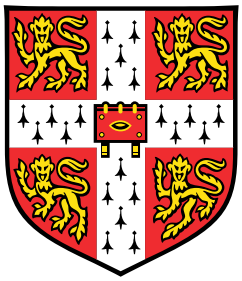
Sparse Circular Gaussian Process Approximation

Insights and Applications



William Tebbutt

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Master of Philosophy

Darwin College

August 2016

Declaration

I, Will Tebbutt of Darwin College, being a candidate for the M.Phil in Machine Learning, Speech and Language Technology, hereby declare that this dissertation and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose. This dissertation contains 9900 words.

William Tebbutt
August 2016

Acknowledgements

I wish to thank my supervisor, Rich Turner, for his guidance and insight. Also my family for their unwavering support.

Abstract

Efficient approximate posterior inference in Gaussian processes makes use of exploitable structure in the covariance function and observed data.

The first chapter reviews one well-known approach to approximate inference in Gaussian processes, based on representing the posterior through a set of pseudo-data, and one which is less so, in which the covariance function is approximated as being circular. The second chapter progresses existing work in which the circulant approximation is leveraged to better understand the properties of the state of the art pseudo-data inference scheme. The primary contribution in this chapter is a derivation of a frequency-domain representation of the low-rank covariance approximation.

The final chapter presents a new method designed to combine the strengths of the two approximation techniques to solve problems that neither could solve individually. This novel technique is developed throughout the chapter, its algorithmic properties analysed and its performance compared to exact inference experimentally.

Contents

List of Figures

Notation

All mathematical objects will be represented using 'unbolded' characters. The type of an object, whether it be a scalar, vector, matrix, function etc, will be stated explicitly.

Probability Theory and Approximate Inference

Probability distributions are specified by their arguments.

$p(x)$	marginal probability of x
$p(x_1, ..., x_N)$	joint probability of $(x_1, ..., x_N)$
$p(x_2 \mid x_1)$	conditional probability of x_2 given x_1
$\mathcal{N}(x \mid \mu, \Sigma)$	probability density of x is that of the multivariate Normal distribution with mean μ and covariance Σ
$x \sim \mathcal{N}(\mu, \Sigma)$	x is distributed according to a multivariate Normal distribution with mean μ and covariance Σ
$\mathbb{E}_{p(x)}[f(x)]$	expectation of some function f of x under the distribution $p(x)$
$\mathrm{Cov}_x[$	the covariance of x
$\mathrm{D}_{KL}(q \mid\mid p)$	Kullback-Leibler divergence between the distributions q and p
$H[p]$	Shannon-entropy of p

p is not used other than to denote a probability distribution, q is not used other than to denote an approximate posterior distribution.

Gaussian processes

X_A	set of inputs
$K_{A,B}$	(cross) covariance matrix between the inputs X_A and X_B
f_A	function values at inputs X_A
y	(noisy) observations

Misc

$\mathcal{O}\left(\right)$	Bachmann-Landau notation for asymptotic complexity
$\odot$	the Hadamard product
$\operatorname{diag}\left(x\right)$	diagonal matrix with vector x on the diagonal
$\operatorname{diag}\left(X\right)$	column-vector whose elements are the diagonal of the matrix X
$\operatorname{circ}\left(x\right)$	circulant matrix specified by vector x
$\operatorname{toep}\left(x\right)$	Toeplitz matrix specified by vector x
$A^\dagger$	the conjugate transpose of the matrix A

Chapter 1

Introduction and Review

1.1 Gaussian Processes for Regression

This thesis concerns itself with the application of Gaussian processes (GPs) to the problem of regression under additive Gaussian white noise, thus it is necessary to define both unambiguously. A probabilistic definition of (univariate) regression is adopted in which it is the task of predicting a distribution over plausible outputs $f \in \mathbb{R}$ given an input $x \in \mathbb{R}$. This is achieved by observing N input-output pairs $(x, y)_{n=0}^{N-1}$, assuming a stochastic procedure governing their generation, and combining these using the laws of probability to uncover a distribution over plausible mappings from inputs to outputs, which is used to solve the regression problem.

The stochastic procedure assumed throughout is that first a function $f : \mathbb{R} \to \mathbb{R}$ is drawn from some distribution, then each output is generated as

$$y_n \sim f(x_n) + \epsilon_n, \quad \epsilon_n \sim \mathcal{N}\left(0, \beta^{-1}\right). \tag{1.1}$$

Thus the key modelling problem is to choose a prior distribution over a family of functions believed to be plausible. One possible approach is to define a parametric family f_θ (eg the polynomials of a particular order) and induce a distribution over functions by placing a prior distribution over θ. An alternative approach is to place a nonparametric prior over functions, in which we have an unbounded number of parameters, and the posterior distribution over functions is represented in terms of the observed data.

Gaussian processes are such a nonparametric device. A GP is formally defined as follows [Rasmussen and Williams, 2006]:

Definition 1.1.1. *A* Gaussian process *is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

A GP is therefore completely specified by its mean and covariance, which are parameterised by functions $m : \mathbb{R} \to \mathbb{R}$ and $k : (\mathbb{R}, \mathbb{R}) \to \mathbb{R}$ respectively. Without loss of generality, I will take the mean function m to be the zero function throughout this thesis, thus the joint probability of a set of function values $f_D \in \mathbb{R}^N$ given a set of inputs $X_D \in \mathbb{R}^N$ is

$$p\left(f \mid X_D\right) = \mathcal{N}\left(f_D \mid 0, K_{D,D}\right) \tag{1.2}$$

where $\left(K_{D,D}\right)_{m,n} = k\left(x_m, x_n\right)$. k is know as the covariance function or 'kernel' and can be thought of as a measure of similarity between a pair of inputs.

The marginal probability of the observations under this model is thus

$$\begin{aligned} p\left(y \mid X_D\right) &= \mathbb{E}_{f_D \sim \mathcal{N}\left(0, K_{D,D}\right)}\left[\mathcal{N}\left(y \,\middle|\, f_D, \beta^{-1}\mathcal{I}\right)\right] \\ &= \mathcal{N}\left(y \,\middle|\, 0, K_{D,D} + \beta^{-1}\mathcal{I}\right). \end{aligned} \tag{1.3}$$

To perform inference (compute the distribution over outputs f for inputs X given the observations y, X_D it is necessary only to compute the posterior predictive distribution

$$p\left(f_\# \mid X_\#, y, X\right) = \mathcal{N}\left(f_\# \,\middle|\, K_{\#,D}\left(K_{D,D} + \beta^{-1}\mathcal{I}\right)^{-1} y, K_{D,D} - Q_{D,D}\right) \tag{1.4}$$

where $Q_{D,D} := K_{\#,D}\left(K_{D,D} + \beta^{-1}\mathcal{I}\right)^{-1} K_{D,\#}$.

It is the covariance function which defines the characteristics of any particular GP. A commonly used covariance function is the exponentiated quadratic (EQ) and is defined as

$$k\left(x_m, x_n\right) = \sigma^2 \exp\left(-\frac{1}{2l^2}(x_m - x_n)^2\right), \tag{1.5}$$

where $\sigma^2, l^2 > 0$ are covariance function parameters that can be learned by maximising the (log) marginal likelihood (equation 1.3). Choosing a covariance function function appropriate for the data set at hand is essential for good regression performance and many approaches have been investigated. For example Wilson and Adams [2013] propose a covariance function which can approximate any stationary covariance function (stationarity will be explained later in the chapter) arbitrarily well under certain conditions, whilst Tobar et al. [2015] take this idea to its logical conclusion and make the covariance function itself nonparametric. Other approaches include the construction of complicated covariance functions through the composition of simple covariance functions [Lloyd et al., 2014] and a Deep GPs [Damianou and Lawrence, 2013], a technique inspired by the recent emergence of 'deep learning', in which multiple Gaussian processes with simple

covariance functions are composed to yield an object which is conditionally a GP with a complicated covariance function.

I do not, however, address this important modelling issue in this thesis. Rather, note that, although the unusually simple form of equations 1.3 and 1.4 make GPs convenient objects in a small-data scenario, the $\mathcal{O}(N^3)$ Cholesky decomposition required to compute the quantities in equations 1.3 and 1.4 prohibits their direct application at scale[1].

Accepting that this computational hurdle is not trivially circumvented, there are essentially two options. We must either abandon the full GP prior over functions and instead make a new set of modelling assumptions that may be less elegant but result in a model in which inference is tractable[2], or instead exploit structure present in the data to perform inference approximately. I opt to pursue the second of these options as, although in practice one tends to take practical considerations into account when defining a model, a good approximation to the model that best reflects one's beliefs about the problem at hand is typically preferable.

The remaining sections of this introductory chapter discuss two complementary techniques for approximate inference which exploit different types of structure in data. The final short section will summarise, and outline the contributions and structure of the rest of the remaining two chapters.

1.2 Exploiting Redundancy

Probably the most well known paradigm for approximating inference in GPs is based upon the observation that commonly more data is observed than is actually necessary to represent the posterior distribution, potentially rendering much of the data redundant. This motivates so-called sparse GP approximations which aim to exploit exactly this property by augmenting the observed data with a smaller set of pseudo-data and representing the GP posterior through these. The approaches to this augmentation process have traditionally been viewed as falling into one of two categories. The first can be viewed as performing exact inference under an approximate GP prior Quiñonero-Candela and Rasmussen [2005], whilst the second as performing approximate inference using Variational Bayes under an exact GP prior Titsias [2009]. Recent work [Bui et al., 2016]

[1] Chapter 2 of [Rasmussen and Williams, 2006] presents a procedure for evaluating equation 1.3 and computing the mean and covariance in equation 1.4 using the Cholesky decomposition. In the appendix they allude to the computational and numerical reasons to prefer this procedure over the naive direct computation of the inverse and determinant of $K_{D,D}$.

[2] I will always discuss the concept of tractability from a practical perspective. Thus whenever a computation is referred to as 'intractable' it is meant that it is not possible to compute given the time and resources available, rather than it having polynomial asymptotic complexity.

has unified these two views using Power Expectation Propagation (Power EP) [Minka, 2004, 2001] by demonstrating that approaches in the first category are in fact performing EP, thus both can be viewed as performing approximate inference under an exact GP prior.

Adopting the approximate-prior view, whilst the first of these two approaches has produced some very successful techniques [Snelson and Ghahramani, 2005] it sacrifices some of the elegant (and highly useful) properties of GPs, in particular the defence against overfitting offered to GPs by the Bayesian Occam's Razor [Ghahramani, 2005; MacKay, 2003]. The approximate GP priors correspond to parametric models which can be susceptible to the problem of overfitting when the pseudo-data input locations are optimised with respect to the log marginal likelihood of the data.

The second of these paradigms, however, retains its defence against overfitting. Of particular interest is [Titsias, 2009] and the methodology by which it can be used to handle the 'big data' problems that are becoming increasingly common [Hensman et al., 2013]. As this technique will be exploited and investigated throughout this thesis, it is explained in some detail below.

As discussed above, [Titsias, 2009] develops a framework for directly approximating the posterior distribution over functions $f_D \mid y, X_D$ of a GP using a small set of pseudo-data. This framework is given a mathematically rigorous treatment in [Matthews et al., 2015]. The derivation that I present here is based on the opening sections of [Matthews et al., 2015] but, at the expense of ignoring that some integrals are taken over infinite dimensional objects, is devoid of measure theory.

Partition the input space $\mathbb{R}$ into three disjoint subsets: a set of pseudo-inputs X_Z of size M, the set of input positions for the observed data excluding any input positions that are also pseudo-inputs $X_{D\setminus Z}$ of size N, and the rest of the index set X_*, therefore $\mathbb{R} \equiv X_Z \cup X_{D\setminus Z} \cup X_*$. Since each element of the index set $\mathbb{R}$ corresponds to exactly one random variable, we define corresponding sets of random variables f_Z, $f_{D\setminus Z}$ and f_* such that $f = \{f_Z, f_{D\setminus Z}, f_*\}$. For convenience I will neglect the $\setminus Z$ notation and write the set of random variables $f_{D\setminus Z}$ as f_D. This allows us to, for example, write the joint distribution over these random variables as $p(f) = p(f_Z, f_D, f_*) = p(f_* \mid f_Z, f_D)\, p(f_D \mid f_Z)\, p(f_Z)$ using the chain rule of probability, where the dependence upon the index set has been neglected.

Choosing to perform variational inference by minimising the KL-divergence between an approximating distribution and the true posterior distribution, we wish to find $q(f \mid X_z)$ such that

$$\mathrm{D}_{KL}\left(q(f) \;||\; p(f \mid y)\right) \tag{1.6}$$

is minimised, where the dependencies on X, X_* and X_z have been dropped for notational convenience. If one is suspicious of the infinite-dimensional integral required to compute the KL-divergence, one could suppose that we are dealing with a GP valued only on a finite subset of $\mathbb{R}$, but one that is nonetheless sufficiently large to be indistinguishable from $\mathbb{R}$ for all practical purposes, in which case the divergence will clearly converge.

Proceeding as is usual in this framework, we re-phrase the problem of minimising the KL-divergence as that of maximising a lower bound to the evidence by noting that, since for any arbitrary distributions a and b we have that

$$\mathrm{D}_{KL}\left(a \;||\; b\right) \geq 0 \tag{1.7}$$

with equality iff a is equal to b over the entire support of both distributions,

$$\log p\left(y\right) = \int q\left(f\right) \log \frac{p\left(f\right)}{q\left(f\right)} \,\mathrm{d}f + \mathrm{D}_{KL}\left(q\left(f\right) \;||\; p\left(f \,|\, y\right)\right) \geq \int q\left(f\right) \log \frac{p\left(f\right)}{q\left(f\right)} \,\mathrm{d}f =: L\left(q\right). \tag{1.8}$$

The evidence lower bound (ELBO) L is tight only when $\mathrm{D}_{KL}\left(q\left(f\right) \;||\; p\left(f \,|\, y\right)\right)$ is minimised. Defining the approximate posterior distribution $q\left(f\right) := p\left(f_* \,|\, f_D, f_Z\right) p\left(f_D \,|\, f_Z\right) q\left(f_Z\right)$ we find that

$$L = \mathbb{E}_{q(f_D)}\left[p\left(y \,|\, f_D\right)\right] - \mathrm{D}_{KL}\left(q\left(f_Z\right) \;||\; p\left(f_Z\right)\right), \tag{1.9}$$

and, noting that for any $x, y, A, \mu, \Sigma_0, \Sigma_1$

$$\mathbb{E}_{\mathcal{N}(x \,|\, \mu, \Sigma_0)}\left[\log \mathcal{N}\left(y \,|\, Ax, \Sigma_1\right)\right] = \log \mathcal{N}\left(y \,|\, A\mu, \Sigma_1\right) - \frac{1}{2}\mathrm{tr}\left(A^T \Sigma_1^{-1} A \Sigma_0\right), \tag{1.10}$$

we find that the first term in L (equation 1.9) is

$$\begin{aligned}\mathbb{E}_{q(f_D)}\left[p\left(y \,|\, f_D\right)\right] &= \mathbb{E}_{q(f_Z)}\left[\mathbb{E}_{p(f_D \,|\, f_z)}\left[p\left(y \,|\, f_D\right)\right]\right] \\ &= \mathbb{E}_{q(f_Z)}\left[\log \mathcal{N}\left(y \,\middle|\, K_{D,Z} K_{Z,Z}^{-1} f_Z, \beta^{-1}\mathcal{I}\right)\right] - \frac{\beta}{2}\mathrm{tr}\left(K_{D,D} - K_{D,Z} K_{Z,Z}^{-1} K_{Z,D}\right). \end{aligned} \tag{1.11}$$

Observe that for arbitrary distributions a, b

$$\mathrm{D}_{KL}\left(a \;||\; b\right) = -H\left[a\right] - \mathbb{E}_a\left[\log b\right] \tag{1.12}$$

where H is the Shannon Entropy. Expanding the second term in equation 1.9 using this identity, the first term using the result in equation 1.11, and defining

$$\Sigma_q := (\beta K_{Z,Z}^{-1} K_{Z,D} K_{D,Z} K_{Z,Z}^{-1} + K_{Z,Z}^{-1})^{-1} \tag{1.13}$$

$$\mu_q := \beta \Sigma_q K_{Z,Z}^{-1} K_{Z,D} y, \tag{1.14}$$

we obtain

$$\begin{aligned}
&\underset{q}{\operatorname{argmax}}\, L \\
&= \underset{q}{\operatorname{argmax}}\, H(q) - \frac{1}{2}\mathbb{E}_{q(f_Z)}\left[\beta(y - K_{D,Z}K_{Z,Z}^{-1}f_Z)^T(y - K_{D,Z}K_{Z,Z}^{-1}f_Z) + f_Z^T K_{Z,Z}^{-1} f_Z\right] \\
&= \underset{q}{\operatorname{argmax}}\, H(q) - \frac{1}{2}\mathbb{E}_{q(f_Z)}\left[f_Z^T \Sigma_q^{-1} f_Z - 2\beta f_Z^T K_{Z,Z}^{-1} K_{Z,D} y\right] \\
&= \underset{q}{\operatorname{argmax}} -\mathrm{D}_{KL}\left(q(f_Z) \;||\; \mathcal{N}(f_Z \,|\, \mu_q, \Sigma_q)\right).
\end{aligned} \tag{1.15}$$

Thus without having assumed a particular form for q it is clear that the optimal approximate posterior distribution is

$$q(f_Z) = \mathcal{N}(f_Z \,|\, \mu_q, \Sigma_q) \tag{1.16}$$

as this uniquely minimises the KL-divergence.

