Circular Pseudo-Point Approximations for Scaling Gaussian Processes

Will Tebbutt
Invenia Labs, Cambridge, UK
will.tebbutt@invenialabs.co.uk

Thang D. Bui
University of Cambridge
tdb40@cam.ac.uk

Richard E. Turner
University of Cambridge
ret26@cam.ac.uk

Abstract

We introduce a new approach for performing accurate and computationally efficient posterior inference for Gaussian Process regression problems that exploits the combination of pseudo-point approximations and approximately circulant covariance structure. We argue mathematically that the new technique has substantially lower asymptotic complexity than traditional pseudo-point approximations and demonstrate empirically that it returns results that are very close to those obtained using exact inference.

1 Introduction

Exact inference in Gaussian Process (GP) regression for large data sets is rendered infeasible by the poor scalability of the method with data set size [1]. This has motivated the development of a large selection of techniques that exploit specific properties of a problem to accelerate the computations required for inference. For example, it is often the case that a rather simple posterior process results from a very large amount of data. Pseudo-point ("sparse") GP approximations (see [2] and [3] for reviews of the available techniques) exploit this property elegantly using a small pseudo-dataset to summarise the true dataset. However, these methods return a terrible approximation if too few pseudo-data are used. This becomes problematic for large complicated problems that necessitate the use of a large number of pseudo-data.

Such large scale problems can be rendered tractable by the presence of special structure in the covariance matrix [4]. For example, if observations are on a regular grid and the covariance function is stationary then the covariance matrix is Toeplitz, and therefore approximately circulant [5], which can be exploited to accelerate inference substantially. The fundamental issue with these methods is their limited domain of applicability requiring special input locations (regular sampling) and special covariance structure that is inherited by the posterior (stationarity).

This paper examines a new approach to the combination of pseudo-point methods and those which exploit special structure, with the goal of obtaining lower asymptotic complexity than pseudo-point methods whilst placing as few restrictions as possible on the types of problems that can be tackled. This work is similar to existing work [6, 7, 8] in that we propose to place the pseudo-points on a regular grid, but dissimilar in the approach taken to exploiting this to accelerate inference. We now briefly review pseudo-point and circulant GP approximations and develop the Pseudo-Circular GP (PCGP) approximation.

2 Pseudo-Point GP Approximation

The variational pseudo-point GP (VFE) approximation [9] is arguably the state-of-the-art approximation method. Its properties are examined in [10], and is rederived from the perspective of the KL divergence between stochastic processes in [11]. In the case of regression, the optimal posterior distribution over the pseudo-data f_Z is shown to be a multivariate normal distribution $\mathcal{N}(f_Z \mid \mu_q, \Sigma_q)$

with parameters

$$\Sigma_q := K_{Z,Z}\left(\beta K_{Z,D}K_{D,Z} + K_{Z,Z}\right)^{-1}K_{Z,Z}, \;\; \mu_q := \beta\Sigma_q K_{Z,Z}^{-1}K_{Z,D}y.$$

where $K_{Z,Z}$, $K_{D,Z}$, β^{-1} and y are the prior pseudo-data covariance, cross-covariance between observed data and pseudo-data, the observation noise and observations respectively. Furthermore the expression for the Evidence Lower Bound (ELBO) at this optimum is

$$L = \log\mathcal{N}\left(y \,\middle|\, K_{D,Z}K_{Z,Z}^{-1}\mu_q, \beta^{-1}\mathcal{I}\right) - \frac{\beta}{2}\text{tr}\left(\hat{K}_{D,D} + R_{D,D}\right) - \mathcal{KL}[\mathcal{N}\left(f_Z \,|\, \mu_q, \Sigma_q\right)||\mathcal{N}\left(f_Z \,|\, 0, K_{Z,Z}\right)],$$

where $\hat{K}_{D,D} := K_{D,D} - K_{D,Z}K_{Z,Z}^{-1}K_{Z,D}$ and $R_{D,D} := K_{D,Z}K_{Z,Z}^{-1}\Sigma K_{Z,Z}^{-1}K_{Z,D}$.

