Spectral Methods in Gaussian Process Approximations

David R. Burt

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Master of Philosophy

Emmanuel College

August 2018

Declaration

I, David R. Burt, being a candidate for the MPhil in Machine Learning, Speech and Language Technology, hereby declare that this report 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.

Word Count (counting appendix): 14,058.

David R. Burt
August 2018

Acknowledgements

I would first like to thank my advisors, Mark van der Wilk and Professor Carl Rasmussen for their guidance in this project. It has been a pleasure to work with both of them, and I have learned a lot from the experience. I would also like to thank my parents and brother for their constant support.

Abstract

Gaussian process models are flexible, robust to overfitting and give good estimates of predictive uncertainty. However, they are computationally expensive and approximations must be made in order to apply them to large datasets. Standard methods for approximating full inference include parametric methods, which utilise a parametric model that is in some sense similar to the full model, or nonparametric methods, which instead define an approximate posterior with additional structure that allow for computational savings.

This work develops a new nonparametric approximation to Gaussian processes that fits within the 'interdomain inducing feature' framework. This approximation is based on the spectral properties of the kernel. This imposes additional structure on the covariance matrices used in inference that can be leveraged for computational benefits when applied to large data sets.

This work additionally investigates the convergence properties of these new features for $M \ll N$ features, obtaining explicit rates of convergence in certain situations for M in this regime. These results give theoretical insight into the difficulty of making sparse approximations to different kernels, and unlike common previous convergence results for inducing points which are known for $M \geq N$, are applicable in the same regime where computational savings is realized. Finally, we show the practical applicability of these features on a large classification task, in which they can outperform optimized inducing points when computational resources are limited.

Table of contents

List of figures xi

List of tables xiii

1 Introduction 1

1.1 Introduction 1

1.2 Contributions and Layout of this Thesis 2

2 Theoretical Framework 5

2.1 Gaussian Processes 5

2.2 Gaussian Process Models 7

2.3 Spectral Properties 8

2.3.1 Mercer Kernels and Gaussian Processes 9

2.3.2 Noncompact Kernel Operators 11

2.4 Parametric Approximations 11

2.5 Nonparametric Approximation 13

2.5.1 Variational Inference 14

2.5.2 Interdomain Inducing Features 15

2.5.3 Variational Fourier Features 16

3 Diagonal Covariance Matrices and Eigenfunction Based Inducing Points 19

3.1 Why are Diagonal Covariance Matrices Desirable? 19

3.2 Orthogonal Features for Stationary Kernels 20

3.3 Eigenfunction Based Inducing Points 23

3.3.1 Covariances 24

3.3.2 Cross covariances 24

3.3.3 Eigenfunction based inducing points and the mean field approximation 24

3.4 Example: Squared Exponential Kernel and Hermite Inducing Points 26

3.5 Example: Exponential Kernel and Variational Fourier Features 29

3.5.1 Relationship to Variational Fourier Features 31
3.6 Optimality and Infinite Principal Component Analysis 32

4 Convergence Properties 35
4.1 Preliminary Results 37
4.1.1 Bounding the KL-divergence based on the Trace Error 37
4.1.2 General Results Based on Eigenvalues 38
4.1.3 Asymptotic Properties of Eigenvalues of the Kernel Matrix 39
4.2 General Bounds for Eigenfunction based Inducing Points 41
4.2.1 Concentration Inequalities 42
4.3 Bounds for Hermite Based Inducing Points and The Squared Exponential Kernel 43
4.3.1 The Number of Inducing Points 48
4.4 The Exponential Kernel 49
4.4.1 Inducing Points Placed on a Grid 49
4.4.2 Using Eigenfunction based Inducing Points 52
4.5 General Bounds Using Optimized Inducing Points 53

5 Experiments 57
5.1 Toy One Dimensional datasets 57
5.1.1 Effect of input Distribution 57
5.1.2 Performance Prior to Convergence 60
5.2 Scaling To Higher Dimensions: Additive and Multiplicative Structures . . . 62
5.3 Kronecker Gaussian Process Models 62
5.4 Additive Kernels 65

6 Conclusion 71

References 73

Appendix A Structure of Covariance Matrices 77
A.1 Matérn Feature Covariance outside Interval 77
A.2 Covariance Matrix of Optimal Approximating Distribution 78

List of figures

3.1 A comparison of the gap between the ELBO with and without the diagonal approximation averaged over 3 random normally distributed datasets (left) as we increase N and convergence of the ELBO on a toy one dimensional data set with 500 inputs sampled from a normal distribution as we increase M (right). 26

3.2 The covariance matrix selected by variational inference with the eigenfunction inducing features (left) is nearly diagonal when the data is normally distributed, while for standard inducing points (right) the off-diagonal entries are not necessarily near zero. 27

3.3 Covariance of first six inducing features for Hermite inducing features with the process for $\ell = 1, s = 1$. 28

3.4 For fixed $\ell^2 = 1$, larger choices of s^2 result in slower decay in eigenvalues. . 29

3.5 Covariance of first 8 eigenfunction inducing features for Matérn kernel on [0,1] with $\ell = 1$, normalized so that the covariance matrix between features is the identity matrix. Even features are shown in the top row, and odd in the bottom. 30

3.6 Comparison of the eigenvalues for the Matérn kernel on $[0,1]$ and the SE-kernel with respect to a Gaussian with variance 1/12, both with length scales 1. (Note the priors are chosen to have the same variance. 31

4.1 Decay of first 20 eigenvalues for data points sampled from a uniform, normal and mixture of gaussian input distributions, with $N = 500$ for a SE kernel with $\ell^2 = 1$. All three input distributions with standard deviation 1. The eigenvalues of the covariance operator for defined with respect to the same normal distribution are also shown. 40

4.2 Actual error on the trace and the bound given in (4.18) (on an absolute scale on the left and log scale on the right) that holds for large N plotted for a synthetic data set with $N = 200$, $x \sim \mathcal{U}[-\sqrt{75}, \sqrt{75}]$ and $s = 5$. 44

4.3 Actual error as measured by KL-divergence as compared to the bound in (4.16) that holds for large N plotted for a synthetic data set with $N = 200$, $x \sim \mathcal{U}[-\sqrt{75}, \sqrt{75}]$ and $s = 5$. Plotted on an absolute scale on the left and log scale on the right. 45
4.4 Actual error on the trace and the bound given in (4.18) that holds for large N plotted for a synthetic data set with $N = 200$, $x \sim \mathcal{N}(0, 5^2)$ and $s = 5$. Plotted on an absolute scale (left) and a log-scale (right). 46
4.5 Actual KL-Divergence and upper bound given by combining (4.18) from Theorem 4.3.3 and Lemma 4.1.2 that holds for large N plotted for a synthetic data set with $N = 200$, $x \sim \mathcal{N}(0, 5^2)$ and $s = 5$. Plotted on an absolute scale (left) and a log-scale (right). 46

5.1 Rate of convergence of approximate log likelihood for optimized inducing points, randomly selected inducing points and eigenfunction based inducing points for $x \sim \mathcal{N}(0, 5^2)$. 58
5.2 Rate of convergence of approximate log likelihood for optimized inducing points, randomly selected inducing points and eigenfunction based inducing points for $x \sim \mathcal{U}(-\sqrt{75}, \sqrt{75})$. 59
5.3 Rate of convergence of approximate log likelihood for optimized inducing points, randomly selected inducing points and eigenfunction based inducing points for $x \sim q$. 59
5.4 Approximate models for $M = 8, 16, 24$ as well as the full model for the 1-dimensional toy dataset with x sampled from the mixture of Gaussian model. 61
5.5 Approximate models for $M = 8, 16, 24$ as well as the full model for the 1-dimensional toy dataset with x sampled from the mixture of Gaussian model, with the kernel hyperparameters fixed to the optimal obtained by the full model. 61
5.6 Decision Boundaries for the normally distributed ‘bananas’ dataset for different numbers of inducing points (top), the full model (top, rightmost) and a Kronecker product of the Hermite Inducing Features (bottom row). 64
5.7 Input distributions for the airline dataset. 66
5.8 Input distributions for the airline dataset. 67
5.9 ELBO for various M for raw dataset. 68
5.10 ELBO for various M for dataset after applying PCA. 68
5.11 Test accuracy for various M for dataset after applying PCA. 69

List of tables

5.1 Comparison of time (minutes) to run 1000 iterations of optimization for various M for Eigenfunction inducing points and Optimized inducing points. 67

Chapter 1

Introduction

1.1 Introduction

In Bayesian machine learning, the goal is to combine prior beliefs with observed data, $\mathcal{D}$, in order to learn a mapping from an input space, $\mathcal{X}$, to an output space, $\mathcal{Y}$, in a manner that takes into account uncertainty. One common approach is *parametric* modelling, which introduces an intermediate parameter space, $\mathcal{P}$ and then splits the task by learning a mapping from $\mathcal{X}$ to $\mathcal{P}$, and then a mapping from $\mathcal{P}$ to $\mathcal{Y}$. These mappings can then be composed by either marginalizing over the possible parameters settings in order to obtain the desired mapping from $\mathcal{X}$ to $\mathcal{Y}$. A key consideration in any such model is the *complexity* of the parameter space $\mathcal{P}$. If $\mathcal{P}$ is not sufficiently large, a great deal of information can be lost when mapping from the input to parameter space, making the model unable to capture complex relationships within the data. On the other hand, if $\mathcal{P}$ is large, marginalization becomes computationally intractable and an approximation, such as simply choosing the most probable parameter values must be used. This makes models with many parameters subject to *overfitting* which leads to poor estimates of uncertainty in the final model and poor generalization of the trained models.

An alternative approach is offered by Bayesian *nonparametrics*, which take a direct approach to learning the desired mapping. This allows for models that can model arbitrarily complex data faithfully. All (or almost all) variables within the model can be treated probabilistically and marginalized over, avoiding the risk of overfitting that arises in complex parametric models. The core concept of these models is to define a *stochastic process*, a collection of random variables indexed by the input domain $\mathcal{X}$, and use this process as a prior over the desired map from $\mathcal{X}$ to $\mathcal{Y}$. Commonly and for the remainder of this thesis, the stochastic process will be a *Gaussian process*, meaning that any finite subset of the random variables in the process is normally distributed. Inference in Gaussian process models scales cubically with the size of the available data. This limits the applicability of Gaussian process models to small data sets unless approximations are made to reduce the computational burden.

1.2 Contributions and Layout of this Thesis

In this thesis we develop a new nonparametric approximation within the variational inducing feature framework based on the spectral properties of Gaussian processes. Chapter 2 provides the necessary theoretical background for the remainder of the thesis. The new inducing features, which we will refer to as *eigenfunction inducing features* are derived in Chapter 3. This approximation has several nice properties:

- The features defined are closely related to the *optimal* (in terms of mean squared reconstruction error) finite linear approximation to a Gaussian process with respect to a specific prior. This gives rise to theoretical questions of the optimality of the resulting inducing features with respect to the same prior in the Kullback-Liebler sense.

- The features defined are orthogonal. This offers a computational savings when trained using stochastic variational inference.

- Under certain assumptions, the optimal approximating distribution within the variational family associated to these features can be shown to have a diagonal covariance matrix for regression tasks. Making this assumption can provide further computational savings in stochastic variational inference.

We additionally show how to derive more general orthogonal inducing features for a stationary kernel in Chapter 3.

In Chapter 4, we examine rates of convergence of inducing feature approximations to full Gaussian process regression, for $M \ll N$ inducing features. We show:

- For the eigenfunction inducing features developed in this work, we derive an explicit upper bound on the KL-divergence between the full and approximate models in the case of the squared exponential kernel. This bound is shown to hold with high probability for large data sets. In contrast to well known bounds in the literature, these are derived *a priori* (prior to actually computing the ELBO) for a given kernel. This can indicate the number of inducing features needed to approximate a specific kernel well.

- In the case of the exponential kernel, we derive an upper bound on the KL-divergence that holds in probability as the amount of data tends to infinity using standard inducing points. This bound is not strong enough to give interesting results in the regime $M \ll N$. This aligns with the intuition that sparse approximation of non-smooth functions is far more challenging than sparse approximation of smooth functions.

In Chapter 5 we show the practical performance of the features defined, including on an 8-dimensional dataset used in related works containing over 5 million data points. We conclude in Chapter 6 with a discussion of the advantages and disadvantage of these features in relation to existing methods and several open questions of theoretical and practical interest.

Chapter 2

Theoretical Framework

This chapter reviews results in the Gaussian process literature that will be needed in later sections. We begin with formally defining a Gaussian process, then discuss Gaussian process models. Next we discuss spectral properties of Gaussian processes. In section 2.4 we discuss several parametric Gaussian process approximation methods based on spectral properties. We conclude in section 2.5 with a brief discussion of variational inference as a nonparametric method for approximate Gaussian process inference.

2.1 Gaussian Processes

Consider an input domain, indexed by a set $\mathcal{X}$. We will generally focus on the case when $\mathcal{X} = \mathbb{R}^d$. The Kolmogorov extension theorem tells us that, subject to certain consistency conditions, a collection of probability distributions defined on every finite $X \subset \mathcal{X}$, can be extended to a stochastic process defined on $\mathcal{X}$, which we will denote as f. This extension is unique when $\mathcal{X}$ is equipped with the product $\sigma-$algebra (the smallest $\sigma-$algebra containing all finite subsets of $\mathcal{X}$). For a precise measure theoretic formulation of Gaussian processes see Matthews (2016). Because the marginals of a Gaussian process are normally distributed, it suffices to define the mean and covariance of every finite dimensional marginal distribution.

This is done choosing a *mean function* $\mu : \mathcal{X} \to \mathbb{R}$ which returns the expected value of the process at a given point, i.e. $\mathbb{E}[\mathbf{x}] = \mu(\mathbf{x})$. For the remainder of this thesis, we will assume that for the prior process, $\mu(\{\mathbf{x}_i\}_i) = \mathbf{0}$ for all finite collections of $\{\mathbf{x}_i\}_i \in \mathcal{X}$, as the general case can be easily recovered.

The covariance structure of the finite dimensional marginal must also be defined. A *kernel function* $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is chosen, and

$$\operatorname{cov}\left(f(\mathbf{x}_i), f(\mathbf{x}_j)\right) = k(\mathbf{x}_i, \mathbf{x}_j). \tag{2.1}$$

In order for consistency conditions to be satisfied, the kernel function must be symmetric and positive semidefinite, i.e.

$$k(\mathbf{x}_i,\mathbf{x}_j) = k(\mathbf{x}_j,\mathbf{x}_i) \qquad \forall \;\; \mathbf{x}_i,\mathbf{x}_j \in \mathcal{X}, \tag{2.2}$$

and

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j k(\mathbf{x}_i,\mathbf{x}_j) \qquad \forall \;\; \mathbf{x}_i,\mathbf{x}_j \in \mathcal{X}, c \in \mathbb{R}. \tag{2.3}$$

The covariance function defines a structure on the random variables that constitute the Gaussian process. This structure allows us to extrapolate or interpolate information about a collection of random variables $\{f(\mathbf{x}_i)\}_{i=1}^{N}$ to make predictions at new input values $\{f(\mathbf{x}_j^*)\}_{j=1}^{T}$. In particular, it defines a particular inner product on the space of random variables. This inner product measures the strength of the relationship between different parts of the input spcae.

Common choices of kernel function include: the *squared exponential* function defined by

$$k_{se}(\mathbf{x}_i,\mathbf{x}_j) = \sigma \exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|_2^2}{2\ell^2}\right), \tag{2.4}$$

the *exponential* kernel

$$k_{exp}(\mathbf{x}_i,\mathbf{x}_j) = \sigma \exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|_1}{\ell}\right), \tag{2.5}$$

and the *linear* kernel

$$k_{linear}(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j, \tag{2.6}$$

all defined on $\mathbb{R}^d$.

A kernel is said to be *stationary* if it is a function of $\mathbf{x}-\mathbf{x}'$ alone. The squared exponential and exponential kernels are stationary, while the linear kernel is not.

For any finite subset of points $\{\mathbf{x}_i\}_{i=1}^{N} \subset \mathcal{X}$, the probability distribution over the corresponding random variables is given by

$$f\left(\{\mathbf{x}_i\}_{i=1}^{N}\right) \sim \mathcal{N}(\mu_n, \mathbf{K}_{n,n}), \tag{2.7}$$

where μ_n is the N-dimensional vector given by $\mu_i = \mu(\mathbf{x}_i)$ and $\mathbf{K}_{n,n}$ is the $N \times N$ covariance matrix with $k_{i,j} = k(\mathbf{x}_i,\mathbf{x}_j)$.

The entire stochastic process is denoted,

$$f \sim \mathcal{GP}(\mu, k). \tag{2.8}$$

In the next section, we will discuss how this Gaussian processes is used to define a prior over functions from $\mathcal{X}$ to $\mathcal{Y}$ in nonparameteric models.

2.2 Gaussian Process Models

Given a set of data, $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$, with $\mathbf{x}_i \in \mathcal{X}$, and $\mathbf{y}_i \in \mathcal{Y}$, our goal is to infer the posterior distribution over functions mapping between inputs and outputs, that is $p(\mathbf{y}_*|\mathbf{x}_*, \mathcal{D})$. In order to define a probabilistic model over these function we need to link the observed inputs and data probabilistically through a *likelihood function*. This is a function of the model given the data.

$$L_{\mathcal{D}}(f) := p(\mathbf{y}|\mathbf{x}, f). \tag{2.9}$$

For regression tasks, it is often assumed that $p(\mathbf{y}|\mathbf{x}, f)$ is normally distributed. In other words, the outputs are given by the latent function plus independent, identically distributed Gaussian noise.

We additionally assume that the inputs are drawn independently according to some (not necessarily Gaussian) distribution, $p(\mathbf{x})$, over the input space although this assumption is generally not explicitly necessary for Gaussian process inference.

The modeling assumptions for a Gaussian process model can then be summarized as,

$$p(\mathbf{x}_i) = \prod_{i=1}^{n} p(x), \tag{2.10}$$

$$f \sim \mathcal{GP}(f; \mu, k), \tag{2.11}$$

$$p(\mathbf{y}|f, \mathbf{x}\ ) = \prod_{i=1}^{n} p(y_i|f(\mathbf{x}_i)) = \prod_{i=1}^{n} L_{\mathbf{y}_i}(f(\mathbf{x}_i)). \tag{2.12}$$

In the case the likelihood function is also Gaussian, i.e. $p(\mathbf{y}|\mathbf{x}, f) \sim \mathcal{N}(f(\mathbf{x}), \sigma^2_{noise}\mathbf{I})$, inference and prediction in the model can be performed exactly. In this case, the marginal likelihood of the model is given (Rasmussen and Williams, 2005, Chapter 2.2) by,

$$p(\mathbf{y}) = \mathcal{N}(y; 0, \mathbf{K}_{n,n} + \sigma^2_{noise}\mathbf{I}), \tag{2.13}$$

and the predictive mean and variance at test $\mathbf{x}^*$ are given by

$$f(\mathbf{x}^*) \sim \mathcal{N}\left(\mathbf{K}_{*,n}\mathbf{K}_{n,n}^{-1}\mathbf{y}, \mathbf{K}_{*,*} - \mathbf{K}_{*,n}\mathbf{K}_{n,n}^{-1}\mathbf{K}_{*,n}^{T}\right), \tag{2.14}$$

where $\mathbf{K}_{*,*}$ is the covariance matrix formed by evaluating the kernel function at pairs of test points and $\mathbf{K}_{*,n}$ is the matrix formed by evaluating the kernel function at pairs formed by taking a single test point and a single training point.

Despite the simple closed forms expression for inference and prediction, both operations are often computationally intractable, as inversion of the $N \times N$ covariance matrix is generally a $O(N^3)$ operation and storing this matrix requires $O(N^2)$ memory. Prediction is also $O(N^2)$. For even moderately large data sets, these operations are computationally intractable, so exact Gaussian process inference is limited to relatively small data sets. Approximations are utilised for applications with large datasets. Usually these approximation induce a low rank structure on the covariance matrix.

An alternative perspective is given by *variational* approximations, (Titsias, 2009) and (Matthews, 2016). Variational methods rely on finding an approximate posterior process that is as close to the true posterior as possible, while satisfying additional structural constraints. These constraints are imposed to allow for computationally efficient inference.

In the case when the likelihood is non-conjugate (for example in classification tasks) inference can be performed via expectation propagation or Laplace approximation. For a review and comparison of these methods, see Kuss and Rasmussen (2005). Straightforward approximate inference in the non-conjugate setting has similar computational burdens as exact inference in the conjugate setting, so additional approximations must also be made in this case for large data sets. Variational sparse approximate inference in the non-conjugate setting has been addressed in Hensman et al. (2015).

2.3 Spectral Properties of Gaussian Processes

Spectral methods allow us to decompose an operator, often a matrix, so that its constituent parts act independently along each direction of a transformed coordinate system. As stated in the previous section, many approximate Gaussian process models utilise a low rank version of the covariance matrix. The efficacy of low rank matrix approximation is closely tied to the spectral properties of the matrix being approximated.

As a motivating example, consider the problem of approximating the full rank positive definite, symmetric matrix

$$\mathbf{K}_{n,n} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T,$$

where $\mathbf{\Lambda}$ is a diagonal matrix with entries $\lambda_1 \geq \lambda_2, \geq \ldots, \geq \lambda_N \geq 0$ and $\mathbf{U}$ a unitary matrix with columns corresponding to the eigenvectors of the $N \times N$ matrix $\mathbf{K}_{n,n}$. Suppose the goal is to approximate $\mathbf{K}_{n,n}$ with a rank M matrix. It is well known, (Horn and Johnson, 1990) that an optimal rank M approximation to $\mathbf{K}_{n,n}$ in terms of any unitarily invariant norm (for example, the Frobenius, operator, or trace norms) is achieved by taking $\mathbf{Q}_{n,n} = \mathbf{U}_m \mathbf{\Lambda}_m \mathbf{U}_m^T$, where $\mathbf{U}_m$ is the $N \times M$ submatrix consisting of the first M columns of $\mathbf{U}$ and $\mathbf{\Lambda}_m$ is the $M \times M$ principle submatrix of $\mathbf{\Lambda}$.