From this all other quantities of interest follow in short order. Substituting equation 1.11 and 1.16 into equation 1.9 and again applying the identity in equation 1.10 we find that the ELBO is

$$\begin{aligned}
L = \; &\log \mathcal{N}\left(y \,\middle|\, K_{D,Z}K_{Z,Z}^{-1}\mu_q, \beta^{-1}\mathcal{I}\right) - \frac{\beta}{2}\mathrm{tr}\left(K_{D,D} - Q_{D,D} + R_{D,D}\right) \\
&- \mathrm{D}_{KL}\left(\mathcal{N}(\mu_q, \Sigma_q) \;||\; \mathcal{N}(0, K_{Z,Z})\right),
\end{aligned} \tag{1.17}$$

where $Q_{D,D} := K_{D,Z}K_{Z,Z}^{-1}K_{Z,D}$ and $R_{D,D} := K_{D,Z}K_{Z,Z}^{-1}\Sigma_q K_{Z,Z}^{-1}K_{Z,D}$.

Let $X_\# \in \mathbb{R}$ be a finite set of input locations with corresponding vector of outputs $f_\#$. The approximate posterior predictive distribution

$$p(f_\# \,|\, y) = \mathbb{E}_{p(f\,|\,y)}\left[p(f_\# \,|\, f)\right] \approx \mathbb{E}_{q(f)}\left[p(f_\# \,|\, f)\right] \tag{1.18}$$

will be Gaussian, and it is therefore sufficient to compute only its mean and covariance. These sufficient statistics can be found quickly by noting that

$$f_{\#} = K_{\#,Z} K_{Z,Z}^{-1} f_Z + \epsilon, \tag{1.19}$$

where

$$f_Z \sim \mathcal{N}\left(\mu_q, \Sigma_q\right), \tag{1.20}$$

$$\epsilon \sim \mathcal{N}\left(0, K_{\#,\#} - Q_{\#,\#}\right), \tag{1.21}$$

$$Q_{\#,\#} := K_{\#,Z} K_{Z,Z}^{-1} K_{Z,\#}. \tag{1.22}$$

The posterior predictive mean

$$\mu_{\#} := \mathbb{E}_{p(f\,|\,y)}\left[f_{\#}\right] \approx K_{\#,Z} K_{Z,Z}^{-1} \mathbb{E}_{q(f_Z)}\left[f_Z\right] = K_{\#,Z} K_{Z,Z}^{-1} \mu_q, \tag{1.23}$$

and similarly the posterior predictive covariance

$$\begin{aligned} \Sigma_{\#} &:= \mathbb{E}_{p(f\,|\,y)}\left[f_{\#} f_{\#}^T\right] - \mathbb{E}_{p(f\,|\,y)}\left[f_{\#}\right] \mathbb{E}_{p(f\,|\,y)}\left[f_{\#}^T\right] \\ &\approx K_{\#,\#} - Q_{\#,\#} + R_{\#,\#} \end{aligned} \tag{1.24}$$

where $R_{\#,\#} := K_{\#,Z} K_{Z,Z}^{-1} \Sigma_q K_{Z,Z}^{-1} K_{Z,\#}$ are defined analogously to $Q_{D,D}$ and $R_{D,D}$.

The most intensive operation required to compute this posterior approximation is the matrix multiplication $K_{Z,D} K_{D,Z}$ which requires $\mathcal{O}\left(NM^2\right)$ operations. Assuming that $M \ll N$ this can result in a dramatic saving over the $\mathcal{O}\left(N^3\right)$ matrix inversion required for exact inference.

Figure 1.1 presents a course grained cartoon of manner in which the discussed sparse approximation breaks. In all four images there are 1000 regularly spaced observations and 10 regularly spaced pseudo-data. The only factor that changes between them is the length scale of the EQ covariance function used to generate the plots. In the first two the length scale is sufficiently large that the posterior distribution is simple enough to be accurately represented through the 10 pseudo-data. However, in the second pair of plots the length scale is shorter, resulting in more complicated ('wiggly') posterior distributions that cannot be accurately represented using only 10 pseudo-data. The primary point that these plots aim to convey is that, in the presence of 'large' amount of data distributed evenly over the input domain of interest, the number of pseudo-data required to accurately approximate the posterior distribution will be a factor of the 'length scale' and smoothness of the covariance function relative to the size of the region

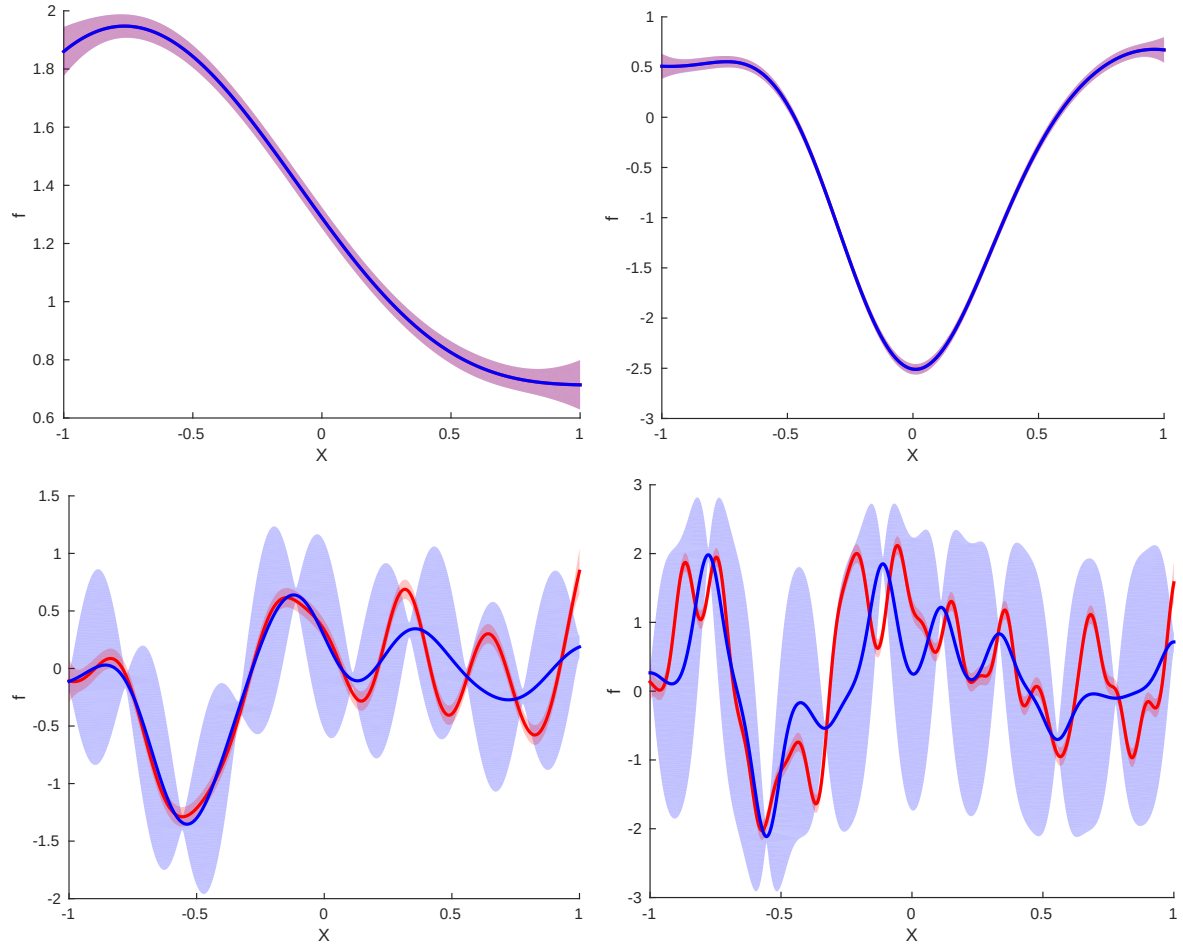


Figure 1.1 Series of figures demonstrating the decay in performance of the sparse variational GP approximation [Titsias, 2009] (blue) to the exact posterior distribution of a GP (red) having observed $N = 1000$ input-output pairs under Gaussian white-noise with variance $\beta^{-1} = 0.1$ and an EQ covariance function with $\sigma^2 = 1$ as $\log l^2$ is varied. Lines represent the (approximate) posterior mean and filled regions represent two marginal (approximate) posterior standard deviations from the mean. Top row: $\log l^2 = 0, -2$, for these longer length scales the posterior is well approximated by the 10 pseudo-data used. Bottom row: $\log l^2 = -4, -6$. For these cases the pseudo-data are insufficient to capture the posterior distribution. This is a very course grained visualisation of the fact that as the GP becomes more complex more pseudo-data are required to represent it if we have more pseudo-data than data.

covered by the observed data. This observation provides extremely strong motivation for the sparse posterior approximation method discussed, but also reveals a short-coming of the formulation presented. To see this consider the first two functions in figure 1.1, in which $N = 1000$ data have been observed. Quite clearly observing another 1000 data in the range $[-1, 1]$ will leave the posterior distribution essentially unchanged, but will roughly double the amount of computational resources required to compute the posterior approximation. Furthermore, we could repeat this process of observing a batch of 1000 new data an unbounded number of times and not really learn anything new, but would now require an unbounded amount of computation to tell us this under the current formulation! This unnecessary state of affairs is addressed by the technique presented in the following section.

1.2.1 Scaling to Large Data Sets

The methodology presented by Hensman et al. [2013] provides exactly the same posterior approximation as that presented above, but finds the optimal variational parameters μ_q, Σ_q using stochastic optimisation to address the issue of computational complexity in the presence of an increasingly large amount of redundant observed data. In particular it observes that the data dependent terms of L can be written as an independent sum over the data:

$$L = \sum_{i=1}^{N} \log \mathcal{N}\left(y_i \,\middle|\, K_{i,Z}K_{Z,Z}^{-1}\mu_q, \beta^{-1}\mathcal{I}\right) - \frac{\beta}{2}\left[(K_{D,D})_{i,i} - (Q_{D,D})_{i,i} + (R_{D,D})_{i,i}\right] - \mathrm{D}_{KL}\left(\mathcal{N}\left(\mu_q, \Sigma_q\right) \,||\, \mathcal{N}\left(0, K_{Z,Z}\right)\right). \tag{1.25}$$

where $K_{i,Z}$ is the $(1 \times M)$ cross-covariance matrix between the i^{th} observation input and each of the pseudo-inputs. A cheap unbiased Monte Carlo estimator for L can now be constructed by sampling a subset of data uniformly at random, and computing

$$\hat{L} \simeq \frac{N}{T}\left[\log \mathcal{N}\left(y_S \,\middle|\, K_{S,Z}K_{Z,Z}^{-1}\mu_q, \beta^{-1}\mathcal{I}\right) - \frac{\beta}{2}\mathrm{tr}\left(K_{S,S} - Q_{S,S} + R_{S,S}\right)\right] - \mathrm{D}_{KL}\left(\mathcal{N}\left(\mu_q, \Sigma_q\right) \,||\, \mathcal{N}\left(0, K_{Z,Z}\right)\right) \tag{1.26}$$

where y_S is the output vector for the subset, $K_{S,Z}$, $K_{S,S}$, $Q_{S,S}$ and $R_{S,S}$ are defined in the same manner as previously and T is the number of samples taken.

Given the estimator for L, to perform inference one could simply compute its gradient with respect to both μ_q and Σ_q, which is trivial to accomplish given modern automatic differentiation technology, and perform stochastic gradient ascent [Robbins and Monro,

1951] in $\hat{L}$. This, however, would not guarantee the symmetry or positive definiteness of Σ_q, meaning that further constraints would need to be placed upon it. A superior approach is investigated in the work on stochastic variational inference by Hoffman et al. [2013], which elucidates how the natural gradient [Amari, 1998] can be used to simultaneously improve the rate of convergence of stochastic gradient ascent, reduce the computational burden associated with 'gradient' computations and ensure the positive definiteness of the resulting covariance matrix. Thus the approach taken in [Hensman et al., 2013] is to work in terms of the natural parameters

$$\theta_1 = \Sigma_q^{-1}\mu_q \tag{1.27}$$

$$\theta_2 = -\frac{1}{2}\Sigma_q^{-1} \tag{1.28}$$

and perform natural gradient ascent as discussed.

The asymptotic complexity of this method is now $\mathcal{O}\left(TM^2 + M^3\right)$, the TM^2 term arising from the need to compute $Q_{S,S}$ and $R_{S,S}$ and the M^3 from still needing to compute the Cholesky decomposition of several matrices of size $M \times M$. Crucially, this method no longer depends explicitly upon the size of the observed data set, but rather on the number of pseudo-data required to represent the function from which the data are assumed to have been drawn.

1.3 Exploiting Approximately-Circular Structure

This second class of approximation concerns the covariance function directly. I will show how, from certain conditions on the covariance function and input locations, exploitable approximately-circular structure will arise in the covariance matrix. This section provides a few different perspectives showing how this structure arises and motivating its exploitation as a good approximation.

1.3.1 Approximately Circulant Toeplitz Matrices

Definition 1.3.1. *A matrix $C \in \mathbb{R}^{M\times M}$ is called* circulant *if $C_{m,n} = c_{(n-m) \bmod M}$ for some $c \in \mathbb{R}^M$.*

This definition states that the first row of the matrix C is exactly c, and each subsequent row is produced by taking the previous row, shifting it one position to the right and wrapping the last element around into the first position. For example the

matrix

$$\begin{bmatrix} 0 & 1 & 2 & 3 \\ 3 & 0 & 1 & 2 \\ 2 & 3 & 0 & 1 \\ 1 & 2 & 3 & 0 \end{bmatrix} \tag{1.29}$$

is circulant. I define the operator $\mathrm{circ}(c)$ to generate a circulant matrix in the manner described above from the vector c. For example, $\mathrm{circ}([0,1,2,3])$ would generate the above circulant matrix.