This technique has asymptotic complexity $\mathcal{O}\left(NM^2 + M^3\right)$, owing to the matrix multiplication $K_{Z,D}K_{D,Z}$ and the Cholesky decomposition required to compute the posterior covariance and KL divergence.

3 Circulant GP Approximation

An alternative approach to GP approximation is to utilise special structure. The circulant GP (CGP) approximation can be applied when the data are regularly sampled and the covariance function is stationary, $k\left(x, y\right) := k\left(x - y\right)$. The method works by transforming the problem from the original Euclidean input space to a ring such that

$$\hat{k}\left(\Delta\right) := k\left(\left[\left(\Delta + d\right) \bmod 2d\right] - d\right), \quad d := (u - l)/2. \tag{1}$$

The covariance matrix resulting from evaluation of $\hat{k}$ at each pairing of the input locations $x_n = n(u-l)/N$ for $n \in \{0, ..., N-1\}$ will be exactly circulant. As the ring, and therefore $u-l$, becomes large the bias introduced into log marginal likelihood computations becomes minimal [12, 5].

Being circulant, this covariance matrix can be expressed as $K = U\Gamma U^\dagger$, where $U \in \mathbb{C}^{N\times N}$ is the Discrete Fourier Transform (DFT) matrix, defined as $U_{m,n} := N^{-\frac{1}{2}}e^{-2\pi imn/N}$, and $\Gamma = \text{diag}\left(\gamma\right)$ is the diagonal matrix whose diagonal γ is the DFT of the first row of K. This means that $\log|K| = \sum_{n=1}^{N}\log\gamma_n$ can be computed in $\mathcal{O}\left(N\log N\right)$ time by using the Fast Fourier Transform (FFT) to obtain γ. Furthermore the quadratic form $xK^{-1}x^T = xU\Gamma^{-1}U^\dagger x^T = \sum_{n=1}^{N}\gamma_n^{-1}\left|Ux\right|_n^2$ can be computed in $\mathcal{O}\left(N\log N\right)$ time by efficiently computing Ux with the FFT.

4 Pseudo-Circular GP Approximation

The Pseudo-Circular GP (PCGP) approximation combines the discussed approximations so that non-regularly sampled input data can be approximated using a large number of regularly spaced pseudo-data. This is achieved by placing the pseudo-data on a regular grid which extends outside the domain on which we observe data and circularising the covariance function. $K_{Z,Z}$ is rendered circulant and, consequently, $K_{Z,Z}^{-1}$ and $|K_{Z,Z}|$ inexpensive to compute. μ_q can be found from the ELBO efficiently using Conjugate Gradients (CG) [13] as L is quadratic in μ_q,

$$L = -\frac{1}{2}\mu_q^T\left(\beta K_{Z,Z}^{-1}K_{Z,D}K_{D,Z}K_{Z,Z}^{-1} + K_{Z,Z}^{-1}\right)\mu_q - 2y^TK_{D,Z}K_{Z,Z}^{-1}\mu_q + \text{const.} \tag{2}$$

The most expensive computation required for CG is $(\beta K_{Z,Z}^{-1}K_{Z,D}K_{D,Z}K_{Z,Z}^{-1} + K_{Z,Z}^{-1})\mu_q$, which can be performed efficiently by summing $\beta K_{Z,Z}^{-1}K_{Z,D}K_{D,Z}K_{Z,Z}^{-1}\mu_q$ and $K_{Z,Z}^{-1}\mu_q$, which require $\mathcal{O}\left(NM\right)$ and $\mathcal{O}\left(M\log M\right)$ operations respectively to leading order.