In this section, we review several foundational results in the spectral theory of stochastic processes. A more detailed discussion is contained in Rasmussen and Williams (2005, Chapter 4).

2.3.1 Mercer Kernels and Gaussian Processes

The spectral theory for finite dimensional Hermitian operators (i.e. symmetric matrices) is simple and well-known in the machine learning literature. Any such operator is necessarily self-adjoint and compact, hence it is diagonalizable. Moreover, all of its eigenvalues are real. Restricting to the case of covariance (positive semidefinite) matrices, the eigenvalues are additionally nonnegative. For any such $N \times N$ matrix, $\mathbf{K}$ we have,

$$\mathbf{K} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \sum_{i=1}^{N} \lambda_i \mathbf{u}_i \mathbf{u}_i^T, \tag{2.15}$$

with $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$ and $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N]$, a unitary matrix with entries given by the eigenvectors of $\mathbf{K}$. Generally, we will assume that eigenvalues are ordered, so that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq 0$. The spectral properties of finite dimensional matrices are commonly used in statistical and machine learning methods such as principal component analysis for finding compact representations of data.

In the case that the input space is infinite dimensional as is the case in Gaussian process models, analogous statements can be made. Usually, these focus around Mercer's theorem and Bochner's theorem.

Consider a Gaussian process defined over $\mathcal{X} = \mathbb{R}$ with kernel function k. The case with $\mathcal{X} = \mathbb{R}^d$ is analogous. Additionally, equip $\mathcal{X}$ with a measure μ. In the case that μ is finite, this has the natural interpretation as defining a prior over input variables.

The corresponding Hermitian operator, which we will refer to as the *covariance operator* is defined on $L^2(\mathcal{X}, \mu)$ by:

$$\mathcal{K} : f(\mathbf{x}) \to \int_{\mathcal{X}} K(\mathbf{x}, \mathbf{x}') f(\mathbf{x}) d\mu(\mathbf{x}). \tag{2.16}$$

The operator $\mathcal{K}$ can be motivated from the perspective of random matrix theory. When using the covariance matrix $\mathbf{K}_{n,n}$, the measure μ used to define $\mathcal{K}$ is replaced by the empirical input locations, $\{\mathbf{x}_i\}_{i=1}^N$. It is known that if $\mathcal{K}$ is compact, the eigenvalues of the normalized covariance matrix, $\frac{1}{n}\mathbf{K}_{n,n}$ converge to the eigenvalues of $\mathcal{K}$, see for example Shawe-Taylor et al. (2002). This connection is discussed in greater detail in Section 4.1.3.

We would like to invoke the spectral theorem in order to diagonalize this operator. As a consequence of the spectral theorem for self-adjoint compact operators, it suffices to show that the operator is compact to obtain the infinite dimensional analogue of (2.15). In this case, we have countably many discrete eigenvalues.

Compactness is implied by the condition:

$$\int_{\mathcal{X}}\int_{\mathcal{X}} K(\mathbf{x},\mathbf{x}')d\mu(\mathbf{x})d\mu(\mathbf{x}') < \infty. \tag{2.17}$$

Assuming (2.17) is satisfied we can decompose the covariance operator as,

$$\mathcal{K} = \mathcal{U}\mathcal{D}\mathcal{U}^T \tag{2.18}$$

where $\mathcal{U}$ is a unitary operator and $\mathcal{D}$ is a diagonal operator. Informally, we can think of each of the above operators as infinite dimensional matrices, with columns of $\mathcal{U}$ being given by eigenfunctions of $\mathcal{K}$ and $\mathcal{D}$ an infinite diagonal matrix with entries $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$.

This decomposition gives rise to Mercer's theorem, which states that the kernel function can be written as:

$$k(\mathbf{x},\mathbf{x}') = \sum_{i=1}^{S} \lambda_i \phi_i(\mathbf{x})\phi_i(\mathbf{x}'), \tag{2.19}$$

with equality holding μ almost everywhere. The right hand side of (2.19) is $L^2(\mathcal{X},\mu)$-convergent, and the eigenvalues are

The Karhunen-Loève expansion states that we can rewrite the entire Gaussian process via a reparameteriztion as,

$$f = \sum_{i=1}^{S} \phi_i(\mathbf{x})z_i, \tag{2.20}$$

with the z_i pairwise independent, zero-mean normal random variables with variance λ_i defined by

$$\int_{\mathcal{X}} \phi_i(\mathbf{x})f(\mathbf{x})d\mu(\mathbf{x}). \tag{2.21}$$

By definition, for a *nondegenerate kernel* we have $S = \infty$ in both Mercer's theorem and the Karhunen-Loève expansion.

Note that (2.17) is satisfied for a stationary kernel that is bounded and not zero almost everywhere if and only if μ is a finite measure. In particular, if $\mathcal{X} = \mathbb{R}^d$ it is not satisfied for any interesting stationary kernel if μ is taken to be the Lebesgue measure on $\mathbb{R}^d$, but it is satisfied if μ is Lebesgue measure on a finite interval or a Gaussian distribution over $\mathbb{R}^d$.

2.3.2 Noncompact Kernel Operators

While we cannot apply Mercer's theorem, (2.19) for stationary kernels with lebesgue measure on $\mathbb{R}^d$, one can intuitively, one can imagine taking the limit as the variance of the input distribution μ tends to infinity. As this limit is taken, the discrete eigenvalues accumulate, and the basis functions oscillate over long regions, becoming nearly sinusoidal.

More formally, let $\kappa(\omega) = k(0, \omega)$. Note the the associated operator, $\mathcal{K}$ is a convolution operator. The spectral theorem in this setting takes the form of Bochner's Theorem, which tells us

$$\kappa(\omega) = \frac{1}{\sqrt{2\pi}} \int s(\omega) e^{i\omega\mathbf{x}} d\mathbf{x}. \tag{2.22}$$

The *spectral measure,* $s(\omega)$ is given by the Fourier transform of $\kappa(\mathbf{x})$,

$$s(\omega) = \mathcal{F}[\kappa](\omega) = \frac{1}{\sqrt{2\pi}} \int \kappa(\mathbf{x}) e^{-i\omega\mathbf{x}} d\mathbf{x}, \tag{2.23}$$

and is nonnegative and integrable (i.e. proportional to a probability density). The spectral measure plays the role of the diagonal operator in (2.18), only the eigenvalues are no longer discrete, and instead form a distribution. The Fourier transform replaces the unitary operator $\mathbf{U}$.

2.4 Parametric Approximations Using Spectral Methods

Many parametric approximations to Gaussian process models involve either modifying the prior kernel function, or modifying the empirical kernel matrix in order to reduce the computational cost of inference and prediction. In general, both approaches lead to low rank structure in the covariance matrix in the approximate model, that reduces the cost of storing and inverting large matrices. We discuss the infinite version of principal component analysis, with respect to a prior input distribution, developed in Zhu et al. (1997). We then compare this method to the well know Random Fourier Feature approximation used in Rahimi and Recht (2008) for general kernel methods and for Gaussian process regression in Lazaro-Gredilla et al. (2010).

Infinite Dimensional PCA

In Zhu et al. (1997), the authors propose the infinite dimensional analogue of PCA. In particular, given a Gaussian process with covariance function k and assuming the inputs come from some distribution with density p, and some fixed M they define the degenerate kernel,

$$k^{(M)}(\mathbf{x},\mathbf{x}') = \sum_{i=1}^{M} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}').$$

λ_i and ϕ_i are the same as those appearing in (2.19). This model is parametric, parameterized by the M eigenfunction-eigenvalue pairs, (λ_i, ϕ_i). The authors of Zhu et al. (1997) show this is the optimal M-dimensional linear model with respect to the prior $p(\mathbf{x})$, where optimality is defined in terms of minimizing expected mean squared approximation error.

Inference can be performed in $O(NM^2)$ in the approximate model, as the covariance matrix is now rank M.

There are several issues with this model. First, it is optimal with respect to the prior $p(\mathbf{x})$, but does not take into account the locations of the actual data. As observed in Ferrari-Trecate et al. (1999), significant modelling improvement can be gained from using information about the actual locations of training data. This is achieved in Ferrari-Trecate et al. (1999) by using information about the empirical kernel matrix evaluated at the inputs, instead of decomposing the kernel operator with respect to a fixed prior. The challenge with methods utilising information about the of the empirical kernel matrix is that even forming the full kernel matrix is $O(N^2)$, both in computation and memory, and finding the first M eigenvalues is generally not significantly faster than inverting the entire kernel matrix, unless M is very small.

Secondly it is unclear what how to choose kernel hyperparameters within this framework. Any good choice of hyperparameters should account for both account for both the quality of the chosen kernel in modelling the data and the ability of features to approximate the kernel. This issue is a primary motivation for nonparametric models discussed later, as they address this tradeoff in a probabilistically consistent manner.

Sparse Spectrum Gaussian Processes

Sparse spectrum Gaussian processes rely on Bochner's theorem, (2.22), to define a degenerate kernel that converges to the original kernel pointwise as the number of basis functions increases.

They construct this approximate kernel via Monte Carlo estimate of the Fourier transform appearing in (2.22). Since this approximation is with respect to Lebesgue measure on $\mathbb{R}^d$, it tends to behave uniformly along the input domain. Monte Carlo sampling is similar to truncating the Karhunen-Loève expansion of a process defined via a compact kernel, in the sense of obtaining a finite approximation to the spectrum of the associated operator.

Elaborating on this connection in the case $\mathcal{X} = \mathbb{R}$, given M samples from s, without loss of generality, assume all of the samples are rational. Define μ_M to be the uniform measure on an interval, $[a,b]$ of length given by the least common multiple of the denominators of these frequencies, so that all of the frequencies are harmonic on $[a,b]$. With respect to μ_M, the M basis functions defined are orthogonal (they are all elements of the standard Fourier basis).

Define a new kernel on the interval $[a,b]$ by,

$$\tilde{k}^{(M)}(x,x') = (b-a)\sum_{i=1}^{M} \phi_i(x),$$

where the ϕ_i are sines and cosines functions with the sampled frequencies. This defines a degenerate Karhunen-Loève expansion, which we can imagine is produced by truncation of a Mercer kernel.

The truncated kernel can be extended periodically to the whole real line, and inference is performed with this kernel. As M grows, the measure converges $(b-a)\mu_M$ to μ pointwise. Via Monte Carlo principles, the average weights of the basis functions in any small interval converge to the spectral weights given by $s(\omega)$ to this same interval. In other words, the approximate kernel converges to the full kernel pointwise as M tends to infinity. A slightly different formulation of this pointwise convergence, with explicit bounds on the rate is given in Rahimi and Recht (2008).

Note that as M is finite, the given approximation is parametric and corresponds to a degenerate Gaussian process. Because of this, the model tends to underestimate the model uncertainty (as compared to the full model) in certain regions of the input space far from the training data. Additionally, when the frequencies of the ϕ_i are chosen to optimize the marginal likelihood of this approximate model, convergence to the full model is no longer guaranteed and overfitting results if the number of optimized frequencies is large. Additionally, the approximate model significantly underestimates uncertainty in regions of the domain far from training examples and, if insufficiently many basis functions are used, can have pathological predictive uncertainties in regions of the domain close to training points, which may be of practical concern. These issues are discussed in detail in Lazaro-Gredilla et al. (2010).

2.5 Nonparametric Approximations to Gaussian Processes

The models discussed in the previous section directly approximated the kernel. In contrast, nonparametric approximations use the original kernel as a prior and approximate inference by finding an approximate posterior that is close to the full model but has a additional structure that allows for more tractable inference.

Parametric models will necessarily underestimate uncertainty in some regions of the input space with little data. To see why this is the case, it is useful to consider the limiting case where the noise variance tends to zero. By a dimensionality argument, for any nondegenerate Gaussian process the parameterization of the input space takes points that are far apart to the same point in parameter space. For many commonly used covariance functions, we expect such points to be essentially independent, yet in the parametric model an observation at one *completely* determines the prediction at the other.

Solutions to this have been suggested, such as ‘healing’ for the relevance vector machine Rasmussen and Quinonero-Candela (2005). However these solutions are somewhat ad hoc and move away from the principled Bayesian framework utilised in full Gaussian process models. In contrast, the approximate posteriors used in variational methods are nondegenerate Gaussian processes. This maintains the infinite dimensional properties of the original model making them robust to overfitting and better able to capture uncertainty.

2.5.1 Variational Inference

The core concept of variational inference as applied to Gaussian processes is to define a new Gaussian process with the same prior as the original model, and a posterior that depends only on the prior and $M \ll N$ random variables. This posterior is used to approximate the posterior process of the original model. The random variables used in variational inference, which we will denote by $\{u_m\}_{m=1}^{M}$, must be somehow correlated with the initial inputs in order to make use of the data. Commonly, they are chosen as ‘pseudo-inputs,’l lying in the initial input space, $\mathcal{X}$ (for a review of other ‘pseduo-input’ methods, see Quinonero-Candela and Rasmussen (2005)). The goal of approximate inference is to make the approximate posterior process as close to the full posterior as possible, through minimizing the KL-divergence between the approximate posterior and the full posterior. As the approximate posterior is restricted by the assumption that it only depends on M random variables, the KL-divergence will generally be greater than zero. A notable exception is if $M = N$ and the $\{u_m\}$ are chosen as the random variables in the input space corresponding to the original datapoints, in which case full inference is recovered.

By placing the task in a variational framework, any parameters in the $\{u_m\}$ (e.g. locations of pseudopoints) are variational parameters, which can be optimized in a manner that accounts for the locations of observed data, while remaining robust to overfitting. The form of the KL-divergence for Gaussian process regression used in this optimization was derived in Titsias (2009) and put on rigorous measure theoretic footing in Matthews et al. (2016). Titsias (2009) showed that minimizing the KL-divergence is equivalent to maximizing the following variational lower bound to the log marginal likelihood of the original model:

$$\mathcal{L}_{lower} = \log\left(\mathcal{N}\left(\mathbf{y}; 0, \mathbf{K}_{n,m}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{n,m}^{T} + \sigma_{noise}^{2}\mathbf{I}\right)\right) - \frac{1}{2\sigma_{noise}^{2}}\mathrm{tr}\left(\mathbf{K}_{n,n} - \mathbf{K}_{n,m}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{n,m}^{T}\right), \tag{2.24}$$

where $\mathbf{K}_{m,m}$ is the covariance matrix associated to the random variables and $\mathbf{K}_{n,m}$ is a matrix made up of the cross-covariances between the training inputs and the inducing variables. The first term in (2.24) can be thought of as an approximate marginal likelihood and the second term is a regularization term. Note that while the full covariance matrix appears in the trace term, only its diagonal elements must be computed, so $\mathcal{L}$ can be computed in $O(NM^2 + M^3)$ with $O(M^2)$ storage, allowing it to scale for relatively large data sets. This lower bound is *collapsed* in that it is derived in Titsias (2009) by choosing the optimal mean and variance for the approximating Gaussian distribution given the current setting of other variational parameters.

A more general variational lower bound that allows for stochastic inference and can be utilised in the non-conjugate case (e.g for classification tasks) is derived in Hensman et al. (2013). The variational distribution is assumed to be Gaussian with $q(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \mathbf{m}, \mathbf{S})$. The expanded variational lower bound is given by:

$$\mathcal{L}_{lower} = \sum_{i=1}^{n}\left(\log\left(\mathcal{N}\left(y_i; \mathbf{k}_i\mathbf{K}_{m,m}^{-1}\mathbf{m}, \sigma_{noise}^{2}\mathbf{I}\right)\right) - \frac{1}{2\sigma_{noise}^{2}}\left(\mathbf{K}_{n,n} - \mathbf{K}_{n,m}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{n,m}^{T}\right)_{i,i} - \frac{1}{2}\mathrm{tr}(\mathbf{S}\boldsymbol{\Lambda}_i)\right) - \mathrm{KL}\left(q(\mathbf{u})\|p(\mathbf{u})\right), \tag{2.25}$$

where

$$p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; 0, \mathbf{K}_{m,m}) \text{ and } \boldsymbol{\Lambda}_i = \frac{1}{\sigma_{noise}^{2}}\mathbf{K}_{m,m}^{-1}\mathbf{k}_i\mathbf{k}_i\mathbf{K}_{m,m}^{-1}.$$

This expanded lower bound allows for stochastic evaluation. This reduce the computational burden of hyperparameter learning to $O(\tilde{N}M^2 + M^3)$, per iteration where $\tilde{N}$ is the size of a minibatch. This savings is crucial when applying inducing point methods to extremely large data sets. It has been extended for classification in Hensman et al. (2015).

2.5.2 Interdomain Inducing Features

While most commonly the $\{u_m\}_{i=1}^{M}$, are selected as ‘inducing points,’ the validity of inference with more general forms of random variables was noted in Titsias (2009, technical report). As shown in Matthews et al. (2016) the $\{u_m\}$ can be taken to be any measurable transformation of the random variables lying in the input space.

Choosing random variables that are not in the initial input domain may lead to more expressive representation of the data for fixed M, or additional structure in approximating distributions that can be leveraged for computational benefit. This idea was investigated in detail in Lázaro-Gredilla and Figueiras-Vidal (2009), which introduced the notion of 'interdomain' features.

A simple form of an inducing feature is given by linear transformations of the original input space,

$$u_m = \int \phi_m(\mathbf{x}) d\mu(\mathbf{x}), \tag{2.26}$$

where μ is a finite measure. This is a particularly natural generalization of the inducing point methods when viewing the (mean-centered) random variables indexed by $\mathcal{X}$ as forming a vector space. A general element of this vector space is then of the form $\sum_{i=1}^{K} w_i f(\mathbf{x}_i)$. The covariance function defines a particular inner product on this space, $\langle f(\mathbf{x}), f(\mathbf{x}')\rangle = \text{cov}(\mathbf{x}, \mathbf{x}')$. This inner product space can be turned into a Hilbert space by taking its completion with respect to the inner product, and general elements in this space then take the form of the u_m in (2.26). The covariance function is well-defined on these random variables through extending it linearly and verifying the necessary limits converge. For a detailed discussion of the connection between Gaussian process methods and Hilbert spaces, see Wahba (1990).

Variational inference with random variable of the form in (2.26) was proposed in Lázaro-Gredilla and Figueiras-Vidal (2009) and placed on rigorous footing in Matthews et al. (2016) (actually, (2.26) is a special case of the features considered in both works). In order to perform approximate inference in the resulting model, it suffices to compute $\text{cov}(u_m, u_n)$ and $\text{cov}(u_m, f(\mathbf{x}))$. Once these are computed, the equations stated in 2.5.1 can be directly applied for inference and prediction.

2.5.3 Towards Combining Spectral and Inducing Point Methods: Variational Fourier Featues

An example of a successful application of the interdomain inducing feature framework for approximating Matèrn kernels was given in Hensman et al. (2016). In this work, the authors worked with respect to Lebesgue measure over a fixed interval, $[a,b]$ in input space, and sought to define features based on trigonometric functions. In particular they defined inducing variables such that

$$\text{cov}(u_m, f(\mathbf{x})) = \cos(\omega_m \mathbf{x}), \tag{2.27}$$

(or $\sin(\omega_m \mathbf{x})$) with ω_m harmonic on the interval $[a,b]$, and suggested Fourier anaylsis as a motivation for these features. They showed that these features result in a nearly diagonal covariance matrix, $\mathbf{K}_{m,m}$, and utilised this to obtain a computational savings during stochastic hyperparameter optimization. In the next chapter, we elaborate more on the structure of this approximation, and develop new interdomain inducing point methods that utilise spectral analysis in order to obtain an exactly diagonal covariance matrix, giving similar computational benefits.

Chapter 3

Diagonal Covariance Matrices and Eigenfunction Based Inducing Points

In this chapter, we utilise ideas from spectral analysis to define orthogonal inducing features. We begin this section with following simple question: given a kernel function, k, can we define inducing features u_m such that $\mathbf{K}_{m,m}$ is exactly diagonal? We investigate this question in two distinct settings. In section 3.2, under the assumption that the kernel function is stationary we obtain a general framework for defining inducing points that are *exactly* orthogonal. This framework allows for analytic computation of the covariance between feature and inducing points in certain cases and Monte Carlo approximations of the covariance more generally. In section 3.3, we show how to obtain orthogonal inducing features based on the Mercer expansion of the kernel with respect to an input distribution. When the Mercer expansion with respect to a parameterized family of input distributions is known for a kernel, all variational parameters in the features and hyperparameters in the kernel can be jointly optimized using variational inference. We give several examples of these inducing features for square exponential kernel with Gaussian inputs and the exponential kernel with uniform inputs over a fixed interval. We discuss the similarity between the eigenfunction inducing features for the exponential kernel and the Variational Fourier Features of Hensman et al. (2016). We conclude with a brief discussion relating relating the eigenfunction inducing feature model to the model of Zhu et al. (1997) discussed in Section 2.4.

3.1 Why are Diagonal Covariance Matrices Desirable?

It is not immediately clear that any real benefit is obtained by avoiding the inversion of the matrix $\mathbf{K}_{m,m}$ through a diagonal structure. After all, in Gaussian process regression using the collapsed marginal likelihood bound (2.24), we still must form the matrix product $\mathbf{K}_{m,n}\mathbf{K}_{m,n}^T$ in each iteration, and invert the $M \times M$ matrix $(\mathbf{K}_{m,m}^{-1} + \frac{1}{\sigma_{noise}^2}\mathbf{K}_{m,n}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{m,n}^T)$ which leads to complexity $O(NM^2 + M^3)$ per iteration. Removing the $O(M^3)$ cost of inverting $\mathbf{K}_{m,m}$ provides only a lower order computational savings in this setting.