These matrices are particularly tractable as their eigenvalues γ are given by the Discrete Fourier Transform (DFT) of c, and their eigenvectors in stacked form are always the Inverse Discrete Fourier Transform (IDFT) matrix, therefore

$$C = \mathrm{FT}\Gamma\mathrm{FT}^\dagger \tag{1.30}$$

where FT and $\mathrm{FT}^\dagger$ are the unitary DFT and IDFT matrices respectively and $\Gamma = \mathrm{diag}(\gamma)$. Observing that

$$C^{-1} = \left(\mathrm{FT}_\Gamma\mathrm{FT}^\dagger\right)^{-1} = \left(\mathrm{FT}^\dagger\right)^{-1}\Gamma^{-1}\mathrm{FT}^{-1} = \mathrm{FT}\Gamma^{-1}\mathrm{FT}^\dagger, \tag{1.31}$$

the inverse of C is therefore found by computing the eigenvalues of C using an FFT operation, taking $\mathcal{O}(M \log M)$ time, and then computing their reciprocal in $\mathcal{O}(M)$ time to find Γ^{-1}[3]. Similarly the determinant is

$$|C| = |\mathrm{FT}|\,|\Gamma|\left|\mathrm{FT}^\dagger\right| = |\Gamma|, \tag{1.32}$$

the determinant of the diagonal matrix Γ which is also computed in $\mathcal{O}(M)$ time. These matrices are members of the more general class of Toeplitz matrices:

Definition 1.3.2 (Toeplitz matrix)**.** *$T \in \mathbb{R}^{M\times M}$ is* Toeplitz *if $T_{m,n} = t_{n-m}$ for some vector $t \in \mathbb{R}^{2M-1}$ (which is indexed from $-(M-1)$ to $(M-1)$). We write $T = \mathrm{toep}(t)$.*

For example,

$$\mathrm{toep}\left(\begin{bmatrix} t_{-3} & t_{-2} & t_{-1} & t_0 & t_1 & t_2 \end{bmatrix}\right) = \begin{bmatrix} t_0 & t_1 & t_2 & t_3 \\ t_{-1} & t_0 & t_1 & t_2 \\ t_{-2} & t_{-1} & t_0 & t_1 \\ t_{-3} & t_{-2} & t_{-1} & t_0 \end{bmatrix}. \tag{1.33}$$

[3]Note that Γ^{-1} should not be represented explicitly, as this would require $\mathcal{O}(M^2)$ memory. Indeed, it is never necessary to do so.

We now consider the conditions under which a Toeplitz matrix can be approximated by a circulant matrix. First consider the band-diagonal Toeplitz matrix with band-width $2B-1 < 2M-1$

$$T = \text{toep}\left(\underbrace{\begin{bmatrix}0 & \dots & 0 & t_{-(B-1)} & \dots & t_0 & \dots & t_{(B-1)} & 0 & \dots & 0\end{bmatrix}}_{2M-1}\right), \tag{1.34}$$

given which, construct the circulant matrix

$$C = \text{circ}\left(\underbrace{\begin{bmatrix}t_0 & \dots & t_{(B-1)} & 0 & \dots & 0 & t_{-(B-1)} & \dots & t_{-1}\end{bmatrix}}_{M}\right). \tag{1.35}$$

Upon computing the matrices T and C it is clear that T is identical to C everywhere except the top right and bottom left corners, in which T is zero and C is non-zero. One can easily imagine that as M becomes large, the difference between the two matrices will tend to zero in some matrix norm. This idea is formalised in the following lemma.

Lemma 1.3.1. *(Asymptotic Equivalence) The matrices T and C are asymptotically equivalent under the weak (Frobenius) norm. That is*

$$\lim_{M\to\infty} \sum_{m,n=0}^{M-1} |C_{m,n} - T_{m,n}|^2 = 0. \tag{1.36}$$

Proof. See [Gray, 2006][4]. □

1.3.2 Approximately Circulant Covariance Matrices

Primed with the knowledge about the close relationship between Toeplitz and circulant matrices from section 1.3, I now present the circumstances under which the covariance matrix $K_{D,D}$ will be Toeplitz, thus motivating its approximation with a circulant matrix.

Note 1.3.1 (Symmetric Toeplitz matrix)**.** *A* symmetric Toeplitz *matrix $T \in \mathbb{R}^{M\times M}$ is specified by a vector $t \in \mathbb{R}^M$ such that $T_{m,n} = t_{|n-m|}$.*

[4] As discussed in [Gray, 2006], it is not necessary to assume that T and C are band-diagonal to obtain this result. It is in fact sufficient to assume that, in the limit as $M \to \infty$, the sequence $t_{-(M-1)}, \dots, t_{(M-1)}$ is either square or absolutely summable (ie. $\sum_{m=-\infty}^{\infty} |t_m|^2$ or $\sum_{m=-\infty}^{\infty} |t_m|$ respectively). These weaker conditions will not be necessary in practice and are therefore not discussed.

Definition 1.3.3 (Stationary covariance function)**.** *A covariance function k is called* stationary *if it is invariant under the simultaneous translation of both of its inputs such that*

$$k(x, y) = k(x + a, y + a). \tag{1.37}$$

Such covariance functions will be denoted more simply by overloading notation [5]*:*

$$k(x, y) = k(x - y). \tag{1.38}$$

The covariance matrix computed using such a covariance function will be symmetric and Toeplitz if the inputs $X \in \mathbb{R}^M$ are regularly spaced such that $x_n = an + b$ for some $a, b \in \mathbb{R}$. To see this note that

$$\begin{aligned}(K_{D,D})_{m,n} &= k(x_m - x_n) \\ &= k(a|n - m|) \\ &= k(a(0 - |n - m|)) \\ &= (K_{D,D})_{0,|n-m|},\end{aligned} \tag{1.39}$$

thus the covariance matrix $K_{D,D}$ can be expressed in terms of the elements of its first row, making it both symmetric (as with all covariance matrices) and Toeplitz. This motivates the construction of a circulant approximation to $K_{D,D}$ using the construct in equation 1.35.

Furthermore, to recover the banded assumption discussed in section 1.3.1 it is necessary to assume that there exists $b > 0$ such that $\forall |x| > b$, $k(x) = 0$. This is not a criterion that common covariance functions (such as the EQ) satisfy exactly, however, many do so approximately in the following sense.

Definition 1.3.4 (Approximately banded covariance function)**.** *A covariance function k is called* approximately banded (AB) *if, $\forall \epsilon > 0$, $\exists b \in [0, \infty)$ such that $\forall |x| > b$, $k(x) < \epsilon$.*

Any finite-precision implementation[6] of an AB covariance function will yield a banded covariance function as there will always exist some $\epsilon > 0$ such that any $\delta \in (0, \epsilon)$ will be truncated to zero by the finiteness of the representation. Much of the work in this dissertation will appeal to a covariance function being banded based on this observation.

[5] All covariance functions discussed in this thesis will be stationary, so this abuse of notation will be used without explicit warning.

[6] Any implementation on a finite digital computer.

1.3.3 Circular Covariance Function Approximation

The above result, that certain covariance matrices are approximately circulant, can be established without appealing to Toeplitz approximations to circulant matrices, but rather by approximating the covariance function directly. This more general approach will turn out to be essential for good posterior prediction in Chapter 3.

The approximation is given by

$$k_c(x) = k(|x|) + k(\Delta - |x|), \tag{1.40}$$

for some $\Delta > 0$ and stationary covariance function k.

Lemma 1.3.2. *Given M data $\{x_m = \delta m : m \in \{0, ..., M-1\}, \delta > 0\}$ and $\Delta = M\delta$,*

$$(K_c)_{m,n} = k_c(x_m - x_n) \tag{1.41}$$

is a symmetric circulant matrix.

Proof. Symmetry is trivial to show:

$$k_c(-x) = k(|-x|) + k(\Delta - |-x|) = k(|x|) + k(\Delta - |x|) = k_c(x). \tag{1.42}$$

Recalling that $x_n = an + b$, that K_c is circulant is similarly simple:

$$(K_c)_{m,n} = k_c(a|m-n|) = k_c(a|0-(n-m)|) = (K_c)_{0,|n-m|}.$$

□

Observing that as $\Delta \to \infty$, $\Delta - |x| \to \infty$, if k is banded then it is recovered from k_c in this limit provided that x is finite.

1.3.4 Circular Input-Domain Approximation

Yet another approach to obtain the approximation is to define a mapping which makes the input domain itself circular. Specifically define

$$g(x) := \mathbb{I}(x \in [0, \Delta/2))\, x + \mathbb{I}(x \in [\Delta/2, \Delta))\,(x - \Delta), \tag{1.43}$$

and assume that $x \in [0, \Delta)$. Equivalence with the circular covariance function approximation k_c is shown by evaluating $k\left(g\left(x\right)\right)$:

$$k\left(g\left(x\right)\right) = k\left(\mathbb{I}\left(x \in [0, \Delta/2)\right) x\right) + \mathbb{I}\left(x \in [\Delta/2, \Delta)\right)\left(x - \Delta\right),$$

and since the two sets $[0, \Delta/2)$ and $[\Delta/2, \Delta)$ are disjoint,

$$k\left(g\left(x\right)\right) = k\left(x\right)\mathbb{I}\left(x \in [0, \Delta/2)\right) + k\left(x - \Delta\right)\mathbb{I}\left(x \in [\Delta/2, \Delta)\right).$$

Further assuming that k is banded with band-width $b < \Delta/2$, such that $k\left(x\right) = 0$ for $|x| > b$, the indicator functions are redundant and k_c is recovered.

This input-domain view, or equivalently the covariance function view, on the circulant approximation will be appealed to frequently throughout this dissertation. With regards to terminology when doing so, I will state that the circulant approximation is to be applied, and subsequently assume that computations are exact under this approximation. For example, let k be a stationary covariance function to which the circulant approximation has been applied, and X a vector of regularly sampled inputs, then

$$K_{D,D} = \mathrm{FT}_D \Gamma_D \mathrm{FT}_D^\dagger. \tag{1.44}$$

1.3.5 Efficient Inference under the Circulant Approximation

The discussed circulant approximation has been exploited previously in the GP literature, for example, embedding the the Toeplitz covariance matrix in a large circulant matrix for efficient sampling in Geostatistical simulations [Dietrich and Newsam, 1997]. More recent work [Cunningham et al., 2008] exploits the approximation to accelerate MAP estimation and the use of the Lapace approximation for inference and learning in Point Process intensity estimation, [Turner, 2010] exploits the approximation directly to accelerate inference in GP models for probabilistic amplitude demodulation, whilst [Ulrich et al., 2015] employs it for use in multi-output GPs.

I now elucidate probably the simplest manner in which the circulant approximation can be applied for accelerated inference in regularly sampled univariate regression. Let the input vector $X \in \mathbb{R}^N$ comprise N data in which, without log of generality, $x_n = n$ and let $y \in \mathbb{R}^N$ be arbitrary noisy observations. Letting k be a stationary covariance function we approximate the prior covariance matrix as

$$K_{D,D} = \mathrm{FT}_D \Gamma_D \mathrm{FT}_D^\dagger, \tag{1.45}$$

where $\Gamma_D := \text{diag}(\gamma_D)$ is the diagonal matrix of eigenvalues of the first row of $K_{D,D}$. It follows that the log marginal likelihood can be approximated as

$$\begin{aligned}\log p\left(y \,\middle|\, X, \beta^{-1}\right) &= -\frac{1}{2}\log\left|2\pi\left(\Gamma_D + \beta^{-1}\mathcal{I}\right)\right| - \frac{1}{2}\tilde{y}^\dagger\left(\Gamma_D + \beta^{-1}\mathcal{I}\right)^{-1}\tilde{y} \\ &= -\frac{N}{2}\log 2\pi - \frac{1}{2}\sum_{n=0}^{N-1}\left[\log\left(\gamma_n + \beta^{-1}\right) + \left(\gamma_n + \beta^{-1}\right)^{-1}|\tilde{y}_n|^2\right], \quad (1.46)\end{aligned}$$

where $\tilde{y} := \text{FT}_D^\dagger y$. Covariance function parameter learning can now be performed efficiently through the application of the chain-rule using one's preferred reverse-mode automatic differentiation tool[7]. The observations can be de-noised using the posterior predictive mean at the input locations

$$\mathbb{E}_{p(f_D \,|\, y, X, \beta^{-1})}\left[f_D\right] = \text{FT}_D\Gamma_D\left(\Gamma_D + \beta^{-1}\mathcal{I}\right)^{-1}\tilde{y}. \quad (1.47)$$

For efficient computation it is imperative the the computations here be performed from right to left. To see this observe that $\Gamma_D\left(\Gamma_D + \beta^{-1}\mathcal{I}\right)^{-1}\tilde{y}$ can be computed in linear time given $\tilde{y}$ (both matrices are diagonal) to yield a new column vector. The final FT_D operation can be performed in $\mathcal{O}(N\log N)$ time. Were the computation performed in the reverse order it would be necessary to multiply FT_D by a diagonal matrix, requiring $\mathcal{O}(N^2)$ operations. This kind of order-dependent asymptotic complexity will frequently be a subject of primary concern in Chapter 3.

Finally, the uncertainty surrounding f_D can be quantified through its marginal variance

$$\begin{aligned}\text{diag}\left(\text{Cov}_{p(f_D \,|\, y, X, \beta^{-1})}\left[f_D\right]\right) &= \text{diag}\left(\text{FT}_D\left(\Gamma_D - \Gamma_D^2\left(\Gamma_D + \beta^{-1}\mathcal{I}\right)^{-1}\right)\text{FT}_D^\dagger\right) \\ &= N^{-1}\text{diag}\left(\Gamma_D - \Gamma_D^2\left(\Gamma_D + \beta^{-1}\mathcal{I}\right)^{-1}\right). \quad (1.48)\end{aligned}$$

1.4 Outline and Contributions

The next chapter expands on previous work investigating the use of the circulant approximation to analyse the properties of the sparse variational inference framework presented in this chapter. The primary contribution of this chapter is to provide an accurate derivation of a quantity of primary concern, and to show that it in doing so several new issues arise.

[7] eg. [Abadi et al., 2015] or Autograd (http://github/com/HIPS/autograd)

The third chapter presents an approach to combining the strengths of the two approximate inference schemes discussed to accelerate inference in a uni-dimensional setting. The methodology developed will require no particular structure in the observation inputs and will have asymptotic complexity far below the $\mathcal{O}(M^3)$ discussed earlier in the chapter.

Chapter 2

Analysis of Sparse Variational GP Approximation

Understanding the properties of the approximate inference scheme presented by Titsias [2009] is essential for its use in practice. Although it is now well understood that increasing the number of pseudo-data will only ever improve the quality of the approximation [Bauer et al., 2016], tightening the ELBO, there is a lack of quantitative understanding of the failure modes of the approximation. As was shown in the introduction, the quality of the approximation can drop quite suddenly when the posterior distribution becomes too complicated for the available pseudo-data to faithfully represent, and it would be very useful to understand at exactly what point this occurs in terms of the properties of the covariance function and observed data. This chapter aims to address exactly this problem.

This chapter builds upon an unpublished technical report produced by Thang Bui entitled 'Efficient Probabilistic Time Frequency Analysis', in which the set up presented below was devised. This work exploited the circulant approximation to obtain simple expressions governing the quality of the posterior approximation found using sparse variational inference for GPs. These are, however, based upon making an assumption about the number of non-zero eigenvalues that will not generally be valid. The focus of this chapter is thus directed towards lifting this assumption to enable a more reliable analysis of the properties of sparse GP approximations.

2.0.1 Background and Preliminaries

Further to the discussion of the variational approximation in the first chapter, it is important to note that a more concise form for the ELBO can be found than that

presented previously [Titsias, 2009]:

$$L = \log \mathcal{N}\left(y \,\middle|\, 0, K_{D,D} + \beta^{-1}\mathcal{I}\right) - \frac{\beta}{2}\mathrm{tr}\left(K_{D,D} - Q_{D,D}\right), \tag{2.1}$$

where $Q_{D,D} := K_{D,Z}K_{Z,Z}^{-1}K_{Z,D}$ is clearly a quantity of importance if an analysis of the properties of the sparse variational approximation is to be made. Noting that the circulant approximation has previously lead to simple expressions for the quantities of interest for GPs, it seems reasonable that one might attempt to construct a scenario in which the circulant approximation could be leveraged to analyse the performance of other approximation techniques. For this to be undertaken for the sparse variational approximation, it will be necessary to represent $Q_{D,D}$ in the frequency domain. Although one can readily apply the circulant approximation to $K_{Z,Z}^{-1}$, the cross-covariance matrix $K_{Z,D}$ is more troublesome. It is thus the primary contribution of this chapter to provide a derivation of a frequency-domain representation for these quantities that requires no additional assumptions beyond those already made by the circulant assumption.

