 978-0-8218-1054-5.

David Burt, Carl Edward Rasmussen, and Mark van der Wilk. Rates of convergence for sparse variational Gaussian process regression. In 36^{th} International Conference on Machine Learning (ICML), 2019.

David R Burt, Carl Edward Rasmussen, and Mark van der Wilk. Variational orthogonal features, 2020. URL https://arxiv.org/abs/2006.13170.

Vincent Dutordoir, Nicolas Durrande, and James Hensman. Sparse Gaussian processes with spherical harmonic features. In 37^{th} International Conference on Machine Learning (ICML), 2020. URL https://proceedings.mlr.press/v119/dutordoir20a.html.

James Hensman, Alexander G de G Matthews, and Zoubin Ghahramani. Scalable variational Gaussian process classification. In 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS), 2015.

James Hensman, Nicolas Durrande, and Arno Solin. Variational Fourier features for Gaussian processes. Journal of Machine Learning Research, 2017.

Miguel Lázaro-Gredilla and Aníbal Figueiras-Vidal. Inter-domain Gaussian processes for sparse inference using inducing features. In 26^{th} Conference on Neural Information Processing Systems (NeurIPS), 2009.

Mikhail A Lifshits. Lectures on Gaussian Processes. Springer, 2012. ISBN 978-3-642-24938-9.

Carl Edward Rasmussen and Christopher K I Williams. Gaussian Processes for Machine Learning. MIT Press, 2006. ISBN 978-0-262-18253-9.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS), 2009.

A Computation

Recall the collapsed objective.

$$\mathcal{F}(\mu_u, \Sigma_u) = \log \mathcal{N}(y|0,\ C_{zx}^* \bar{K}_{zz}^{-1} C_{xz} + \sigma^2 I) - \frac{1}{2}\sigma^{-2}\mathrm{tr}(K_{xx} - C_{zx}^* \bar{K}_{zz}^{-1} C_{xz}) \tag{13}$$

$$= -\frac{1}{2}\log|C_{zx}^* \bar{K}_{zz}^{-1} C_{xz} + \sigma^2 I| - \frac{1}{2} y^\top (C_{zx}^* \bar{K}_{zz}^{-1} C_{xz} + \sigma^2 I)^{-1} y - \frac{1}{2}\sigma^{-2}\mathrm{tr}(K_{xx} - C_{zx}^* \bar{K}_{zz}^{-1} C_{xz}) \tag{14}$$

With $\tilde{C}_{zx} = \bar{S}_z^{-\frac{1}{2}} C_{zx}$, $\bar{y} = \bar{S}_z^{-\frac{1}{2}} C_{zx} y$, $\sum_n y_n^2 = \nu^2$, we apply the Woodbury identity to the inverse in the quadratic form.

$$\begin{aligned}
(C_{zx}^* \bar{K}_{zz}^{-1} C_{xz} + \sigma^2 I)^{-1} &= \sigma^{-2} I - \sigma^{-4} C_{zx}^* (\bar{K}_{zz} + \sigma^{-2} C_{zx} C_{zx}^*)^{-1} C_{zx} \\
&= \sigma^{-2} I - \sigma^{-4} \tilde{C}_{zx}^* (\varepsilon^{-D} S_z^{-1} + \sigma^{-2} \tilde{C}_{zx} \tilde{C}_{zx}^*)^{-1} \tilde{C}_{zx} \\
\implies y^\top (C_{zx}^* \bar{K}_{zz}^{-1} C_{xz} + \sigma^2 I)^{-1} y &= \sigma^{-2}\nu^2 - \sigma^{-4} \bar{y}^\top (\varepsilon^{-D} S_z^{-1} + \sigma^{-2} \tilde{C}_{zx} \tilde{C}_{zx}^*)^{-1} \bar{y}
\end{aligned}$$

For the log determinant, we can use the matrix determinant lemma.

$$\begin{aligned}
|\sigma^2 I + C_{zx}^* \bar{K}_{zz}^{-1} C_{zx}| &= |\bar{K}_{zz} + \sigma^{-2} C_{zx}^* C_{zx}||\bar{K}_{zz}|^{-1} |\sigma^2 I_n| \\
&= |\varepsilon^D S_z||\varepsilon^{-D} S_z^{-1} + \sigma^{-2} \tilde{C}_{zx}^* \tilde{C}_{zx}||\sigma^2 I_n|
\end{aligned}$$

Finally, we write down the trace directly using the fact that $\bar{K}_{zz}$ is diagonal.

$$\mathrm{tr}(K_{xx} - C_{zx}^* \bar{K}_{zz}^{-1} C_{xz}) = \sum_n (k(x_n, x_n) - \varepsilon^D \sum_m s(z_m) = N(1 - \varepsilon^D \sum_m s(z_m))$$