The computational benefit is of a diagonal covariance matrix is realized when using the expanded bound, (2.25) in conjunction with stochastic optimization. In this case, the cost per iteration of hyperparameter learning is $O(\tilde{N}M^2 + M^3)$, where $\tilde{N}$ is the size of a minibatch. In this case, it is possible that $M \gg \tilde{N}$ in which case the computational bottleneck per iteration *is* the inversion of $\mathbf{K}_{m,m}$. Choosing inducing features that are orthogonal reduces the cost of each iteration of parameter learning from $O(\tilde{N}M^2 + M^3)$ to $O(\tilde{N}M^2)$ in this setting. If additionally the approximating covariance matrix is assumed to be diagonal, a commonly used approximation that we show can be made *with no loss of performance* for large regression tasks with certain training input distributions, this is further reduced to $O(\tilde{N}M)$. Computational benefits for diagonal and nearly diagonal $\mathbf{K}_{m,m}$ matrices were also observed with MCMC inference in Hensman et al. (2016), though we do not explore MCMC methods for inference here.

3.2 Orthogonal Features for Stationary Kernels

The main result of this section is the following theorem,

Theorem 3.2.1. *Let $k(x - x') = \kappa(\omega)$ be a stationary kernel with spectral measure $s(\omega)$ such that $s(\omega)$ is strictly positive for all ω. Let ϕ_m be an orthonormal basis for $L^2(\mathbb{R})$, such that $\phi_m(-x) = (-1)^m\phi_m(x)$ and both $\frac{\phi_m(x)}{\sqrt{s(\omega)}}$ and $\phi_m(x)\sqrt{s(\omega)}$ are absolutely integrable with absolutely integrable Fourier transforms. Define*

$$u_m := \frac{1}{(2\pi)^{1/4}} \int_{\mathbb{R}} f(x)\mathcal{F}^{-1}\left(\frac{\psi_m(x)}{\sqrt{s(\omega)}}\right) dx, \tag{3.1}$$

where $\psi_m = \phi_m$ if m is even and $\psi_m = i\phi_m$ if m is odd. Then the u_m are valid inducing features satisfying $cov(u_m, u_n) = \delta_{m,n}$.

The novelty of this result is that it gives us a recipe for defining inducing features that are *exactly* orthogonal for any stationary kernel.

Remark 3.2.2. *There is some redundancy in the condition $\frac{\phi_m(x)}{\sqrt{s(\omega)}}$ and $\phi_m(x)\sqrt{s(\omega)}$ are absolutely integrable, as the first condition implies the second. More generally, this condition in conjunction with the condition on the Fourier transforms should be thought of as saying* *the chosen basis functions decay rapidly and have rapidly decaying derivatives, with the precise rate of decay needed depending on the kernel function.*

Proof. We begin by showing that the covariance matrix between the inducing features defined in (3.1) is diagonal:

$$\begin{aligned}\operatorname{cov}(u_m,u_n) &= \mathbb{E}\left[\int_{\mathbb{R}} f(x)\mathcal{F}^{-1}(\psi_m s^{-1/2})(x)\int_{x'} f(x')\mathcal{F}^{-1}(\psi_n s^{-1/2})(x')dxdx'\right]\\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\int_{\mathbb{R}} \mathcal{F}^{-1}(\psi_m s^{-1/2})(x)\mathcal{F}^{-1}(\psi_n s^{-1/2})(x')\mathbb{E}[f(x)f(x')]dxdx'\\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\int_{\mathbb{R}} \mathcal{F}^{-1}(\psi_m s^{-1/2})(x)\mathcal{F}^{-1}(\psi_n s^{-1/2})(x')k(x,x')dxdx'\\ &= \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} \mathcal{F}^{-1}(\psi_m s^{-1/2})(x)\mathcal{F}^{-1}(\psi_n s^{-1/2})(x')\int_{\mathbb{R}} e^{-i\omega(x-x')}s(\omega)d\omega dxdx'.\end{aligned}$$

By our assumption that the first two integrals are absolutely convergent and since s is a probability density, (implying the third integral is absolutely convergent), we may apply Fubini's theorem to rearrange the order of integration:

$$\begin{aligned}\operatorname{cov}(u_m,u_n) &= \frac{1}{2\pi}\int_{\mathbb{R}}\int_{\mathbb{R}} \mathcal{F}^{-1}(\psi_m s^{-1/2})(x)\mathcal{F}^{-1}(\psi_n s^{-1/2})(x')\int_{\mathbb{R}} e^{-i\omega(x-x')}s(\omega)d\omega dxdx'\\ &= \frac{1}{2\pi}\int_{\mathbb{R}}\left(\int_{\mathbb{R}} \mathcal{F}^{-1}(\psi_m s^{-1/2})(x)e^{-i\omega x}dx\right)\left(\int_{\mathbb{R}} \mathcal{F}^{-1}(\psi_n s^{-1/2})(x')e^{i\omega x'}dx'\right)s(\omega)d\omega\end{aligned} \tag{3.2}$$

We can now recognize that the first term is $\sqrt{2\pi}$ times the Fourier transform of $\mathcal{F}^{-1}(\psi_m s^{-1/2})(x)$ evaluated at ω and the second is $\sqrt{2\pi}$ times the complex conjugate of the Fourier transform $\mathcal{F}^{-1}(\psi_n s^{-1/2})$ evaluated at ω. This simplifies considerably to

$$\begin{aligned}\operatorname{cov}(u_m,u_n) &= \frac{1}{2\pi}\int_{\mathbb{R}}\left(\int_{\mathbb{R}} \mathcal{F}^{-1}(\psi_m s^{-1/2})(x)e^{-i\omega(x)}dx\right)\left(\int_{\mathbb{R}} \mathcal{F}^{-1}(\psi_n s^{-1/2})(x')e^{i\omega(x')}dx'\right)s(\omega)d\omega\\ &= \int_{\mathbb{R}} \psi_m(\omega)s^{-1/2}(\omega)(-1)^n\psi_n(\omega)s^{-1/2}(\omega)s(\omega)d\omega\\ &= (-1)^n\int_{\mathbb{R}} \psi_m(\omega)\psi_n(\omega)d\omega\\ &= \delta_{m,n}.\end{aligned}$$

In particular, the matrix $\mathbf{K}_{m,m}$ is the identity matrix as claimed.

It remains to derive a formula for the covariance between inducing features and the original points in the process.

$$\begin{aligned}
\mathrm{cov}(u_m, f(x)) &= \mathbb{E}\left[\int_{\mathbb{R}} f(x')\mathcal{F}^{-1}(\psi_m s^{-1/2})(x') f(x) dx'\right] \\
&= \frac{1}{(2\pi)^{1/4}} \int_{\mathbb{R}} \mathcal{F}^{-1}(\psi_m s^{-1/2})(x') \mathbb{E}\left[f(x'), f(x)\right] dx' \\
&= \frac{1}{(2\pi)^{1/4}} \int_{\mathbb{R}} \mathcal{F}^{-1}(\psi_m s^{-1/2})(x') k(x', x) dx' \\
&= \frac{1}{(2\pi)^{1/4}} \int_{\mathbb{R}} \mathcal{F}^{-1}(\psi_m s^{-1/2})(x') \int_{\mathbb{R}} e^{-i\omega(x-x')} s(\omega) d\omega dx'.
\end{aligned}$$

Applying Fubini's theorem,

$$\mathrm{cov}(u_m, f(x)) = \frac{1}{(2\pi)^{1/4}} \int_{\mathbb{R}} \int_{\mathbb{R}} \mathcal{F}^{-1}(\psi_m s^{-1/2})(x') e^{i\omega x'} dx' e^{-i\omega x} s(\omega) d\omega$$

The first term is (up to a constant) the Fourier transform of the inverse Fourier transform of ψ evaluated at ω, giving

$$\mathrm{cov}(u_m, f(x)) = (2\pi)^{1/4} \int_{\mathbb{R}} \psi_m(\omega) s^{-1/2}(\omega) e^{-i\omega x} s(\omega) d\omega \quad = (2\pi)^{1/4} \int_{\mathbb{R}} \psi_m(\omega) e^{-i\omega x} \sqrt{s(\omega)} d\omega.$$

An unbiased estimator for this covariance can therefore be obtained via the Monte Carlo estimate,

$$\mathrm{cov}(u_m, f(x)) = \frac{(2\pi)^{1/4}}{T} \sum_{t=1}^{T} \psi_m(\omega_t) e^{-i\omega_t x}.$$

with ω_t sampled from a probability distribution proportional to $\sqrt{s(\omega)}$. Additionally, since $\sqrt{s(\omega)}$ is even, we can sample the ω_t in pairs, $\pm\omega_t$, to ensure the resulting approximate covariance is real valued. □

Remark 3.2.3. *More generally, from* (3.2) *we see that for a stationary kernel, the covariance between features is an inner product of the Fourier transform of the features taken in* $L^2(\mathbb{R}, s(\omega))$. *It follows that features that are well-defined (i.e. have finite variance) are orthogonal if and only if their Fourier transforms are orthogonal in* $L^2(\mathbb{R}, s(\omega))$.

Remark 3.2.4. *While in notation we have assumed one-dimensional inputs, the proof generalizes to the mulitdimensional case.*

The only requirement for applying this method is to choose a basis for $L^2(\mathbb{R})$ consisting of sufficiently well-behaved functions so that the necessary Fourier transforms converge. In section 3.4 we choose the Hermite functions as such a basis, which are infinitely differentiable with rapidly decaying derivatives. Smooth wavelet bases, for example the Meyer wavelet, would provide an interesting choice of local inducing features that satisfy this criteria. Wavelet based Gaussian process approximations have been previously proposed, in Zhu et al. (1997), though it does not appear such approximations have been practically implemented. We do not investigate these in the remainder of this work.

3.3 Eigenfunction Based Inducing Points

In the previous section, we focused on obtaining orthogonal inducing features. While the resulting features make use of the Fourier transform the relationship between the resulting features and the spectrum of the kernel is unclear and relies on a seemingly arbitrary choice of basis. In this section, we define orthogonal inducing features directly based on the spectral expansion of a compact kernel. Compactness is achieved for stationary kernels through assuming the existence of some prior probability measure on the inputs μ. Once a particular prior on inputs is chosen, the features are fully determined, making the resulting features more easily interpretable.

Recall, Mercer's theorem, (2.19):

$$k(\mathbf{x},\mathbf{x}') = \sum_{n=1}^{\infty} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{x}'),$$

with the λ_i absolutely summable and ϕ_i orthonormal eigenfunctions of the kernel operator $\mathcal{K}$ with respect to the input measure μ.

Define the following inducing variables:

$$u_m := \int_{\mathcal{X}} \phi_m(\mathbf{x}) f(\mathbf{x}) d\mu(\mathbf{x}).$$

These are exactly the same as the random variables utilised in the Karhunen–Loéve expansion in (2.20), so the entire process can be written as

$$f(\mathbf{x}) = \sum_{m=1}^{\infty} \lambda_m u_m \phi_m(\mathbf{x}),$$

where the equality indicates convergence in distribution in $L^2(\mathcal{X},\mu)$.

3.3.1 Covariances

Recall that since the ϕ_m are the eigenfunctions of a linear operator with respect to μ the ϕ_m and ϕ_n are pairwise μ-orthogonal if $\lambda_m \neq \lambda_n$, and a basis can be chosen to be orthogonal when eigenvalues are equal. This yields a diagonal structure to the matrix $\mathbf{K}_{m,m}$.

$$\begin{aligned}
\operatorname{cov}(u_m, u_n) &= \mathbb{E}\left[\int_{\mathcal{X}}\int_{\mathcal{X}} \phi_m(\mathbf{x})\phi_n(\mathbf{x}')f(\mathbf{x})f(\mathbf{x}')d\mu(\mathbf{x}')d\mu(\mathbf{x})\right] \\
&= \int_{\mathcal{X}} \phi_m(\mathbf{x}) \left(\int_{\mathcal{X}} \phi_n(\mathbf{x}')k(\mathbf{x},\mathbf{x}')d\mu(\mathbf{x}')\right) d\mu(\mathbf{x}) \\
&= \lambda_n \int_{\mathcal{X}} \phi_m(\mathbf{x})\phi_n(\mathbf{x})d\mu(\mathbf{x}) \\
&= \lambda_n \delta_{m,n}.
\end{aligned}$$

3.3.2 Cross covariances

We additionally must compute the covariance between the inducing features and random variables in the untransformed input domain. Using the defining property of eigenfunctions:

$$\begin{aligned}
\operatorname{cov}(u_m, f(f\mathbf{x}')) &= \mathbb{E}\left[\int_{\mathcal{X}} \phi_m(\mathbf{x})f(\mathbf{x})f(\mathbf{x}')d\mu(\mathbf{x})\right] \\
&= \int_{\mathcal{X}} \phi_m(\mathbf{x})k(\mathbf{x},\mathbf{x}')d\mu(\mathbf{x}) \\
&= \lambda_m \phi_m(\mathbf{x}').
\end{aligned}$$

3.3.3 Eigenfunction based inducing points and the mean field approximation

Additional structure emerges in the covariance matrix $\mathbf{S}$ appearing in the expanded variational bound, (2.25), in the case of conjugate likelihood if the training data is actually distributed according to the probability measure μ. A popular approximation that can be used to dramatically increase the speed of variational inference is to restrict the variational distribution to have a diagonal covariance function. Titsias (2009) showed that the optimal choice of $\mathbf{S}$ for conjugate inference has the closed form,

$$\mathbf{S}_{opt}^{-1} = \mathbf{K}_{m,m}^{-1} + \frac{1}{\sigma_{noise}^2}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{m,n}\mathbf{K}_{m,n}^{T}\mathbf{K}_{m,m}^{-1}. \tag{3.3}$$

In the case of eigenfunction based inducing points, since $\mathbf{K}_{m,m}$ is diagonal and without loss of generality can be be taken to be the identity matrix, this simplifies to

$$\mathbf{S}_{opt}^{-1} = \mathbf{I} + \frac{1}{\sigma_{noise}^2}\mathbf{K}_{m,n}\mathbf{K}_{m,n}^T. \tag{3.4}$$

Theorem 3.3.1. *If the training data comes from the assumed input distribution $p(\mathbf{x})$, then as $N \to \infty$ for fixed M, the optimal variational distribution in Gaussian process regression when using M eigenfunction inducing points has a diagonal covariance matrix.*

Proof. Let $\mathbf{D} = \frac{1}{\sigma_{noise}^2}\mathbf{K}_{m,n}\mathbf{K}_{m,n}^T$ so that $\mathbf{S}_{opt}^{-1} = \mathbf{I} + \mathbf{D}$. With this normalization, $\mathrm{cov}(u_i, f(\mathbf{x}_k)) = \sqrt{\lambda_i}\phi_i(\mathbf{x}_k)$, so

$$\mathbf{D}_{i,j} = \sqrt{\lambda_i\lambda_j}\sum_{i=1}^{N}\phi_i(\mathbf{x}_k)\phi_j(\mathbf{x}_k). \tag{3.5}$$

If we assume that the $\mathbf{x}_i$ are in fact distributed according to the prior measure μ, the right hand side of (3.5) is a Monte Carlo estimate of the integral

$$N\sqrt{\lambda_i\lambda_j}\int_{\mathcal{X}}\phi_i(\mathbf{x})\phi_j(\mathbf{x})d\mu(\mathbf{x}) = N\sqrt{\lambda_i\lambda_j}\delta_{i,j} + o(N). \tag{3.6}$$

(The error term can be made more precise under the mild assumption that the variance of $\phi_i(\mathbf{x})\phi_j(\mathbf{x})$ under $p(\mathbf{x})$ exists.) Then,

$$\mathbf{S}_{opt} = (\mathbf{\Lambda} + \mathcal{E})^{-1}$$

where the entries in $\mathcal{E}$ are all $o(N)$, and $\mathbf{\Lambda}$ is a diagonal matrix with $\mathbf{\Lambda}_{i,i} = \lambda_i N$. Intuitively, it seems as though if the precision matrix is essentially diagonal, then the covariance matrix should be as well. We show in appendix section A.2, that this is the case in this setting, and $\mathbf{S}_{opt}$ is approximately diagonal with on-diagonal entries on the order of $\frac{1}{N}$ and off-diagonal entries $o(1/N)$.

Under the assumption that the sum that the variance of $\phi_i(\mathbf{x})\phi_j(\mathbf{x})$ under $p(\mathbf{x})$ exists, the off diagonal entries are $O(N^{-3/2})$. □

It follows that if the features are defined with respect to a parameterized family of measures, and there exists a selection of parameters so that the x_i are approximately distributed according μ_θ, approximating $\mathbf{S}$ with a diagonal matrix should introduce almost no additional error in regression tasks. This is shown in Figure 3.1, which compares the additional error introduced by applying this diagonal approximation to eigenfunction inducing points as opposed to standard inducing point for training data sampled according $x_i \sim p(x)$, the

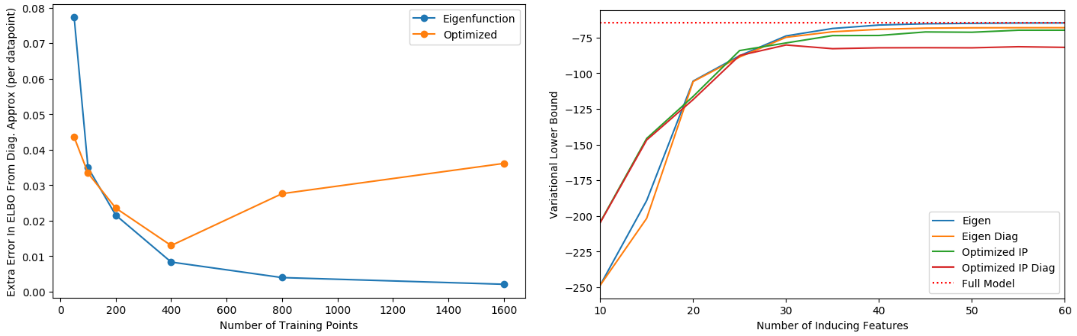

Fig. 3.1 A comparison of the gap between the ELBO with and without the diagonal approximation averaged over 3 random normally distributed datasets (left) as we increase N and convergence of the ELBO on a toy one dimensional data set with 500 inputs sampled from a normal distribution as we increase M (right).

distribution over which the features are defined. In this case, the features used are the eigenfunction inducing features defined in the next section, which are defined for a SE-kernel and assume a Gaussian input distribution.

As N increases, the additional approximation error goes to zero for the eigenfunction features, while no such guarantee is known for the inducing points. For a fixed dataset with 500 normally distributed inputs, we see that the diagonal approximation leads to almost no loss in performance for the eigenfunction function inducing features, while the gap for standard inducing points is noticeable.

Figure 3.2 compares the covariance matrices selected by eigenfunction based and optimized inducing points for a normally distributed data set. The covariance matrix of the variational distribution selected is nearly diagonal for the eigenfunction covariance matrix, while the matrix selected for optimized inducing points has several relatively large off-diagonal entries.

3.4 Example: Squared Exponential Kernel and Hermite Inducing Points

We now introduce an example of inducing features that are orthogonal and can be derived in the frameworks of both Sections 3.2 or 3.3, although we only present the derivation in the latter case, as it is significantly simpler and gives rise to more natural interpretations of variational parameters. Fix a squared exponential kernel, k with variance v_k and lengthscale

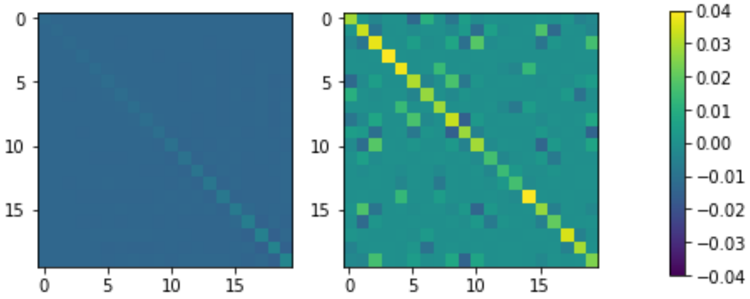

Fig. 3.2 The covariance matrix selected by variational inference with the eigenfunction inducing features (left) is nearly diagonal when the data is normally distributed, while for standard inducing points (right) the off-diagonal entries are not necessarily near zero.