The DFT matrix $\mathrm{FT}_D \in \mathbb{R}^{N\times N}$, where the subscript D indicates that it is the same size as the observed data covariance matrix, is

$$\left(\mathrm{FT}_D\right)_{m,n} = \frac{1}{\sqrt{N}}e^{-2\pi imn/N} \tag{2.2}$$

and the corresponding inverse transform $\mathrm{FT}_D^\dagger \in \mathbb{R}^{N\times N}$ is its conjugate transpose

$$\left(\mathrm{FT}_D^\dagger\right)_{m,n} = \frac{1}{\sqrt{N}}e^{2\pi imn/N}. \tag{2.3}$$

Analogously the DFT matrices associated with the pseudo-data covariance matrix are

$$\left(\mathrm{FT}_Z\right)_{m,n} = \frac{1}{\sqrt{M}}e^{-2\pi imn/M}, \qquad \left(\mathrm{FT}_Z^\dagger\right)_{m,n} = \frac{1}{\sqrt{M}}e^{2\pi imn/M}. \tag{2.4}$$

2.1 Approximating $Q_{D,D}$

Consider the situation in which we observe N regularly sampled data $\{x_n\}_{n=0}^{N-1}$ such that, without loss of generality, $x_n = n$. Now generate M pseudo-inputs by sub-sampling the observation-inputs regularly with period $P = N/M \in \mathbb{N}$. An expression for the circulant approximation to $Q_{D,D}$ of the form $\mathrm{FT}_D\Lambda\mathrm{FT}_D^\dagger$ is found following using the following lemma.

Lemma 2.1.1. *The cross-covariance matrix $K_{Z,D} \in \mathbb{R}^{M\times N}$ between the $M = N/P$ pseudo-inputs X_Z and N observations X_D under the circulant approximation is*

$$K_{Z,D} = G\,\Gamma_D F T_D^\dagger \tag{2.5}$$

where G is an $M \times N$ matrix whose $(m,n)^{th}$ element is

$$(G)_{m,n} = N^{-\frac{1}{2}} e^{-2\pi imn/M}, \tag{2.6}$$

and $\Gamma_D = diag(\gamma_D)$ is the matrix whose diagonal contains the eigenvalues γ_D of the circulant approximation to $K_{D,D}$.

Proof. First consider the matrices A, B, C whose shapes are $M \times N$, $N \times S$ and $M \times M$ respectively, and denote as $A_{\setminus i}$, $C_{\setminus i}$ the matrices resulting from removing the i^{th} row of A and C respectively. It is readily verified that if $C = AB$ then $C_{\setminus i} = A_{\setminus i}B$. We now apply this result iteratively to construct $K_{Z,D}$ from $K_{D,D}$, specifically by removing each row whose row number is not in the set $\{0, P, 2P, ..., (M-1)P\}$. Recalling that approximation $K_{D,D} = \mathrm{FT}_D\Gamma_D\mathrm{FT}_D^\dagger$ under the circulant approximation, it follows that we can approximate $K_{Z,D}$ by removing the same rows from FT_D as we would $K_{D,D}$ to yield the matrix $G \in \mathbb{R}^{M\times N}$, whose $(m,n)^{th}$ element is given by

$$(G)_{m,n} = N^{-\frac{1}{2}} e^{-2\pi i(mP)n/N} = N^{-\frac{1}{2}} e^{-2\pi imn/M}, \tag{2.7}$$

thus completing the proof. □

Equipped with lemma 2.1.1, the circulant approximation to $Q_{D,D}$ can be written as

$$\begin{aligned} Q_{D,D} &= K_{D,Z}K_{Z,Z}^{-1}K_{Z,D} \\ &= \mathrm{FT}_D \left(\mathrm{FT}_Z^\dagger\, G\,\Gamma_D\right)^\dagger \Gamma_Z^{-1} \left(\mathrm{FT}_Z^\dagger\, G\,\Gamma_D\right) \mathrm{FT}_D^\dagger. \end{aligned} \tag{2.8}$$

Considering this expression in stages, beginning with $\mathrm{FT}_Z G$, we see that

$$\begin{aligned} \left(\mathrm{FT}_Z^\dagger G\right)_{m,n} &= (NM)^{-\frac{1}{2}} \sum_{k=0}^{M-1} e^{2\pi ik(m-n)/M} \\ &= \begin{cases} \sqrt{M/N} & : \; n = pM + m, \; p \in \{0, ..., P-1\} \\ 0 & : \; \text{otherwise.} \end{cases} \end{aligned} \tag{2.9}$$

A more intuitive representation of this matrix can be found by expressing it as a block-matrix. Denote by $\mathcal{I}_M$ the identity matrix of size $M \times M$, then noting that

$\sqrt{M/N} = 1/\sqrt{P}$, it follows that

$$\mathrm{FT}_Z^\dagger G = \frac{1}{\sqrt{P}} \Big[\underbrace{\mathcal{I}_M \quad ... \quad \mathcal{I}_M}_{P \text{ blocks}} \Big] \tag{2.10}$$

is the $M \times N$ matrix formed by the horizontal concatenation of P copies of $\mathcal{I}_M/\sqrt{P}$.

Now consider the following division of Γ_D into P^2 blocks, in which 0_M denotes the zero-matrix of size $M \times M$ and $\Gamma_D^{(p)} = \mathrm{diag}\left(\left[(\gamma_D)_{pM} \ \ldots \ (\gamma_D)_{pM+M-1}\right]\right)$ is the p^{th} block of M eigenvalues:

$$\Gamma_D = \begin{bmatrix} \Gamma_D^{(0)} & 0_M & \ldots & 0_M \\ 0_M & \Gamma_D^{(1)} & & \\ \vdots & & \ddots & \\ \underbrace{0_M}_{P \text{ blocks}} & & & \Gamma_D^{(P-1)} \end{bmatrix}. \tag{2.11}$$

In this form it is clear that right-multiplying $\mathrm{FT}_Z^\dagger G$ by Γ_D yields the $M \times N$ matrix

$$\mathrm{FT}_Z^\dagger G \Gamma_D = \frac{1}{\sqrt{P}} \Big[\underbrace{\Gamma_D^{(0)} \quad \ldots \quad \Gamma_D^{(P-1)}}_{P \text{ blocks}} \Big], \tag{2.12}$$

comprising P diagonal blocks. It is now possible to find a block-matrix representation for $\left(\mathrm{FT}_Z^\dagger G \Gamma_D^\dagger\right)^\dagger \Gamma_Z^{-1} \left(\mathrm{FT}_Z^\dagger G \Gamma_D\right)$. First left-multiply the previous result by Γ_Z^{-1} to obtain

$$\Gamma_Z^{-1} \left(\mathrm{FT}_Z^\dagger G \Gamma_D\right) = \frac{1}{\sqrt{P}} \Big[\underbrace{\Gamma_Z^{-1}\Gamma_D^{(0)} \quad \ldots \quad \Gamma_Z^{-1}\Gamma_D^{(P-1)}}_{P \text{ blocks}} \Big], \tag{2.13}$$

and finally right-multiply by $\mathrm{FT}_Z^\dagger G \Gamma_D^\dagger$ to obtain

$$\begin{aligned}
&\left(\mathrm{FT}_Z^\dagger G \Gamma_D^\dagger\right)^\dagger \Gamma_Z^{-1} \left(\mathrm{FT}_Z^\dagger G \Gamma_D\right) \\
&= \frac{1}{P} \Big[\underbrace{\Gamma_D^{(0)} \quad \dots \quad \Gamma_D^{(P-1)}}_{P \text{ blocks}} \Big]^\dagger \Big[\underbrace{\Gamma_Z^{-1}\Gamma_D^{(0)} \quad \dots \quad \Gamma_Z^{-1}\Gamma_D^{(P-1)}}_{P \text{ blocks}} \Big] \\
&= \frac{1}{P} \underbrace{\begin{bmatrix} \Gamma_D^{(0)}\Gamma_Z^{-1}\Gamma_D^{(0)} & \Gamma_D^{(0)}\Gamma_Z^{-1}\Gamma_D^{(1)} & \dots & \Gamma_D^{(0)}\Gamma_Z^{-1}\Gamma_D^{(P-1)} \\ \Gamma_D^{(0)}\Gamma_Z^{-1}\Gamma_D^{(1)} & \Gamma_D^{(1)}\Gamma_Z^{-1}\Gamma_D^{(1)} & & \vdots \\ \vdots & & \ddots & \\ \Gamma_D^{(P-1)}\Gamma_Z^{-1}\Gamma_D^{(0)} & \dots & & \Gamma_D^{(P-1)}\Gamma_Z^{-1}\Gamma_D^{(P-1)} \end{bmatrix}}_{P \text{ blocks}} =: \Lambda, \qquad (2.14)
\end{aligned}$$

in which each block is diagonal. Λ is restricted to be zero everywhere except for elements on the diagonal strips which appear with period M throughout the matrix, however, this matrix is clearly not diagonal in the traditional sense.

2.2 Discussion

In this chapter I have presented a derivation of a method by which to obtain a frequency-domain representation of $Q_{D,D}$, the quantity of central interest when using sparse GP approximations. Unfortunately, it turns out that it is not diagonal in the frequency domain (and consequently not circulant in the time domain) meaning that the use of this representation to analyse the quality of a GP posterior approximation is not going to be straightforward. For example it is not obvious how to invert $Q_{D,D}$ using this representation, or to compute its determinant. However, since it is extremely sparse in the frequency domain, it would seem reasonable that such computations might be tractable.

Chapter 3

Sparse Circular Approximate Inference

3.1 Outline

In this chapter I present and analyse a methodology for performing approximate inference in GPs which leverages both circulant and sparse GP approximations to reduce the asymptotic complexity of approximate inference in a class of problems which neither can solve individually. This formulation simultaneously corrects for the circular covariance introduced at either end of the observed data by the circulant approximation, allows for the inputs of observed data to be arbitrarily located, and drastically increases the number of pseudo-data that can be used to represent the posterior distribution relative to the number that can practically be used in existing variational GP approximations. As in [Hensman et al., 2013], these gains are achieved at the expense of being able to perform approximate inference analytically; an iterative optimisation procedure with low asymptotic complexity is instead required.

Consider again a one dimensional input / output regression problem in which we have $X_D \in \mathbb{R}^N$ inputs and $y \in \mathbb{R}^N$ outputs with no particular structure present in the values of X. We assume the standard GP regression model in which latent function values $f_D \sim \mathcal{N}(0, K_{D,D})$ and $y \sim \mathcal{N}(f_D, \beta^{-1}\mathcal{I})$. Clearly the covariance function cannot be directly approximated as being circulant as it is not Toeplitz. However, if pseudo-data are introduced into the problem their input locations can be chosen such that Toeplitz structure is induced.

Work on Tree-Structured GPs (TSGPs) [Bui and Turner, 2014] also addresses exactly this problem. It is reasonable to ask, therefore, what this new work will add. There are principally two properties of TSGPs that could be improved upon by the methodology

that I will present. Firstly TSGPs do not provide a lower bound to the marginal likelihood or preserve the GP prior. Although this is not necessarily a problem in many scenarios, it would be desirable to show that approximate inference can be performed in large scale problems whilst preserving exactly the GP prior over functions. Secondly, learning covariance function parameters using the TSGP could be problematic for large problems as each iteration of an optimisation procedure requires a pass over the entire data set, whereas I will present a methodology in which data can be mini-batched to provide cheap gradient estimates for the covariance function parameters.

I will now outline the toy regression problem that will be used throughout the remainder of this chapter and discuss the issue of posterior predictive inference under the approximate posterior distributions that result. The rest of the chapter will be devoted to the development of approximate inference techniques, beginning with the posterior mean and followed by the more thorny issue of the posterior covariance.

3.1.1 Toy Experiment Setup

Figure 3.1 illustrates the toy set up that will be used repeatedly throughout this section for proof-of-concept experimentation. 75 pseudo-inputs are spaced regularly on $[-15, 15)$. 750 observed data, drawn jointly from a single GP whose kernel is an Exponentiated Quadratic with length scale $l^2 = 1$ and variance $\sigma^2 = 1$, are split into three groups of 250 observations whose input locations are drawn from uniform distributions on $[-10, 5]$, $[-3, 3]$ and $[6, 10]$ respectively. Observation noise with variance $\beta^{-1} = 10^{-1}$ is added.

The pseudo-input locations in the toy scenario are defined such that coverage of all of the observed data is provided, but also so that there are regions at either end of the region covered by the pseudo-data where there is no observed data. This is intentional as it will mean that, when the prior covariance matrix for the pseudo-data $K_{Z,Z}$ is circularly approximated, the correlations will have minimal impact on the posterior distribution. This point will be discussed further later in the chapter.

I will perform approximate inference in this problem set up using the variational approaches discussed in the preceding chapters, with modifications to exploit the circulant structure introduced in this section. Specifically, an optimal approximate posterior distribution $q(f_Z) = \mathcal{N}(f_Z \,|\, \mu_q, \Sigma_q)$ will be found in such a manner that the asymptotic complexity of inference and posterior prediction is held well below than $\mathcal{O}(M^2)$ and only depends linearly upon the number of observed data.

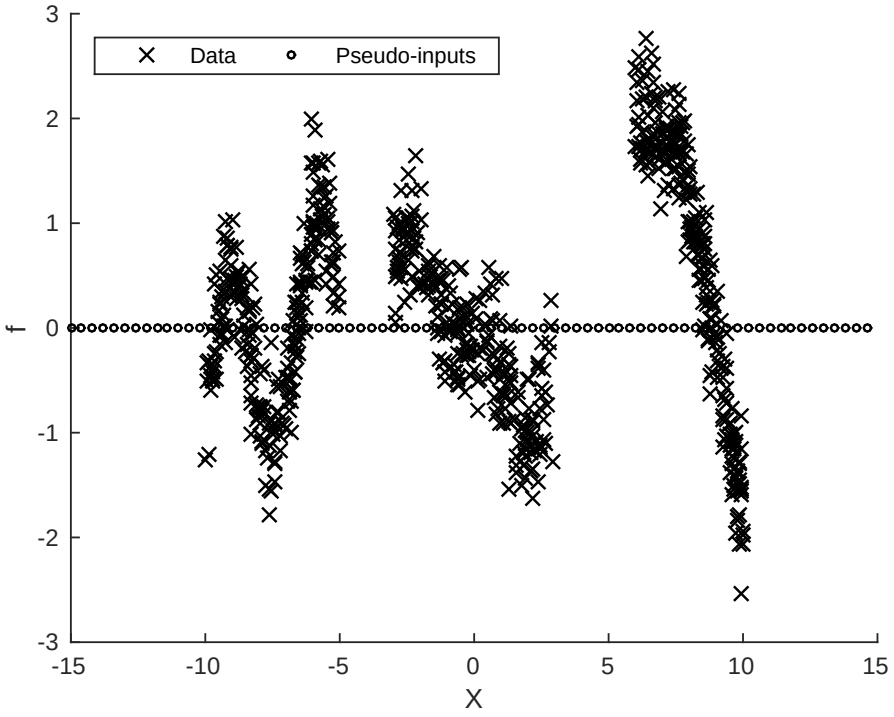


Figure 3.1 Toy problem setup. Pseudo-inputs are positioned such that coverage of the regions containing observed data is provided.

3.1.2 Posterior Predictive Inference

Posterior predictive inference in the mean is particularly efficient when $K_{Z,Z}$ is circulant. Recall that the posterior predictive mean is

$$\mathbb{E}_{f\sim q}\left[f_{\#}\right] = K_{\#,Z}K_{Z,Z}^{-1}\mu_q. \tag{3.1}$$

Diagonalising $K_{Z,Z}$ we obtain

$$\mathbb{E}_{f\sim q}\left[f_{\#}\right] = K_{\#,Z}\mathrm{FT}_Z\Gamma_Z^{-1}\mathrm{FT}_Z^{\dagger}\mu_q \tag{3.2}$$

which, when computed from right to left, requires two $\mathcal{O}\left(M\log M\right)$ FFT operations, one $\mathcal{O}\left(M\right)$ multiplication of a vector by a diagonal matrix and a final $\mathcal{O}\left(MN_{\#}\right)$ operation, where $N_{\#}$ is the number of observations at which we wish to make predictions.