Σ_q is more problematic as it contains $M(M+1)/2$ parameters, meaning that an arbitrary positive definite matrix quite clearly cannot be used. We propose to use $\Sigma_q := K_{Z,Z}^{\frac{1}{2}}VV^TK_{Z,Z}^{\frac{1}{2}}$ for some lower-triangular band-diagonal matrix $V \succeq 0$, with bandwidth b, and $K_{Z,Z}^{\frac{1}{2}} := U\Gamma^{\frac{1}{2}}U^\dagger$. The matrix V can be interpreted as the Cholesky decomposition of a positive definite band-diagonal matrix. Critically, this parameterisation ensures that the variational posterior can be non-stationary. The terms in L dependent upon Σ_q

$$L_\Sigma := -\frac{1}{2}\text{tr}\left(K_{D,Z}K_{Z,Z}^{-\frac{1}{2}}VV^TK_{Z,Z}^{-\frac{1}{2}}K_{Z,D}\right) - \frac{1}{2}\text{tr}\left(VV^T\right) + \log|V| \tag{3}$$

can now be computed efficiently. The matrix multiplication $A := K_{D,Z}K_{Z,Z}$ can be computed in $\mathcal{O}\left(NM(\log M + b)\right)$ time, the multiplication $B := AV$ can be performed in $\mathcal{O}\left(NMb\right)$ time, and $\text{tr}\left(BB^T\right) = \sum_{m,n=1}^{M,N} B_{m,n}$ can also be computed in $\mathcal{O}\left(MN\right)$ time. Furthermore, since V is lower-triangular $\log|V| = \sum_{m=1}^{M} \log L_{m,m}$. These savings are significant over the standard VFE approach.

Under the parametrisation the prior is recovered when $V = \mathcal{I}$, and the posterior covariance is reduced globally by setting elements of V less than 1. This parametrisation can be thought of as decorrelating $K_{Z,Z}$ using a band-diagonal matrix, rather than attempting to construct the posterior from scratch using a band-diagonal matrix. Therefore, although V is band-diagonal, the posterior covariance approximation is dense. ELBO evaluation now has $\mathcal{O}\left(NM(b + \log M)\right)$ asymptotic complexity, however, it is not clear how to solve directly for V (or equivalently for $W := VV^T$) owing to the band-diagonal constraint except in the case that V is diagonal. As such, the gradient-based optimisation method Adagrad [14] is used to find the optimal solution for $b > 0$.

5 Parametrisation Experiments

We examine the performance of the PCGP approximation as the bandwidth b is varied on a toy problem in which 750 data are drawn from a GP with an Exponentiated Quadratic covariance function with length scale $l^2 = 1.0$ and variance $\sigma^2 = 1.0$ under observation noise $\beta^{-1} = 0.1$. Figure 1 displays this toy data set, along with the mean and marginal variance for the exact GP posterior and several band-widths with $M = 50$ pseudo-data. This clearly demonstrates the reasonable performance of a diagonal V and the improvements attained by increasing the bandwidth.