ℓ^2,

$$k(x,x') = v_k \exp\left(-\frac{\|x-x'\|^2}{2\ell^2}\right). \tag{3.7}$$

Suppose $x_i \sim \mathcal{N}(x;0,s^2) =: p(x)$. Consider the associated covariance operator

$$\mathcal{K} : f \to \int_{\mathbb{R}} f(x)k(x,x')p(x)dx.$$

The eigenbasis for $\mathcal{K}$ is given (up to reparameterization) in Zhu et al. (1997) as:

$$\lambda_m = v_k \sqrt{\frac{2a}{A}} B^m, \tag{3.8}$$

and

$$\phi_m(x) = \exp(-(c-a)x^2)H_m(\sqrt{2c}x), \tag{3.9}$$

with $a = 1/(4s^2)$, $b = 1/(2\ell^2)$, $c = \sqrt{a^2+2ab}$, $A = a+b+c$, $B = b/A$, and H_m the m^{th} Hermite polynomial. The Hermite polynomials are defined by the differential equation:

$$H_m(x) = (-1)^m \exp(x^2)\frac{d^m}{dx^m}\exp(-x^2). \tag{3.10}$$

The corresponding inducing features, normalized so that $\text{cov}(u_m, u_m) = 1$, are given by:

$$u_m := \frac{(4cs^2)^{1/4}\sqrt{\lambda_m}}{2^{n/2}\sqrt{n!}} \int H_m(\sqrt{2c}x)\exp(-(c-a)x^2)f(x)d\exp\left(-\frac{x}{2s^2}\right). \tag{3.11}$$

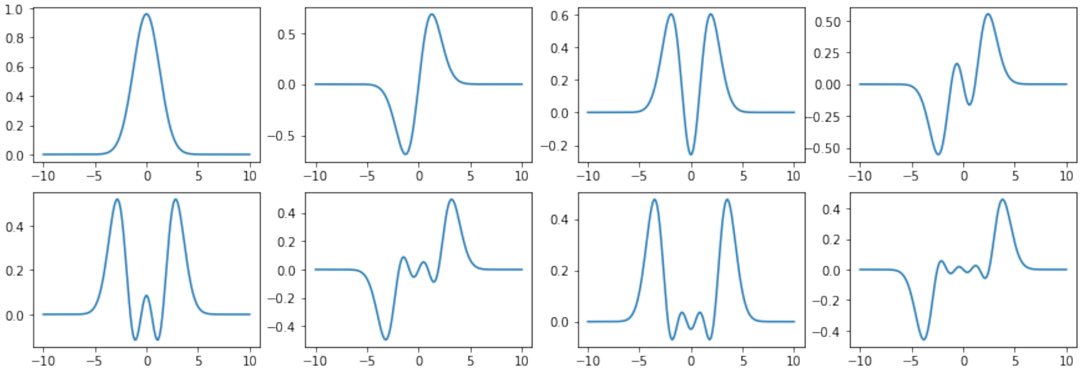

Fig. 3.3 Covariance of first six inducing features for Hermite inducing features with the process for $\ell = 1, s = 1$.

With this normalization,

$$\operatorname{cov}(u_m, f(x)) = \sqrt{\frac{\sqrt{4cs^2}\lambda_m v_k}{n!2^n}} H_m(\sqrt{2c}x)\exp\left(-(c-a)x^2\right). \tag{3.12}$$

To relate this to the framework given in Section 3.2, the Hermite functions defined by $\psi_n := H_n(x)\exp(-x^2/2)$, form an orthogonal basis for $L^2(\mathbb{R})$. Using this fact the u_m defined above can be derived via computing the necessary Fourier transforms with the proof being simplified by noting the Hermite functions are eigenfunctions of the Fourier transform.

The decay of the eigenvalues, plotted on a log-linear scale in figure 3.4 for different choices of s^2, determines the rate of convergence of truncations of the sum appearing in Mercer's theorem to the full covariance function. As we will discuss in the next chapter, this rate of convergence is also closely related to the convergence of variational approximations, with convergence measured by the KL-divergence between the full and approximate posteriors. Rapidly decaying eigenvalues indicate that extremely sparse approximations are possible.

For practical implementations, we must consider the computational cost of evaluating the first M Hermite functions at each data point, as this is necessary when forming the matrix $\mathbf{K}_{n,m}$. This can be done in $O(\tilde{N}M)$ time by exploiting the second order recursion relation,

$$\phi_{m+1}(x) = x\sqrt{\frac{2}{m+1}}\phi_m(x) + \sqrt{\frac{m}{m+1}}\phi_{m-1}(x). \tag{3.13}$$

While this allows for $O(\tilde{N}M)$ computation of $\mathbf{K}_{n,m}$, this recursion does not allow for easy parallelization over the features. Additionally, care must be given to normalize the inducing features (e.g. by computing this recursion in a manner so that $\phi_m(x)$ is uniformly bounded in

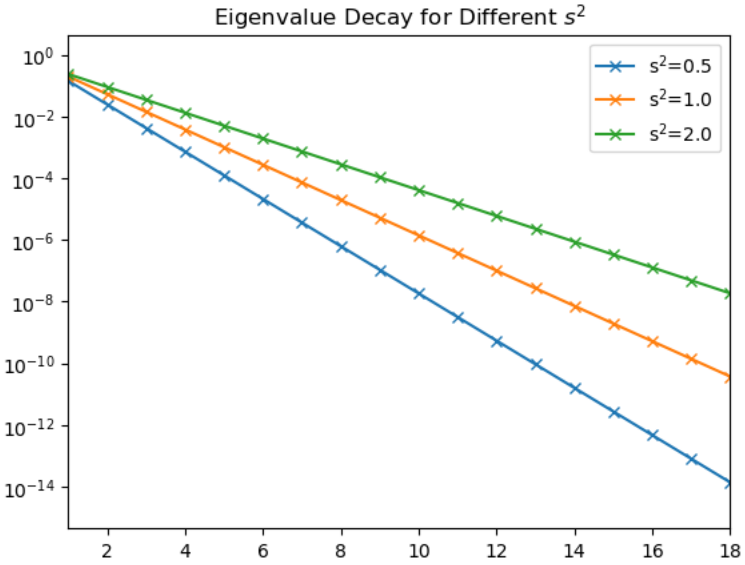

Fig. 3.4 For fixed $\ell^2 = 1$, larger choices of s^2 result in slower decay in eigenvalues.

both x and m) to make this recursion sufficiently stable to allow for optimization of feature hyperparameters.

3.5 Example: Exponential Kernel and Variational Fourier Features

In Hensman et al. (2016), the authors derived nearly orthogonal inducing features for Matèrn Kernels over a fixed interval $[a,b]$ based on trigonometric functions. In this section, we consider the relationship between their "Variational Fourier Features" and the eigenfunction inducing features derived in section 3.3. We restrict to the case of the Matérn 1/2 kernel, which is up to a reparameterization the same as the exponential kernel. The kernel considered is,

$$k(x,x') = v_k \exp\left(-\frac{|x-x'|}{\ell}\right).$$

Suppose $x_i \sim \mathcal{U}[a,b]$ so $\mathcal{K} : f \to \int_a^b f(x)k(x,x')dx$. The corresponding eigenbasis is given in Le Maître and Knio (2010, Chapter 2, Equation 2.2) by:

$$\lambda_m = v_k \frac{2\ell}{1+(\omega_m \ell)^2}, \tag{3.14}$$

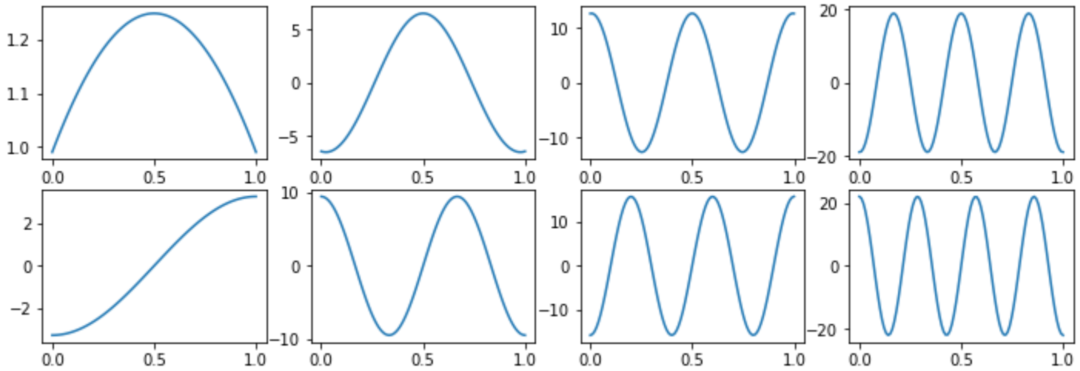

Fig. 3.5 Covariance of first 8 eigenfunction inducing features for Matérn kernel on [0,1] with $\ell = 1$, normalized so that the covariance matrix between features is the identity matrix. Even features are shown in the top row, and odd in the bottom.

and

$$\phi_m(x) = \begin{cases} \frac{\cos(\omega_m(x-(a+b)/2))}{\sqrt{(b-a)/2+\sin(\omega_m(b-a))/2\omega_m}} & m \text{ even}, \\ \frac{\sin(\omega_m(x-(a+b)/2))}{\sqrt{(b-a)/2-\sin(\omega_m(b-a))/2\omega_m}} & m \text{ odd}. \end{cases} \tag{3.15}$$

The odd $\omega_m > 0$ are given by the ordered solutions to the transcendental equation:

$$\ell\omega\tan((b-a)\omega/2) - 1 = 0, \tag{3.16}$$

and the even $\omega_m > 0$ are given by solutions to the equation:

$$\tan((b-a)\omega/2) + \ell\omega = 0. \tag{3.17}$$

Remark 3.5.1. *To fully define the approximate inference model it is also necessary to compute the covariance with the inducing features and the process at outputs outside* $[a,b]$*. This computation is given in the section A.1 of the appendix.*

The first 8 eigenfunctions are shown in figure 3.5.

Since $\omega\tan(\omega/2)$ and $\frac{1}{\omega}\tan(\omega/2)$ are one-to-one with $\mathbb{R}$ on the interval $[0, 2\pi]$, there is exactly one solution to each of these equation on each interval of the form $\left[\frac{2k\pi}{b-a}, \frac{2(k+1)\pi}{b-a}\right]$. From this, we conclude $\omega_m = \pi m + O(1)$ and $\lambda_m = \frac{2\nu_k(b-a)^2}{\ell\pi^2 m^2} + O(1/m^3)$. Note that the eigenvalues for this kernel decay much more slowly than in the case of the squared exponential kernel, as shown in figure 3.6 In general, the rate of decay of the eigenvalues of $\mathcal{K}$ is closely related to the smoothness of sample functions, as mentioned in Rasmussen and Williams (2005, Chapter 4).

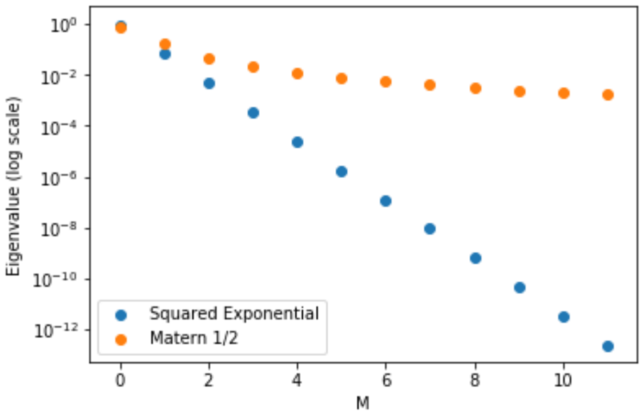

Fig. 3.6 Comparison of the eigenvalues for the Matérn kernel on $[0,1]$ and the SE-kernel with respect to a Gaussian with variance 1/12, both with length scales 1. (Note the priors are chosen to have the same variance.

3.5.1 Relationship to Variational Fourier Features

Aesthetically, these features are very similar to the Matérn 1/2 Variational Fourier Features. Both result in covariances that are sinusoidal, with the difference being that the VFF features have harmonic frequencies over $[a,b]$ while these inducing features have frequencies given by the roots of a transcendental equation. To expand on the similarity between these eigenfunction inducing points and VFF, consider the integral equation

$$\lambda_m g_m(x') = \int_a^b k(x,x')\phi_m(x)dx. \tag{3.18}$$

Note that this is the covariance between the process at x' and $u_m := \int_a^b \phi_m(x)f(x)dx$. Incorporating all normalizing constants into the constant term λ_m, defining eigeninducing features involves finding solutions to $g_m(x) = \phi_m(x)$. If we instead find solutions to (3.18) of the form $\phi_m(x) = \cos(\omega_m(x - \frac{a+b}{2}))$ (or $\sin(\omega_m(x - \frac{a+b}{2}))$) with $\omega_m = \frac{2\pi m}{(b-a)}$, this corresponds to the even (odd) L^2 Fourier features derived in Hensman et al. (2016). Alternatively, solving (3.18) for $g_m(x) = \cos(\omega_m(x - \frac{a+b}{2}))$ (or $\sin(\omega_m(x - \frac{a+b}{2}))$) with ω_m harmonic corresponds to the RKHS version of VFF.

The RKHS VFF features of Hensman et al. (2016) have several significant advantages for approximating this specific kernel over using the eigenfunction inducing features defined in (3.15) directly. First, in order to use (3.15), we must find the zeros of the transcendental equations (3.16) and (3.17). The locations of these zeros depends on the kernel hyperparameters, so zeros must be recomputed in each iteration of optimization of kernel hyperparameters.

Additionally, if (3.18) is solved for fixed g_m independent of the kernel hyperparameters *and* any variational parameters, as is done in the RKHS VFF features then there is no need to compute $\mathbf{K}_{n,m}$ in each iteration of optimization for regression tasks. Hensman et al. (2016) precompute this matrix and related quantities and optimize kernel parameters with cost $O(M^3)$ per iteration completely independent of the size of the dataset.

The same pre-computation cannot be performed in optimizing eigenfunction based inducing features, as (up to normalization) *all* of the variational parameters are in the matrix $\mathbf{K}_{n,m}$.

3.6 Optimality and Infinite Principal Component Analysis

The parametric model of Zhu et al. (1997) can be considered the infinite dimensional analogue of PCA. The authors show that it minimizes expected mean squared modelling error with respect to the prior on the input distribution $p(x)$, among all linear M-dimensional approximations. As the inducing features we define are essentially identical to the basis used in defining this parametric model, it is natural to ask whether these features can be said to be optimal with respect to the KL-divergence. For any fixed data set, this will not be the case, as the KL-divergence between full and approximate inference depends on the empirical distribution of $\mathbf{x}$ through both terms in (2.24). It additionally depends on $\mathbf{y}$ through the likelihood term.

In order to define a quantity related to the KL-divergence that is independent of the observed data, as is the case for the modelling error in Zhu et al. (1997), we can average the log marginal likelihood of the full model minus the lower bound (2.24), over the priors for both $\mathbf{x}$ and $\mathbf{y}$. Let $\mathbf{Q}_{n,n} = \mathbf{K}_{n,m}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{n,m}^T$,

$$\mathbb{E}_{\mathbf{y}}[KL(p(\mathbf{f}|\mathbf{u})\|p(\mathbf{f}|\mathbf{u},\mathbf{y}))] = \int_{\mathcal{X}}\int_{\mathcal{Y}} \log\left(\frac{\mathcal{N}\left(\mathbf{y};0,\sigma_{noise}^2\mathbf{I}+\mathbf{K}_{n,n}\right)}{\mathcal{N}\left(\mathbf{y};0,\sigma_{noise}^2\mathbf{I}+\mathbf{Q}_{n,n}\right)}\right) p(\mathbf{y})d\mathbf{y}p(\mathbf{x})d\mathbf{x} + \frac{1}{2\sigma_{noise}^2}\int_{\mathcal{X}} \operatorname{tr}\left(\mathbf{K}_{n,n}-\mathbf{Q}_{n,n}\right)p(\mathbf{x})d\mathbf{x}.$$

For any fixed $\mathbf{x}$, our prior on $\mathbf{y}$ is $\mathcal{N}\left(\mathbf{y};0,\mathbf{K}_{n,n}+\sigma^2_{noise}\mathbf{I}\right)$ so the integral over $\mathbf{y}$ is a KL-divergence between two mean-centered Gaussian distributions. This yields,

$$\mathbb{E}[KL(p(f|\mathbf{u})\|p(\mathbf{f}|\mathbf{u},\mathbf{y}))] = -\frac{N}{2}+\frac{1}{2}\int_{\mathcal{X}}\log\left(\frac{|\sigma^2_{noise}\mathbf{I}+\mathbf{Q}_{n,n}|}{|\sigma^2_{noise}\mathbf{I}+\mathbf{K}_{n,n}|}\right)$$
$$+\operatorname{tr}\left(\left(\sigma^2_{noise}\mathbf{I}+\mathbf{Q}_{n,n}\right)^{-1}\left(\sigma^2_{noise}\mathbf{I}+\mathbf{K}_{n,n}\right)\right)+\frac{1}{\sigma^2_{noise}}\operatorname{tr}\left(\mathbf{K}_{n,n}-\mathbf{Q}_{n,n}\right)d\mathbf{x}. \quad (3.19)$$

Minimizing this entire quantity over low rank $\mathbf{Q}_{n,n}$ appears to be a challenging problem.

The problem of minimizing the final term in the integral in (3.19) for any $\mathbf{x}$ is equivalent to finding the optimal rank M approximation to the positive semidefinite matrix $\mathbf{K}_{n,n}$. As discussed in section 2.3 one solution to this minimization is achieved by taking the approximating matrix to have column space spanned by the lead M eigenfunctions $\mathbf{K}_{n,n}$. The optimal error is given by the sum:

$$\sum_{i=M+1}^{N}\lambda_i\left(\mathbf{K}_{n,n}\right), \quad (3.20)$$

with $\lambda_i\left(\mathbf{K}_{n,n}\right)$ denoting the i^{th} eigenfunction of the matrix $\mathbf{K}_{n,n}$. For $\mathbf{x}_i \sim p(\mathbf{x})$ the eigenvalues of $\frac{1}{N}\mathbf{K}_{n,n}$ approach the eigenvalues of the corresponding operator $\mathcal{K}$, so optimal bounds on this term are closely related to the eigenvalues utilised in defining eigenfunction inducing features. In the next chapter, we focus on obtaining bounds on this trace term, as it seems to be the most tractable of the terms and previous work of Titsias in Titsias (2014) in conjunction with (2.24) enables us to obtain explicit bounds on the KL-divergence between the approximate and full models based on the trace term alone.

Chapter 4

Eigenfunction Based Inducing Points and the Convergence of Approximate Gaussian Process Models

In this chapter, we motivate the eigenfunction inducing features introduced in section 3.3 from the perspective of obtaining theoretical guarantees on the rate of convergence of the approximate likelihood to the full likelihood.

Much of the existing literature on the convergence of approximate Gaussian process models focuses on proving that the approximate model will agree with full inference model for either $M = N$ (e.g. variational inducing point methods) or as M tends to infinity (e.g. with parametric methods such as Sparse Spectrum Gaussian Processes). However, computational savings in sparse models only occur when $M \ll N$ and empirical evidence suggests that good approximations are obtainable with M much smaller than N.

A natural method for defining convergence (in M) between the full and approximate models is to consider the gap between the full marginal likelihood and the ELBO, which is given (Hensman et al., 2013) by

$$KL\left[p\left(\mathbf{f}|\mathbf{u}\right), p\left(\mathbf{f}|\mathbf{u},\mathbf{y}\right)\right],$$

with $\mathbf{f} = \{f(\mathbf{x}_i)\}_{i=1}^N, \mathbf{u} = \{u_j\}_{j=1}^M$ and $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^N$.

If this gap is small for all reasonable kernel hyperparameter settings, we can be sure the approximate model has the same global optima as the full model. More precisely, if the globally optimal choice of hyperparameters has a log marginal likelihood that exceeds all other local optima by $\delta > 0$ and for all reasonable kernel hyperparameter settings, this KL divergence is less than δ, then the global optima of the approximate inference objective function is near that of the full model, so the approximate posterior will behave similarly to the full posterior after optimization. In this chapter we obtain both explicit and asymptotic results about the rate of convergence of the variational lower bound obtained via inducing feature methods to the marginal likelihood of the full model (equivalently rates of convergence of the KL-divergence between the full and approximate models to zero) in the case of GP regression.

Bounds appearing in the literature allow us to assess the convergence of the approximate model to the full model *after* computing the full and approximate log likelihood. In particular, by comparing the upper and lower bounds on the marginal likelihood provided in Titsias (2009) and Titsias (2014), one can get some sense of the degree to which the approximation has converged. In contrast the bounds in this chapter allow us to bound the rate of convergence *a priori* for a given kernel, giving insight into the number of features needed to be confident of recovering similar hyperparameter settings as full inference.

The main result in this chapter is the following for Gaussian process regression:

Theorem 4.0.1. *Given a SE-kernel, suppose $x_i \sim N(0, s'^2)$ or $x_i \sim \mathcal{U}(-r, r)$ and that $\|\mathbf{y}\|^2 = O(N^2)$ almost surely (this is the case if the true process generating the data has finite mean and variance, as the likelihood is assumed to be Gaussian). Let K denote the KL-divergence between the full and approximate processes. Then we can find an M on the order of* $\log(N)$ *such that as $N \to \infty$, almost surely $K = o(1)$.*

This says that if N is sufficiently large, we can find a very sparse (relative to N) model that approximates the full model with arbitrary accuracy. Along the way, we derive explicit probabilistic upper bounds on the KL-divergence for the squared exponential kernel under the assumptions that the data is generated according to either a uniform or normal distribution in Theorem 4.3.1 and 4.3.3, in conjunction with Lemma 4.1.2.

All of the proofs rely on obtaining bounds on trace of the covariance matrix used in computation of the “approximate likelihood” relative to the trace of the full covariance matrix. We additionally discuss the optimality properties of this error in the asymptotic regime when $N \to \infty$.

Later in the chapter, we consider the Matérn kernel, and show that we can obtain bounds on the ‘trace term’ in the ELBO, (2.24) in this setting for both standard inducing points placed on a grid and for eigenfunction inducing points in this setting. These bounds imply that we would need to take on the order of N inducing points to obtain a bounded trace error, and due to the slackness of Lemma 4.1.2, we would need to take M significantly larger than N to have a provably bounded KL-divergence within the framework derived in this chapter.

We conclude this chapter with theoretical results on convergence for standard inducing points with the points subsampled from observed data. These rely on existing bounds in the literature for the error of Nyström approximation. While not practically applicable due to slackness in the constants in these bounds, they provide an interesting area for further investigation.

4.1 Preliminary Results

Let $\mathbf{Q}_{n,n} = \mathbf{K}_{n,m}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{n,m}^{T}$. In the case of Gaussian process regression, Titsias (2009) and Titsias (2014) establish a strong link between the trace of the matrix $\mathbf{E}_{n,n} := \mathbf{K}_{n,n} - \mathbf{Q}_{n,n}$, and the KL-divergence between the approximate and exact posterior. Titsias (2014) showed that for a fixed data set, if $\text{tr}(\mathbf{E}_{n,n}) = 0$ then exact inference is recovered. We begin by stating this result and leveraging it to obtain an explicit bound on the KL-divergence based on $\mathbf{E}_{n,n}$.