The overall tractability of the posterior predictive covariance

$$\mathrm{Cov}_{f\sim q}\left[f_{\#}\right] = K_{\#,\#} - Q_{\#,\#} + R_{\#,\#}, \tag{3.3}$$

where $Q_{\#,\#} := K_{\#,Z}K_{Z,Z}^{-1}K_{Z,\#}$ and $R_{\#,\#} := K_{\#,Z}K_{Z,Z}^{-1}\Sigma_q K_{Z,Z}^{-1}K_{Z,\#}$, is dependent upon the particular parameterisation chosen for Σ_q. However, the first two terms are independent of Σ_q and $K_{\#,\#}$ is easily evaluated, as is $Q_{\#,\#}$ by diagonalising $K_{Z,Z}$.

3.2 Inference in the Posterior Mean

A key observation is that the optimisation of the ELBO (equation 1.17) with respect to the mean μ_q can be performed independently of the covariance Σ_q. This problem is equivalent to maximising

$$L_\mu := -\frac{\beta}{2}\left(y - K_{D,Z}K_{Z,Z}^{-1}\mu_q\right)^T\left(y - K_{D,Z}K_{Z,Z}^{-1}\mu_q\right) - \frac{1}{2}\mu_q^T K_{Z,Z}^{-1}\mu_q. \tag{3.4}$$

As discussed in section 1.2, the solution to this maximisation problem is

$$\mu_q = \sigma_n^{-2}K_{Z,Z}\left(\sigma_n^{-2}K_{Z,D}K_{D,Z} + K_{Z,Z}\right)^{-1}K_{Z,D}\,y, \tag{3.5}$$

however, since $K_{D,Z}$ is not Toeplitz and therefore not approximately circulant, the multiplication $K_{Z,D}K_{D,Z}$ remains intractable for large (pseudo) data sets as the operation has asymptotic complexity $\mathcal{O}(NM^2)$. Furthermore, as a consequence of not being able to make a circulant approximation to each term in the inversion $(\beta K_{Z,D}K_{D,Z} + K_{Z,Z})^{-1}$, it retains the usual $\mathcal{O}(M^3)$ asymptotic complexity associated with matrix inversion. It is therefore not practical to find the solution in an analytic manner and we must resort to an iterative optimisation.

[Hensman et al., 2013] utilised stochastic gradient ascent in the natural parameters to find the optimal q-distribution. Unfortunately this will not be possible in our case as the covariance Σ_q will be further constrained, beyond simply being symmetric and positive semidefinite (PSD), through the efficient parameterisations discussed in subsequent sections. As such different optimisation procedures must be considered.

3.2.1 Conjugate Gradients

The method of Linear Conjugate Gradients (CG) [Fletcher and Reeves, 1964] is formulated to solve minimisation problems of the form

$$\min_x \quad x^T Ax - b^T x + c. \tag{3.6}$$

By expanding out equation 3.4, it is immediately clear that Linear CG can be directly applied to the minimisation of $-L_\mu$, and thus the maximisation of L_μ, with

$$A := K_{Z,Z}^{-1} + \sigma_n^{-2} K_{Z,Z}^{-1} K_{Z,D} K_{D,Z} K_{Z,Z}^{-1} \tag{3.7}$$

and

$$b := \sigma_n^{-2} K_{Z,Z}^{-1} K_{Z,D} \mathbf{y}. \tag{3.8}$$

Algorithm 3.1 The Linear Conjugate Gradients algorithm as presented in [Nocedal and Wright, 2006].

$r_0 \leftarrow b - Ax_0$
$p_0 \leftarrow r_0$
$k \leftarrow 0$
loop
  $\alpha_k \leftarrow \frac{r_k^T r_k}{p_k^T A p_k}$
  $x_{k+1} \leftarrow x_k + \alpha_k p_k$
  $r_{k+1} \leftarrow r_k - \alpha_k A p_k$
  if r_{k+1} is sufficiently small then exit loop.
  $\beta_k \leftarrow \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}$
  $p_{k+1} \leftarrow r_{k+1} + \beta_k p_k$
  $k \leftarrow k + 1$
end loop
  return x_{k+1}

Algorithm 3.1 shows the computation required to compute the Linear CG procedure. The most computationally intensive terms are clearly $p_k^T A p_k$ and Ap_k, which would both have asymptotic complexity $\mathcal{O}(M^2 N)$ if implemented naively by first computing A and then computing the multiplication Ap_k. A more efficient procedure can be obtained by exploiting the structure present in A. Algorithm 3.2 presents a procedure by which the required computations can be performed using only four $\mathcal{O}(M \log M)$ (inverse) FFT operations, two $\mathcal{O}(MN)$ operations ($K_{D,Z}\lambda_1$ and $K_{Z,D}\lambda_2$) and a handful of $\mathcal{O}(M)$ operations.

3.2.2 Optimisation with Stochastic Gradients

The Linear CG algorithm discussed previously has the disadvantage of scaling linearly with the amount of data observed, which is less than ideal when handling large modern data sets. Although the $\mathcal{O}(MN)$ operations are readily parallelised, an ideal procedure

Algorithm 3.2 Algorithm to compute $p_k^T A p_k$ and Ap_k efficiently by exploiting the structure present in A as it is defined in equation 3.7.

Require: p_k, $K_{D,Z}$ and Γ_Z where Γ_Z is the diagonal matrix of eigenvalues of the circulant approximation to $K_{Z,Z}$.
$\lambda_1 \leftarrow \mathrm{FT}_Z \Gamma_Z^{-1} \mathrm{FT}_Z^\dagger p_k$
$\lambda_2 \leftarrow K_{D,Z} \lambda_1$
$\lambda_3 \leftarrow \mathrm{FT}_Z \Gamma_Z^{-1} \mathrm{FT}_Z^\dagger K_{Z,D} \lambda_2$
$p_k^T A p_k \leftarrow p_k^T \lambda_3 + p_k^T \lambda_1$
$Ap_k \leftarrow \lambda_3 + \lambda_1$
return Ap_k, $p_k^T A p_k$

would utilise the stochastic gradients obtained that can be obtained when mini-batches of the data are processed.

Algorithm 3.3 The AdaGrad optimisation algorithm [Duchi et al., 2011] applied to maximise L_μ.

Require: Initial posterior mean $\mu_0 \in \mathbb{R}^M$ and learning rate $\eta > 0$.
$G \leftarrow \{0\}^M$
$\mu \leftarrow \mu_0$
for $t = 1, 2, ...$ until convergence **do**
  $g \leftarrow \nabla_\mu L_\mu(\mu)$
  for $j \in \{0, ..., M-1\}$ **do**
    $G_j \leftarrow G_j + g_j^2$
    $\mu_j \leftarrow \mu_j + \eta g_j / \sqrt{G_j}$
  end for
end for
$\mu_q \leftarrow \mu$

The 'vanilla' stochastic decent (SGD) procedure is one such procedure, but its performance is known to rely heavily on a good choice of learning rate and the particular decay schedule chosen to satisfy the Robbins-Monro conditions [Robbins and Monro, 1951]. Several algorithms have been developed in recent years to attempt to deal with these problems including AdaGrad [Duchi et al., 2011], AdaDelta [Zeiler, 2012] and Adam [Kingma and Ba, 2014]. Each algorithm requires access only to an unbiased estimate of the gradient of the function of interest w.r.t. the parameters being optimised, and each explicitly avoids the need to access any higher order derivatives.

The gradient of L_μ w.r.t. μ_q is

$$\nabla L_\mu = K_{Z,Z}^{-1} \left(\beta K_{Z,D} \left(y - K_{D,Z} K_{Z,Z}^{-1} \mu_q \right) - \mu_q \right) \tag{3.9}$$

which can again be computed with four FFT operations, two $\mathcal{O}(MN)$ operations and a few $\mathcal{O}(M)$ operations. The difference in complexity between this procedure and CG is that the data may be mini-batched, resulting in the two $\mathcal{O}(MN)$ terms becoming dramatically faster to compute as N is much smaller. The results presented below use the AdaGrad algorithm as it was found to perform better than AdaDelta or Adam with little tuning of the parameters of the procedures. More attention to optimising the parameters of AdaDelta or Adam may have resulted in better performance than AdaGrad.

3.2.3 Toy Experiments and Results

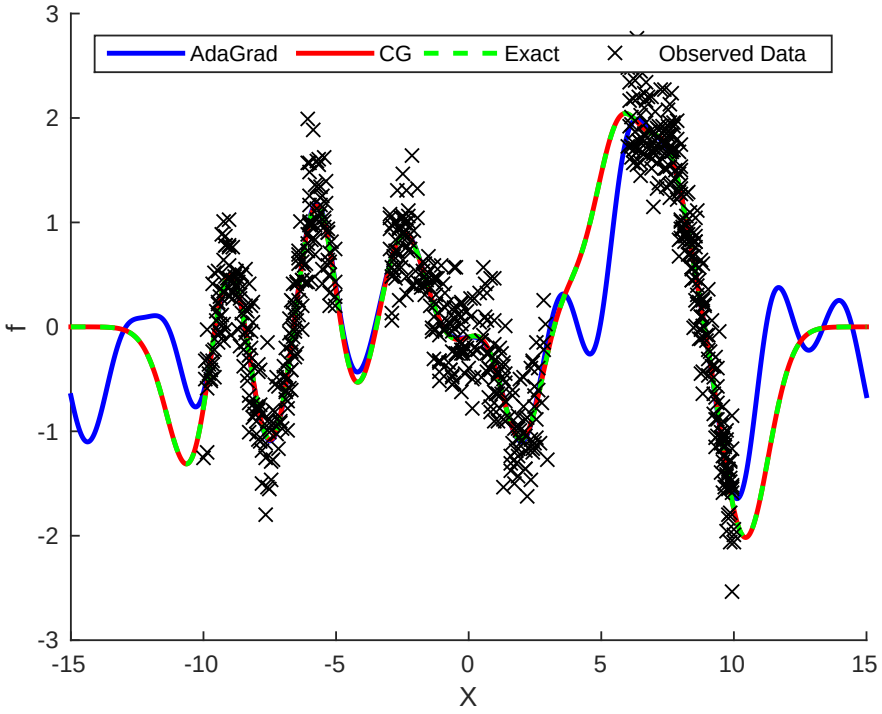


Figure 3.2 Comparison of the posterior mean found using exact inference, AdaGrad and CG after 150×750 data observations each. The results found with CG are (visually) indistinguishable from those found using exact inference. AdaGrad clearly has yet to converge in sections of the input domain without observed data.

Figure 3.2 shows the posterior mean converged to by both AdaGrad (with mini-batches comprising 75 data) and CG. The approximate posterior mean computed using 150 iterations of CG is visually indistinguishable from the exact posterior mean. The results using AdaGrad are, however, less impressive; although the approximate posterior mean obtained is good in sections of the input domain containing observations, inference has broadly failed where there is missing data. The quality of extrapolation, however,

appears to depend strongly upon the width of the section of missing data. For example, the AdaGrad posterior mean is substantially more accurate on $[-5, -3]$ than on $[3, 6]$

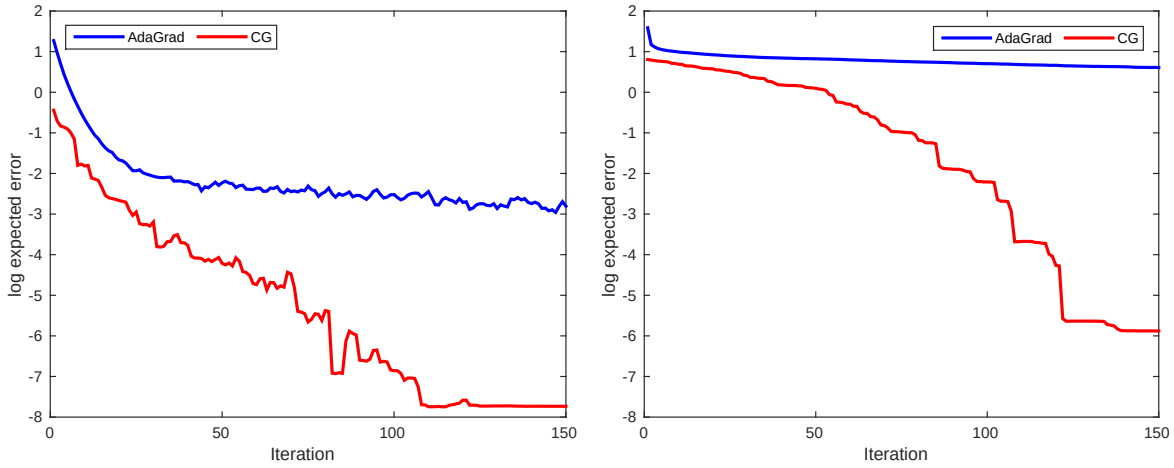


Figure 3.3 The log of the expected absolute error in sections of input domain containing data (left) and sections not containing data (right). Graphs computed after each algorithm has seen 750 data (one iteration for CG, 10 iterations of AdaGrad with 75 data in each mini-batch). Expectations approximated using Simpson's rule quadrature.

Figure 3.3 shows how the performance of inference varies depending upon whether we consider the in-sample ($x \in \{[-10, -5], [-3, 3], [6, 10]\}$) or out-of-sample ($x \in \{[-15, 10), (-5, -3), (3, 6), (10, 15]\}$) regions of the input domain. The left hand plot shows that CG obtains substantially better performance, even on the in-sample regions, but that AdaGrad does manage to obtain small errors in these regions of the input domain. The right hand plot reinforces quantitatively what is discerned qualitatively from visual inspection of figure 3.2, that AdaGrad performs poorly in out-of-sample regions of the input domain. Experiments in which optimisation was performed using AdaGrad with full batches showed similar results to those with mini-batches.

3.3 Stationary Posterior Approximation

I now move on to consider how to efficiently approximate the posterior covariance. Were Σ_q simply constrained to be symmetric and PSD, it would be necessary to store $M(M+1)/2$ unique real numbers. This is undesirable as it would induce a component of our approximate inference scheme with $\mathcal{O}(M^2)$ asymptotic complexity. A parameterisation of the covariance matrix must therefore be chosen in which the number of parameters scales linearly with M. There are a couple of obvious initial choices; that the covariance matrix chosen to be (band) diagonal or circulant. The first of these choices is clearly

going to yield a poor approximation to the posterior covariance in all situations in which the covariance function is not highly localised. The second option, however, merely implies stationarity and some circular covariance at either end of the pseudo-data. The added circular covariance is inconsequential as we have already prescribed to have extra pseudo-data at either end of the observed data to prevent the circular approximation to the prior covariance interfering with the approximate posterior distribution around the observed data. The assumption of stationarity is in general a poor one as posterior covariances are highly non-stationary; for example missing data in a regularly-sampled time series should induce higher uncertainty in the value of the pseudo-outputs whose input locations are close to those of the missing data than is present in the rest of the time series. By definition this cannot be captured by a stationary posterior distribution.

However, for example if only one observation is occasionally missing, it may be the case that the amount of extra uncertainty is minimal meaning that a stationary Σ_q is reasonable. The other issue with a circular Σ_q is that, as we have introduced extra pseudo-data outside of the observed data, there are essentially blocks of 'missing data' at either end of the time-series. If we assume that the time-series is long in comparison to the length of the overhanging pseudo-data this may not be a problem from the perspective of learning kernel parameters as it should introduce only a small amount of bias into the ELBO.

Despite the potential problems associated with a circulant Σ_q, I will proceed to investigate its properties as it may be useful in certain cases and results in particularly tractable forms for the ELBO and its gradients.