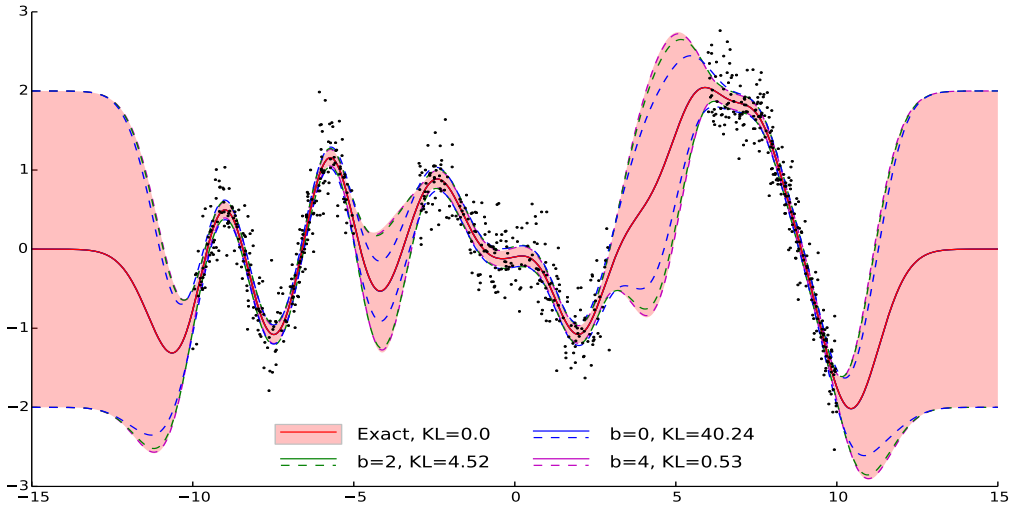


Figure 1: Mean and marginal variances for the toy problem. The KL divergence between each approximation and the exact solution is displayed in the appropriate legend entry. Owing to the independence of the mean optimisation on the bandwidth of V, the means recovered for bandwidth $b = 0, 2, 4$ are identical and, as $M = 50$ is sufficient to represent the function accurately over the input domain chosen, are indistinguishable from the true posterior mean. The posterior covariance for band width $b = 0$ (diagonal V) does a reasonable job of recovering the posterior, although appears to underestimate the marginal statistics in short regions of high posterior variance. Bandwidths $b = 2, 4$ converge to the posterior marginals more convincingly.

Figure 2 shows the performance of the proposed PCGP approximation for a range of bandwidths b and pseudo-data counts M. The left hand image shows that the performance for small M or b is relatively poor, however, the performance quickly improves to yield close to 0 KL divergence between the approximation and true posterior for roughly $M = 40$, $b = 3$.

A moderately large experiment with $M = 10^4$ pseudo-data and $N = 2 \times 10^4$ observations was conducted on audio sub-band data for bandwidth $b = 0$. Table 3 shows that at this scale, the RMSE for

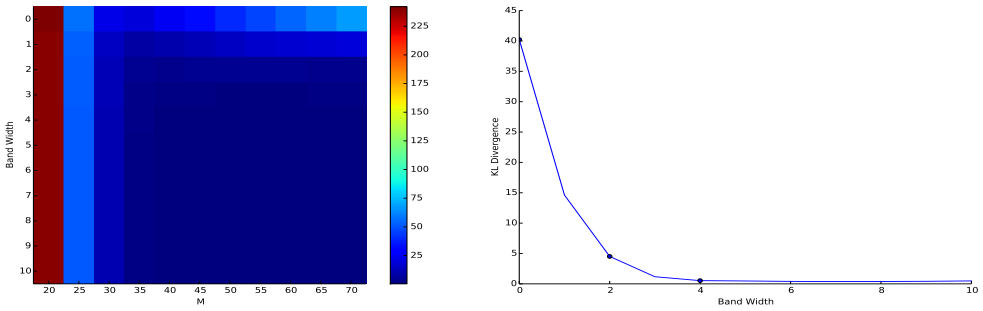
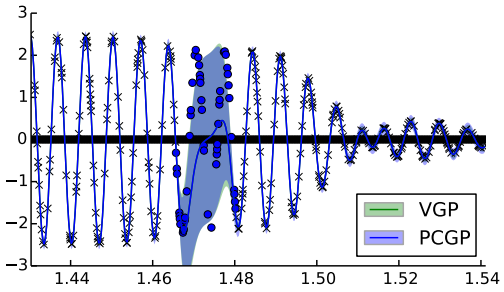


Figure 2: (Left) The KL divergence achieved after 1000 iterations of Adagrad for a range of band widths and numbers of pseudo-data. (Right) The $M = 50$ column of the image on the left, shown for clarity. Performance decreases when pseudo-data are too close together for too narrow a band width, as for $b = 0$, $M = 70$, because the difference between prior and posterior covariance cannot be represented. Highlighted point (circled) correspond to the approximations shown in figure 1.

both in and out of training domain data is the same for VFE and PCGP to within a reasonable tolerance. However, PCGP achieves this performance in almost half the time taken by VFE, with reasonable recovery of the posterior marginal variance as shown in figure 3. Note also that experiments involving $M > 10^4$ quickly become infeasible for VFE due to the $\mathcal{O}\left(M^2\right)$ memory requirements, whereas PCGP can handle a very large number of pseudo-data as the memory requirement is linear in M.