4.1.1 Bounding the KL-divergence based on the Trace Error

Define $t := \text{tr}(\mathbf{E}_{n,n})$, (here and throughout, we suppress the dependence of this error on M in notation). We begin by recalling the variational lower bound of Titsias (2009), (2.24),

$$\mathcal{L}_{lower} = \log\mathcal{N}(\mathbf{y}; 0, \mathbf{Q}_{n,n} + \sigma_{noise}^2\mathbf{I}) - \frac{t}{2\sigma_{noise}^2} \leq \log\mathcal{N}(\mathbf{y}; 0, \mathbf{K}_{n,n} + \sigma_{noise}^2\mathbf{I}) = \mathcal{L},$$

where $\mathcal{L}$ denotes the log marginal likelihood of the full model.

On the other hand, Titsias (2014) provides the following upper bound on $\mathcal{L}$,

$$\begin{aligned}\mathcal{L}_{upper} := \log\left(\mathcal{N}(\mathbf{y}; 0, \mathbf{Q}_{n,n} + t\mathbf{I} + \sigma_{noise}^2\mathbf{I})\right) + \frac{1}{2}\left(\log\left(|\mathbf{Q}_{n,n} + t\mathbf{I} + \sigma_{noise}^2\mathbf{I}|\right)\right.\\ \left. - \log\left(|\mathbf{Q}_{n,n} + \sigma_{noise}^2\mathbf{I}|\right)\right). \quad (4.1)\end{aligned}$$

Remark 4.1.1. *The proof provided in Titsias (2014) indicates that the trace term appearing in the first two terms in* (4.1) *could be replaced by the operator norm (largest eigenvalue) of* $\mathbf{E}_{n,n}$, *which could be useful in refining convergence results presented in Section 4.5.*

Our goal is to bound $\mathcal{L} - \mathcal{L}_{lower}$ in terms of M and kernel hyperparameters alone. It suffices to bound $\mathcal{L}_{upper} - \mathcal{L}_{lower}$. To this end we establish the following lemma:

Lemma 4.1.2. *With the notation established above,*

$$\mathcal{L}_{upper} - \mathcal{L}_{lower} \leq \frac{t}{2\sigma_n^2} + \frac{t\|\mathbf{y}\|^2}{2\sigma_n^4 + 2t\sigma_n^2}. \quad (4.2)$$

Proof. The proof has a similar spirit to that of (4.1) provided in Titsias (2014). Let $\mathbf{R} = \mathbf{Q}_{n,n} + \sigma_n^2\mathbf{I}$.

$$\begin{aligned}\mathcal{L}_{upper} - \mathcal{L}_{lower} &= \frac{t}{2\sigma_n^2} + \frac{1}{2}\left(\mathbf{y}^T\mathbf{R}^{-1}\mathbf{y} - \mathbf{y}^T(\mathbf{R}+t\mathbf{I})^{-1}\mathbf{y}\right) \\ &= \frac{t}{2\sigma_n^2} + \frac{1}{2}\left(\mathbf{y}^T\left(\mathbf{R}^{-1} - (\mathbf{R}+t\mathbf{I})^{-1}\right)\mathbf{y}\right). \qquad (4.3)\end{aligned}$$

Since $\mathbf{Q}_{n,n}$ is symmetric positive semidefinite, $\mathbf{R}$ is positive definite with eigenvalues bounded below by σ^2_{noise}. Write, $\mathbf{R} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{U}$ is unitary and $\mathbf{\Lambda}$ is a diagonal matrix with non-increasing diagonal entries $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_N \geq \sigma^2_{noise}$. (For $M < N$ the last $M - N$ eigenvalues of $\mathbf{R}$ must equal $\sigma_n^2\mathbf{I}$.)

We can rewrite the second term (ignoring the factor of one half) in (4.3) as,

$$\mathbf{y}^T\left(\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^T - \mathbf{U}(\mathbf{\Lambda}+t\mathbf{I})^{-1}\mathbf{U}^T\right)\mathbf{y} = (\mathbf{U}^T\mathbf{y})^T\left(\mathbf{\Lambda}^{-1} - (\mathbf{\Lambda}+t\mathbf{I})^{-1}\right)(\mathbf{U}^T\mathbf{y}).$$

Now define, $\mathbf{z} = (\mathbf{U}^T\mathbf{y})$. Since $\mathbf{U}$ is unitary, $\|\mathbf{z}\| = \|\mathbf{y}\|$.

$$\begin{aligned}(\mathbf{U}^T\mathbf{y})^T\left(\mathbf{\Lambda}^{-1} - (\mathbf{\Lambda}+t\mathbf{I})^{-1}\right)(\mathbf{U}^T\mathbf{y}) &= \mathbf{z}^T\left(\mathbf{\Lambda}^{-1} - (\mathbf{\Lambda}+t\mathbf{I})^{-1}\right)\mathbf{z} \\ &= \sum_i z_i^2 \frac{t}{\lambda_i^2 + \lambda_i t} \\ &\leq \|\mathbf{y}\|^2 \frac{t}{\lambda_N^2 + \lambda_N t}. \qquad (4.4)\end{aligned}$$

The last inequality comes from noting that the fraction in the sum attains a maximum when λ_i is minimized. Since σ^2_{noise} is a lower bound on the smallest eigenvalue of $\mathbf{R}$, we can also write,

$$\mathbf{y}^T\left(\mathbf{R}^{-1} - (\mathbf{R}+t\mathbf{I})^{-1}\right)\mathbf{y} \leq \frac{t\|\mathbf{y}\|^2}{\sigma^4_{noise} + \sigma^2_{noise}t},$$

from which Lemma 4.1.2 follows. □

4.1.2 General Results Based on Eigenvalues

Let $\mathbf{K}_{n,n}$ have eigendecomposition

$$\mathbf{K}_{n,n} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T.$$

where $\mathbf{U}$ a unitary $N \times N$ matrix and $\mathbf{\Lambda}$ a diagonal matrix with decreasing eigenvalues $\lambda_1(\mathbf{K}) \geq \lambda_2(\mathbf{K}) \geq \ldots \geq \lambda_N(\mathbf{K}) \geq 0$. Such an eigendecomposition necessarily exists as the matrix is symmetric, positive semidefinite. As mentioned in section 2.3 an optimal rank M approximation $\mathbf{K}_{n,n}$ in terms of any unitarily invariant matrix norm (including the trace norm) is given by $\mathbf{U}_m\mathbf{\Lambda}_m\mathbf{U}_m^T$, where $\mathbf{U}_m$ is given by the the first M columns of $\mathbf{U}$. The resulting approximation leads to $t = \sum_{i=M+1}^{N} \lambda_i(\mathbf{K})$. It follows, as noted by Titsias (2014), that,

$$t \geq \sum_{i=M+1}^{N} \lambda_i(\mathbf{K}). \tag{4.5}$$

In general, computing this optimal lower bound is computationally intractable, as it involves finding the first M eigenvalues of $\mathbf{K}_{n,n}$, which most commonly used algorithms require at least $\Omega(N^2M)$ computation to achieve.

4.1.3 Asymptotic Properties of Eigenvalues of the Kernel Matrix

While finding a precise lower bound for the error of the trace term is computationally intractable for large kernel matrices, understanding its asymptotic properties if the training data comes from a known input distribution $p(\mathbf{x})$ is simpler. In this case, eigenvalues of $\frac{1}{N}\mathbf{K}_{n,n}$ converge to the eigenvalues of the operator, $\mathcal{K} : f \to \int_{\mathcal{X}} f(\mathbf{x})k(\mathbf{x},\mathbf{x}')p(\mathbf{x})d\mathbf{x}$. Moreover, concentration results can be used to obtain explicit rates of convergence in probability, see for example Shawe-Taylor et al. (2002). Additionally, Shawe-Taylor et al. (2002) show that the empirical covariance function has smaller tail sums in expectation than the covariance operator, so we should expect that the optimal trace term for a finite N is somewhat lower than an estimate derived from working with the corresponding infinite dimensional operator. This trend can be seen in Figure 4.1 for a SE-kernel with the operator eigenvalues plotted for normally distributed inputs, and the empricial eigenvalues plotted for covariance matrices with different input distributions.

As an example, consider the squared exponential kernel given by

$$k(x,x') = v_k \exp\left(\frac{\|x-x'\|^2}{2\ell^2}\right), \tag{4.6}$$

and suppose $x_i \sim \mathcal{N}(0,s^2) =: p(x)$. Recall the eigenvalues of the corresponding operator, $\mathcal{K}$, are given by (3.8):

$$\lambda_m = v_k\sqrt{\frac{2a}{A}}B^m, \tag{4.7}$$

with $a = 1/(4s^2)$, $b = 1/(2\ell^2)$, $c = \sqrt{a^2+2ab}$, $A = a+b+c$, $B = b/A$. Let $\lambda_i(\mathbf{K})$ denote the i^{th} eigenvalue of the empirical covariance matrix (we have suppressed the dependence on n for notational convenience.

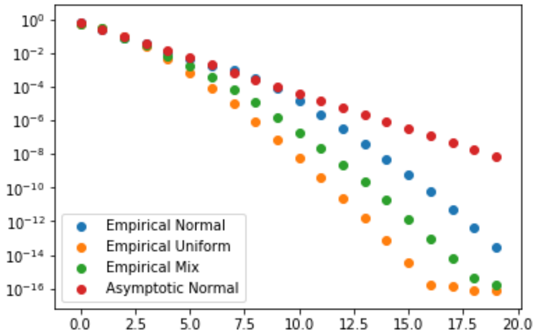

Fig. 4.1 Decay of first 20 eigenvalues for data points sampled from a uniform, normal and mixture of gaussian input distributions, with $N = 500$ for a SE kernel with $\ell^2 = 1$. All three input distributions with standard deviation 1. The eigenvalues of the covariance operator for defined with respect to the same normal distribution are also shown.

For fixed i, $\lim_{N\to\infty} \frac{1}{N}\lambda_i(\mathbf{K}) \to \lambda_i$. So for any fixed M,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=0}^{M-1} \lambda_i(\mathbf{K}) \to \sum_{i=0}^{M-1} \lambda_i.$$

In this case, the right hand side is a finite geometric series,

$$\sum_{i=0}^{M-1} \lambda_i = v_k \sqrt{2a/A}\frac{1-B^M}{1-B}. \tag{4.8}$$

Additionally, $\frac{1}{N}\sum_{i=0}^{N-1} \lambda_i(\mathbf{K}) = v_k$, as each diagonal entry in $\mathbf{K}_{n,n}$ for this kernel is v_k. Write

$$v_k = \sum_{i=0}^{N-1} \frac{1}{N}\lambda_i(\mathbf{K}) = \sum_{i=0}^{M-1} \frac{1}{N}\lambda_i(\mathbf{K}) + \sum_{i=M}^{N-1} \frac{1}{N}\lambda_i(\mathbf{K}),$$

So,

$$\begin{aligned}
\frac{1}{N} \lim_{N\to\infty} \sum_{i=M}^{N-1} \lambda_i(\mathbf{K}) &= v_k - \sum_{i=0}^{M-1} \lambda_i \\
&= v_k - v_k\sqrt{2a/A}\frac{1-B^M}{1-B} \\
&= \frac{v_k\sqrt{2a}B^M}{\sqrt{A}(1-B)}.
\end{aligned} \tag{4.9}$$

In other words, for fixed M sending to N infinity, we should expect that under the above assumptions the trace term will be bounded below by a quantity on the order of B^M for *any inducing feature*. In the next section, we show a matching bound holds almost surely for eigenfunction based inducing points as $N \to \infty$.

4.2 General Bounds for Eigenfunction based Inducing Points

The primary obstacle to obtaining bounds on the trace term in order to apply Lemma 4.1.2 with standard inducing points is understanding the the matrix $\mathbf{K}_{m,m}^{-1}$. As this matrix is diagonal for the eigenfunction inducing points bounding the trace term is significantly simplified. Additionally, since the spectrum of the empirical covariance matrix converges to that of $\mathcal{K}$ under the assumption $x_i \sim p(x)$ these bounds will be quite good if this condition is satisfied.

Recall from Mercer's Theorem that

$$k_{i,j} = k(x_i, x_j) = \sum_{m=0}^{\infty} \lambda_m \phi_m(x_i) \phi_m(x_j),$$

holds μ a.e., with the right hand side convergent in $L^2(\mathcal{X}, \mu)$. Note the small change of notation, indexing the eigenvalues starting at 0.

From the definition of the covariance matrix $\mathbf{K}_{n,m}$ and the computation of covariances in Section 3.3,

$$q_{i,j} = \sum_{m=0}^{M-1} \lambda_m \phi_m(x_i) \phi_m(x_j),$$

Bounding the termwise error in the estimated covariance matrix is the equivalent to bounding the sum,

$$|k_{i,j} - q_{i,j}| = \Big| \sum_{m=M}^{\infty} \lambda_m \phi_m(x_i) \phi_m(x_j) \Big|.$$

Consider the normalized trace error,

$$\frac{1}{N} t = \frac{1}{N} \sum_{i=1}^{N} |k_{i,i} - q_{i,i}| = \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{m=M}^{\infty} \lambda_m \phi_m(x_i)^2 \right). \tag{4.10}$$

Suppose $x_i \sim q$, with $q(x)$ a probability density such that $q(x) \leq cp(x)$, $\forall x$ where p is the density with respect to which the features are defined.

The right hand side of (4.10) is a sum of N i.i.d random variables, $\frac{1}{N}\sum_{i=1}^{N} A_i$, with $A_i := \left(\sum_{m=M}^{\infty} \lambda_m \phi_m(x_i)^2\right)$. By the strong law of large numbers, as $N \to \infty$, $\frac{1}{N} t \to \mathbb{E}_q[A_i]$ almost surely if the expectation exists.

$$\begin{aligned}\mathbb{E}_q[A_i] &= \int_{\mathcal{X}} \sum_{m=M}^{\infty} \lambda_m \phi_m(x)^2 q(x)dx \\ &\leq c \int_{\mathcal{X}} \sum_{m=M}^{\infty} \lambda_m \phi_m(x)^2 p(x)dx \\ &= c \sum_{m=M}^{\infty} \lambda_m \int_{\mathcal{X}} \phi_m(x)^2 p(x)dx \\ &= c \sum_{m=M}^{\infty} \lambda_m. \end{aligned} \tag{4.11}$$

The interchanging of the sum and integral is justified from the Tonelli-Fubini theorem as the terms are nonnegative.

Hence

$$\frac{1}{N} t \leq c \sum_{m=M}^{\infty} \lambda_m, \tag{4.12}$$

almost surely. As p and q are both probability densities, c is greater or equal to 1 with equality if and only if $p = q$ almost everywhere. Additionally, while we have shown this convergence almost surely for any fixed M, since the countable union of almost sure events holds almost surely, this estimate holds almost surely for all M.

For fixed M, as $N \to \infty$, this matches the asymptotic lower bound on the trace term of the empirical matrix discussed in the previous section, though this lower bound does not necessarily hold as $M \to \infty$.

4.2.1 Concentration Inequalities

While the strong law of large numbers tells us that for sufficiently large N we can be confident the bound in (4.12) holds, it is desirable to obtain a more explicit result. This can be achieved using concentration inequalities.

Theorem 4.2.1 (Hoeffding's Inequality, Hoeffding (1963))**.** *Suppose $A_1, A_2, \dots, A_N$ are i.i.d random variable with $0 \leq A_i \leq v$, and $\mathbb{E}[A_i] = \mu \leq \infty$. Define $S_N = \frac{1}{N}\sum_{i=1}^{N} A_i$. Then,*

$$Pr(S_n - \mu > r) \leq \exp\left(\frac{-2Nr^2}{v^2}\right), \tag{4.13}$$

for $r > 0$.

Using Hoeffding's inequality for the trace with $r = \delta\mu$, $\mu \leq \sum_{m=M}^{\infty} \lambda_m$ and under the supposition that $k_{i,i} - q_{i,i} \leq v_M$,

$$\Pr\left(t < (1+\delta)cN \sum_{m=M}^{\infty} \lambda_m\right) \leq \exp\left(\frac{-2Nc^2\delta^2 \left(\sum_{m=M}^{\infty} \lambda_m\right)^2}{v_M^2}\right). \tag{4.14}$$

This bound will be used for the squared exponential kernel when it is assumed that the data is uniformly distributed, as in this setting the individual terms in the trace can be bounded above by a quantity that decays at the same rate as the sum of the tail of the eigenvalues.

When we consider the squared exponential kernel with $x_i \sim \mathcal{N}(0, s'^2)$, this bound will not be applicable, individual terms along the trace cannot be bounded uniformly in x at a rate that decreases with M. Instead, we will use a generalized Chebyshev bound based on the the r^{th} moment.

Theorem 4.2.2 (von Bahr and Esseen (1965))**.** *Let $\{A_i\}_{i=1}^N$ be i.i.d random variables with $\mathbb{E}[A_i] = 0$ and $\beta_r = \mathbb{E}\left[|M_i|^r\right]$, for $1 < r \leq 2$. Take $S_N := \frac{1}{N}\sum_{i=1}^{N} A_i$, then*

$$P(S_N > a) \leq \left(2 - \frac{1}{N}\right)\beta_r a^{-r} N^{1-r}. \tag{4.15}$$

Remark 4.2.3. *A slight refinement using a one sided inequality, e.g. a generalized Cantelli inequality, is likely possible though this will not improve the rate of convergence significantly in the interesting cases where we are able to show β_r is small.*

Remark 4.2.4. *For r large (greater than 1.6), von Bahr and Esseen (1965) give a refinement in to the constant $\left(2 - \frac{1}{N}\right)$. For $r = 2$ the standard Chebyshev inequality allows us to replace the constant $\left(2 - \frac{1}{N}\right)$ with 1.*

In section 4.3 we consider the special case of the squared exponential kernel with features defined with respect to a normal distribution, in two settings: first when the inputs are uniformly distributed and then when inputs are distributed according to a normal distribution with lengthscale less than the prior input distribution used to define the features.

4.3 Bounds for Hermite Based Inducing Points and The Squared Exponential Kernel

Throughout this sections we consider the problem of approximate inference with a squared exponential kernel

$$k(x, x') = v_k \exp\left(-\frac{\|x - x'\|_2^2}{2\ell^2}\right)$$

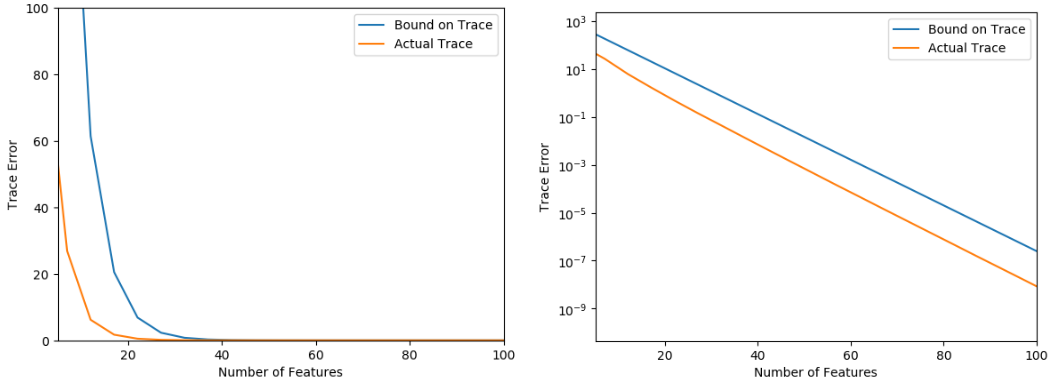

Fig. 4.2 Actual error on the trace and the bound given in (4.18) (on an absolute scale on the left and log scale on the right) that holds for large N plotted for a synthetic data set with $N = 200$, $x \sim \mathcal{U}[-\sqrt{75}, \sqrt{75}]$ and $s = 5$.

using Hermite features defined with respect to the input distribution $p(x) \sim \mathcal{N}(0, s^2)$.

Theorem 4.3.1. *Suppose $x_i \sim \mathcal{U}[-R, R]$ then for all $M \geq 0$, let $c = \frac{\sqrt{2\pi s^2}}{R} \exp(\frac{R^2}{2s^2})$. Then,*

$$\frac{1}{N} t \leq c \sum_{m=M}^{\infty} \lambda_m \tag{4.16}$$

holds almost surely for all $M \geq 0$ as $N \to \infty$. Moreover, we have the estimate

$$Pr\left(t > (1+\delta)cN \sum_{m=M}^{\infty} \lambda_m\right) \leq \exp\left(\frac{-2N\delta^2}{v_M^2}\right) \tag{4.17}$$

with $v_M = 1.44(1 + s^2/\ell^2)^{1/4} \frac{R}{\sqrt{2\pi s^2}}$.