Firstly define

$$L_\Sigma := -\frac{\beta}{2}\mathrm{tr}\left(K_{D,Z}K_{Z,Z}^{-1}\Sigma_q K_{Z,Z}^{-1}K_{Z,D}\right) - \frac{1}{2}\mathrm{tr}\left(K_{Z,Z}^{-1}\Sigma_q\right) + \frac{1}{2}\log|\Sigma_q|, \tag{3.10}$$

the components of the ELBO L which depend upon Σ_q. Now constraining Σ_q to be circulant such that

$$\Sigma_q := \mathrm{FT}_Z\Gamma_q\mathrm{FT}_Z^\dagger \tag{3.11}$$

for some diagonal Γ_q with positive elements $\{\gamma_m\}_{m=0}^{M-1}$, and applying the circulant approximation to the prior covariance

$$K_{Z,Z} = \mathrm{FT}_Z\Gamma_Z\mathrm{FT}_Z^\dagger, \tag{3.12}$$

we find that L_Σ simplifies significantly:

$$\begin{aligned} L_\Sigma &\approx -\frac{1}{2\beta^{-1}}\mathrm{tr}\left(K_{D,Z}\mathrm{FT}_Z\Gamma_Z^{-1}\Gamma_q\Gamma_Z^{-1}\mathrm{FT}_Z^\dagger K_{Z,D}\right) - \frac{1}{2}\mathrm{tr}\left(\Gamma_Z^{-1}\Gamma_q\right) + \frac{1}{2}\sum_{m=0}^{M}\log\gamma_m \\ &= -\frac{1}{2\beta^{-1}}\sum_{m=0}^{M-1}\sum_{n=0}^{N-1}\frac{(\gamma_q)_m}{(\gamma_Z)_m^2}\left|\left(\tilde{K}_{Z,D}\right)_{m,n}\right|^2 - \frac{1}{2}\mathrm{tr}\left(\Gamma_Z^{-1}\Gamma_q\right) + \frac{1}{2}\sum_{m=0}^{M}\log\gamma_m \end{aligned} \tag{3.13}$$

where $\tilde{K}_{Z,X} := \mathrm{FT}_Z^\dagger K_{Z,D}$. Here use has been made of the identities

$$\begin{aligned} \mathrm{FT}_Z\mathrm{FT}_Z^\dagger &= \mathcal{I}, \\ \mathrm{tr}\left(\mathrm{FT}_Z\Gamma\mathrm{FT}_Z^\dagger\right) &= \mathrm{tr}\left(\Gamma\right), \\ \left|\mathrm{FT}_Z\Gamma\mathrm{FT}_Z^\dagger\right| &= \left|\Gamma\right|, \end{aligned}$$

where $A \in \mathbb{R}^{M\times M}$ is some arbitrary conformal matrix.

The second and third terms in equation 3.13 clearly have asymptotic complexity $\mathcal{O}\left(M\right)$, whereas the first term has asymptotic complexity $\mathcal{O}\left(NM\log M\right)$ as there are N DFT computations required to compute both $\tilde{K}_{Z,X}$.

Noting that the first term in equation 3.13 contains a sum over n, we can obtain a cheap unbiased estimator by mini-batching the data. I use the following single-sample Monte Carlo estimator of the quantity

$$L_\Sigma \approx -\frac{N}{2\beta^{-1}}\sum_{m=0}^{M-1}\frac{(\gamma_q)_m}{(\gamma_Z)_m^2}\left|\left(\tilde{K}_{Z,D}\right)_{m,n}\right|^2 - \frac{1}{2}\mathrm{tr}\left(\Gamma_Z^{-1}\Gamma_q\right) + \frac{1}{2}\sum_{m=0}^{M}\log\gamma_m \tag{3.14}$$

where the index n is drawn from a uniform distribution over $\{0, ..., N-1\}$.

The gradient of L_Σ w.r.t. Γ_q is

$$\nabla_{\Gamma_q}L_\Sigma = -\frac{1}{2}\left[\beta\Gamma_Z^{-1}\tilde{K}_{Z,D}\tilde{K}_{D,Z}\Gamma_Z^{-1} + \Gamma_Z^{-1} - \Gamma_q^{-1}\right]\odot\mathcal{I}, \tag{3.15}$$

where $\odot\mathcal{I}$ denotes the Hadamard product with the identity matrix (this element-wise product appears as Γ_q is constrained to be diagonal). Since only the diagonal of $\nabla_{\Gamma_q}L_\Sigma$ is required, only the diagonal of the first term need be computed:

$$\beta\left(\Gamma_Z^{-1}\tilde{K}_{Z,D}\tilde{K}_{D,Z}\Gamma_Z^{-1}\right)_m = \beta(\gamma_Z)_m^{-2}\sum_{n=0}^{N-1}\left|\left(\tilde{K}_{Z,D}\right)_{m,n}\right|^2. \tag{3.16}$$

The most computationally intensive operation required, with asymptotic complexity $\mathcal{O}\left(NM\log M\right)$, is the N inverse DFTs needed to compute $\tilde{K}_{Z,D}$. Again the complexity

can be reduced dramatically to $\mathcal{O}(M \log M)$ by considering a single-sample Monte Carlo estimate

$$L_\Sigma \approx -\frac{1}{2}\left[N\beta\Gamma_Z^{-1}A\Gamma_Z^{-1} + \Gamma_Z^{-1} - \Gamma_q^{-1}\right], \tag{3.17}$$

where $A \in \mathbb{R}^{M\times M}$ is a diagonal matrix whose m^{th} element is

$$A_{m,m} = (\gamma_Z)_m^{-2}\left|\left(\tilde{K}_{Z,D}\right)_{m,n}\right|^2. \tag{3.18}$$

Furthermore, noting that the element-wise product distributes over the sum operator, we can immediately see from equation 3.15 that the exact solution for the eigenvalues of the posterior covariance matrix is

$$\Gamma_q = \left(\sigma_n^{-2}\Gamma_Z^{-1}\tilde{K}_{Z,D}\tilde{K}_{D,Z}\Gamma_Z^{-1} \odot \mathcal{I} + \Gamma_Z^{-1}\right)^{-1} \tag{3.19}$$

which again has asymptotic complexity $\mathcal{O}(NM \log M)$ as all of the inversion operations are taken over diagonal matrices and therefore only require computation time linear in M.

3.3.1 Toy Experiments and Results

To highlight the nature of the posterior covariance approximation that a circulant approximation provides, I investigate the solution on the toy data set described at the start of the chapter.

Figure 3.4 shows the convergence of the optimisation procedure is swift, only around fifteen data are processed before convergence to the exact solution is reached. It is particularly interesting that the variance of the estimator appears to be negligible as there is minimal optimisation noise (the 'learning curve' of the optimiser is smooth and apparently noise-free).

Figure 3.5 shows the circulant solution converged to by the optimiser, which fails to adjust to the heteroscedasticity in the exact posterior. [1] As shown in the right hand image, the circulant posterior slightly overestimates the marginal variance in regions of the input domain where data has been observed, while severely underestimating the marginal variance where there are no observations.

In summary, a circulant approximation to the covariance of the approximation posterior results in tractable inference with $\mathcal{O}(M \log M)$ complexity when the mini-

[1] The exact circulant solution is not depicted as it is visually indistinguishable from the solution converged to by the optimiser.

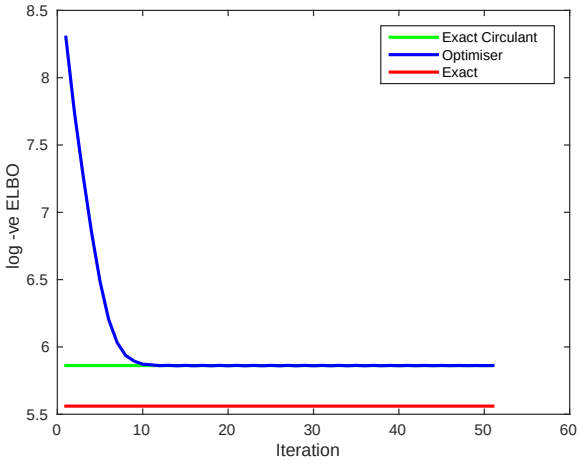


Figure 3.4 Convergence of the log of the negative ELBO as AdaGrad is iterated with a mini-batch size of 1 and learning rate 2.0. Convergence to the exact solution, whose ELBO is depicted as a constant line at its solution, is achieved after approximately 15 iterations. Interestingly the variance introduced by mini-batching appears to be negligible.

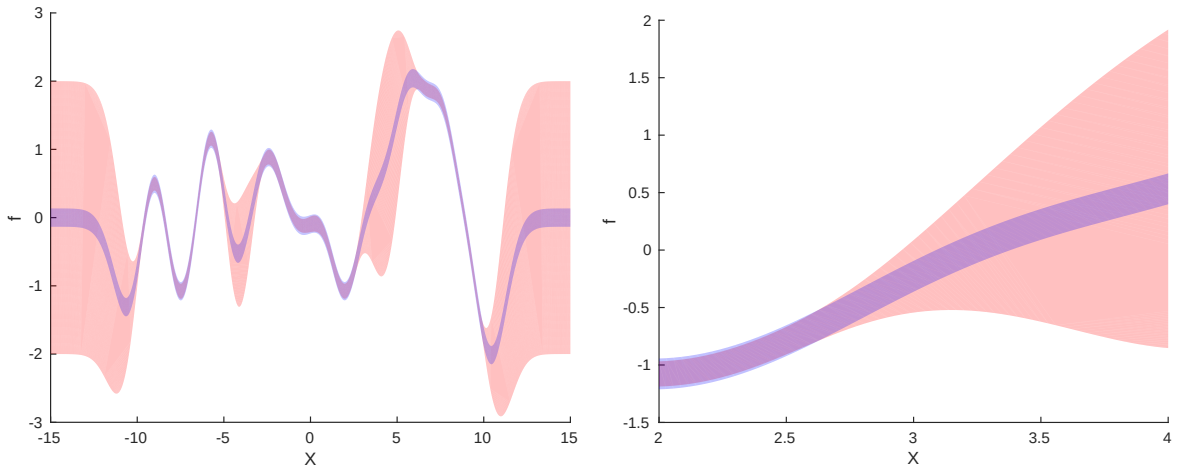


Figure 3.5 Filled regions represent marginal posterior variance. Red fill is computed under the exact posterior whilst blue is computed under the circulant approximate posterior. The mean was found using the Conjugate Gradients algorithm described in the previous section. Left: The marginal variance computed using the circulant covariance appears to be accurate in regions of high data density but as expected fails to accurately adapt to the additional uncertainty where there is no data. Right: Zoomed in to the input range $[2, 4]$. The circular approximation has very slightly over-estimated the marginal variance where there is a large amount of data (roughly, inputs < 3), and severely underestimated it where there is none (roughly, inputs > 3).

batches are used, but undesirably low posterior variance estimates wherever there is missing data.

3.4 Non-Stationary Posterior Approximation

Although a circulant parameterisation appears to be a computationally convenient choice for Σ_q, it is clearly insufficient to capture important aspects of the posterior distribution of a GP in most scenarios, namely variations in posterior variance. It is therefore necessary to consider other types of parameterisation for Σ_q that are more expressive but remain tractable.

The obvious way forward is to combine tractable matrices such that the resulting Σ_q is non-stationary. I will consider two types of matrices; diagonal and circulant. The matrix resulting from the combination of these matrices must be symmetric, PSD, and both its inverse and determinant must be easily computed. This unfortunately precludes the summation of two such matrices as, although the result will be symmetric, computing the determinant and inverse becomes difficult. To see this, consider that for some circulant matrix K with positive eigenvalues Γ and diagonal matrix W whose diagonal elements are $\{w_m\}_{m=0}^{M-1}$,

$$(K+W)^{-1} = \left(\mathrm{F}\mathrm{T}\Gamma\mathrm{F}\mathrm{T}^\dagger + W\right)^{-1} = \mathrm{F}\mathrm{T}\left(\Gamma + \mathrm{F}\mathrm{T}W\mathrm{F}\mathrm{T}^\dagger\right)^{-1}\mathrm{F}\mathrm{T}^\dagger. \tag{3.20}$$

Intuitively, the problem is that the manner in which we have previously been able to invert a circulant matrix easily is to perform the inverse in the frequency domain, in which everything that we have needed to invert is diagonal and therefore simple to invert. The diagonal matrix W is not diagonal in the frequency domain as it is not circulant. A similar argument holds for computing the determinant of the summation:

$$|K+W| = \left|\mathrm{F}\mathrm{T}\Gamma\mathrm{F}\mathrm{T}^\dagger + W\right| = \left|\Gamma + \mathrm{F}\mathrm{T}W\mathrm{F}\mathrm{T}^\dagger\right| \tag{3.21}$$

Thus it appears to be the case that we cannot exploit the particularly simple structure of the two matrices if we sum them.

Multiplication of a circulant matrix by a pair of diagonal matrices, however, is much more tractable. Considering $\Sigma_q := WKW$, we can easily see that it is symmetric by noting that K is symmetric, from which it follows that

$$(WKW)_{m,n} = w_m w_n K_{m,n} = w_n w_m K_{n,m} = (WKW)_{n,m}. \tag{3.22}$$

It must also be PSD, for which it is necessary to show that $x^T(WKW)x \geq 0$ for any vector x [Rasmussen and Williams, 2006]. This can be seen to be the case by first recalling that K has positive eigenvalues by definition, and is therefore itself PSD. Thus we know that for any vector z, $z^T K z > 0$, hence $x^T(WKW)x = (Wx)^T K(Wx) > 0$.

Furthermore, the inverse can be computed as

$$(WKW)^{-1} = W^{-1}K^{-1}W^{-1} \tag{3.23}$$

and the determinant

$$|WKW| = |K|\,|W|^2\,. \tag{3.24}$$

Having established the candidacy of the proposed form, it is necessary now to consider further constraints on the values taken by W and K. By noting that observing additional data should never increase the magnitude of the covariance, the largest values admissible in the posterior covariance should be those given by the prior. One way to enforce this is to allow the circulant matrix to take on arbitrary values and constrain W based on them to prevent larger values than the prior covariance being present in WKW. However, a simpler approach is simply to define $K := K_{Z,Z}$, the prior covariance between the pseudo-inputs, and to require that $\forall m \in \{0, ..., M-1\},\ w_m \in [0, 1]$[2].

3.4.1 Efficient ELBO Evaluation

Under this parameterisation, the ELBO w.r.t. terms involving Σ_q is

$$L_\Sigma := -\frac{1}{2\beta^{-1}}\mathrm{tr}\left(K_{D,Z}K_{Z,Z}^{-1}WK_{Z,Z}WK_{Z,Z}^{-1}K_{Z,D}\right) - \frac{1}{2}\mathrm{tr}\left(K_{Z,Z}^{-1}WK_{Z,Z}W\right) + \log|W| \tag{3.25}$$

The first and third terms are easy to evaluate efficiently by applying the circulant decomposition to $K_{Z,Z}$, resulting in leading order asymptotic time-complexities of $\mathcal{O}(NM\log M)$ and $\mathcal{O}(M)$ respectively, where N is either the total number of data or the size of the mini-batches used.

It is not immediately clear, however, how to efficiently compute the second trace term. To devise a procedure for the efficient computation of this quantity, consider the following slightly more general problem. [3]

[2] The problem is converted back into an unconstrained optimisation via the monotonic transform $w = \log(1 + \exp(\hat{w}))$. $\hat{w}\mathbb{R}^M$ is then optimised in place of w

[3] I would like to thank Wessel Bruinsma for some discussions that helped lead towards my derivation of lemma 3.4.1. After deriving it I was disappointed to discover that it is, in fact, a widely known property of the Hadamard product.

Let $A, B \in \mathbb{R}^{M\times M}$ be arbitrary circulant matrices specified by vectors $a, b \in \mathbb{R}^M$. Furthermore let $V, W \in \mathbb{R}^{M\times M}$ be diagonal matrices defined by vectors $v, w \in \mathbb{R}^M$ respectively. We wish to evaluate the trace

$$\operatorname{tr}(AVBW). \tag{3.26}$$

Naively this has asymptotic time-complexity $\mathcal{O}(M^3)$, but this can be reduced by simply exploiting the circulant structure of A and B to express the computation as

$$\operatorname{tr}(AVBW) = \operatorname{tr}\left(\mathrm{FT}\Gamma_a\mathrm{FT}^\dagger V\mathrm{FT}\Gamma_b\mathrm{FT}^\dagger W\right), \tag{3.27}$$

which has asymptotic time-complexity $\mathcal{O}(M^2 \log M)$. With some additional insight, it emerges that the computation can in fact be achieved in $\mathcal{O}(M \log M)$ time. To see this a couple of simple lemmas are required.