Combining the above, we get an easy to evaluate expression, with $A = \varepsilon^D S_z^{-1} + \sigma^{-2} \tilde{C}_{zx} \tilde{C}_{zx}^*$.

$$\mathcal{F} = -\frac{1}{2}\left(\sum_m \log s(z_m) - MD\log\varepsilon + N\log\sigma^2 + \log|A| - \sigma^{-2}\nu^2 + \sigma^{-4}\bar{y}^\top A^{-1}\bar{y} + N\sigma^{-2}(v - \sigma^{-2}\varepsilon^D \sum_m s(z_m))\right) \tag{15}$$

Notably, $\nu^2 \in \mathbb{R}$ depends only on y, so can be precomputed and stored with only $O(N)$ cost, and $\bar{y} \in \mathbb{R}^M$ depends only on x, y and z, so can also be precomputed and stored with $O(NM)$ cost, as with A. For large N, we split the data into chunks of 10 000 to save memory. For direct inversion, rotating by $S_z^{\frac{1}{2}}$ is advisable to improve the condition number.

The prediction equation is

$$q(f(x_*)|x_*, x, y) = \int p(f(x_*)|x, z, u) q(u) du = \mathcal{N}(f(x_*)|C_{z*}^* \bar{K}_{zz}^{-1} \mu_u,\ K_{**} - C_{z*}^* \bar{K}_{zz}^{-1} C_{*z} + C_{z*}^* \bar{K}_{zz}^{-1} \Sigma_u \bar{K}_{zz}^{-1} C_{z*}) \tag{16}$$

where $*$ stands for x_* in the subscripts, and μ_u, Σ_u are given in Equation (7).

B Convergence

Let

$$t = \mathrm{tr}(K_{xx} - C_{zx}^* \bar{K}_{zz} C_{xz}).$$

We use bounds on the average KL based on t, which converges to zero, and then apply Markov's inequaity for the final result. The KL bounds are that, if y is distributed according to Equation (1), then

$$\mathbb{E}_y[D_{KL}(q(f)\|p(f|y))] \leq \frac{t}{\sigma^2} \tag{17}$$

and if instead we consider y to be fixed and assume that $x \sim p(x)$,

$$\mathbb{E}_x[D_{KL}(q(f)\|p(f|y))] \leq \frac{\mathbb{E}_x[t]}{2\sigma^2}\left(1 + \frac{\|y\|_2^2}{\sigma^2}\right). \tag{18}$$

Throughout the treatment below, we assume $k = v\sigma^2\tilde{k}$ with $\tilde{k}(x,x) = 1$ with corresponding spectral density $\tilde{s}$ and $v > 0$, that $\int_\rho^\infty ... \int_\rho^\infty \tilde{s}(\xi)d\xi_1...d\xi_D \leq \beta\rho^{-q}$ for some $\beta, q > 0$ for $\rho > 0$, and that $M^{1/D}$ is an even integer. Moreover, we assume that the derivative of the spectral density has a lower bound

$$\frac{ds(\xi)}{d\xi} \leq 2L\sqrt{\tilde{s}(\xi)} \implies \frac{d\sqrt{s(\xi)}}{d\xi} \leq L$$

where the latter follows wherever $s(\xi) \neq 0$, and that the second derivative is also bounded. Finally, let $z_m = (-(M+1)/2 + m)\varepsilon$ in one dimensions, and an analogous regular grid in higher dimensions. In the rest of this section, we refer to these are the standard assumptions.

Lemma B.1. *Under the standard assumptions, there exists $\varepsilon_0 > 0$ such that for all $\varepsilon < \varepsilon_0$*

$$c(z_m, x_n)c^*(z_m, x_n) \geq s(z_m)(1 + O(\varepsilon^{2D})). \tag{19}$$

Proof. We consider the 1D case. By Taylor's theorem, $\sqrt{s(\xi)} \geq \sqrt{s(z_m)} + L|\xi - z_m|$. Then,

$$\begin{aligned} c(z_m, x_n) &= \varepsilon^{-1}\int_{z_m-\varepsilon/2}^{z_m+\varepsilon/2} \sqrt{s(\xi)}e^{-i2\pi\xi x_n}d\xi & (20)\\ &\geq \varepsilon^{-1}\int_{z_m-\varepsilon/2}^{z_m+\varepsilon/2} (\sqrt{s(z_m)} + L|\xi - z_m|)e^{-i2\pi\xi x_n}d\xi & (21)\\ &= \sqrt{s(z_m)}e^{-i2\pi z_m x_n}\left(\mathrm{sinc}(\pi\varepsilon x_n) + \frac{\varepsilon L}{2}\left(\mathrm{sinc}(\pi\varepsilon x_n) - \frac{1}{2}\mathrm{sinc}^2\left(\pi\frac{\varepsilon}{2}x_n\right)\right)\right) & (22)\end{aligned}$$