Remark 4.3.2. *The proof of 4.3.1 generalizes to $x_i \sim q$ with q any bounded input distributions on a fixed interval with different constants c_q and $v_{M,q}$.*

Theorem 4.3.3. *Suppose $x_i \sim \mathcal{N}(0, s'^2)$ with $s' \leq s$ then for all $M \geq 0$,*

$$t \leq cN \sum_{m=M}^{\infty} \lambda_m \tag{4.18}$$

for $c = \frac{s}{s'}$ holds almost surely for all $M \geq 0$ as $N \to \infty$. Moreover, for $s' < s$ we have the estimate

$$Pr\left(t \geq (c+\delta)N\left(\sum_{m=M}^{\infty} \lambda_m\right)\right) \leq \min_{1<r<\min(s^2/s'^2,2)} \alpha_r \delta^{-r} N^{1-r}, \tag{4.19}$$

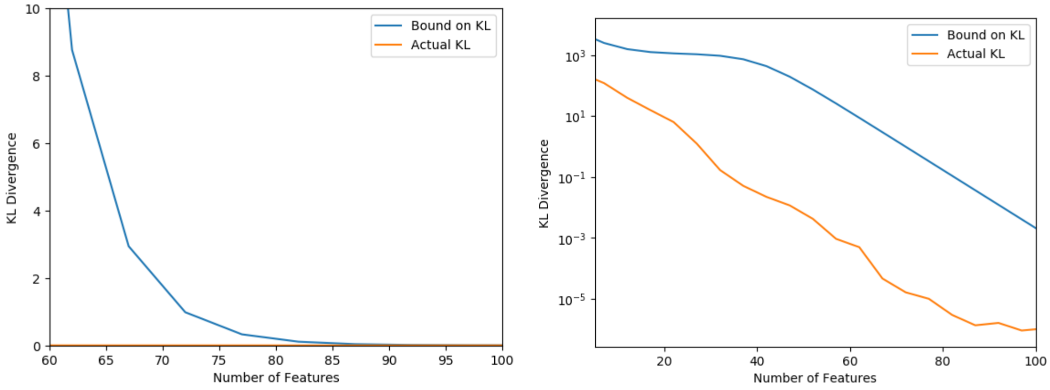

Fig. 4.3 Actual error as measured by KL-divergence as compared to the bound in (4.16) that holds for large N plotted for a synthetic data set with $N = 200$, $x \sim \mathcal{U}[-\sqrt{75}, \sqrt{75}]$ and $s = 5$. Plotted on an absolute scale on the left and log scale on the right.

with

$$\alpha_r = \left(2 - \frac{1}{N}\right) 1.44^r (1 + s^2/\ell^2)^{r/4} \sqrt{\frac{1}{1 - r\left(\frac{s'}{s}\right)^2}}.$$

Remark 4.3.4. *Analogous results to Theorem 4.3.3 can be achieved for $r > 2$ obtaining a rate of convergence of $O(N^{1-r})$, but we consider the case when s' and s are close as we expect this to generally be the result of optimization of the feature lengthscale parameter.*

Remark 4.3.5. *In Theorems 4.3.1 and 4.3.3, decreasing s^2 increases the rate of convergence of* (4.16) *and* (4.18) *to* 0*, as B becomes closer to zero. At the same time, decreasing s^2 leads to lower probabilities of convergence, and increases the constant c in Theorem 4.3.1. The rates of convergence also depend on the spread of the data, and align with the intuition that we should choose inducing features corresponding to the input distribution being very wide if the empirical data covers a large range, and inducing features corresponding to a normal distribution with low variance if the data is tightly clustered.*

Figure 4.4 indicates that the bound on the trace that can be show to hold on average is very tight if $p(x) = q(x)$, although in figure 4.5 we see that the resulting KL-bound is somewhat weaker. The bounds in the uniform setting, shown in figures 4.2 and 4.3 are, perhaps unsurprisingly, somewhat weaker. Despite this, as the bound on the KL-divergence converges to zero exponentially, it shows the approximation has essentially converged for M small enough to be within the realm of practical application.

The key result in the proofs is the following term-wise bound on the entries of the error matrix $\mathbf{E}_{n,n}$:

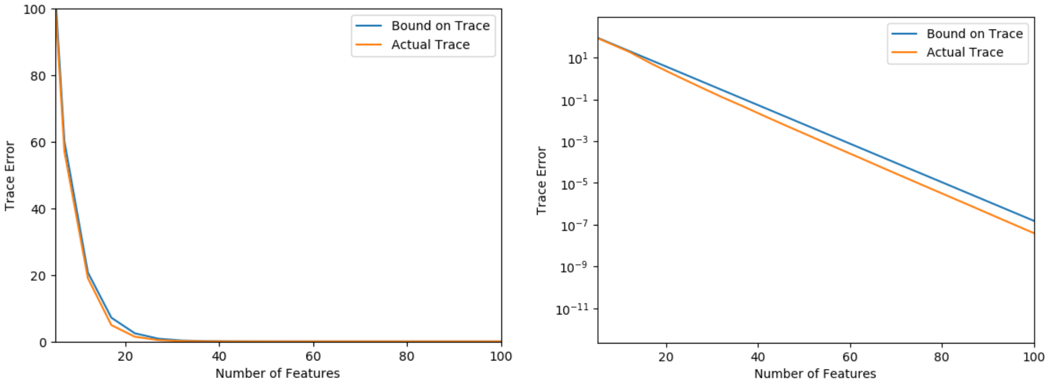

Fig. 4.4 Actual error on the trace and the bound given in (4.18) that holds for large N plotted for a synthetic data set with $N = 200$, $x \sim \mathcal{N}(0, 5^2)$ and $s = 5$. Plotted on an absolute scale (left) and a log-scale (right).

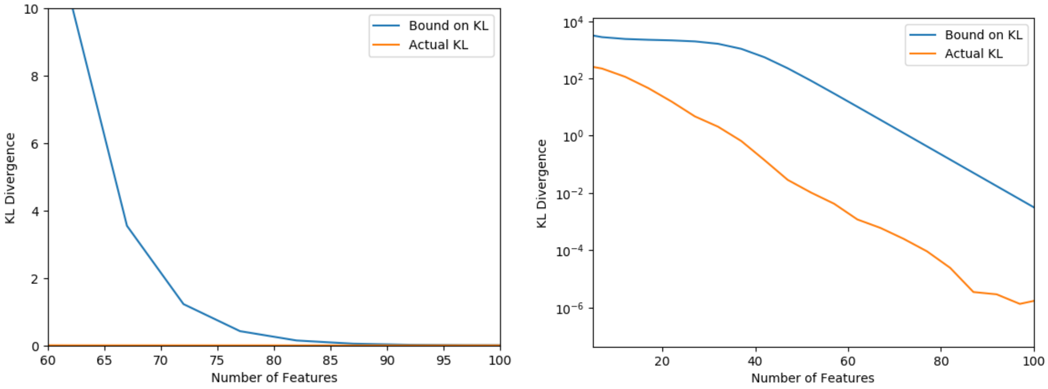

Fig. 4.5 Actual KL-Divergence and upper bound given by combining (4.18) from Theorem 4.3.3 and Lemma 4.1.2 that holds for large N plotted for a synthetic data set with $N = 200$, $x \sim \mathcal{N}(0, 5^2)$ and $s = 5$. Plotted on an absolute scale (left) and a log-scale (right).

Lemma 4.3.6. *Let $\mathbf{Q}_{n,n} := \mathbf{K}_{n,m}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{n,m}^{T}$, $q_{i,j}$ denote the i,j^{th} entry in $\mathbf{Q}_{n,n}$ and $k_{i,j} = k(x_i,x_j)$ denote the i,j^{th} entry in $\mathbf{K}_{n,n}$, then*

$$|k_{i,j} - q_{i,j}| \leq \frac{2.4}{\sqrt{2}} v_k \frac{\lambda_M (1+s^2/\ell^2)^{1/4}}{1-B} \exp\left(\frac{x_i^2+x_j^2}{4s^2}\right). \tag{4.20}$$

holds for $x_i, x_j \sim q$ almost surely.

Proof of Lemma 4.3.6. As in earlier proofs, we have

$$|k_{i,j} - q_{i,j}| = \Big|\sum_{m=M}^{\infty} \lambda_m \phi_m(x_i)\phi_m(x_j)\Big|,$$

We need to now take into account the location of the x_i in this bound. For this we use the following bound on Hermite functions, obtained by squaring the bound in Gradshteyn and Ryzhik (2014),

$$|H_n(x_i)||H_n(x_j)| < 1.44 n! 2^{n/2} e^{(x_i^2+x_j^2)/2}.$$

Expanding into the definition of the (normalized) ϕ_m we obtain

$$\begin{aligned}
|k_{i,j} - q_{i,j}| &= \sqrt{4cs^2}\exp(a(x_i^2+x_j^2))v_k\sqrt{\left(\frac{2a}{A}\right)}\left|\sum_{m=M}^{\infty}\frac{B^m H_m(\sqrt{2c}x_j)e^{-cx_j^2}H_m(\sqrt{2c}x_i)e^{-cx_i^2}}{2^m m!}\right| \\
&\leq \sqrt{4cs^2}\exp(a(x_i^2+x_j^2))v_k\sqrt{\left(\frac{2a}{A}\right)}\sum_{m=M}^{\infty}\frac{B^m\left|H_m(\sqrt{2c}x_j)e^{-cx_j^2}\right|\left|H_m(\sqrt{2c}x_i)e^{-cx_i^2}\right|}{2^m m!} \\
&\leq 1.44\sqrt{4cs^2}\exp(a(x_i^2+x_j^2))v_k\sum_{m=M}^{\infty}\lambda_m \\
&\leq 1.44(1+s^2/\ell^2)^{1/4}\exp\left(\frac{x_i^2+x_j^2}{4s^2}\right)\sum_{m=M}^{\infty}\lambda_m.
\end{aligned}$$

The first inequality (triangle inequality) is sharp on the terms effecting the trace since when $x_i = x_j$ the sum must be term-wise positive. □

This gives us that for diagonal terms, under the assumption that $x_i \leq R$, we have

$$v_M \leq 1.44(1+s^2/\ell^2)^{1/4}\exp\left(\frac{R^2}{2s^2}\right)\sum_{m=M}^{\infty}\lambda_m.$$

Theorem 4.3.1 then follows from Hoeffding's Theorem, 4.2.1.

For the proof of Theorem 4.3.3, we must show that the terms satisfy a bound on the r^{th} centered moment. We focus on the uncentered moment, which gives an upper bound on the centered moment as the mean is nonnegative. Using Lemma (4.3.6) in the first line:

$$\begin{aligned}\beta_r &\leq 1.44^r(1+s^2/\ell^2)^{r/4}\left(\sum_{m=M}^{\infty}\lambda_m\right)^r \frac{1}{\sqrt{2\pi s'^2}}\int_{\mathbb{R}}\exp\left(\frac{rx^2}{2s^2}\right)\exp\left(\frac{-x^2}{2s'^2}\right)dx\\ &= 1.44^r(1+s^2/\ell^2)^{r/4}\left(\sum_{m=M}^{\infty}\lambda_m\right)^r \frac{1}{\sqrt{2\pi s'^2}}\int\exp\left(\frac{(s^2-rs'^2)}{s^2s'^2}\left(\frac{-x^2}{2}\right)\right)dx\\ &= 1.44^r(1+s^2/\ell^2)^{r/4}\left(\sum_{m=M}^{\infty}\lambda_m\right)^r\sqrt{\frac{1}{1-r\left(\frac{s'}{s}\right)^2}}.\end{aligned} \tag{4.21}$$

The last line follows from noting that the integral is proportional to the integral of Gaussian distribution with standard deviation $\sqrt{s'^2/(1-r\frac{s'^2}{s^2})}$ Take $a=\delta\left(\sum_{m=M}^{\infty}\lambda_m\right)$, then

$$Pr\left(t\geq(c+\delta)N\left(\sum_{m=M}^{\infty}\lambda_m\right)\right)\leq\left(2-\frac{1}{N}\right)1.44^r(1+s^2/\ell^2)^{r/4}\sqrt{\frac{1}{1-r\left(\frac{s'}{s}\right)^2}}\delta^{-r}N^{1-r} \tag{4.22}$$

for $1<r<\min(2,(s^2/s'^2))$. We can minimize this bound with respect to valid r.

4.3.1 The Number of Inducing Points

We now restate and prove the main result of this section, Theorem 4.0.1.

Theorem 4.3.7. *Given a SE-kernel, suppose $x_i\sim N(0,s'^2)$ or $x_i\sim\mathcal{U}(-r,r)$ and that $\|\mathbf{y}\|^2\leq aN^2$ almost surely for some constant a (this is the case if the true process generating the data has finite mean and variance, as the likelihood is assumed to be Gaussian). Let K denote the KL-divergence between the full and approximate processes. Then we can find an M on the order of* $\log(N)$ *such that as $N\to\infty$, almost surely $K=o(1)$.*

Proof. Substituting the assumption that $\|\mathbf{y}\|^2\leq aN^2$ almost surely into Lemma 4.1.2,

$$K=O\left(t+\frac{tN^2}{1+t}\right), \tag{4.23}$$

holds almost surely, with the implicit constant depending on both a and σ^2_{noise}.

It suffices to show that for some $M=O(\log(N))$, the trace is almost surely $o(1/N^2)$. Take $M=(4+\varepsilon)\log(N)=\log(N^{4+\varepsilon})$. The result then follows in the case of uniformly distributed inputs from Theorem 4.3.1 and in the case of Gaussian distributed inputs from Theorem 4.3.3. In particular, from (4.9), $\sum_{i=M}^{\infty}\lambda_i=O(B^M)$, so

$$t = O\left(N^2 B^M\right) = O(1/N^{2+\varepsilon}).$$

almost surely. □

4.4 The Exponential Kernel

We now investigate the exponential kernel,

$$k_{exp}(x_i, x_j) = v_k \exp\left(\frac{-|x_i - x_j|}{\ell}\right)$$

We begin by analysing the trace term in the case of standard inducing points placed on a grid. Understanding this case has two benefits. First, while the bound on the trace term proven for inducing points placed on a grid might not bound the trace term of optimized inducing points, any bound on the KL-divergence derived from this bound must hold as a bound on the actual KL-divergence, because this is the quantity with which optimality is defined with respect to.

Additionally, we will assume the actual data is uniformly distributed over some interval. In this case, for large data sets it is reasonable to expect that amongst any collection of inducing points method, placing the inducing points on an evenly spaced grid over the interval is near optimal.

4.4.1 Inducing Points Placed on a Grid

Analysing grid structured data for the exponential kernel is particularly tractable since the exponential kernel corresponds to a first order Markov process, so the inverse covariance matrix $\mathbf{K}_{m,m}^{-1}$ has a particularly simple form.

The bound we are able to obtain in this case is the following:

Theorem 4.4.1. *Given an exponential kernel k with length scale and variance* 1, *suppose* $x_i \sim \mathcal{U}[0,1]$. *Consider M inducing points placed so that $z_m = \frac{m}{M}$. Then,*

$$\frac{1}{N} t = \frac{1}{3M} + O(1/M^2) \tag{4.24}$$

holds almost surely as $N \to \infty$.

Remark 4.4.2. *It would likely be slightly better to place the inducing points at locations so as to divide the interval into $M+1$ regions by not placing z_M at the end of the interval, but the* *improvement in the bound obtained is $O(1/M^2)$, so for ease of exposition, this technicality is avoided.*

Proof. Define, $r := \exp(-1/(M\ell))$. Then, $\operatorname{cov}(z_m, z_{m+k}) = r^k$. The covariance matrix $\mathbf{K}_{m,m}$ is then a *Kac-Murdock-Szegö* or *Markovian* matrix, meaning it has the form

$$\mathbf{K}_{m,m} = \begin{pmatrix} 1 & r & r^2 & \dots & r^M \\ r & 1 & r & \dots & r^{M-1} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ r^M & r^{M-1} & r^{M-2} & \dots & 1 \end{pmatrix}.$$

The inverse of such a matrix is tridiagonal, see (Horn and Johnson, 1990, Exercise 7.2.P13), given by

$$K_{m,m}^{-1} = \frac{1}{1-r^2} \begin{pmatrix} 1 & -r & 0 & \dots & 0 \\ -r & 1+r^2 & -r & \dots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & -r & 1+r^2 & -r \\ 0 & 0 & \dots & -r & 1 \end{pmatrix}.$$

We assume $0 \le x_i \le 1$, and write $x_i = \frac{m_x - \delta}{M}$, with $0 \le \delta < 1$. Define $d = \exp(\delta/(M\ell))$. Then,

$$\begin{aligned} \operatorname{cov}(z_m, x_i) &= \exp(-|m - m_x + \delta|/(M\ell)) \\ &= \begin{cases} r^k d & m_x - m = k > 0 \\ r^k d^{-1} & m - m_x = k \ge 0. \end{cases} \end{aligned} \tag{4.25}$$

If $x_i = z_m$ for some m, then $d = 1$. The quadratic form $= \mathbf{k}^{(i)T} K_{m,m}^{-1} \frac{1+r^2-2r^2}{1-r^2} = 1$ (the boundary cases $i = 1, n$) can also be easily verified. More generally, suppose $m = m_{x_i}$, then

$$\begin{aligned}
\mathbf{k}^{(i)T} K_{m,m}^{-1} \mathbf{k}^{(i)} &= \frac{1}{1-r^2}\left(k_1^2 + k_M^2 + (1+r^2)\sum_{j=2}^{M-1} k_j^2 - 2r\sum_{j=1}^{M-1} k_j k_{j+1}\right) \\
&= \frac{1}{1-r^2}\left(d^2 r^{2m_x} + d^{-2} r^{2M-2m_x} + (1+r^2) d^2 \sum_{j=2}^{m_x-1} r^{2m_x-2j} + (1+r^2) d^{-2} \sum_{j=m_x}^{M-1} r^{2j-2m_x}\right. \\
&\left. - 2rd^2 \sum_{j=1}^{m_x-2} r^{2m_x-2j-1} - 2r^2 - 2rd^{-2} \sum_{j=m_x}^{M-1} r^{2j-2m_x+1}\right) \\
&= \frac{1}{1-r^2}\left(d^2 r^{2m_x} + d^{-2} r^{2M-2m_x} + (1-r^2) d^2 \sum_{i=1}^{m_x-1} r^{2i} + d^{-2}(1-r^{2M-2m_x}) - 2r^2\right) \\
&= \frac{1}{1-r^2}\left(d^2 r^{2m_x} + d^{-2} r^{2M-2m_x} - (1-r^2) d^2 + d^2(1-r^{2m_x}) + d^{-2}(1-r^{2M-2m_x}) - 2r^2\right) \\
&= \frac{r^2 + d^{-4} - 2r^2 d^{-2}}{d^{-2}(1-r^2)}. \qquad (4.26)
\end{aligned}$$

The corresponding entry in $\mathbf{K}_{n,n}$ is one, so the trace term consists of elements of the form,

$$1 - \frac{r^2 + d^{-4} - 2r^2 d^{-2}}{d^{-2}(1-r^2)} = \frac{(1-d^{-2})(d^{-2}-r^2)}{d^{-2}(1-r^2)}.$$

Via Taylor expansion,

$$r = \exp(-1/(M\ell)) = 1 - 1/(M\ell) + O(1/M^2).$$

So

$$d^{-2} = 1 - 2\delta/(M\ell) + O(1/M^2) \text{ and } r^2 = 1 - 2/(M\ell) + O(1/M^2).$$

Substituting these expansions into (4.26)

$$k_{i,i} - \mathbf{k}^{(i)T} K_{m,m}^{-1} \mathbf{k}^{(i)} = \frac{\frac{4\delta - 4\delta^2}{(M\ell)^2} + O(1/M^3)}{2/(M\ell) + O(1/M^2)} = \frac{2-2\delta}{(M\ell)} + O(1/M^2).$$

Since $x_i \sim U[0,1]$ implies $\delta \sim U[0,1]$, $2E[\delta - \delta^2] = 1 - 2E[\delta^2] = 1/3$, so for large M as $N \to \infty$, we have

$$\frac{1}{N}\sum_{i=1}^{N}(k_{i,i} - \mathbf{k}^{(i)T}\mathbf{K}_{m,m}^{-1}\mathbf{k}^{(i)}) = \frac{1}{3M} + O(1/M^2).$$

holds almost surely by the strong law of large numbers. □

Unlike with the Hermite based inducing points and the squared exponential function in which the trace error decayed exponentially in the number of inducing points, with the exponential kernel and standard inducing points placed on a grid, we observe only linear decay in the number of inducing points. It is natural to ask whether this is a result of the different kernel or the features. In the next section, we apply eigenfunction based inducing features to help resolve this.

Remark 4.4.3. *Summing the termwise bound in the worst case, where we have* $d = \sqrt{\frac{1}{r}}$ *for all* x_i *gives the trivial bound on the trace*

$$\frac{1}{N}t \leq \frac{(1-r)^2 r}{r(1-r^2)} = 1. \tag{4.27}$$

On the other hand, when optimizing inducing points such a situation cannot occur, since this bound results from all of the actual data falling on a grid consisting of the midpoints of the inducing points, and if the M inducing points were instead selected to fall on these points, we would recover exact inference.

4.4.2 Using Eigenfunction based Inducing Points

We now consider the bounds obtained by instead using the first M eigenfunction based inducing points defined according to exponential kernel with uniform input distribution. In this setting, we prove the bound:

Theorem 4.4.4. *Under the same assumptions as in Theorem 4.4.1, but using eigenfunction inducing points defined with respect to the exponential kernel defined with respect to the uniform input distribution* $[0,1]$ *(see* (3.15)*), we have*

$$\frac{1}{N}t = \frac{2}{M\pi^2} + O(1/M) \tag{4.28}$$

holds almost surely.