Lemma 3.4.1. *Let $v, w \in \mathbb{R}^M$ be the diagonal elements of $V := diag(v)$ and $W := diag(w)$. Given the matrices $C, D \in \mathbb{R}^{M\times M}$ with arbirary values,*

$$tr(CVDW) = tr(WCVD) = w^T(C \odot D^T)v. \tag{3.28}$$

Proof.

$$\operatorname{tr}(CVDW) = \operatorname{tr}(WCVD) \tag{3.29}$$

holds as a consequence of the invariance of the trace under cyclic permutations. From here the proof follows by expanding the trace elementwise.

$$\begin{aligned}
\operatorname{tr}(WCVD) &= \sum_{m=0}^{M-1} w_m \sum_{n=0}^{M-1} C_{m,n}v_nD_{n,m} \\
&= \sum_{m=0}^{M-1} w_m \sum_{n=0}^{M-1} v_nC_{m,n}\left(D^T\right)_{m,n} \\
&= \sum_{m=0}^{M-1} w_m \sum_{n=0}^{M-1} v_n\left(C \odot D^T\right)_{m,n} \\
&= w^T\left(C \odot D^T\right)v
\end{aligned}$$

□

Lemma 3.4.2. *Let $a, b \in \mathbb{R}^M$ define the circulant matrices $A := circ(a)$, $B := circ(b)$. $A \odot B^T$ is another circulant matrix $C := circ(c)$ where the m^{th} element of $c \in \mathbb{R}^M$ is $a_m b_{M-m-1}$.*

Proof. The proof requires simply that we show that circulant matrices are closed under transposition and Hadamard product and that c is defined correctly.

Closure under transposition can be shown constructively by first defining $f \in \mathbb{R}^M$ with elements $\{f_0\ f_1\ ...\ f_{M-1}\}$. Now construct $g \in \mathbb{R}^M$ such that $g_m = f_{M-m-1}$. It is then simple to verify that

$$\operatorname{circ}(f)^T = \operatorname{circ}(g).$$

Closure under the Hadamard product is shown by observing that for circulant matrices $\operatorname{circ}(f)$, $\operatorname{circ}(g)$

$$\begin{aligned}(\operatorname{circ}(f) \odot \operatorname{circ}(g))_{m,n} =& \operatorname{circ}(f)_{m,n} \operatorname{circ}(g)_{m,n} \\ =& f_{(n-m) \bmod M} g_{(n-m) \bmod M}.\end{aligned}$$

Define $h_m := f_m + g_m$, then

$$(\operatorname{circ}(f) \odot \operatorname{circ}(g))_{m,n} = h_{(n-m) \bmod M} = \operatorname{circ}(h)_{m,n}.$$

□

Now applying Lemma 3.4.1 to equation 3.26 yields

$$\operatorname{tr}(AVBW) = w^T \left(A \odot B^T\right) v. \tag{3.30}$$

Recalling that we defined A and B to be circulant, it follows from lemma 3.4.2 that $A \odot B^T$ must also be circulant. Thus it follows that

$$\operatorname{tr}(AVBW) = w^T \mathrm{F} \mathrm{T} \Gamma \mathrm{F} \mathrm{T}^\dagger v, \tag{3.31}$$

where $\Gamma := \operatorname{diag}(\gamma_{a \odot b})$ and $(\gamma_{a \odot b})_m = a_m b_{M-m-1}$. This quantity now requires two $\mathcal{O}(M \log M)$ (I)FFT operations and two $\mathcal{O}(M)$ operations (multiplication of a vector by a diagonal matrix and inner product between two vectors), thus the asymptotic complexity is $\mathcal{O}(M \log M)$ to leading order.

Applying this result to the (previously awkward) trace term in the KL-divergence yields

$$\operatorname{tr}\left(K_{Z,Z}^{-1} W K_{Z,Z} W\right) = w^T \left(K_{Z,Z}^{-1} \odot K_{Z,Z}\right) w = w^T \mathrm{F} \mathrm{T}_Z \Lambda \mathrm{F} \mathrm{T}_Z^\dagger w \tag{3.32}$$

where $\Lambda := \text{diag}(\lambda)$ and $\lambda \in \mathbb{R}^M$ contains the eigenvalues of the circulant matrix $K_{Z,Z}^{-1} \odot K_{Z,Z}$.

Combining this with the asymptotic complexities for the other terms in L_Σ, the leading order asymptotic complexity is $\mathcal{O}(NM \log M)$. With single-sample mini-batching this drops to $\mathcal{O}(M \log M)$.

3.4.2 Efficient Gradient Computation

Having established that the all terms on the ELBO depending upon w can be computed efficiently, it is important to establish that gradients can also be computed efficiently. Considering each terms equation 3.25 as it appears,

$$-\frac{1}{2}\beta\nabla_{W\odot\mathcal{I}}\text{tr}\left(K_{D,Z}K_{Z,Z}^{-1}WK_{Z,Z}WK_{Z,Z}^{-1}K_{Z,D}\right) = \left[K_{Z,Z}WK_{Z,Z}^{-1}K_{Z,D}K_{D,Z}K_{Z,Z}^{-1}\right]\odot\mathcal{I}. \tag{3.33}$$

Observing that due to the diagonal constraint on W we require only the diagonal elements of the result, the required computations simplify to

$$\text{diag}\left(K_{Z,Z}WK_{Z,Z}^{-1}K_{Z,D}K_{D,Z}K_{Z,Z}^{-1}\right) = \text{diag}\left(\left[K_{Z,Z}WK_{Z,Z}^{-1}K_{Z,D}\right]\left[K_{D,Z}K_{Z,Z}^{-1}\right]\right), \tag{3.34}$$

where bracketing has been introduced to indicate the most efficient manner in which to perform the computations. We can now see that the vector of gradients w.r.t. w is simply the following Hadamard product between column vectors:

$$\left[K_{Z,Z}WK_{Z,Z}^{-1}K_{Z,D}\right]^T \odot \left[K_{D,Z}K_{Z,Z}^{-1}\right], \tag{3.35}$$

where each vector requires $\mathcal{O}(M \log M)$ time to compute if single-sample mini-batching is used such that $K_{Z,D} \in \mathbb{R}^{M\times 1}$.

The gradients of the trace term in equation 3.25 can be computed efficiently by exploiting the result in equation 3.32 to obtain

$$\text{diag}\left(\nabla_{W\odot\mathcal{I}}\text{tr}\left(K_{Z,Z}^{-1}WK_{Z,Z}W\right)\right) = 2\text{FT}_Z\Lambda\text{FT}_Z^\dagger w, \tag{3.36}$$

and finally

$$\text{diag}\left(\nabla_{W\odot\mathcal{I}}\log|W|\right) = \begin{bmatrix} w_0^{-1} & \dots & w_{M-1}^{-1}\end{bmatrix}^T. \tag{3.37}$$

To leading order, the gradient computations are therefore require $\mathcal{O}(M \log M)$ operations.

3.4.3 Concavity of the Objective Function

It is important to know whether or not the solution found to an optimisation problem is likely to be sensitive to the manner in which the problem is initialised. The previous two problems involved quadratic forms in which the maximisation of the ELBO is well known to be concave, meaning that there is only one local maximum in the objective function, the global maximum. It is not obvious, however, that the objective function is concave in this instance. To show concavity it is necessary to show that the Hessian H of the objective function w.r.t. the parameters being optimised is negative semidefinite [Nocedal and Wright, 2006], such that for any vector x

$$x^T H x \leq 0. \tag{3.38}$$

To show this, I first appeal again to lemma 3.4.1 to re-write equation 3.25 as

$$L_\Sigma = -\frac{\beta}{2} w^T \left(\left[K_{Z,Z}^{-1} K_{Z,D} K_{D,Z} K_{Z,Z}^{-1} \right] \odot K_{Z,Z} \right) w - \frac{1}{2} w^T \left(K_{Z,Z}^{-1} \odot K_{Z,Z} \right) w + \log |W| \,. \tag{3.39}$$

First note that the *Schur Product Theorem* states that the Hadamard product of two positive definite matrices is also positive definite, which implies an analogous relationship between PSD matrices and the Hadamard product. To show that the Hessians of the first two terms $-\frac{\beta}{2} w^T \left(\left[K_{Z,Z}^{-1} K_{Z,D} K_{D,Z} K_{Z,Z}^{-1} \right] \odot K_{Z,Z} \right) w$ and $-\frac{1}{2} w^T \left(K_{Z,Z}^{-1} \odot K_{Z,Z} \right) w$ are negative semidefinite, it is therefore sufficient to show that each of the three matrices $K_{Z,Z}$, $K_{Z,Z}^{-1}$ and $K_{Z,Z}^{-1} K_{Z,D} K_{D,Z} K_{Z,Z}^{-1}$ are PSD (due to the negation of each of the terms). $K_{Z,Z}$ is trivially PSD as it is defined to be a valid covariance matrix, whilst $K_{Z,Z}^{-1}$ must also be PSD as its eigenvalues are the reciprocal of the eigenvalues of $K_{Z,Z}$, whose eigenvalues must be positive as it is PSD. Also for any arbitrary vector x

$$x^T \left(K_{Z,Z}^{-1} K_{Z,D} K_{D,Z} K_{Z,Z}^{-1} \right) x = \left(K_{D,Z} K_{Z,Z}^{-1} x \right)^T \left(K_{D,Z} K_{Z,Z}^{-1} x \right) \geq 0. \tag{3.40}$$

Finally, the Hessian of $-\log |W| = \text{diag}\left([w_0^{-2} \; ... \; w_{M-1}^{-2}] \right)$, and its negation is therefore negative semidefinite.

It has been shown that each term individually is log concave owing to the corresponding Hessians being PSD, thus establishing that their summation, L_Σ, is also concave. This result is important as it removes the need to perform experiments in the following section to establish sensitivity of the final solution obtained to the initialisation.

3.4.4 Toy Experiments and Results

As there is not a closed form solution available w, the only variational parameter that considered in this section, it is not possible to compare against the ground truth as found in the previous two sections. However, a good approximation to the ground truth can be found using the BFGS optimisation procedure [Broyden, 1970; Fletcher, 1970; Goldfarb, 1970; Shanno, 1970]. Although the asymptotic complexity of BFGS is super-linear in the parameters, and is therefore not suitable for the large-scale inference to which this method will ultimately be applied, it is useful in this small-scale experiment to judge whether the procedures which are more scalable converge to a near-optimal solution.

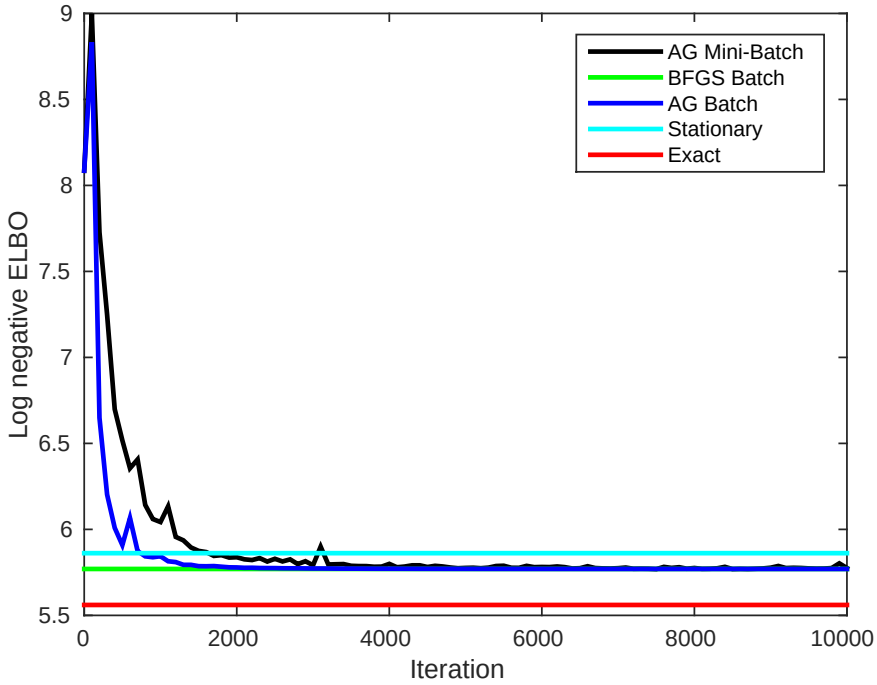


Figure 3.6 Optimisation convergence results for non-stationary Σ_q. The constant red and teal lines are the log marginal likelihood of the GP and ELBO of the stationary Σ_q approximation (from the previous section) respectively. The constant green line is the solution found by BFGS. The blue line is the learning curve associated with learning using AdaGrad and full-batches of data, whilst the black line is AdaGrad with mini-batches of size 1. AdaGrad converges to the solution found by BFGS whether full or mini-batches are used. The solution found by the non-stationary posterior distribution is substantially better in terms of the ELBO than the solution found using a stationary approximate posterior.

Figure 3.6 shows the results, in terms of the convergence in ELBO, of an experiment in which the approximate posterior mean was first found using the CG procedure described

in section 3.2. Each element of w was then initialised to 0.9 and optimised w.r.t. the ELBO using BFGS, AdaGrad with full batches and AdaGrad with mini-batches of size 1 and step-size 0.5.

BFGS converged such that the gradient w.r.t. each of the elements of w was less than 10^{-5} after 236 iterations. AdaGrad was run for 10000 iterations with both full batches and mini-batches, meaning that the full-batch optimisation was run for 10000 epochs whereas the mini-batch optimisation was run for just over 13 epochs. There certainly appears to be little advantage to computing the gradient w.r.t. all of the data in this scenario, as the batched optimisation converges after approximately 2000 iterations / epochs, whereas the mini-batched takes 4000 iterations or just over 5 epochs. As the mini-batched optimisation with AdaGrad is the procedure that is ultimately of most interest (as it is most scalable), and does not appear to suffer from any major deficiencies relative to batch optimisation with BFGS / AdaGrad, the rest of the analysis in this section focusses on that approach.

Figure 3.7 indicates that the posterior marginal variance (PMV) estimates provided by the non-stationary Σ_q are, at least qualitatively, closer to those of the exact posterior distribution than those provided by the stationary Σ_q. Where the optimal stationary Σ_q is unable to represent location-dependent PMV, and thus erred on the side of correctly estimating the posterior variance in regions of high data-density, the non-stationary Σ_q finds a solution that provides much more realistic PMV estimates in regions with no data. It has the particularly pleasing property of reverting to the prior marginal variance in the tails of the input space (approximately $-12.5 < x < 12.5$).

Figure 3.8 indicates that the expected KL-divergence between the posterior marginals of the non-stationary approximate posterior and the exact posterior has different properties from those of the divergence between the stationary approximation posterior marginals and the exact posterior marginals. Whereas the expected in-sample KL-divergence is lower than the expected out-of-sample KL-divergence for the stationary approximation, the order is reversed for the non-stationary approximation.

Figure 3.9 clarifies the nature of the approximate posterior by decomposing $\Sigma_q := WK_{Z,Z}W = K_{Z,Z} \odot (W\,\mathbf{1}W)$ and displaying the component matrices. Intuitively, the effect of W can be thought of as decorrelating the prior covariance $K_{Z,Z}$ in the manner which best reflects the posterior, subject to the constraint that the extent to which the covariance between the m^{th} and n^{th} observations is reduced is a product of a factor only depending on the m^{th} (w_m) and a separate factor depending only upon the n^{th} (w_n), yielding $(\Sigma_q)_{m,n} = w_m w_n \, (K_{Z,Z})_{m,n}$.