Here $\mathrm{sinc}x = \sin(x)/x$. Hence,

$$\begin{aligned} c(z_m, x_n)c^*(z_m, x_n) \geq s(z_m)\Bigg(\left(1 + \varepsilon L + \frac{1}{4}\varepsilon^2L^2\right)\mathrm{sinc}^2(\pi\varepsilon x_n) - \frac{\varepsilon L}{2}(1 + \frac{\varepsilon L}{2})\mathrm{sinc}(\pi\varepsilon x_n)\mathrm{sinc}^2(\pi\frac{\varepsilon}{2}x_n)\\ + \frac{\varepsilon^2L^2}{16}\mathrm{sinc}^4(\pi\frac{\varepsilon}{2}x_n)\Bigg). \end{aligned} \tag{23}$$

But for $x < 1$, $\mathrm{sinc}(x) \in [1 - x^2/6, 2^{1/3}(1 - x^2/6)]$. We apply a lower bound to the positive terms and an upper bound to the negative terms. If $L \geq 0$ use $1 - x^2/6, 2^{1/3}$ respectively. If $L < 0$ use $(1 - x^2/6)/2, 1$. Then the ε^1 terms cancel, and indeed

$$c(z_m, x_n)c^*(z_m, x_n) \geq s(z_m)(1 + O(\varepsilon^2)) \tag{24}$$

for $\varepsilon < 1/(\pi x_n)$. In higher dimensions, $c(z_m, x_n) = \varepsilon^{-D}\int_{z_{m1}-\varepsilon/2}^{z_{m1}+\varepsilon/2} ... \int_{z_{mD}-\varepsilon/2}^{z_{mD}+\varepsilon/2} \sqrt{s(\xi)}e^{-i2\pi\xi^\top x_n}d\xi_D...d\xi_1$. By following the same argument, we arrive at

$$c(z_m, x_n) = \sqrt{s(z_m)}e^{-i2\pi z_m^\top x_n}\left(\mathrm{sinc}^D(\pi\varepsilon x_n) + \frac{\varepsilon DL}{2}\left(\mathrm{sinc}^D(\pi\varepsilon x_n) - \frac{1}{2}\mathrm{sinc}^{2D}\left(\pi\frac{\varepsilon}{2}x_n\right)\right)\right) \tag{25}$$

and hence the result (with $\varepsilon_0 = \min_d 1/(\pi x_{nd})$). □

Lemma B.2. *Under the standard assumptions, there exists a choice of $\varepsilon \to 0$ as $M \to \infty$ such that*

$$\frac{t}{N\sigma^2} \in O(M^{-\frac{2q}{q+3D}}).$$

Proof. When M is sufficiently large that $\varepsilon < \varepsilon_0 = \min_{n,d} 1/(\pi x_{nd})$,

$$\frac{t}{N\sigma^2} = \frac{1}{N\sigma^2}\sum_n (k(x_n,x_n) - \sum_{m,m'} c(x_n,z_m)c^*(x_n,z_{m'})\bar{k}^{-1}(z_m,z_{m'})) \tag{26}$$

$$= \frac{1}{N\sigma^2}\sum_n (v\sigma^2 - \sum_{m,m'} c(x_n,z_m)c^*(x_n,z_{m'})\varepsilon^D\delta_{m-m'}) \tag{27}$$

$$= \frac{1}{N\sigma^2}\sum_n (v\sigma^2 - \varepsilon^D\sum_m s(z_m)(1+O(\varepsilon^2 D))) = \frac{1}{N\sigma^2}\sum_n (v\sigma^2 - v\sigma^2\varepsilon^D\sum_m \tilde{s}(z_m)(1+O(\varepsilon^2 D))) \tag{28}$$

$$= v(1-\varepsilon^D\sum_m \tilde{s}(z_m)(1+O(\varepsilon^2 D))) \tag{29}$$

$$= v(1 - \int_{-M^{1/D}\varepsilon/2}^{M^{1/D}\varepsilon/2} ... \int_{-M^{1/D}\varepsilon/2}^{M^{1/D}\varepsilon/2} \tilde{s}(\xi)d\xi_1...d\xi_D) + vE_1 \tag{30}$$

$$= 2^D v \int_{M^{1/D}\varepsilon}^{\infty} ... \int_{M^{1/D}\varepsilon}^{\infty} \tilde{s}(\xi)d\xi + vE_1 \tag{31}$$