Remark 4.4.5. *Comparing the bounds in Theorem 4.4.1 and Theorem 4.4.4 shows that at least asymptotically the eigenfunction based inducing points outperform grid based inducing points in terms of minimising the trace, as* $2/\pi^2 = .203 \leq 1/3$, *but only by a constant factor.*

Proof. From the general results on eigenfunction inducing points derived in section 4.2, it suffices to bound the sum,

$$\sum_{m=M}^{\infty} \lambda_M. \tag{4.29}$$

Recall, $\lambda_m = \frac{2}{1+(\omega_m)^2}$, with $\omega_m = \pi m + O(1)$.

$$\sum_{m=M}^{\infty} \lambda_M = \sum_{m=M}^{\infty} \frac{2}{1+(\omega_m)^2} \tag{4.30}$$

$$= \frac{2}{\pi^2} \sum_{m=M}^{\infty} \frac{1}{m^2} + O(1/M^2) \tag{4.31}$$

$$= \frac{2}{M\pi^2} + O(1/M^2). \tag{4.32}$$

□

We can additionally obtain a termwise bound on the trace using that $\phi_m^2(x_i) \leq \frac{2}{1-\frac{1}{w_m}}$,

$$\sum_{m=M+1}^{\infty} \lambda_m \phi_m^2(x_i) \leq 4 \sum_{m=M+1}^{\infty} \left(\frac{1}{1-\frac{1}{w_m}} \right) \frac{1}{1+\omega_m^2} \tag{4.33}$$

$$= \frac{4}{\pi^2} \sum_{m=M+1}^{\infty} \frac{1}{m^2} + O(1/M^2) \tag{4.34}$$

$$= \frac{4}{\pi^2} \frac{1}{M} + O(1/M^2). \tag{4.35}$$

This could be used in conjunction with concentration results. However, as M needs to be on the order of N in order for the trace to be $O(1)$ in expectation, we cannot obtain an interesting result (i.e. one that holds for $M \leq N$) for the number of inducing features needed to obtain convergence in KL-divergence using Lemma 4.1.2.

4.5 General Bounds Using Optimized Inducing Points

For optimized inducing points, we have from Lemma 4.1.2 with K denoting the KL-divergence,

$$K \leq \frac{t}{2\sigma_{noise}^2} + \frac{t\|\mathbf{y}\|^2}{2\sigma_{noise}^4 + 2t\sigma_{noise}^2}. \tag{4.36}$$

with $t = \min_{\{\mathbf{z}_m\}_{m=1}^M : \mathbf{z}_m \in \mathbb{R}^d} \mathbf{K}_{n,n} - \mathbf{K}_{n,m}\mathbf{K}_{m,m}^{-1}\mathbf{K}_{n,m}^T$. Note that the $\mathbf{z}_m$ actually chosen may not minimize this trace error, but the bound on the KL-divergence still applies as the lefthand side of (4.36) is the target of the optimization. In general it is not clear how to find this minimum.

However, bounds are known for specific choices of $\mathbf{z}_m$, which can be used in place of the optimal selection to derive upper bounds on K. For example, random selection of a subset of $M \ll N$ points (i.e. a subset of regressors model) leads to bounding the error of a standard

Nyström approximation of the matrix $\mathbf{K}_{n,n}$. The Nyström method has been directly applied to Gaussian process approximation, in Williams and Seeger (2001) and connections between this approximation and inducing points methods have been previously observed in Williams et al. (2002). This observation can be used in conjunction with the following bound:

Lemma 4.5.1. *Gittens and Mahoney (2013, Lemma 4) Let E_k^* denote the error of the optimal rank M approximation to $\mathbf{K}_{n,n}$ by any rank k matrix, i.e. $E_k^* = \sum_{i=k+1}^{N} \lambda_i(\mathbf{K}_{n,n})$. Let $\mu_{\mathbf{K}} = \frac{n}{k} \max_{i \leq n} \|\mathbf{U}_i\|_2^2$, where $\mathbf{U}_i$ is an $n \times k$ unitary matrix consisting of the top k eigenvectors of $\mathbf{K}_{n,n}$. Fix δ and $\varepsilon \in (0,1)$. If $M \geq \frac{2\mu k}{\varepsilon} \log(k/\delta)$ then*

$$t \leq \left(1 + \frac{1}{\delta^2(1-\varepsilon)}\right) E_k^* \tag{4.37}$$

with probability at least $1-4\delta$.

Remark 4.5.2. *Gittens and Mahoney (2013) also give bounds on the operator norm of the Nyström approximation that could be used with a refined version of Lemma 4.1.2 to offer an improvement on the upper bound on the KL-divergence.*

Remark 4.5.3. *The constant $\mu_{\mathbf{K}}$, known as the coherence of the matrix is generally difficult to compute. Good bounds on this constant in the setting of kernel matrices with inputs coming from a specified distribution would be of interest for obtaining better bounds on the convergence rates of non-optimized inducing point methods.*

Remark 4.5.4. *As we are generally interested in optimized inducing points, we can maximize the bound over any choice of δ, as the randomness in the bound comes from the sampling scheme and not the matrix being approximated*

An alternative, based on a *leveraged based* sampling scheme of columns is also given in Gittens and Mahoney (2013). Again, as we are interested primarily in optimized points, we do not focus on the sampling scheme and instead view it as a method to show the existence of a subset of columns yielding a good low rank approximation.

Lemma 4.5.5. *Gittens and Mahoney (2013, Lemma 3) There exists a stochastic method for sampling columns of the covariance matrix such that for δ and ε in $(0,1]$. For $M \geq 3200k/\varepsilon^2 \log(4k/\delta)$*

$$t \leq \left(1 + \varepsilon^2\right) E_k^*, \tag{4.38}$$

with probability at least $0.6 - \delta$.

Remark 4.5.6. *We can take any $\delta < 0.6$ and be sure the resulting bound on the KL-divergence holds, as we optimize the selection.*

While theoretically interesting, in our setting this bound is not useful, as we would need to utilise over 3200 inducing points for it to apply and at least an order of magnitude more to obtain interesting bounds on the KL-divergence.

Bounds on the quality of Nyström approximation in trace error, specifically tailored to kernel matrices, and perhaps under the assumption that points are drawn from a given prior distribution would be of interest in terms of obtaining general, practical and explicit bounds on the error introduced by performing approximate inference with optimized inducing points.

Chapter 5

Experiments

In this chapter we explore the practical applicability of eigenfunction based inducing points. We begin the chapter looking at their performance on several toy one-dimensional datasets, and show the effect of the empirical distribution of training data on the number of features needed to achieve convergence. Additionally, we discuss potential issues with the approximation prior to convergence. We then discuss methods for scaling the features to approximate multidimensional data and show the practical applicability on a large 'airline' dataset that has been used to benchmark prior work in the scalable Gaussian process literature. All models are implemented in Python using the 'GPFlow' library (Matthews et al., 2017).

5.1 Toy One Dimensional datasets

In defining the eigenfunction based inducing points, we needed to choose an input distribution, $p(x)$. In the case of the Hermite based features, this is assumed to be a normal distribution, with mean, μ and variance s^2, where μ and s are variational parameters that can be optimized along with kernel hyperparameters. While we have justified this choice by assuming the training inputs are distributed according to this distribution and shown convergence strong guarantees if the training inputs are in fact drawn from a normal distribution, the inducing features defined are valid regardless of the locations of the actual training data. However, the number of features needed to achieve a good model depends heavily on the locations of the training inputs.

5.1.1 Effect of input Distribution

The one-dimensional datasets considered are created by randomly sampling a function from a zero-mean Gaussian process with squared exponential kernel with variance and lengthscale

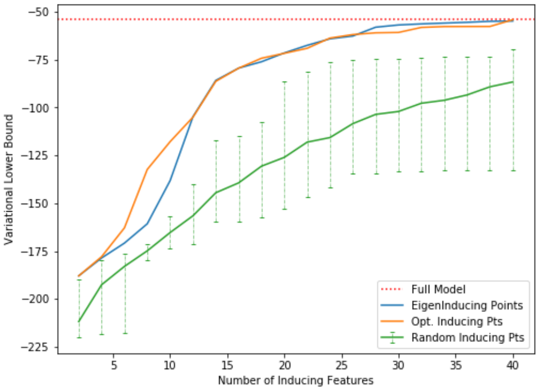

Fig. 5.1 Rate of convergence of approximate log likelihood for optimized inducing points, randomly selected inducing points and eigenfunction based inducing points for $x \sim \mathcal{N}(0,5^2)$.

parameters both equal to 1, and adding gaussian noise. Two-hundred training input locations are selected according to:

- $x \sim \mathcal{N}(0,5^2)$,
- $x \sim \mathcal{U}(-\sqrt{75},\sqrt{75})$ (note this lead to a standard deviation of 5, for sake of comparison to the normal distribution),
- $x \sim q$ where q_i is a mixture of two Gaussian with equal weight, means ± 10 and standard deviations $\sigma = 1$.

q is chosen to show an adversarial example when the empirical input distribution is multi-modal and supported on a region many lengthscales long.

Figures 5.1, 5.2 and 5.3 show the ELBO for various choices of M for each set of training inputs.

As optimized inducing points are prone to getting trapped in local optima, to ensure monotonicity of convergence and a good solution, we initialize the inducing points for the optimized models with $M+1$ inputs using the same randomly selected points as was used in the approximation with M inputs, and take the best approximation of 3 such initializations with different seeds. This ensures monotonicity in the bound, as well as a good optima.

The eigenfunction inducing inputs appear to avoid such optimization errors in general as they involve far fewer parameters. These are initialized with $p(x)$ set to have the same mean and standard deviation as the empirical training data.

The performance for inducing points randomly selected from training data is also plotted, along with error bars showing the best and worst performance for 10 trials.

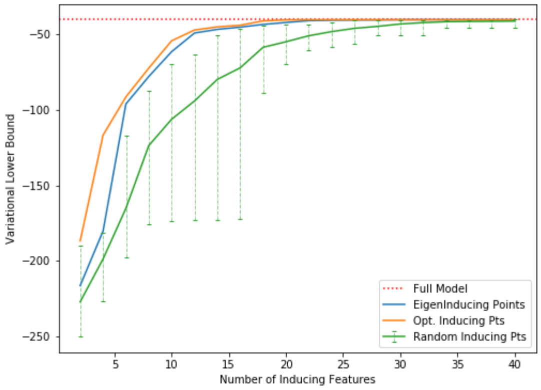

Fig. 5.2 Rate of convergence of approximate log likelihood for optimized inducing points, randomly selected inducing points and eigenfunction based inducing points for $x \sim \mathcal{U}(-\sqrt{75}, \sqrt{75})$.

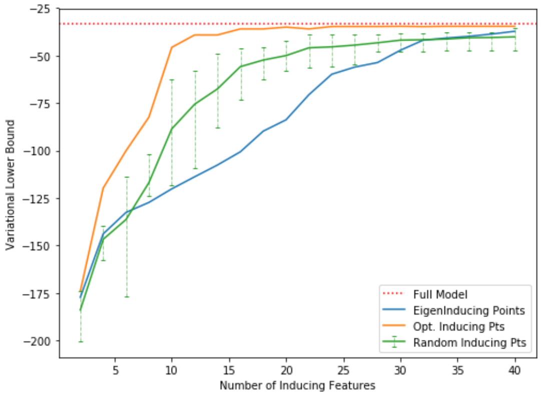

Fig. 5.3 Rate of convergence of approximate log likelihood for optimized inducing points, randomly selected inducing points and eigenfunction based inducing points for $x \sim q$.

Comparing the performance of the three methods, we see that the eigenfunction inducing point features perform comparably well to the optimized inducing inputs for a given M for inputs that are normally distributed or uniformly distributed on a short interval. At the same time they require the optimization of fewer variational parameters and offer a computational savings for large M.

For inputs coming from the multimodal input distribution, the quality of approximation of M eigenfunction inducing points is inferior to both optimized and randomly selected inducing points for moderate M. For larger M the eigenfunction based inducing points outperforms the randomly selected inducing points, while introducing only two additional variational parameters. In general, the eigenfunction inducing points struggle to recover the full model in sections of the domain in the tail of the input distribution. While optimized inducing points can handle these training inputs by placing a single or handful of points in these regions, the eigenfunction inducing features must cover the entire support of the data in order to model inputs contained in the tails.

5.1.2 Performance Prior to Convergence

In practical applications we may not be able to use sufficiently many inducing features in order for the model to have fully converged to the full model. For small M, the optimized model tends to select kernel hyperparameters that make the kernel *easier to approximate* sparsely. This is presumably in order to achieve a reasonably small trace error term in the ELBO and leads to the underfitting observed in previous papers for optimized inducing point methods (e.g. Bauer et al. (2016)) This means selecting a kernel with larger lengthscale and variance than the full model.

When used in conjunction with nonlocal basis functions, this can lead to pathological mean predictions away from the data, although these prediction come with large uncertainty. Figure 5.4 shows model behaviour prior to convergence, when all parameters are jointly optimized so the model chooses an easy to approximate kernel. This corresponds to the while figure 5.5 shows uses the same number of features to approximate a fixed kernel with the same hyperparameters at the full model, in order to isolate the impact of nonlocal basis functions.

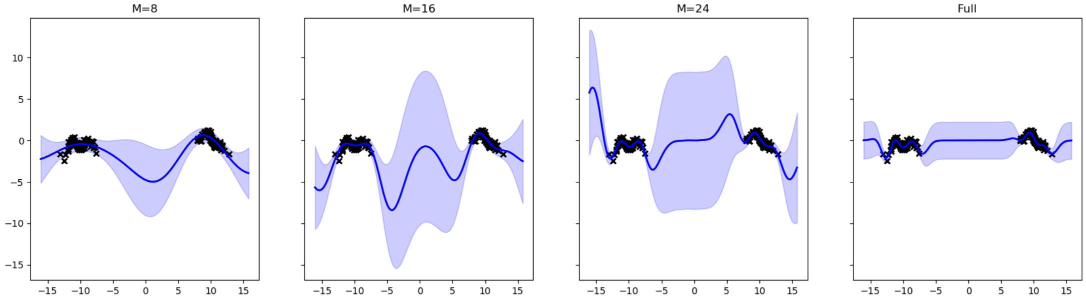

Fig. 5.4 Approximate models for $M = 8, 16, 24$ as well as the full model for the 1-dimensional toy dataset with x sampled from the mixture of Gaussian model.

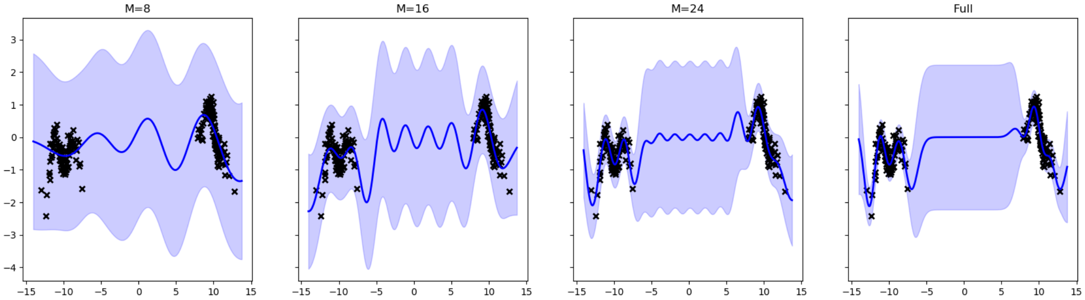

Fig. 5.5 Approximate models for $M = 8, 16, 24$ as well as the full model for the 1-dimensional toy dataset with x sampled from the mixture of Gaussian model, with the kernel hyperparameters fixed to the optimal obtained by the full model.

5.2 Scaling To Higher Dimensions: Additive and Multiplicative Structures

While the eigenfunction inducing point framework does not depend on the dimensionality of input data, we have not given specific examples indicating the eigenfunctions of higher dimensional operators. Two methods for defining a kernel on D-dimensions from a basic univariate kernels are *additive* kernels and *Kronecker* kernels, both applied to Variational Fourier Features in Hensman et al. (2016). In this section, both models will be reviewed considering theoretical implications from a spectral perspective as well as practical scaling considerations. We then apply each method to a different classification task.

5.3 Kronecker Gaussian Process Models

Commonly kernels on multidimensional inputs are defined by assuming that the kernel factors as a product of kernels over each dimension. If $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d$. and $\mathbf{x}' = (x'_1, \ldots, x'_d) \in \mathbb{R}^d$. we define a kernel on $\mathbb{R}^d$ by

$$k(\mathbf{x}, \mathbf{x}') = \prod_{i=1}^{d} k_i(x_i, x'_i). \tag{5.1}$$

It is then natural to define features on $\mathbb{R}^d$ by

$$u_{\mathbf{m}} = \int_{\mathbb{R}^d} \phi_{\mathbf{m}}(\mathbf{x}) k(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') d\mathbf{x}, \tag{5.2}$$

where $\mathbf{m} = (m_1, \ldots, m_d)$ is a multi-index and,

$$\phi_{\mathbf{m}}(\mathbf{x}) = \prod_{i=1}^{d} \phi_{m_i}(x_i). \tag{5.3}$$

Then,

$$\begin{aligned}
\operatorname{cov}(u_{\mathbf{m}}, u'_{\mathbf{m}}) &= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \phi_{\mathbf{m}}(\mathbf{x})\phi'_{\mathbf{m}}(\mathbf{x})k(\mathbf{x},\mathbf{x}')d\mathbf{x}'d\mathbf{x} \\
&= \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} \prod_{i=1}^{d} \phi_{m_i}(x_i)\phi_{m'_i}(x'_i)k_i(x_i,x'_i)d\mathbf{x}'d\mathbf{x} \\
&= \prod_{i=1}^{d} \int_{\mathbb{R}} \phi_{m_i}(x_i)\phi_{m'_i}(x'_i)k_i(x_i,x'_i)dx_i \\
&= \prod_{i=1}^{d} \operatorname{cov}_i(u_{m_i}, u_{m'_i}). \qquad (5.4)
\end{aligned}$$

and

$$\begin{aligned}
\operatorname{cov}(u_{\mathbf{m}}, f(\mathbf{x})) &= \int_{\mathbb{R}^d} \phi'_{\mathbf{m}}(\mathbf{x}')k(\mathbf{x},\mathbf{x}')d\mathbf{x}' \\
&= \prod_{i=1}^{d} \int_{\mathbb{R}} \phi_{m_i}(x'_i)k_i(x_i,x'_i)dx'_i \\
&= \prod_{i=1}^{d} \operatorname{cov}(u_{m_i}, f(x_i)). \qquad (5.5)
\end{aligned}$$

The latter equation indicates that the multidimensional features defined in this way are eigenfunctions of the product kernel, with respect to an input distribution that factors across each dimension. The eigenvalues are given by the product of the eigenvalues along each dimension. This is perhaps the most natural generalization of the eigenfunction framework, but forces axis aligned priors on inputs. Practically, this can be addressed by first performing PCA on training data.

Scaling considerations

While product kernels are often able to express complex relationships in the data, the number of inducing features needed will tend to scale exponentially in the number of dimensions. If we have M features defined along each dimension, this leads to M^d choices for $\mathbf{m}$, each corresponding to a feature. This leads to an inference cost of $O(NM^{2d})$, with a per iteration cost of $O(\tilde{N}M^{2d})$, and a memory cost of $O(NM^d)$. The total number of hyperparameters is $HD + 2MD$ where H is the number of kernel hyperparameters of the base kernel.

The total cost of inference and training can be reduced by imposing additional constraints on the covariance matrix and mean vector used to represent the variational distribution. Such constrained variational distributions are explored in Nickson et al. (2015) and Izmailov et al. (2017) for grid structured inducing points. Both approximations come at a cost in terms of

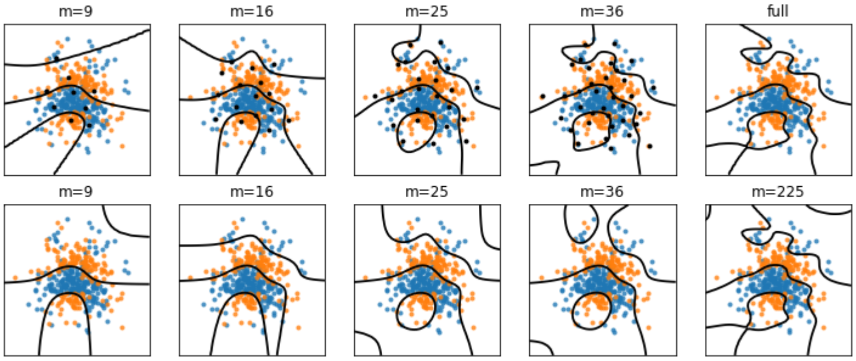

Fig. 5.6 Decision Boundaries for the normally distributed ‘bananas’ dataset for different numbers of inducing points (top), the full model (top, rightmost) and a Kronecker product of the Hermite Inducing Features (bottom row).

the ability of the variational posterior to approximate the true posterior, and can result in the model not converging to full inference, even as the number of features tends to infinity.

A Simple Example of Kronecker Inference

In order to show the eigenfunction inducing features potential applicability to multidimensional tasks, we consider a modified version of the toy ‘banana’ 2D classification experiment first performed in Hensman et al. (2015). We first train a full SE-ARD (Kronecker) kernel on a subset of 400 data points, then resample points from posterior according to a normal distribution with diagonal covariance function, and assign each new sample either a 0 or 1 with probability equal to the samples value. As the resampled points are normally distributed, this should be a favorable test for eigenfunction inducing points.

In figure 5.6 we see that for a fixed number of inducing features, the optimized inducing points outperform the eigenfunction based features in approximating a SE-ARD kernel on this classification task. However, for larger datasets and large M we should expect a significant computational savings for the eigenfunction based features (the dataset we use here is too small for minibatching to be practical).

5.4 Additive Kernels

An alternative method for extending one dimensional kernels to high dimensions is additive modelling, see Durrande et al. (2011) for a review. In an additive model, we have,

$$k(\mathbf{x},\mathbf{x}') = \sum_{i=1}^{d} k_i(x_i, x_i'). \tag{5.6}$$

Following Hensman et al. (2016), we then define inducing features by,

$$u_{m,i} = \int_{\mathbb{R}} \phi_m(x_i) k(x_i', x_i) f_i(x_i') dx_i, \tag{5.7}$$

so that $u_{m,i}$ is the m^{th} eigenfunction of the kernel defined along the i^{th} dimension. In an additive model, the Gaussian processes corresponding to each dimension are independent. This can be seen by noting that a Gaussian process is uniquely characterized by its covariance function and mean, and the covariance function for independent Gaussian processes is additive.

$$\text{cov}(u_{m,i}, u_{n,j}) = \delta_{i,j}\text{cov}(u_m, u_n), \tag{5.8}$$

and

$$\text{cov}(u_{m,i}, f(\mathbf{x})) = \text{cov}(u_m, f(x_i)). \tag{5.9}$$

This leads to a total of Md inducing features (M per dimension) and a corresponding inference cost of $O(NM^2d^2)$ inference cost.