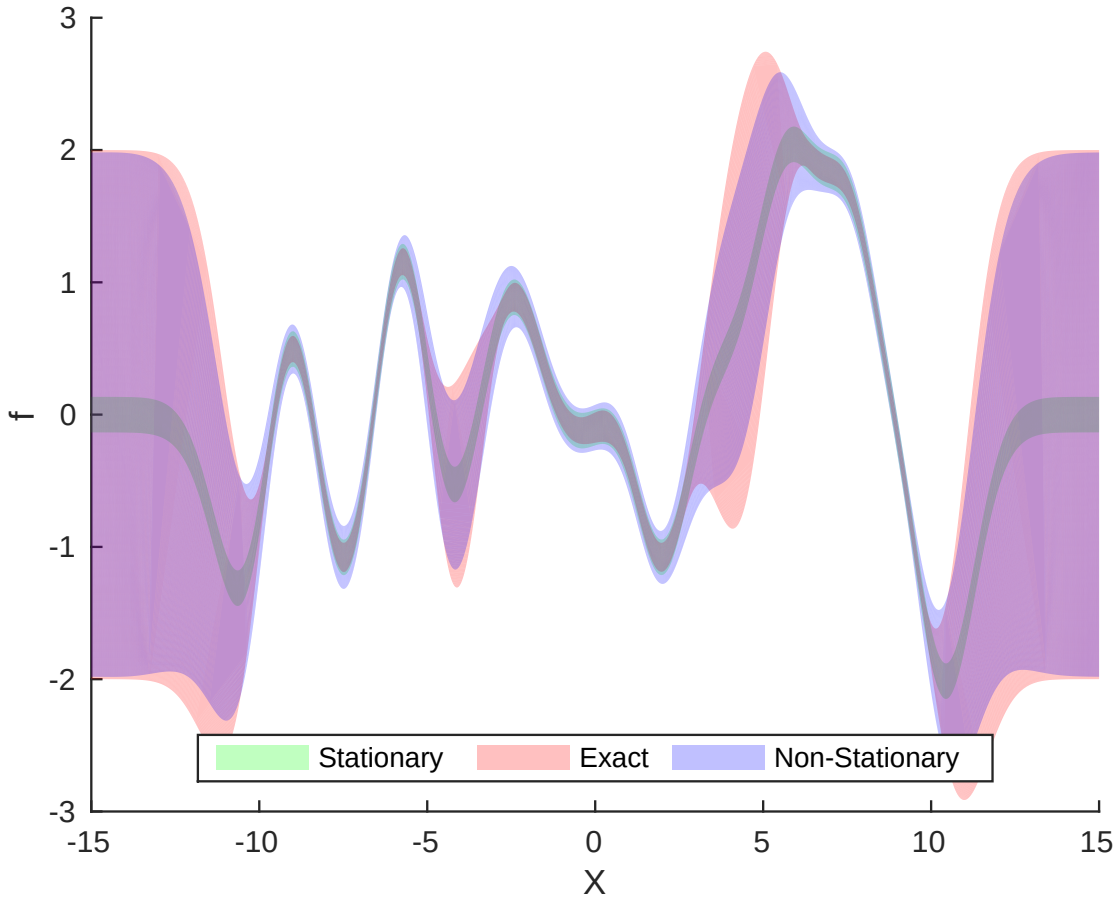


Figure 3.7 Posterior marginal variances of the stationary and non-stationary approximations overlaid on the exact posterior marginal variance (PMV). The non-stationary approximation over-estimates the PMV more than the stationary approximation in regions where there is a lot of data, however, in regions where there is no data the non-stationary approximation vastly out-performs the stationary approximation, yielding PMVs that are broadly consistent with the exact posterior marginals.

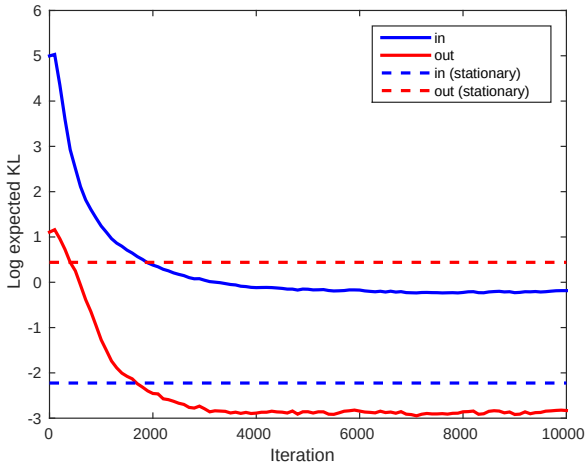


Figure 3.8 Log of the expectation of the KL-divergence between the approximate posterior marginal distribution and the exact posterior marginal distribution as optimisation using single-sample mini-batches proceeds. Results are separated between regions containing data ('in') and regions not containing data ('out'). Dashed lines show the final expected KL-divergence achieved by the stationary approximation as a point of reference.

In summary, this experiment demonstrates the superiority of the proposed non-stationary approximation over the stationary approximation in terms of the ELBO and marginal posterior variance.

3.5 Conclusions and Future Work

In this chapter I have presented an approach to approximate inference in Gaussian Processes which has asymptotic complexity $\mathcal{O}\left(NM + M\log M\right)$ (or simply $\mathcal{O}\left(M\log M\right)$ with mini-batching), which is a substantial improvement over the $\mathcal{O}\left(NM^2 + M^3\right)$ (or $\mathcal{O}\left(M^3\right)$ with mini-batching) associated with the state-of-the-art sparse GP approximations. This acceleration is achieved by applying the same variational framework presented in [Titsias, 2009], constraining the pseudo-data to lie on a grid and exploiting the approximate circulant structure this imposes. Crucially, this approach does not impose any constraints on the input locations of the observed data. Due to the necessarily constrained parameterisation of Σ_q in which the number of parameters is linear in the number of pseudo-data, we sacrifice the tightness of the original ELBO. However, for complicated posterior distributions it may be infeasible to use an unconstrained Σ_q in conjunction with sufficiently many pseudo-data to retain a tight ELBO. It is in such a scenario that the presented approximation method with flourish, as it will allow the use of sufficiently many pseudo-data and thus provide a tighter bound.

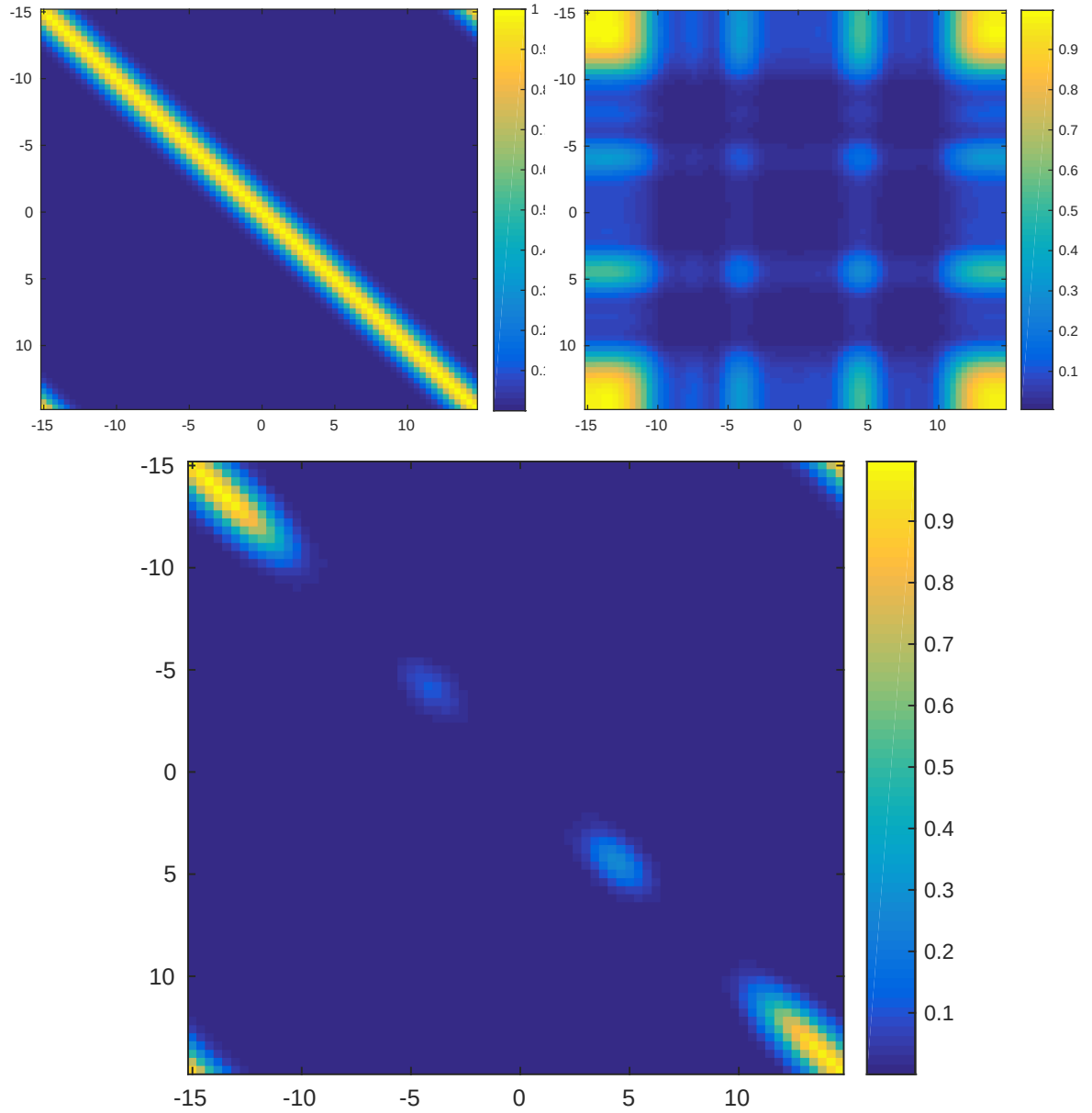


Figure 3.9 Top left: Prior covariance matrix $K_{Z,Z}$. Top right: $W\,\mathbf{1}\,W$. Bottom: Approximate posterior covariance matrix $WK_{Z,Z}W = K_{Z,Z} \odot (W\mathbf{1}W)$.

Furthermore, an annoying property of the circulant approximation, namely that it introduces unwanted covariance structure at either end of the data set, has been resolved by extending the domain on which the pseudo-inputs are located outside of the domain on which data is observed. Whilst the additional covariance structure has not vanished, it no longer affects posterior predictions made where there is observed data, and predictions can, in principle, be made accurately arbitrarily far away from the observed data by extending the pseudo-data grid out further.

There are several directions in which this work could be taken. The most obvious is to investigate different parameterisations of Σ_q which remain tractable whilst offering a better approximation to the posterior distribution. One can, for example, introduce extra parameters into the approximation for Σ_q by multiplying either side by circulant / diagonal matrices. Specifically let W_2 and K_2 be diagonal and circulant matrices respectively, then

$$\Sigma_q := W_2 K_2 W K_{Z,Z} W K_2 W_2 \tag{3.41}$$

remains a tractable posterior covariance in the sense that its determinant and inverse are easily evaluated (although the trace term in the KL-divergence must now be approximated using Monte Carlo, which may hinder inference). It would also be useful to resolve the problems found with out-of-sample inference in the mean when optimising the posterior mean approximation with mini-batched AdaGrad in section 3.2. Although CG resolved these problems and still runs on $\mathcal{O}(NM)$ time, it is not designed to be used with mini-batches and is therefore unsuitable for heavily over-sampled data sets.

Additionally, one of the primary advantages of the TSGP is that it scales 'truly' linearly with the number of pseudo-data by exploiting the local nature of the covariance functions to which it is applied. Such scaling is not achieved here as the primary focus of the work has been directed towards achieving computations in the GP prior that scale almost linearly in M. The only terms that have $\mathcal{O}(NM)$ asymptotic complexity are those involving the cross-covariance matrix $K_{D,Z}$. For AB covariance functions such as the EQ and Spectral Mixture, however, this will be very sparse meaning that much of the computation that is currently performed is redundant and can be avoided quite simply.

Another useful avenue of investigation would be found in determining how to exploit this inference scheme in higher dimensions. The number of pseudo-data necessarily increases exponentially with the dimensionality of the input space (if no additional constraints are imposed upon the covariance function) owing to the requirement that they lie on a regular grid, thus it is unlikely that high-dimensional input spaces could be handled directly. If, however, a low-dimensional representation of a data set can be found, for example using the methodology presented in [Snelson and Ghahramani, 2012]

or [Calandra et al., 2014], it may then be feasible to construct a grid of pseudo-data. [4]. Tangentially, if there is additive structure present in the covariance function then pseudo-data can be placed separately in each dimension, circumventing the exponentially poor scaling. Furthermore, extensions to non-Gaussian likelihoods such as those used in GP classification or the Chained GP framework [Saul et al., 2016] would also be useful and readily implemented.

[4]The first of these approach preserves stationarity, whereas the latter would make the resulting kernel non-stationary

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Amari, S.-I. (1998). Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276.

Bauer, M. S., van der Wilk, M., and Rasmussen, C. E. (2016). Understanding Probabilistic Sparse Gaussian Process Approximations. *arXiv preprint arXiv:1606.04820*.

Broyden, C. G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, 6(1):76–90.

Bui, T. D. and Turner, R. E. (2014). Tree-structured Gaussian process approximations. In *Advances in Neural Information Processing Systems*, pages 2213–2221.

Bui, T. D., Yan, J., and Turner, R. E. (2016). A Unifying Framework for Sparse Gaussian Process Approximation using Power Expectation Propagation. *arXiv preprint arXiv:1605.07066*.

Calandra, R., Peters, J., Rasmussen, C. E., and Deisenroth, M. P. (2014). Manifold Gaussian processes for regression. *arXiv preprint arXiv:1402.5876*.

Cunningham, J. P., Shenoy, K. V., and Sahani, M. (2008). Fast gaussian process methods for point process intensity estimation. In *Proceedings of the 25th international conference on Machine learning*, pages 192–199. ACM.

Damianou, A. C. and Lawrence, N. D. (2013). Deep Gaussian Processes. In *AISTATS*, pages 207–215.

Dietrich, C. and Newsam, G. N. (1997). Fast and exact simulation of stationary Gaussian processes through circulant embedding of the covariance matrix. *SIAM Journal on Scientific Computing*, 18(4):1088–1107.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159.

Fletcher, R. (1970). A new approach to variable metric algorithms. *The computer journal*, 13(3):317–322.

Fletcher, R. and Reeves, C. M. (1964). Function minimization by conjugate gradients. *The computer journal*, 7(2):149–154.

Ghahramani, I. M. Z. (2005). A note on the evidence and Bayesian Occam's razor.

Goldfarb, D. (1970). A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109):23–26.

Gray, R. M. (2006). *Toeplitz and circulant matrices: A review*. now publishers inc.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. *arXiv preprint arXiv:1309.6835*.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. W. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Lloyd, J. R., Duvenaud, D., Grosse, R., Tenenbaum, J. B., and Ghahramani, Z. (2014). Automatic construction and natural-language description of nonparametric regression models. *arXiv preprint arXiv:1402.4304*.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

Matthews, A. G. d. G., Hensman, J., Turner, R. E., and Ghahramani, Z. (2015). On Sparse variational methods and the Kullback-Leibler divergence between stochastic processes. *arXiv preprint arXiv:1504.07027*.

Minka, T. (2004). Power ep. Technical report, Technical report, Microsoft Research, Cambridge.

Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*, pages 362–369. Morgan Kaufmann Publishers Inc.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Quiñonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959.

Rasmussen, C. E. and Williams, C. K. (2006). Gaussian processes for machine learning. *the MIT Press*, 2(3):4.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Saul, A. D., Hensman, J., Vehtari, A., and Lawrence, N. D. (2016). Chained Gaussian Processes. *arXiv preprint arXiv:1604.05263*.

Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Mathematics of computation*, 24(111):647–656.

Snelson, E. and Ghahramani, Z. (2005). Sparse Gaussian processes using pseudo-inputs. In *Advances in neural information processing systems*, pages 1257–1264.

Snelson, E. and Ghahramani, Z. (2012). Variable noise and dimensionality reduction for sparse Gaussian processes. *arXiv preprint arXiv:1206.6873*.

Titsias, M. K. (2009). Variational learning of inducing variables in sparse Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 567–574.

Tobar, F., Bui, T. D., and Turner, R. E. (2015). Learning stationary time series using Gaussian processes with nonparametric kernels. In *Advances in Neural Information Processing Systems*, pages 3501–3509.

Turner, R. E. (2010). *Statistical models for natural sounds*. PhD thesis, UCL (University College London).

Ulrich, K. R., Carlson, D. E., Dzirasa, K., and Carin, L. (2015). GP kernels for Cross-Spectrum analysis. In *Advances in Neural Information Processing Systems*, pages 1999–2007.

Wilson, A. and Adams, R. (2013). Gaussian Process Kernels for Pattern Discovery and Extrapolation. In *Proceedings of The 30th International Conference on Machine Learning*, pages 1067–1075.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.