$$\leq 2^D v\beta(M\varepsilon)^{-q} + vE_1 \tag{32}$$

where $E_1 = \int_{-M^{1/D}\varepsilon/2}^{M^{1/D}\varepsilon/2} ... \int_{-M^{1/D}\varepsilon/2}^{M^{1/D}\varepsilon/2} \tilde{s}(\xi)d\xi_1...d\xi_D - \varepsilon^D\sum_m \tilde{s}(z_m) + O(M\varepsilon^{3D})$. Using the standard bound for the midpoint rule for numerical integration, since the curvature of s is bounded, there exists $\gamma > 0$

$$E_1 \leq \gamma\varepsilon^{3D}M. \tag{33}$$

Let $\varepsilon = \varepsilon_1 M^{-p/D}$. Then, in order that $E_1 \to 0$ and $M\varepsilon \to \infty$ (so that $(M\varepsilon)^q \to 0$), we must have $p \in (1/3, 1)$. From the overall bound

$$\frac{t}{N\sigma^2} \leq v\left(2^D\beta\varepsilon_1 M^{-(1-p)q/D} + \gamma\varepsilon_1^3 M^{-(3p-1)}\right) \tag{34}$$

for sufficiently large M, say $M > M_0 > (\varepsilon_1/\varepsilon_0)^{D/p}$, there exists an $\alpha > 0$ such that

$$\frac{t}{N\sigma^2} \leq M^{-\min\{(1-p)q/D, 3p-1\}}. \tag{35}$$

The optimal value of p is attained when the two exponents are equal, that is $p = (q/D+1)/(q/D+3)$, which yields an exponent of $2q/(q+3D)$. That is, there exists $M_0, \alpha > 0$ such that for all $M > M_0$

$$\frac{t}{N\sigma^2} \leq \alpha M^{-\frac{2q}{q+3D}}. \tag{36}$$

□

Theorem B.3. *Under the standard assumptions, if y is sampled from Equation* (1)*, then for any $\Delta, \delta > 0$, there exists $M_0, \alpha > 0$ such that*

$$Pr\left[\frac{D_{KL}(q(f)\|p(f|y))}{N} \geq \frac{\Delta}{N}\right] \leq \delta$$

for all $M \geq_0$ with

$$M \leq \left(\frac{2\alpha}{\Delta\delta}(1 + \|y\|_2^2/\sigma^2)N\right)^{\frac{q+3D}{2q}}.$$

Proof. Set ε as in Theorem B.2. Then,

$$\Pr\left[\frac{D_{KL}(q(f)\|p(f|y))}{N} \geq \frac{\Delta}{N}\right] = \delta \leq \frac{\mathbb{E}_y[D_{KL}(q(f)\|p(f|y))]/N}{\Delta/N} \tag{37}$$

$$\delta \leq N\frac{\alpha}{\Delta}M^{-\frac{2q}{q+3D}} \tag{38}$$

$$M \leq \left(\frac{\alpha}{\Delta\delta}N\right)^{\frac{q+3D}{2q}}. \tag{39}$$

□

Theorem B.4. *Under the standard assumptions, then for any $\Delta, \delta > 0$, there exists $M_0, \alpha > 0$ such that*

$$Pr\left[\frac{D_{KL}(q(f)||p(f|y))}{N} \geq \frac{\Delta}{N}\right] \leq \delta$$

for all $M \geq M_0$ with

$$M \leq \left(\frac{2\alpha}{\Delta\delta}(1 + ||y||_2^2/\sigma^2)N\right)^{\frac{q+3D}{2q}}.$$

Proof. Since t does not depend on x, $\mathbb{E}_x[t] = t$. Then, if we treat y as fixed and let $x \sim p(x)$, then

$$\Pr\left[\frac{D_{KL}(q(f)||p(f|y))}{N} \geq \frac{\Delta}{N}\right] = \delta \leq \frac{\mathbb{E}_x[D_{KL}(q(f)||p(f|y))]/N}{\Delta/N} \tag{40}$$

$$\delta \leq N\frac{\alpha}{2\Delta}(1 + ||y||_2^2/\sigma^2)M^{-\frac{2q}{q+3D}} \tag{41}$$

$$M \leq \left(\frac{2\alpha}{\Delta\delta}(1 + ||y||_2^2/\sigma^2)N\right)^{\frac{q+3D}{2q}}. \tag{42}$$

□

Corollary B.5. *We can formally equate the case of band-limited or exponential tails with $q \to \infty$, so the bound is $O(\sqrt{N})$ (dominated by E_1).*