An Example of Additive Inference on A Large dataset

As an example of the practical applicability of the features to large dataset, we focus on a classification task. While with stochastic optimization for regression tasks we expect large computational savings as compared to standard inducing points, if it is not essential to use a SE-kernel, using VFF and the collapsed likelihood will generally converge far faster than the eigenfunction features even with the a diagonal approximation to the covariance matrix of the approximating distributions, **S**, discussed in Section 3.3.3. Therefore, we expect the practical use for these features to occur with non-conjugate likelihoods, where they have the same computational complexity per iteration as VFF and lower than optimized inducing points.

The ‘airline’ dataset (Bureau of Transportation Statistics, 2008) has been widely used as a benchmark in Gaussian Process literature, for example in Hensman et al. (2016). The dataset contain 5.7 million US flight records, that has become a standard test for scalable Gaussian

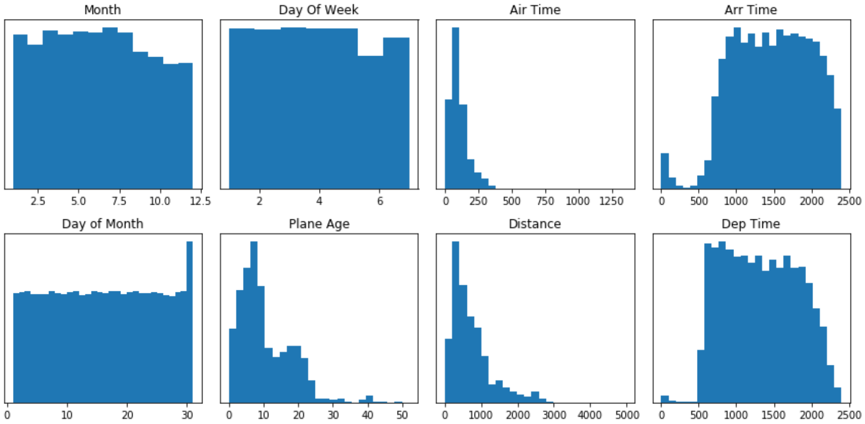

Fig. 5.7 Input distributions for the airline dataset.

process models. The response variable is the delay of the aircraft landing (minutes). We binarize this response, making it one if this delay is greater than zero (the plane is delayed) and zero otherwise. 43.6% of flights are delayed, while 56.4% of flights leave on time or early. The input space is 8-dimensional, with marginal distributions shown in figure 5.7. Several inputs have long tails and that may be difficult to model with the eigenfunction inducing feature approximation.

Several variables are highly correlated, notably arrival and departure time and time in the air and distance. Although it is not clear that independence of input variable distributions is useful for additive eigenfunction models (although it certainly should be the case for the Kronecker features) we fit several models after first applying PCA to orthogonalise the input variables (we keep all components, so this is simply a linear transformation of the input space). Note that an additive kernel defined on a space after a linear transformation defines a different model than an additive kernel defined on the initial space. The distributions of the components are shown in figure 5.8.

Inducing points are initialized via K-means on a random subset of the 10000 training data points. Two thirds of the flight records were split for training and one third was used for testing. The minibatch size was taken to be 500, and ADAM optimization was used with default parameters. All experiments are run on a single cpu. Times per iteration computed in a separate experiment from metrics and rescaled, as the cost of computing the metrics for such a large dataset is far larger than the per iteration cost.

Table 5.1 indicates that the per iteration cost for training the Eigenfunction inducing points is substantially lower than training standard inducing points, particularly for large M. (Here an throughout this section, M is the number of features per dimension, so the total number of features is $8M$), On the other hand, figure 5.9 shows that for this dataset,

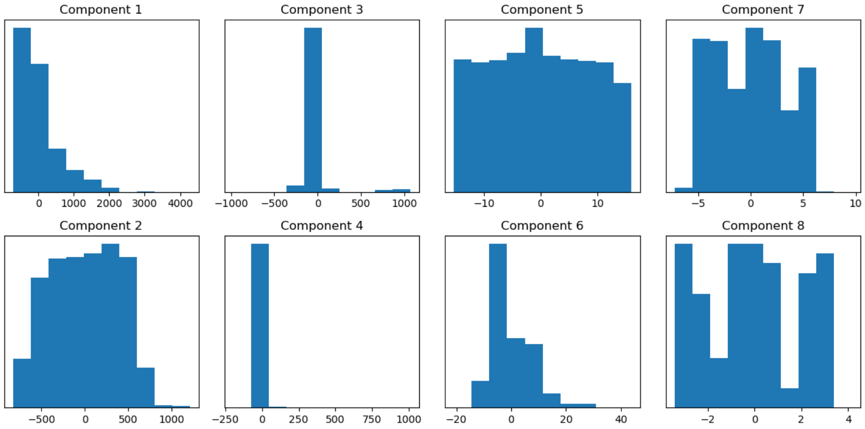

Fig. 5.8 Input distributions for the airline dataset.

Table 5.1 Comparison of time (minutes) to run 1000 iterations of optimization for various M for Eigenfunction inducing points and Optimized inducing points.

Approximation Method	M		
	60	90	120
Optimized Inducing Point	4.44	8.74	15.66
Eigenfunction Inducing Feature	2.92	5.57	8.74

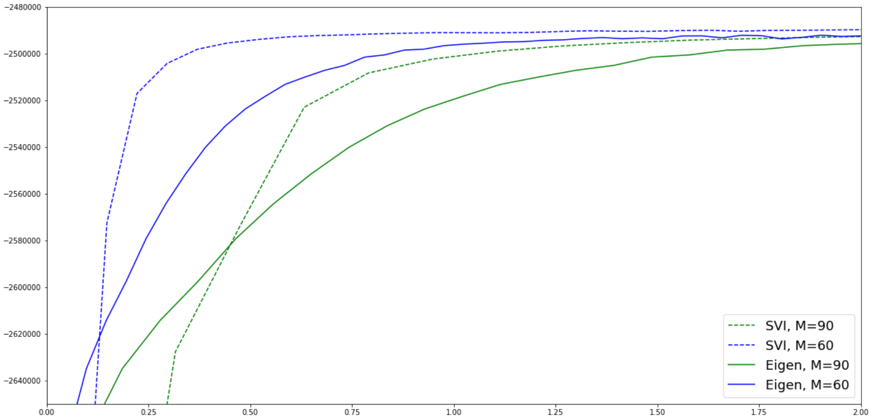

Fig. 5.9 ELBO for various M for raw dataset.

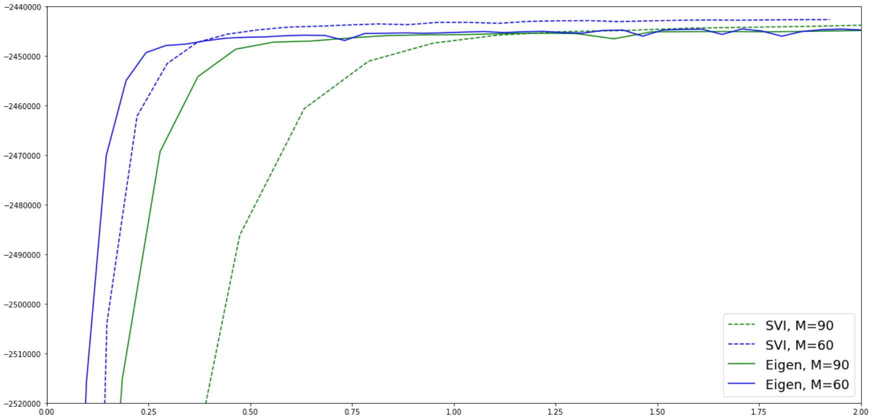

Fig. 5.10 ELBO for various M for dataset after applying PCA.

substantially more iterations of optimization are needed for the parameters to converge for the eigenfunction inducing points. The net effect of this is inferior performance for the eigenfunction inducing points for the untransformed data.

When working with the transformed data, the eigenfunction inducing points appear to need many fewer iterations than with the raw data, though still more than the optimized inducing points. The net effect is that $M = 60$ and $M = 90$ with limited computation (under 15 minutes for $M = 60$), the eigenfunction features outperform the optimized inducing points by a significant margin. This advantage reverses if longer training times are allowed. The ELBO and accuracy over time are shown in figures 5.10 and 5.11. The transformed data also leads to a model that is substantially better than the additive model over the untransformed outputs, as indicated by the comparing the ELBO (y-axis) in figures 5.10 and 5.10. After many iterations of training, there is still a gap between the performance of the eigenfunction

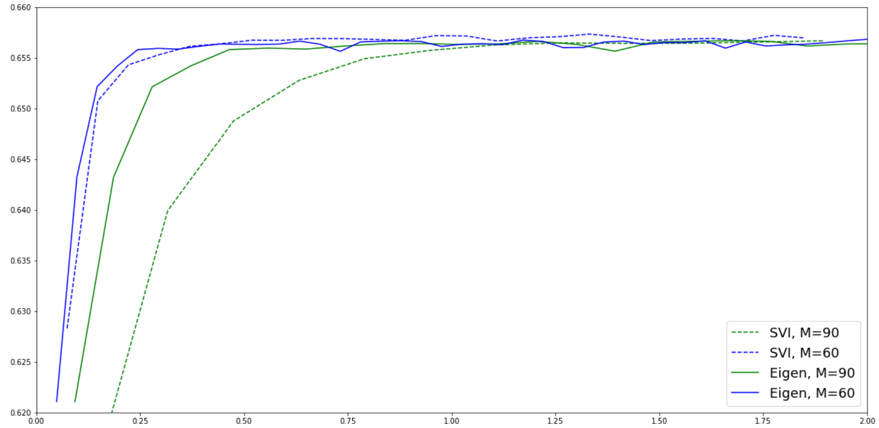

Fig. 5.11 Test accuracy for various M for dataset after applying PCA.

inducing points and the optimized inducing points. It is possible that even for $M = 90$ features along each dimension the eigenfunction inducing points are unable to capture the tails of the input distribution for this dataset and kernel.

The improved performance of the eigenfunction inducing points when applied for approximate inference after a linear transformation of the data space raises interesting questions about whether greater benefits may be observed when applying in conjunction with deep transformations of the input space, as in Wilson et al. (2016).

Chapter 6

Conclusion and Future Work

Spectral properties of Gaussian processes have been used for approximation methods for decades. In this work, we showed that these techniques can be applied within the variational framework, inheriting many of its theoretical and practical advantages. Expanding the scope of this intersection to allow for compact, flexible nonparametric approximations is a promising area for future work.

A key insight from the derivation of the eigenfunction inducing points is viewing variational optimization of inducing features as finding a measure over the input domain that gives sufficient weight to the training data locations. Deriving inducing features parametrized over larger classes of input distributions, and for more diverse kernels would be useful for improving the flexibility of this model. As analytic solutions to the eigenfunction are generally known only in special cases, finding efficient methods to approximate solutions for use in the covariance matrix, $\mathbf{K}_{m,n}$ within the variational framework is desirable.

While theoretical bounds on the rate of convergence for $M \ll N$ optimized inducing points are given, these bounds are only valid for M much larger than those values typically used for approximate inference. Obtaining stronger bounds for the Nyström approximation in this setting, perhaps by leveraging additional structure present in specific kernel matrices, is a promising direction for future work that could lead practical bounds on the convergence of approximate inducing point methods. Additionally, finding bounds on the KL-divergence that take into account the distance between the full and approximate likelihood directly, as opposed to only through the trace error could be an avenue for improving bounds on the convergence of inducing feature methods more generally.

In this work, a new class of 'interdomain features' was derived. The features are in some sense canonical, as they are closely related to spectral properties of the covariance matrix and leverage Mercer's Theorem and the Karhunen-Loéve expansion to yield features with easy to interpret variational parameters. These features were first motivated have computational benefits resulting from orthogonality properties. Additionally, convergence guarantees were obtained in certain cases for very sparse approximations using these features. Finally, the practical performance of these features was shown on a large data set.

References

Bauer, M., van der Wilk, M., and Rasmussen, C. E. (2016). Understanding probabilistic sparse gaussian process approximations. In *Advances in neural information processing systems*, pages 1533–1541.

Bureau of Transportation Statistics (2008). Airline on-time statistics and delay causes.

Durrande, N., Ginsbourger, D., Roustant, O., and Carraro, L. (2011). Additive covariance kernels for high-dimensional gaussian process modeling. *arXiv preprint arXiv:1111.6233*.

Ferrari-Trecate, G., Williams, C. K., and Opper, M. (1999). Finite-dimensional approximation of gaussian processes. In *Advances in neural information processing systems*, pages 218–224.

Gittens, A. and Mahoney, M. (2013). Revisiting the nystrom method for improved large-scale machine learning. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 567–575, Atlanta, Georgia, USA. PMLR.

Gradshteyn, I. S. and Ryzhik, I. M. (2014). *Table of integrals, series, and products*. Academic press.

Hensman, J., Durrande, N., and Solin, A. (2016). Variational fourier features for gaussian processes. *arXiv preprint arXiv:1611.06740*.

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

Hensman, J., Matthews, A. G. d. G., and Ghahramani, Z. (2015). Scalable variational gaussian process classification.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Horn, R. and Johnson, C. (1990). *Matrix Analysis*. Cambridge University Press.

Izmailov, P., Novikov, A., and Kropotov, D. (2017). Scalable gaussian processes with billions of inducing inputs via tensor train decomposition. *arXiv preprint arXiv:1710.07324*.

Kuss, M. and Rasmussen, C. E. (2005). Assessing approximate inference for binary gaussian process classification. *Journal of machine learning research*, 6(Oct):1679–1704.

Lázaro-Gredilla, M. and Figueiras-Vidal, A. (2009). Inter-domain gaussian processes for sparse inference using inducing features. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 1087–1095. Curran Associates, Inc.

Lazaro-Gredilla, M., Quinonero-Candela, J., Rasmussen, C. E., Figueiras-Vidal, A. R., et al. (2010). Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11(Jun):1865–1881.

Le Maître, O. and Knio, O. M. (2010). Spectral expansions. In *Spectral Methods for Uncertainty Quantification*, pages 17–44. Springer.

Matthews, A. G. d. G. (2016). *Scalable Gaussian process inference using variational methods*. PhD thesis.

Matthews, A. G. d. G., Hensman, J., Turner, R., and Ghahramani, Z. (2016). On sparse variational methods and the kullback-leibler divergence between stochastic processes. *Journal of Machine Learning Research*, 51:231–239.

Matthews, A. G. d. G., van der Wilk, M., Nickson, T., Fujii, K., Boukouvalas, A., León-Villagrá, P., Ghahramani, Z., and Hensman, J. (2017). GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 18(40):1–6.

Nickson, T., Gunter, T., Lloyd, C., Osborne, M. A., and Roberts, S. (2015). Blitzkriging: Kronecker-structured stochastic gaussian processes. *arXiv preprint arXiv:1510.07965*.

Quinonero-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.

Rahimi, A. and Recht, B. (2008). Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184.

Rasmussen, C. E. and Quinonero-Candela, J. (2005). Healing the relevance vector machine through augmentation. In *Proceedings of the 22nd international conference on Machine learning*, pages 689–696. ACM.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Shawe-Taylor, J., Williams, C., Cristianini, N., and Kandola, J. (2002). On the eigenspectrum of the gram matrix and its relationship to the operator eigenspectrum. In *International Conference on Algorithmic Learning Theory*, pages 23–40. Springer.

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

Titsias, M. K. (2014). Variational inference for Gaussian and determinantal point processes. Workshop on Advances in Variational Inference (NIPS 2014).

von Bahr, B. and Esseen, C.-G. (1965). Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Siam.

Williams, C. K. I., Rasmussen, C. E., Schwaighofer, A., and Tresp, V. (2002). Observations on the nyström method for Gaussian processes. Technical report.

Williams, C. K. I. and Seeger, M. (2001). Using the nyström method to speed up kernel machines. In Leen, T. K., Dietterich, T. G., and Tresp, V., editors, *Advances in Neural Information Processing Systems 13*, pages 682–688. MIT Press.

Wilson, A. G., Hu, Z., Salakhutdinov, R., and Xing, E. P. (2016). Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378.

Zhu, H., Williams, C. K. I., Rohwer, R., and Morciniec, M. (1997). Gaussian regression and optimal finite dimensional linear models. In *Neural Networks and Machine Learning*, pages 167–184. Springer-Verlag.

Appendix A

Structure of Covariance Matrices

A.1 Matérn Feature Covariance outside Interval

In order to apply the Matérn eigenfunction inducing features, we also must consider covariance with the process for x outside the interval, as the ϕ are no longer eigenfunctions so the covariance calculation does not apply. For $x' > a$,

$$\begin{aligned}\frac{1}{\lambda_i}\int_a^b &\frac{\sin\left(w\left(x-\frac{a+b}{2}\right)\right)\exp\left(-\frac{x'-x}{\ell}\right)}{\sqrt{\frac{b-a}{2}-\frac{\sin(w(b-a))}{2w}}}\,dx \\ &= \frac{e^{-\frac{x'}{\ell}}\left(\ell w\left(e^{a/\ell}-e^{b/\ell}\right)\cos\left(\frac{b-a}{2}w\right)+\left(e^{a/\ell}+e^{b/\ell}\right)\sin\left(\frac{(b-a)}{2}w\right)\right)}{\sqrt{2}\sqrt{-\frac{\sin(w(b-a))}{w}-a+b}}\end{aligned}$$

For the odd roots, we have $\ell\cos(\omega(b-a)/2) = -\frac{\sin(\omega(b-a)/2)}{\omega} =: v$. Then

$$\mathrm{cov}(u_m, f(x')) = \lambda_m \frac{ve^{-\frac{x'}{\ell}}\left(\ell\left(e^{a/\ell}-e^{b/\ell}\right)-\left(e^{a/\ell}+e^{b/\ell}\right)\right)}{\sqrt{2}\sqrt{v+b-a}},$$

for m odd. Similarly, if $x' < a$ we arrive at

$$\mathrm{cov}(u_m, f(x')) = \lambda_m \frac{ve^{-b-a+\frac{x'}{\ell}}\left(\ell\left(e^{b/\ell}-e^{a/\ell}\right)+\left(e^{a/\ell}+e^{b/\ell}\right)\right)}{\sqrt{2}\sqrt{v+b-a}}.$$

For even m, noting that we have $\cos(\omega(b-a)/2) = \ell\omega\sin(\omega(b-a)/2) =: v_m$, a similar calculation gives:

$$\operatorname{cov}(u_m, f(x')) = \begin{cases} \lambda_m \frac{v_m e^{-\frac{x'}{\ell}} \left(\ell\left(e^{b/\ell} - e^{a/\ell}\right) - \left(e^{b/\ell} + e^{a/\ell}\right)\right)}{\sqrt{2}\sqrt{v_m/(\ell\omega^2) + b - a}} & x' > b, \\ \lambda_m \frac{v_m e^{-\frac{b+a-x'}{\ell}} \left(\ell\left(e^{b/\ell} - e^{a/\ell}\right) + \left(e^{b/\ell} + e^{a/\ell}\right)\right)}{\sqrt{2}\sqrt{v_m/(\ell\omega^2) + b - a}} & x' < a. \end{cases}$$

A.2 Covariance Matrix of Optimal Approximating Distribution

In section 3.3.3 we showed that the precision matrix of the approximating distribution has the form,

$$\mathbf{S}_{opt} = (\mathbf{\Lambda} + \mathcal{E})^{-1}$$

where the entries in $\mathcal{E}$ are all $o(N)$, and $\mathbf{\Lambda}$ is a diagonal matrix with $\mathbf{\Lambda}_{i,i} = \lambda_i N$. Define $\mathbf{L} = \frac{1}{N}\mathbf{\Lambda}$ and $\mathbf{E} = \frac{1}{N}\mathcal{E}$. Then, expanding the matrix $\left(\mathbf{I} + \mathbf{E}\mathbf{L}^{-1}\right)^{-1}$ via a geometric series,

$$(\mathbf{L} + \mathbf{E})^{-1} = \mathbf{L}^{-1}\left(\mathbf{I} + \mathbf{E}\mathbf{L}^{-1} + \ldots\right) = \mathbf{L}^{-1} + \mathbf{E}', \tag{A.1}$$

where $\mathbf{E}' := \mathbf{L}^{-1}\mathbf{E}\mathbf{L}^{-1} + \mathbf{L}^{-1}\left(\mathbf{E}\mathbf{L}^{-1}\right)^2 + \ldots$ is a matrix with entries tending to zero with N. The geometric series used converges for N sufficiently large and $\mathbf{E}'$ has entries tending towards zero, as the largesr eigenvalue of $\mathbf{E}\mathbf{L}^{-1}$ is bounded above $\frac{M}{\lambda_M}$ times the largest entry in $\mathbf{E}$, which tends to zero as $N \to \infty$. Using the right hand side of (A.1)

$$\mathbf{S}_{opt} = (\mathbf{\Lambda} + \mathcal{E})^{-1} = \frac{1}{N}\left(\mathbf{L}^{-1} + \mathbf{E}'\right),$$

so $\mathbf{S}_{opt}$ is approximately diagonal with the entries on the diagonal on the order of $\frac{1}{N}$ and off-diagonal entries $o(\frac{1}{N})$. Under the assumption of bounded variance mentioned earlier, we have the off-diagonal entries are $O(N^{-3/2})$.