Method	Time (s)	RMSE (in)	RMSE (out)
PCGP	603	9.03×10^{-3}	1.77
VFE	1045	9.02×10^{-3}	1.77

Figure 3: Results on audio sub-band data comprising $N = 20000$ irregularly sampled observations. Inference was undertaken using $M = 10000$ pseudo-data. The plots show a small section of the audio sub-band data with a region of missing data. 50 observations removed between $t = 1.44$ and $t = 1.46$ form a small held-out data set in this case. The reconstruction results are shown in the table. Note despite the narrow band-width $b = 0$, the recovered marginal variances are very similar between the PCGP and VFE approximations.

6 Conclusion and Future Work

The long term aim of this work is to intelligently embed special covariance structure in pseudo-point approximations to scale GP models to large complicated problems. The Pseudo-Circular GP (PCGP) approximation is an efficient method in this vein for performing approximate inference in a univariate GP regression task. PGCP can in principle be used in conjunction with any stationary covariance function, or non-stationary covariance function of the form $k(x, y) = \hat{k}(g(x), g(y))$ for some stationary covariance function $\hat{k}$ by placing the pseudo-data regularly in the space mapped to by g. PCGP also generalises to arbitrary likelihood functions, such as those used for classification.

The lack of efficient implementations for particular operations involving band-diagonal matrices, notably matrix-matrix multiplication, currently hinders the performance of this method in practice. Furthermore, the sparsity induced by the local nature of pseudo-points is not currently exploited to handle cross-covariances efficiently. To achieve truly linear scaling in N this will need to be addressed.

References

[1] Carl Edward Rasmussen and Christopher KI Williams. Gaussian processes for machine learning. *the MIT Press*, 2(3):4, 2006.

[2] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research*, 6(Dec):1939–1959, 2005.

[3] Thang D Bui, Josiah Yan, and Richard E Turner. A Unifying Framework for Sparse Gaussian Process Approximation using Power Expectation Propagation. *arXiv preprint arXiv:1605.07066*, 2016.

[4] Yunus Saatçi. *Scalable inference for structured Gaussian process models*. PhD thesis, Citeseer, 2012.

[5] Richard E Turner. *Statistical models for natural sounds*. PhD thesis, UCL (University College London), 2010.

[6] Andrew Gordon Wilson and Hannes Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). *arXiv preprint arXiv:1503.01057*, 2015.

[7] Andrew Gordon Wilson, Christoph Dann, and Hannes Nickisch. Thoughts on massively scalable gaussian processes. *arXiv preprint arXiv:1511.01870*, 2015.

[8] Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Stochastic variational deep kernel learning. *arXiv preprint arXiv:1611.00336*, 2016.

[9] Michalis K Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 567–574, 2009.

[10] Matthias Stephan Bauer, Mark van der Wilk, and Carl Edward Rasmussen. Understanding Probabilistic Sparse Gaussian Process Approximations. *arXiv preprint arXiv:1606.04820*, 2016.

[11] Alexander G de G Matthews, James Hensman, Richard E Turner, and Zoubin Ghahramani. On Sparse variational methods and the Kullback-Leibler divergence between stochastic processes. *arXiv preprint arXiv:1504.07027*, 2015.

[12] Robert M Gray. *Toeplitz and circulant matrices: A review*. now publishers inc, 2006.

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

[14] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.